A high-resolution protein architecture of the budding yeast genome

Matthew J. Rossi1, Prashant K. Kuntala1, William K. M. Lai1,2, Naomi Yamada1, Nitika Badjatia1, Chitvan Mittal1,2, Guray Kuzu1, Kylie Bocklund1, Nina P. Farrell1, Thomas R. Blanda1, Joshua D. Mairose1, Ann V. Basting1, Katelyn S. Mistretta1, David J. Rocco1, Emily S. Perkinson1, Gretta D. Kellogg1,2, Shaun Mahony1 & B. Franklin Pugh1,2

The genome-wide architecture of chromatin-associated proteins that maintains chromosome integrity and gene regulation is not well defined. Here we use chromatin immunoprecipitation, exonuclease digestion and DNA sequencing (ChIP–exo/seq)1,2 to define this architecture in Saccharomyces cerevisiae. We identify 21 meta-assemblages consisting of roughly 400 different proteins that are related to DNA replication, centromeres, subtelomeres, transposons and transcription by RNA polymerase (Pol) I, II and III. Replication proteins engulf a nucleosome, centromeres lack a nucleosome, and repressive proteins encompass three nucleosomes at subtelomeric X-elements. We find that most promoters associated with Pol II evolved to lack a regulatory region, having only a core promoter. These constitutive promoters comprise a short nucleosome-free region (NFR) adjacent to a +1 nucleosome, which together bind the transcription-initiation factor TFIIID to form a preinitiation complex. Positioned insulators protect core promoters from upstream events. A small fraction of promoters evolved an architecture for inducibility, whereby sequence-specific transcription factors (ssTFs) create a nucleosome-depleted region (NDR) that is distinct from an NFR. We describe structural interactions among ssTFs, their cognate cofactors and the genome. These interactions include the nucleosomal and transcriptional regulators RPD3-L, SAGA, NuA4, Tup1, Mediator and SWI–SNF. Surprisingly, we do not detect interactions between ssTFs and TFIIID, suggesting that such interactions do not stably occur. Our model for gene induction involves ssTFs, cofactors and general factors such as TBP and TFIIB, but not TFIIID. By contrast, constitutive transcription involves TFIIID but not ssTFs engaged with their cofactors. From this, we define a highly integrated network of gene regulation by ssTFs.

Genomes regulate genes so as to achieve homeostasis—the maintenance of cellular components in proper balance. They also adapt, making adjustments in rapidly changing environments, so as to regain homeostasis3. Achieving these tasks has necessitated the evolution of constitutive and inducible gene control. Whether or not these controls are fundamentally different at the molecular level is unknown. A classical view posits a single basic regulatory paradigm for genes (Extended Data Fig. 1a): environmental signals toggle ‘on’ ssTFs that recruit cofactors and assemble a preinitiation complex (PIC) consisting of Pol II and general transcription factors (GTFs) such as TBP, TFIIID and TFIIIB at core promoter transcription start sites (TSSs)3. However, the extent to which constitutive gene expression involves ssTFs is unclear, as ssTF-binding sites and their cofactors remain unidentified at most promoters. ssTFs, cofactors, chromatin and PICs play into any distinction between inducible and constitutive mechanisms, but their interrelationships remain enigmatic.

Genome-wide protein meta-assemblages

Here we used ChIP–exo (Extended Data Fig. 1b)1,2, an ultra-high-resolution version of ChIP–seq, to map genome-wide binding. We selected target proteins on the basis of Gene Ontology (GO) annotations related to chromosomal function (Extended Data Fig. 1c and Supplementary Data 1 (1BY); characters in parentheses refer to the worksheet number and column letter). In total, we collected 1,229 datasets on 791 targets, of which 400 targets had reproducibly significant data (Supplementary Data 2 (1A)). The interaction pattern of all 1,229 datasets around individual and broad classes of genomic features (Fig. 1a) can be visualized and downloaded at yeastepigenome.org (an example is given in Extended Data Fig. 2). We also developed and provide ScriptManager, a platform for customized analysis of these data (see Methods).

Binarized colocation counts among targets were hierarchically clustered (Fig. 1b). The three largest clusters (yellow) correspond to three
major aspects of gene expression: first, promoter regulation; second, PIC assembly; and third, transcription elongation. Thus, the vast majority of chromatin-associated proteins are dedicated to gene regulation. We used uniform manifold approximation and projection (UMAP) to represent each dataset as a single point in a two-dimensional projection (Fig. 1c and Extended Data Fig. 3). Points in close proximity reflect a population-based composite colocalization of targets (‘meta-assemblages’). We performed K’means clustering on the projection and derived 21 meta-assemblages that correspond largely to known interacting biochemical complexes, or related gene ontologies (Fig. 1c, outer pie, and Supplementary Data 2 (1F, 1H, 2G–2I)). This probably represents a comprehensive predominant protein architecture of the yeast genome (‘epigenome’) in rich media (see Supplementary Data 2 (1–8) for a deeper analysis).

Overall, the organization defined by UMAP represents a remarkable degree of concordance and mutual validation of biochemically purified and functionally annotated complexes with their architectural organization across a genome, particularly from an unsupervised approach. For example, the promoter cofactors Mediator, SWI–SNF, SAGA, NuA4 and their cognate ssTFs each formed tight meta-assemblages that were located near each other but far from gene-body elongation factors (Fig. 1c). Proteins of replication origins, subtelomeres and centromeres also formed distinct tight meta-assemblages that were far from each other and from gene meta-assemblages. This provided strong validation of the ChIP–exo/seq approach and epitope tagging. Notably, we can now link most ssTFs with their cognate cofactors and promoter architecture.

Protein architecture at genomic features

DNA replication initiates at 253 autonomously replicating sequence consensus sequence (ACS) elements that are constitutively bound by origin recognition complexes (ORCs)⁶. The ‘ORC’ meta-assemble contained six measured targets (Fig. 2a and Extended Data Fig. 4), which gave highly structured ChIP–exo patterns based on ORCs and the DNA helicase MCM, spread over roughly 300 base pairs. ORCs at nucleosome-free ACSs engulfed a neighbouring nucleosome. The binding of Mcm5 from ORCs was offset by 50–100 bp, consistent with a recently published model based on cryo-electron microscopy⁷.

Subtelomeric X-elements represent a heterochromatic environment that is repressed by silent information regulators (SIRs), functionally supporting telomeres⁸. Indeed, we found that SIR proteins formed a structurally robust meta-assemble on a single nucleosome, centred on roughly 300-bp X-core elements (XCEs), along with ORC/MCMs and insulator ssTFs at two flanking nucleosomes (Fig. 2b). KU (Yku70) and RIF (Rif1) complexes, along with ssTFs Fkh1, Abf1 and Reb1, were present at the vast majority of mappable X-elements. However, a Sko1-mediated Tup1 repression complex was present at only half, perhaps reflecting variable repression capabilities of subtelomeric regions. Thus, XCEs appear to create a well-structured triple nucleosome ensemble comprising major repressor proteins.

The centromere meta-assemble (‘CEN’) contained 12 targets at 16 centromeres (Fig. 2c), which are responsible for proper chromosomal segregation during cell division. They included site-specifically bound Cbf1 at the centromere centre (CDE I) and kinetochore components offset by roughly 100 bp towards the AT-rich CDE III elements⁹. These factors generated strong and well positioned crosslinks covering roughly 170 bp of DNA, suggesting that they are positionally fixed to CDEs. Condensin and cohesin play a part in chromosomal condensation and segregation. They were absent from the centromere and instead overlapped the surrounding nucleosomes, suggesting that they interact with nucleosomes. In contrast with lower-resolution maps⁹, histones were not detected at centromeres, despite robust detection of histone-like Cse4 and kinetochore components there, and robust detection of histones (H2A, H2A.Z, H2B, H3 and H4) in the immediate flanking regions¹⁰. Thus, yeast centromeres appear to lack the histone components of a nucleosome in vivo. The resident kinetochore complex protects a nucleosome-sized region of DNA from nucleases, which was a basis for a nucleosome originally being called there¹¹. Nonetheless, Cse4-containing nucleosomes have been defined biochemically and structurally in vitro¹², so the question remains open.

The Pol I complex produces ribosomal RNA (rRNA) from a single highly repeated gene. It contained TBP anchored near the rRNA TSS, the broad extension of Pol I downstream into the rRNA gene body around 200 bp of rRNA promoter DNA, with an intervening 100 bp or so. The Pol I initiation complex has a fixed bipartite engagement that covers around 200 bp of rRNA promoter DNA, with an intervening 100 bp or so. The broad extension of Pol I downstream into the rRNA gene body with less occupancy at promoters indicates that Pol I dissociates rapidly from its PIC into an elongating state.

Pol III of the ‘POL3’ meta-assemble transcribes 272 highly similar genes encoding transfer RNAs (tRNAs). It contained 18 targets that could be separated into TFIIB/C and Pol III meta-assembles (Fig. 3b). Their organization matched locations modelled from atomic structures of the TFIIB/Pol III promoter complex¹⁴, but with the TBP component of TFIIB crosslinking approximately 30 bp upstream of the TSS. The ChIP–exo pattern further demonstrated that TFIIB and Pol III make crosslinks not only at the internal A and B boxes, but also at coincident locations roughly 40 bp upstream of TBP. Owing to DNA bending

---

**Fig. 1 | Genome-wide meta-assemblages.** a. Classes of genomic features, with N memberships analysed (Supplementary Data 1 (ID)). Pol II classes are from this study (see Methods), along with relative PIC occupancy levels (green dots). CUTs, cryptic unstable transcripts; SUTs, stable unannotated transcripts; XUT, Xrn1-sensitive unstable transcripts; NCR, noncoding RNA.

b. Hierarchical clustering showing the genome-wide colocalization of 371 targets (Supplementary Data 3). c. UMAP projection showing the colocalizations of 371 targets (coloured on the basis of K’means; Supplementary Data 2 (IC, ID)). AU, arbitrary units.
TBP, this region is in close proximity to TFIIIB/C and Pol III within gene bodies. Equivalent positions of crosslinking points were observed across all TFIIIB/C/Pol III subunits. This suggests that a single predominant structure envelopes entire Pol III genes and approximately 70 bp upstream, as it makes a short (roughly 80 bp) transcript.

There are around 7,500 distinct Pol II transcription units (defined by a TSS/PIC), of which approximately 80% code for proteins. Targets that are associated with transcription elongation generally matched Pol II occupancy across gene bodies, but unlike Pol II (Rpb3) were not present at promoters (Fig. 3c and Extended Data Fig. 5). Instead, occupancy within genes increased in the 5′ region and decreased in the 3′ region, with many having distinct ‘entry/exit’ points, consistent with other studies. Whether these are true cotranscriptional entry/exit points or are simply crosslinkable retention sites is not clear. Termination factors such as Pcf11 were found primarily at sites of termination, along with nearby cohesin. There was little evidence of the binding of

---

**Fig. 2** Architecture at nontranscribed features. a–c. Averaged distribution of strand-separated 5′ ends of ChIP–exo sequencing tags (exonuclease stop sites; see Extended Data Fig. 1b), showing representative targets around strand-oriented annotated features. The diagrams at the top of each panel are cartoon representations of DNA, nucleosomes and protein factors that bind to DNA replication origins (ACS), subtelomeric X-elements (XCE) or centromeres (CEN). The relevant start sequences (coloured As, Ts, Cs and Gs) are also shown in DNA replication origins (ACS) or centromeres (CEN). Underneath are composite data showing the distribution of the protein factors. Same-strand data are oriented with 5′ to 3′ to be read from left to right. Opposite-strand data are inverted (right to left is 5′ to 3′). The y-axes show linear arbitrary units (AU), which are not comparable in magnitude across different datasets. Nucleosome dyads were derived from MNase-digested chromatin that was assayed by H3/H2B ChIP–seq (strands averaged).

---

**Fig. 3** Architecture at transcribed features. a–d. Experiments were carried out as in Fig. 2, but for transcribed features. In a, UCE is an upstream control element at Pol I promoters. In b, A and B are box elements at Pol III promoters. In c, the protein architecture for RP genes is shown (not strand separated); Ser2P and Ser52P are phosphorylated serines 2 and 5 of heptad repeats; grey arrows show nucleosome dyads.
Fig. 4 | Classification of inducible, insulatable and constitutive Pol II promoters. a. Individual promoters (rows) can be grouped into four architectural themes (coloured boxes) and sorted by PIC occupancy level. Targets are listed at the top of columns, with arrows denoting Abf1, Reb1 and H2A.Z. Black lines denote target binding (Supplementary Data 2 (3b)). b, Top, diagram, and bottom, example composite data for the STM, TFO and UNB classes. ‘ssTFs and cofactors’ represents a combined set of target locations determined by ChExMix for those targets labelled as such in Supplementary Data 2 (1K), including ssTFs, SAGA, TUP and Mediator. c, Composite data show that STM promoters have NDRs, whereas TFO and UNB promoters have NFRs. In vitro nucleosomes assembled with purified genomic DNA and histones (black filled areas) had ATP plus either purified RSC (yellow) or INO80 (purple) added (data from ref. 19). Poly (TFα)-regions are sense-strand tracts (larger than 5 bp) of As (red) or Ts (green). d. Insulator ssTFs uncouple divergent transcription. Data on nascent transcription (CRAC data20) for control strain or strains depleted of Rapi or Reb1 by the anchor-away (AA) technique were collected for N divergent gene pairs sharing the same promoter region, then correlated between the gene pairs. To the right are diagrams of divergent gene pairs, with the differential size of each green arrow reflecting the extent of insulator e. The termination factor Pcf11 accumulates at insulator ssTFs. Shown is the architecture at promoters adjacent to an upstream termination region (tandem genes) and having (TFO) or lacking (UNB) an insulator ssTF.

elongation/termination-associated factors being restricted to specific sets of genes, except that Nrd1 of the early termination pathway was enriched at noncoding transcription (ncRNA) units (Extended Data Fig. 5a, lower left). In addition, RNA splicing factors (such as Smd1) were largely limited to the ‘third’ half of intronic genes encoding ribosomal proteins (RPs; Extended Data Fig. 5b, upper right). The data are consistent with one predominant elongation entourage at most Pol II genes that changes in composition at fixed distances from the TSS or transcription end site (TES) (rather than at a percentage of gene length).

Consistent with some other reports24-26, although not all27-29, we found no evidence for Mediator being stably associated with the Pol II core initiation or elongation entourage, despite its detection in upstream promoter regulatory regions (for example, Med2 in Extended Data Fig. 5b). Equivalence binding in gene bodies may be related to approximately 100 genes that produced relatively high and variable background in ChIP assays (see Methods).

The long terminal repeats (LTRs) of certain classes of Ty transposons are transcribed by Pol II as part of retroviral-like transposition30. However, most lacked a PIC, except a subset of full-length Ty1,2,6 (Extended Data Fig. 6). At Ty3(α), the Pol II pheromone factors Ste12, Dlg1 and Kar4 were assembled and had nearly identical points of crosslinking (Fig. 3d). However, instead of Pol II, we detected the Pol III machinery associated with adjacent divergent tRNA genes. This suggests that Pol II ssTFs may work with Pol III at some tRNA genes to integrate mating and Ty3 transposition31.

**Inducible versus constitutive promoters**

In classifying Pol II promoters, we opted against an unsupervised approach, as it treats binding events equivalently, without considering that certain targets have a more central role in defining specific regulatory architectures. Four fundamentally distinct architectural themes emerged (see Methods, Fig. 4a and Supplementary Data 1 (1D)): first, an RP theme, as seen for 137 RP promoters with unique architectures (examined separately32); second, an STM theme, as for 984 promoters that had properties associated with inducibility, and characteristically bound by ssTFs and major cofactor meta-assembles SAGA, TUP and/or Mediator/SWI–SNF; third, a TFO theme, from 1,783 promoters with a ssTF organization that lacked STM cofactors (but typically had the insulator ssTFs Abf1 or Reb1); and fourth, a UNB theme, as seen with 2,474 promoters that were unbound by anything except a PIC. Notably, as detailed in the Supplementary Information, the consensus architecture at TFO/UNB promoters indicates that two-thirds of all promoters evolved to lack regulation by ssTFs and their cofactors under any condition (not just in rich media). This is an architecture suitable for constitutively low gene expression. RP and STM represent the architecture of inducible promoters that have upstream activator sequences (UASs). The roughly 1,300 ncRNA promoters were similarly classified (Supplementary Data 1 (E)), indicating that they are governed by the same regulatory mechanisms.

Assembly of Pol II PICs occurs in the context of chromatin, where the TSS resides on the inside edge of a downstream +1 nucleosome (Fig. 4b). Most promoters have a constitutive NFR. The seemingly interchangeable term ‘NDR’—applying to nucleosome depletion mediated by ssTFs—is problematic. As ssTFs are absent from UNB promoters, they should lack ssTF-regulated nucleosome depletion and an NDR. We therefore considered whether NFRs and NDRs are distinct.

NFRs at TFO/UNB promoters were short (less than 150 bp) and bisected by a pair of oppositely stranded, nucleosome-disfavouring...
poly(dA:dT) tracts (Fig. 4c, red/green). NRs have been biochemically reconstituted on genomic DNA with purified histones and chromatin remodellers38. When applied to our promoter classes, we found that histones alone partially reconstituted NRs in vitro at TFO/UNB promoters, but less effectively at STM promoters (Fig. 4c, compare black-filled dips with in vivo plots, and Extended Data Fig. 7a). TFO/UNB NRs were widened by the RSC remodeler (Fig. 4c, compare black-filled wider dip with the black-filled areas) and had their −1/+1 nucleosomes widened by the RSC remodeler (Fig. 4c, compare black-filled dips with in vivo plots, and Extended Data Fig. 7a). TFO/UNB NRs were widened by the RSC remodeler (Fig. 4c, compare black-filled wider dip with the black-filled areas) and had their −1/+1 nucleosomes widened by the RSC remodeler (Fig. 4c, compare black-filled dips with in vivo plots, and Extended Data Fig. 7a).

**ssTF–cofactor interactions and circuits**

A comprehensive set of 78 ssTFs were detectably bound to promoters in rich media (Supplementary Data 2 (IK)). A search of the JASPAR database of interactions between ssTFs and sequence motifs independently confirmed proper motif specificity for 90% of the ssTFs (Supplementary Data 2 (IM)). Some ssTFs had robust ChiP–exo patterning around their cognate motif (Extended Data Fig. 8a; for example, Cup9 and Cin5), which reflects their site-specific structural interactions with DNA on a genomic scale. Remarkably, most ssTFs had relatively diffuse ChiP–exo patterning flanking their motif (Extended Data Fig. 8a; for example, Nrg1, Bas1 and Yrr1). As exemplified by Yrr1 in Fig. 4a (magenta versus cyan areas), the diffuse patterning of ssTFs was particularly pronounced at sites with multiple STM cofactors present (for example, SAGA, TUP, Mediator, SWI–SNF and RPD3-L), and less diffuse at other sites that bind the same ssTFs but lack STM cofactors. STM cofactors may impart a distinct local environment that results in more dispersed crosslinking. The same diffuse patterning occurred with STM cofactors which were anchored at ssTF sites (Fig. 5a and Extended Data Fig. 8b). As they tend to co-occupy the same set of promoters (Extended Data Fig. 9a, Supplementary Data 2 (IK)), ssTFs might coexist with multiple positive/negative cofactors of chromatin accessibility and Pol II recruitment. This diffuse patterning is consistent with the notion of condensates that are anchored by ssTFs.

In contrast to STM cofactors, we detected essentially no ChiP–exo patterning of TFID, TBPs or any GTFs at a consolidated set of ssTF sites, despite identifying these GTFs in the periphery where TSSs reside (Fig. 5b and Extended Data Fig. 9b). Thus, although using the same paradigm for detecting ssTF–cofactor interactions, our results in yeast do not support the long-standing model that ssTFs stably engage TFID at promoters. PIC assembly is driven by TFID at nearly all genes37,38, although at inducible genes it is augmented through SAGA independently of TFID35,37. Although the gene specificity of SAGA has been enigmatic and controversial31,33, the ChiP–exo assay detects SAGA at only a subset of genes. The discrepancy may reside in the low specificity of other assays34.

We addressed the specificity of SAGA further. As a direct readout of TFID–independent PIC assembly, we expected high levels of GTFs relative to TFID where SAGA is bound. However, we found that most SAGA-bound promoters (RP/STM/SAGA-bound) lacked high ratios of GTFs to TFID, although a smaller fraction did have high ratios (equivalent modes and rightward tail in Fig. 5c and Extended Data Fig. 9c). Thus, SAGA binding is not always concomitant with TFID–independent PIC assembly, and may reflect a poised state. Instead, promoters having multiple STM cofactors displayed high GTF/TFID ratios (STM-bound and ‘RSTM-bound’ in Fig. 5c). Thus, maximal TFID–independent PIC assembly is achieved under conditions in which there is maximal engagement of a wide variety of negative and positive ssTFs and cofactors with NDRs, including but not limited to SAGA.

Promoters bound by ssTFs included both cognate (motif-based) and noncognate interactions (Extended Data Fig. 10). In assessing cognate interactions, we found that most ssTFs bound to the promoters of ssTFs stably interact with STM cofactors but not GTFs. A, Architecture at Yrr1 motifs in two classes of Yrr1-bound promoters: ‘STM-bound’ (cyan and black labels on right) (Methods). The arrow points to where cofactor crosslinking permeates Yrr1 crosslinking. B, Representative architecture of STM cofactors or PIC components at a consolidated set of ssTF-binding motifs at RSTM promoters (strand averaged; see Methods and Supplementary Data 1 (I1A)), and oriented by TSS. Taf12 is in SAGA and TFID; bkd, background that was generated from a strain lacking a TAP tag. C, Frequency distribution of promoters having the indicated PIC/TFIID ratios (average of six GTFs; three-bin moving average), separated by promoter class (RP, STM, TFO or UNB) or promoter sets based on cofactor enrichment. ‘SAGA-bound’ excludes RP promoters, which are highly enriched with SAGA and shown separately. The ‘STM-bound’ promoter set required all of the following to be present: SAGA, Mediator/SWI/SNF and TUP; ‘RSTM-bound’ also required the presence of the RPD3-L complex. The x axis is in arbitrary units.

---

**Fig. 5** ssTFs stably interact with STM cofactors but not GTFs. A, Architecture at Yrr1 motifs in two classes of Yrr1-bound promoters: ‘STM-bound’ (cyan and black labels on right) (Methods). The arrow points to where cofactor crosslinking permeates Yrr1 crosslinking. B, Representative architecture of STM cofactors or PIC components at a consolidated set of ssTF-binding motifs at RSTM promoters (strand averaged; see Methods and Supplementary Data 1 (I1A)), and oriented by TSS. Taf12 is in SAGA and TFID; bkd, background that was generated from a strain lacking a TAP tag. C, Frequency distribution of promoters having the indicated PIC/TFIID ratios (average of six GTFs; three-bin moving average), separated by promoter class (RP, STM, TFO or UNB) or promoter sets based on cofactor enrichment. ‘SAGA-bound’ excludes RP promoters, which are highly enriched with SAGA and shown separately. The ‘STM-bound’ promoter set required all of the following to be present: SAGA, Mediator/SWI/SNF and TUP; ‘RSTM-bound’ also required the presence of the RPD3-L complex. The x axis is in arbitrary units.
motifs have been described. About half of all these ssTF-encoding genes lacked bound ssTFs (42 of 78 were UNBs), and thus are expected to be constitutive and at the start of their regulatory circuit. Of note, about half (43 of 78) of the ssTFs existed within a single highly integrated circuit, suggesting that the regulation of ssTFs is highly interconnected. Eleven ssTFs bound to multiple ssTF-encoding genes (multi-output archetype), suggesting that they have the potential to diversify their control through other such factors. Most ssTFs (47 of 78) bound only one other ssTF gene (single output), thereby propagating the circuit. There were long regulatory series with as many as seven ssTFs in series that bifurcated and/or looped (Extended Data Fig. 11a). About one-third of the ssTFs bound to their own promoter (in a simple loop), indicating that direct feedback control of these factors is common (an autoregulation archetype). Nine promoters of ssTF-encoding genes had multiple ssTFs bound (multi-input archetype; Extended Data Fig. 11b). In most cases, each bound ssTF was a member of a different meta-assemble (for example, RPD, SAGA, TUP and MED). Thus, numerous regulatory mechanisms/meta-assembles may converge at promoters through distinct ssTFs. One-quarter (21 of 78 ssTFs) bound to no other ssTF gene and thus are likely to be at the end of their circuit. 

Conclusions

Consistent with published studies, we have found that the vast majority of Pol II promoters share the same basic constitutive architecture. Local DNA sequence and chromatin remodelers create a constitutive NFR that is flanked by stable and well positioned nucleosomes. This is recognized by TFIIID and is configured for constitutively low gene expression. ssTFs and cofactors are generally required for transcriptional activation, except that some ssTFs (such as Abf1 and Reb1) organize nucleosomes and insulate against nearby genomic events. ssTFs and cofactors that directly regulate PIC assembly define roughly 20% of all genes, with an architecture that supports inducibility. Our data support a dynamic ‘futile cycle’ of nucleosome acetylation (by SAGA and NuA4) and deacetylation (by Rpd3-L), coupled to nucleosome eviction (mediated by SWI/SNF) and stabilization (by Tup1 and Cyc8), which produces an NDR. In this inducible environment, the assembly of a PIC is augmented beyond what TFIIID delivers. The stage is then set for enhanced recruitment of Pol II via ensembles of ssTFs and the Mediator complex. Much of this induced transcription may exist in hubs in which numerous induced promoters coalesce, perhaps for the purposes of efficiently recycling the transcription machinery. Once transcription has cleared the promoter, most genes appear to encounter the same Pol II ensemble, whose architecture changes at fixed distances from either the TSS or the TES.

This comprehensive high-resolution view of genomic chromatin architecture into our understanding of the post-initiation global regulatory control of constitutive genes, and raises questions as to how environmental signalling directly influences the control of ssTFs and cofactors. A clear view of epigenomic architecture should provide a better context for understanding how it integrates with other layers of gene regulation that occur during RNA processing, transport and translation. Moreover, as most of the key proteins examined here are evolutionarily conserved, their architectural themes are likely to exist in other eukaryotes.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03314-8.

1. Rossi, M. J., Lai, W. K. M. & Pugh, B. F. Simplified ChIP-exo assays. Nat. Commun. 9, 2842 (2018).
2. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147, 1408–1419 (2011).
3. Hahn, S. & Young, E. T. Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. Genetica 169, 705–736 (2011).
4. Levine, M., Cattoglio, C. & Tjian, R. Looping back to leap forward: transcription enters a new era. Cell 157, 13–25 (2014).
5. Verrijzer, C. P. Organization and regulation of gene transcription. Nature 573, 45–54 (2019).
6. Eaton, M. L., Galani, K., Kang, S., Bell, S. P. & MacAlpine, D. M. Conserved nucleosome positioning defines replication origins. Genes Dev. 24, 748–753 (2010).
7. Li, N. et al. Structure of the origin recognition complex bound to DNA replication origin. Nature 559, 217–222 (2018).
8. Wellingr, R. J. & Zakian, V. A. Everything you ever wanted to know about Saccharomyces cerevisiae telomeres: beginning to end. Genetica 191, 1073–1105 (2012).
9. Biggins, S. The composition, functions, and regulation of the budding yeast kinetochore. Genetics 194, 817–846 (2013).
10. Camahort, R. et al. Cse4 is part of an octameric nucleosome in budding yeast. Mol. Cell 35, 794–805 (2009).
11. Henikoff, S. et al. The budding yeast centromere DNA element II wraps a stable Cse4 hermin in either orientation in vivo. elife 3, e01861 (2014).
12. Rhee, H. S., Bataille, A. R., Zhang, L. & Pugh, B. F. Subnucleosomal structures and nucleosome asymmetry across a genome. Cell 159, 1377–1388 (2014).
13. Furuyama, S. & Biggins, S. Centromere identity is specified by a single centromeric protein in budding yeast. Proc. Natl Acad. Sci. USA 104, 10764–10769 (2007).
14. Yan, K. et al. Structure of the inner kinetochore CCAN complex assembled onto a centromeric nucleosome. Nature 574, 278–282 (2019).
15. Han, Y., Yan, C., Fishbain, S., Ivanov, I. & He, Y. Structural visualization of RNA polymerase II transcription machineries. Cell Discov. 4, 40 (2018).
16. Mayer, A. et al. Uniform transitions of the general RNA polymerase II transcription complex. Nat. Struct. Mol. Biol. 17, 1272–1278 (2010).
17. Petrenko, N., Jin, Y., Wong, K. H. & Struhl, K. Evidence that Mediator is essential for Pol II transcription, but is not a required component of the preinitiation complex in vivo. elife 6, e28447 (2017).
18. Jeronimo, C. et al. Tail and kinase modules differently regulate core Mediator recruitment and function in vivo. Mol. Cell 64, 455–466 (2016).
19. Androu, J. C. et al. Genome-wide location of the coactivator mediator: binding without activation and transient Cdk8 interaction on DNA. Mol. Cell 22, 179–192 (2006).
20. Paul, E., Zhu, Z. I., Landsman, D. & Morse, R. H. Genome-wide association of mediator and RNA polymerase II in wild-type and mediator mutant yeast. Mol. Cell. Biol. 35, 331–342 (2015).
21. Zhu, X. et al. Genome-wide occupancy profile of mediator and the Srb11-11 module reveals interactions with coding regions. Mol. Cell 22, 169–178 (2006).
22. Kranatsova, O., Hadzhidodorov, M. & Pesheva, M. Ty elements of the yeast Saccharomyces cerevisiae. Biotechnol. Biotechnol. Equ. 19, 19–26 (2005).
23. Reja, R., Vinayachandran, V., Ghosh, S. & Pugh, B. F. Molecular mechanisms of ribosomal protein gene coregulation. Genes Dev. 29, 1942–1954 (2015).
24. Krettenstein, N. et al. Genomic nucleosome organization reconstituted with pure proteins. Cell 167, 709–721 (2016).
25. Chereji, R. V., Ocampo, J. & Clark, D. J. Nras-sensitive complexes in yeast: nucleosomes and non-histone barriers. Mol. Cell 65, 565–577 (2017).
26. Candelli, T. et al. High-resolution transcription maps reveal the widespread impact of roadblock termination in yeast. EMBO J. 37, e90490 (2018).
27. Brzovic, P. S. et al. The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a simple protein interface forming a fuzzy complex. Mol. Cell 44, 942–953 (2011).
28. Huisenga, K. L. & Pugh, B. F. A genome-wide housekeeping role for TFIIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. Mol. Cell. 13, 573–585 (2004).
29. Dudley, A. M., Rougeulle, C. & Winston, F. The Spt components of SAGA facilitate TBP binding to a promoter at a post-activator-binding step in vivo. Genes Dev. 13, 2940–2945 (1999).
30. Moqtaderi, Z., Bai, Y., Poon, D., Weil, P. A. & Struhl, K. TBP-associated factors are not generally required for transcriptional activation in yeast. Nature 383, 188–191 (1996).
31. Baptista, T. A. et al. SAGA is a general cofactor for RNA polymerase II transcription. Mol. Cell 68, 130–143 (2017).
32. Mittal, C., Rossi, M. J. & Pugh, B. F. High similarity among ChEC-seq datasets. Preprint at https://www.biorxiv.org/content/10.1101/2021.02.04.429774v1 (2021).
33. Verrijzer, C. P. et al. Transcriptional regulatory code of a eukaryotic genome. Nature 431, 99–104 (2004).
34. Boja, A. et al. Transcription factors activate genes through the phase-separation capacity of their activation domains. Cell 175, 1842–1855 (2018).
35. Badjuta, N. et al. Acute stress drives global repression through two independent RNA polymerase II stalling events in Saccharomyces. Cell Rep. 34, 108640 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2021
Strains and antibodies

The vast majority of data for this study were collected from tandem affinity purification (TAP)-tagged S. cerevisiae strains (originally purchased from Dharmacon; now available from Horizon Inspired Cell Solutions, Cambridge, UK). The background strain for this collection was BY4741 (a derivative of S288-C; Matα his3Δ1 leu2Δ0 met15Δ0 ura3Δ0). Negative control ChIPs and ChIPs with specific antibodies were performed with BY4741. If the TAP-tagged strain for a particular target was unavailable, we instead used a haemagglutinin (HA)-tagged strain (originally purchased from Dharmacon; now available from Horizon Inspired Cell Solutions). The background strain for the HA-tagged collection was diploid, derived from BY4741 and designated Y8000 (Matα leu2-31Δ1 his3Δ1 Δ2000 ρ2-3 osc1-2 trpl-1 TRPI1 cit0 carrying pGAL-CRE (amp, ori, CEN, LEU2)).

Rabbit IgG (Sigma, catalogue number IS006, various lot numbers) conjugated to Dynabeads was used to immunoprecipitate chromatin from TAP-tagged strains. Santa Cruz Biotechnology sc-7392 antibody was used to immunoprecipitate chromatin from HA-tagged strains.

Millipore antibodies 04-1570-1, 04-1571-1 or 04-1572-1 were used to immunoprecipitate Pol II having its carboxy-terminal domain phosphorylated at positions serine 7, 2 or 5, respectively, of the heptad repeats. Millipore antibody 07-322 was used to immunoprecipitate histone H3 with acetylated lysine 9 (H3K9ac). Cell Signaling antibody S546S was used to immunoprecipitate histone H2B with ubiquitinated lysine 123 (H2BS123ub). Cse4 antibody from C. Wu (Johns Hopkins Univ., Baltimore, MD) was used to immunoprecipitate Cse4. Heat shock factor 1 (Hsf1) antibody from D. Gross (Louisiana State Univ., Baton Rouge, LA) was used to immunoprecipitate Hsf1. ChIP-seq experiments using micrococcal nuclease (MNase) to identify nucleosomes were performed for the following histones and histone modifications: H3 (detected using Abcam antibody ab1971), H3K27ac (ab4729), H3K36me3 (ab9050), H3K4me3 (ab8580), H3K79me3 (ab2621), H3K12ac (ab46983) and H2B (Active Motif 39237).

Cell growth and ChIP–exo

S. cerevisiae strains were grown in 67 ml of yeast peptone dextrose (YPD) media to an optical density at 600 nm (OD600) of 0.8 at 25 °C. Cells were crosslinked with formaldehyde at a final concentration of 1% for 15 min at 25 °C, and quenched with a final concentration of 125 mM glycine for 5 min. Cells were collected by centrifugation, and washed in 1 ml of ST buffer at 4 °C. The cells were pelleted again, the supernatant was removed, and the pellet was flash frozen.

As STM classification criteria included promoters that became bound by SAGA upon acute heat shock (as described (36)), we carried out equivalent heat-shock experiments but using the workflow of this study. We used these new data to assign heat-shock-induced binding locations of SAGA (which correlated highly with binding locations in ref. (36)). For these heat-shock samples, yeast was grown in 67 ml of YPD to an OD600 of 0.8 at 25 °C; an equal volume of YPD medium at 55 °C was added to raise the temperature of the culture to 37 °C and incubated at 37 °C for 6 min. Then, cells were crosslinked with formaldehyde at a final concentration of 1% for 15 min at room temperature by adding a 50 ml solution of ice-cold 3.7% formaldehyde in water. Note that protein–DNA crosslinks occur rapidly. Crosslinking was quenched with a final concentration of 125 mM glycine for 5 min. Cells were collected by centrifugation, and washed in 1 ml of ST buffer at 4 °C. The cells were pelleted again, the supernatant was removed, and the pellet was flash frozen.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment. Chromatin preparations are based on modifications of a prior protocol (37). Frozen cell pellets were resuspended and lysed in 1 ml of FA lysis buffer (50 mM Hepes-KOH, pH 7.5, 150 mM NaCl, 2 mM EDTA, 1% Triton, 0.1% sodium deoxycholate and complete protease inhibitor (CPI)) and a 500 μl volume of 0.5 mM zirconia/silica beads by bead beating in a Mini-Beadbeater-96 machine (Biospec) for three cycles each of three minutes on/seven minutes off (samples were kept in a -20 °C freezer during the off cycle). The lysates were transferred to a new tube and microcentrifuged at maximum speed for 3 min at 4 °C to pellet the chromatin. The supernatants were discarded; the pellets were resuspended in 600 μl of FA lysis buffer and transferred to 15 ml polystyrene conical tubes containing 300 μl of 0.1 mm zirconia/silica beads. The samples were then sonicated in a Bioruptor Pico (Diagenode) for 8 cycles (15 s on/30 s off) to obtain DNA fragments of 100–500 bp in size. Each ChIP–exo assay processed the equivalent of 33 ml of cell culture (roughly 8 × 10⁸ cells). The remaining half of the processed chromatin was flash frozen and stored at –80 °C in case a technical replicate was desired.

A culture equivalent of 33 ml (roughly 630 million cells) of yeast was fragmented to produce solubilized chromatin (roughly 190 μl). This was incubated overnight (roughly 16 h) at 4 °C with the appropriate antibody. A 10 μl bed volume of conjugated IgG–Dynabeads (0.83 mg ml⁻¹ IgG and 5 mg ml⁻¹ Dynabeads) or 3 μg of specific antibodies with a 10 μl slurry-equivalent of Protein A Mag Sepharose (GE Healthcare) was used in each reaction.

ChIP–exo 5.0 was performed as described (38). Essentially, ChIP libraries were partially constructed on the immunoprecipitated resin, and then λ exonuclease was used to trim nucleotides in the 5’ to 3’ direction until stopped by a protein–DNA crosslink. The DNA was then eluted and library construction completed.

In a typical experiment with TAP-tagged yeast strains, 48 ChIP–exo experiments were performed concurrently. Each set included 46 unique targets, a Rebl–TAP sample as a positive control, and a BY4741 sample (from a parental strain lacking the TAP tag) as a negative control. Following 18 cycles of polymerase chain reaction (PCR), all 48 samples were pooled equally by volume. Library concentration was quantified by quantitative PCR (qPCR). Equivalent workflows occurred with other strains.

Using paired-end Illumina sequencing and cellular conditions identical to those used to produce ChIP–exo data, we generated a genome-wide nucleosome map (MNase histone H3 and H2B ChIP–seq) with improved accuracy over our prior maps. MNase ChIP–seq was performed as described (38). Briefly, formaldehyde-crosslinked chromatin was digested with MNase to achieve roughly 80% of mononucleosomes. After H3 or H2B ChIP and library construction, libraries were size selected by agarose gel electrophoresis, and sequenced.

Sequencing and mapping

High-throughput DNA sequencing was performed with an Illumina NextSeq 500 or 550 in paired-end mode, producing a 40 bp Read_1 and a 36 bp Read_2. Additional previously published ChIP–exo datasets (33, 34) for Hsf1, Msn2, Spt15, Spt16, Ifh1 and Fhl1 were included in data processing and analysis for our study. Data were managed, quality controlled, and processed through a custom automated workflow control called PEGRE (Platform for Epi-Genomic Research) (38). Sequence reads were aligned to the yeast (sacCer3) genome using bwa-mem (version 0.7.17). Aligned reads were filtered using Picard (version 2.7.1) and samtools (version 0.1.18) to remove PCR duplicates (that is, where the 5’ coordinates of strand Read_1 and Read_2 were identical to another read pair) and non-uniquely mapping reads. For ChIP–exo, the resulting mapped 5’ end of Read_1 (the exonuclease stop site) is defined as a tag. For MNase, the resulting mapped midpoint of Read_1 and Read_2 is defined as a tag.

Data quality, statistics and reproducibility

We tested many targets that were not expected to bind directly to DNA, and thus could not assume that every target would produce a

---

**Note:** The above text is a partial transcription of the document and may not capture all the details or context due to its formatting and style. It is intended to provide a readable version of the document content for further analysis or use. Detailed computational analysis and interpretation would require access to the full, formatted document.
positive ChIP signal. We empirically determined that a minimum of 200,000 deduplicated tags were required to assess the quality of an individual dataset. If a dataset received less than 200,000 tags, then we required the tag duplication level (number of reads discarded by PICARD)/(number of input reads) of the sample to be less than 70% before we sequenced it more deeply. For example, if a dataset had 100,000 mappable deduplicated tags (unique Read_1 and Read_2 combination), but a total of 1 million mappable tags before filtering, then the duplication level was 90% and it was assumed that the library was insufficiently complex to warrant additional sequencing. If a library was insufficiently complex, we performed a technical replicate with the remainder of the chromatin preparation. Following this procedure, we produced a sufficiently complex library for more than 95% of targets tested from a single yeast culture. In practice, pooling equivalent proportions of 48 barcoded libraries (in terms of reaction volumes) provided similar sequencing depth across all samples. All analysed dataset were confirmed with independent biological replicates that passed our quality-control metrics. A dataset was considered successful if significant locations (binomial, 1.5-fold, \( P < 0.01 \)) were identified by ChExMix (see below) and these locations were not in regions that produce highly variable data. \( N \) values are reported for the number of target datasets (hierarchical clustering and UMAP) or the number of genomic features (composite plots and heatmaps) analysed.

Raw FASTQ reads for each sample were aligned against the known TAP or HA FASTA sequence and nearby genomic sequence to confirm the presence and location of the epitope in each strain. See 03_EpitopeID at https://github.com/CEGRCODE/2021-Rossi_Nature.

Mapping statistics for each dataset are available at yeastepigenome.org, along with mapped data downloads. Analyses shown at yeastepigenome.org can be reproduced or further custom analysed using ScriptManager (https://github.com/CEGRCODE/scriptmanager), which provides a simple user-friendly interface. It includes straightforward instructions for installation and for data analysis. Data values from the paper's composite plots can be found in 01_Composite_Files at https://github.com/CEGRCODE/2021-Rossi_Nature.

ChExMix locations

ChExMix\(^{41}\) version 0.31 was run with the following non-default parameters: --noread2 --scawlin 1000 --minmodelupdateevents 50 --fixedalpha 0 --mememinw 8 --mememaxw 21 --minmodelupdaterefs 25 --noread2 --scalewin 1000 --minmodelupdateevents 50 --fixedalpha 0 --mememinw 8 --mememaxw 21 --minmodelupdaterefs 25 --lenientplus. We also used the --excludebed option to exclude from analysis a custom set of hypervariable regions (ChExMix_Peak_Filter_List_190612.bed), including the rDNA locus, tRNA genes and telomere regions (this list is available in 02_References_and_Features_Files at https://github.com/CEGRCODE/2021-Rossi_Nature). The negative control for ChExMix peak calling, termed ‘masterNoTag_20180928’, was created by merging 15 individual BY4741 (parent strain containing no epitope tag) ChIP–exo experiments into a single BAM file. These negative controls were generated over an 18-month period during the main phase of data collection. The file ‘masterNoTag_20180928.bam’ comprises the following SampleIDs: 11851, 11946, 12094, 12860, 13484, 13822, 14202, 14408, 14637, 14825, 15256, 15818, 16073, 17814 and 18504, and is available at https://doi.org/10.26208/rykf-6050.

Meta-assemblages

Meta-assemblages are based on cell populations. Thus, their member targets tend to bind the same genomic locations, although not necessarily at the same time or above a preset algorithmic threshold. Owing to parameter constraints placed on clustering, significant \( (P < 0.01 \) but rare (for example, HIR) and/or highly isolated (for example, Vid22/Tbf1) binding events tended to cluster near each other in UMAP, and so were placed in a single miscellaneous meta-assemblage (ISO) without further analysis.

Using bedtools intersect (bedtools version 2.27.1), all ChExMix peaks (regardless of whether they were associated with the Pol II sector, defined above) for each of 384 validated input targets were intersected in a 100-bp window around themselves. This produced a symmetrical matrix of counts representing the frequency of peak overlap between all samples. 2D hierarchical clustering\(^{41}\) was then performed, using average linkage and uncentred correlation as the metric.

The interaction matrix was further filtered to remove 13 targets with fewer than five total ChExMix peaks (for example, Pol II targets having only two binding locations that are annotated in the reference yeast genome, despite the rDNA locus being highly repetitive). This produced a symmetrical matrix of 371 samples (Fig. 1b and Supplementary Data 3). The matrix was then used as the input into the UMAP algorithm (version 0.3.7)\(^{47}\) using the following parameters: umap.UMAP(n_neighbors = 5, min_dist = 0.0, n_components = 2, metric = ‘correlation’, random_state = ‘RS’).fit_transform(X). K-means clustering was performed on the resulting 2D projection at a variety of \( K \) values (5, 10, 20, 25, 30, 35, 40, 100, 145). No new biologically distinct clusters appeared beyond \( K = 40 \).

Reference features and intervals

Coordinates for 253 replication origins (ACS sequences, for ‘autonomously replicating sequence (ARS) consensus sequences’) were obtained from ref. \(^{48}\). Note that ACS\_6\_32973 has a duplicate entry on the yeastepigenome.org website, resulting in 254 features. Coordinates for X-core elements (XCEs), centromeres (CENs), RNA polymerase I (PolI) TSS, Pol III TSS, NCR (SGD-defined noncoding RNA annotated as ncRNA_gene, snRNA_gene, and snRNA_gene) and Ty transposon LTRs were obtained from the Saccharomyces Genome Database (SGD: https://www.yeastgenome.org) on 3 March 2017 (available as SGD_features_03-170331.tab in 02_References_and_Features_Files at https://github.com/CEGRCODE/2021-Rossi_Nature). TSS and TES coordinates for Pol II were obtained from ref. \(^{49}\). They were matched to each SGD coding feature through their systematic GenEd. These coordinates were based on microarrays. For TSS, the most 3′-enriched sense-strand coordinate in each promoter is reported. When no transcript was reported for an SGD feature, the TSS and TES were imputed from the SGD coordinates by moving 70 bp upstream of the start ATG (SGD start) for TSSs and 70 bp downstream of the stop codon (SGD end) for TESs. This imputation was based on the empirical observation that the median distance from the TSS defined in ref. \(^{41}\) and the start codon was 70 bp. 'Dubious ORFs' were initially considered and then excluded from further analysis.
because we and others found no validating evidence. Noncoding RNAs (ncRNAs) were from SGD annotations; cryptic unstable transcripts (CUTs) and stable unannotated transcripts (SUTs) were from ref. and Xrn1-sensitive unstable transcripts (XUTs) were from ref. Reference datasets are available in 02_References_and_Features Files at https://github.com/CEGRcode/2021-Rossi_Nature: SGD features (SGD_features_170331.tab), ORF TSS (Xu_2009.ORF-Ts_V64.gff3), CUT (Xu_2009_CUTs_V64.gff3), SUT (Xu_2009_SUTs_V64.gff3), and XUT (van Dijk_2011_XUTs_V64.gff3).

Nucleosome maps at Pol II promoter regions

MNase H3 and H2B ChIP–seq paired-end reads were bioinformatically filtered to fragment sizes of 100–160 bp, and then nucleosome dyads (peaks) were called from the mapped midpoint location of Read 1 and Read 2 5′ ends using GeneTrack (v1) (parameters: s40e80F1). Peaks were required to overlap within a 75-bp window in at least 4 of 6 datasets (H2B and H3 MNase ChIP–seq: SampleIDs 10951, 10952, 10967; 10947, 10948 and 10966) to call a consensus nucleosome (N = 6). The average location of overlapping peaks defined the dyad coordinate of a consensus nucleosome.

The +1 nucleosome was defined as the nucleosome dyad peak that was closest to a TSS in a window −60 bp to +140 bp. If no nucleosome was found, then an additional search was performed −80 bp to −61 bp relative to the TSS. If none was found, then the region was viewed in Integrated Genome Viewer version 2.5.2 (IGV; http://software.broadinstitute.org/software/igv/) and manually assigned. If no nucleosomes could visually be assigned to a TSS in IGV, then a +1 nucleosome dyad coordinate was imputed as the SGD ATG start coordinate (which is the consensus location of +1 nucleosomes). This placed the TSS at the genome-wide canonical location relative to the imputed +1 dyad.

We previously defined consensus –1 nucleosome positions of all genes transcribed by Pol II, regardless of whether a nucleosome had low occupancy or was even detectable. However, here our intent was to define the region encompassing NRs and NDRs, and so we chose to ignore nucleosome positions that were highly depleted of nucleosomes. Our goal was to manually determine the location of the most robust algorithmic nucleosome position (upstream stable nucleosome, USN) that was located closest to a TSS and in a window −500 bp to −60 bp from the TSS, as long as that nucleosome was not already called a +1 nucleosome. If one of the following criteria was met, then the nucleosome landscape was visualized in IGV, and the USN and/or +1 nucleosomes were manually (re)assigned (N = 753): 1) either the USN or a +1 was not present in the original algorithmically defined set; 2) the USN-to(+1) dyad-to-dyad distance was calculated to be smaller than 187 bp (the size of a nucleosome (147 bp) and two linkers (2 × 20 bp)); 3) a ssTF peak was, first, located less than 600 bp upstream of the TSS, and second, upstream (more 5′ to the nearest TSS) of a nucleosome call having an occupancy score that was in the bottom 5% of all nucleosomes (that is, an algorithmically called nucleosome that was in fact highly depleted in the vicinity of a ssTF). If no nucleosomes could visually be assigned, the USN nucleosome coordinate was imputed as 750 bp upstream of the +1 nucleosome dyad (99th percentile of calculated NDR/NFR lengths). The NDR/NFR length at these features was reported as ‘9999’ in Supplementary Data 1 (1S) (N = 297). As the promoter regions defined in this study include arbitrary limits and do not consider limits defined by insulation, there will be some inaccuracies in relation to actual biological promoter boundaries. This is expected to result in some promoter misclassifications.

In total, 59,002 nucleosomes were called across the S. cerevisiae genome. Nucleosome occupancy and fuzziness scores were calculated as described. All nucleosome calls with their median occupancy and fuzziness scores are available as Nucleosome_calls_and_stats.xlsx in 02_References_and_Features Files at https://github.com/CEGRcode/2021-Rossi_Nature.

ChExMix locations at filtered Pol II genes

The initial list of all compiled features totalled 11,112 (Supplementary Data 1). Numerous quality-control metrics were calculated for each Pol II transcribed feature to assess their validity and mappability. We used two GTFs (Sua7 (SampleID = 11743) and Ssl2 (11747)) and a negative control (masterNoTag_20180928.bam), with total tags set to be equal across all three in order to assess the enrichment around each candidate coding and noncoding Pol II TSS (N = 9,844; feature class level: 01–12, 14, 24 and 25 in Supplementary Data 1 (1D)), as described below.

A region of the genome was defined for each transcribed feature that included the transcribed sequence (TSS to TES) and the surrounding regulatory region. The upstream (promoter) regulatory region was defined as the inclusive interval between the dyad coordinate of the USN (see above) and the TSS. When no USN was called for a feature, then the upstream boundary was defined as 750 bp upstream (5') of the TSS. Note that the upstream boundary does not consider boundaries defined by insulators, as they have not yet been fully defined. This may result in unwarranted attachment of ssTF/cofactor locations to some promoters. The downstream regulatory region was defined as the inclusive interval from TES to 100 bp downstream (3'). This boundary was based on the consensus position of the termination machinery relative to the TES. The genomic region from the USN dyad to 100 bp downstream of TES was defined as a 'Pol II sector'.

ChExMix peaks for all datasets here were intersected with each Pol II sector using Bedtools. A protein was defined to be located within a feature if at least one ChExMix peak overlapped with any portion of the sector. If a ChExMix peak intersected two overlapping sectors (that is, the peak exists in the promoter region of two genes in a head-to-head orientation), then that protein was located in both sectors. Consequently, the number of ChExMix peaks and the number of bound features (or sectors) is not equal.

Pol II sectors were excluded as 'hypervariable' if any of the following conditions were met: 1) the TSS was in the highest 1% of masterNoTag_20180928 tag counts (negative control) in a 1,000-bp window centred over the TSS; 2) the TSS was in the highest 5% of masterNoTag_20180928 tag counts in a 200-bp window centred over the TSS and the occupancy ratios of both Sua7/NoTag and Ssl2/NoTag were less than 2 (based on total tag normalization). The rationale for these criteria was that if the signal in the negative control was too high, and the signal-to-noise ratios of robust GTFs such as Sua7 and Ssl2 were not well above the high background, then we did not have confidence in locations called at these sites. The sector was retained if it overlapped with a peak call from any dataset in this study. We assumed that the peak indicated enough dynamic range to have useable data in this region. We excluded N = 75 Pol II sectors by this metric ('08_Hyper-variable' in Supplementary Data 1 (1D)).

Pol II sectors were excluded for having 'poor mappability' if any of the following conditions were met: 1) the TSS was in the lowest 1% of MasterNoTag_20180928 tag counts in a 1,000-bp window centred over the TSS; 2) the TSS was in the lowest 5% of masterNoTag_20180928 tag counts in a 200-bp window centred over the TSS and the occupancy ratios of both Sua7/NoTag and Ssl2/NoTag were less than 2 (based on total tag normalization). Visual inspection of heatmaps confirmed that these segments of the genome were not uniquely mappable, and thus had low intrinsic tag counts. We excluded N = 116 Pol II sectors by this metric ('24_Hyper-variable_noncoding' in Supplementary Data 1 (1D)).

Pol II sectors were excluded as 'Quiescent-NoPIC' if the occupancy ratios of both Sua7/NoTag and Ssl2/NoTag were less than 1. The sector was retained if it overlapped with a peak call from any dataset in this study. The rationale here was that if there were no peaks in the sector vicinity and no enrichment of GTFs, then this feature was relatively quiescent. Thus, it was uninformative to analyse it further. We do not exclude the possibility that these features had low subthreshold
activity. We excluded $N = 251$ Pol II sectors by this metric ('05_NoPIC' in Supplementary Data 1 (ID)).

Pol II sectors were excluded as ‘tRNA proximal’ if peaks from Tfc3 (J1835) – a component of the RNA polymerase III transcription initiation factor complex – overlapped with the region between the +1 nucleosome dyad and the USN dyad of the sector. tRNA genes produced high levels of background owing to strong crosslinking of the Pol III machinery, which digestion by a exonuclease then focuses into high background peaks. Although this background is present in all samples, it is most problematic or evident where the target foreground signal is close to background. We excluded $N = 135$ Pol II sectors by this metric: ('06_tRNAprox' in Supplementary Data 1 (ID)).

Pol II sectors were excluded as ‘ChExMix extreme’ if they overlapped with an unusually high number of peaks. These features contained many gene-body peaks for targets that, across the rest of the genome, were bound primarily in promoter regions. Further analysis revealed that the density of tags across the gene body in the masterNoTag_20180928 negative control was abnormally high or low, relative to the rest of the genome, thereby creating statistical anomalies of bound locations. Consequently, ChExMix produced many false-positive peak calls in unrelated datasets at these extreme regions where the background model appears to break down. The peak calls at these extreme features are still included in the ChExMix peak files but should not be considered valid locations unless validated by orthogonal methods. The number of Pol II sectors given this label was empirically capped at $N = 25$ ('07_ChExMix_extreme' in Supplementary Data 1 (ID)). The value of this filter is that it decreased the number of potentially artefactual locations occurring in noncanonical places, particularly for ssTFs that bind to few genes. However, we do not exclude the possibility of noncanonical extreme, yet still biological, behaviour occurring at these genes. For example, large condensates might behave in this way.

Our analysis of the ncRNA features reported in refs. 16,47 found that many of these calls were not supported by evidence of GTF binding (Sua7) in the TSS vicinity, suggesting that many were false positives. Noncoding Pol II sectors were excluded if no Sua7 peak was found within 80 bp of the TSS. We excluded $N = 2,161$ ncRNA Pol II sectors by this metric ('25_excluded_ncRNA' in Supplementary Data 1 (ID)).

**Pol II promoter classes**

Our unsupervised approach to chromatin organization genome-wide produced meta-assemblages that reflect predominant architectural themes. Meta-assemblages are computed ensembles of many genome-wide locations averaged across millions of cells, and thus do not necessarily correspond to biochemically stable complexes. There are cases in which a meta-assemble such as ORC would appear to have a corresponding biochemical ensemble at replication origins. This makes meta-assemblages and real ensembles seemingly the same. However, as expected, there was no single promoter architecture that emerged from our unsupervised approach. Instead, meta-assembles reflected predominant architectural themes that ranged along a compositional spectrum from relatively heterogeneous (ssTFs/MED/SAGA/TUP) to relatively homogeneous (PIC). Meta-assemblages could be merged or subdivided to achieve levels of granularity, but also levels of uncertainty. They permeated promoters to varying extents.

The variation in the types of meta-assemblies within and across promoter classes gives them their unique regulatory properties, but also makes promoter classification fluid. Classification depends on input criteria that reflect subjective concepts. For example, prior work created SAGA-dominated and TFIIID-dominated gene groups on the basis of functional criteria (relative sensitivity to SAGA and TFIIID mutants)23. This helped to produce a genome-wide concept of inducible versus constitutive genes, but could not address other concepts such as insulation, or the fact that some themes may not be manifested through SAGA and TFIIID, or that there may be more granularity in each of those classes. We attempt here to provide more granularity, but recognize that simplifying overarching concepts are best served with fewer groups. To this end, we created promoter classes that arose in part from our unsupervised learning approach. However, we also injected additional a priori knowledge. This knowledge considers the functionality of each factor that contributes to distinctive regulatory archetypes.

The 137 RP promoters (defined by SGD) encode subunits of the ribosome. They comprise the largest known set of genes that are thought to be coregulated under all conditions. This may be due to the fact that they are predominantly regulated by the ssTF Rap1. They are highly expressed and well studied by ChIP-exo as a group32, and so form a distinct gene set.

SAGA, Mediator and Tup1 (‘STM’) are major cofactor complexes that, along with other ssTFs and cofactors (listed in Supplementary Data 2 (1K)), co-occur at highly expressed genes and formed major UMAP clusters. We therefore defined a set of non-RT-SP promoters (using the Bedtools intersect) if the region between the +1 nucleosome and USN dyads had at least one SAGA, Mediator or TUP ChExMix call (Supplementary Data 2 (10A)) if in YPD at 25°C or a SAGA call upon acute heat shock32 (6 min at 37°C) ($N = 984$ in the STM group; see Supplementary Data 1 (1E)). Most STM promoter regions ($N = 854$, or 87%) also bound at least one of 80 ssTFs site-specifically (Supplementary Data 2 (10C)). The majority of these ssTF peaks overlapped positionally with STM cofactor peaks. Applicable to Fig. 5b, we labelled each ssTF-bound motif as a ‘consolidated ssTF motif’ if it overlapped with a STM peak. This consolidated motif set was considered the organizing centre of that promoter. When a ssTF motif was absent, the ssTF peak call was used in instead. When numerous ssTFs were bound to the same promoter, the ssTF closest to the STM peak was used (Supplementary Data 1 (1Y–1AI)).

Of the remaining promoters (non-RP, non-STM), a subset had ssTF ChExMix peaks (whether site-specifically bound or not) or other cofactor ChExMix peaks in the region between the +1 nucleosome and the USN. This list of ssTFs and cofactors did not include the core transcription machinery (initiation, elongation or termination), which nevertheless were present. We therefore defined these as ‘TFO’ ($N = 1,783$). About one-quarter of TFO promoters had a bound ssTF that was more associated with STM promoters, and thus presumably capable of recruiting cofactors (Supplementary Data 2 (8)). These TFO promoters may have been algorithmically misclassified, perhaps being expressed under other environmental conditions. Those non-RP, non-STM and non-TFO promoters that remained constituted 2,474 promoters whose promoter regions lacked evidence of a binding event beyond a PIC or a nucleosome, and thus formed the largest of all groups, the ‘unbound’ (‘UNB’). These classifications are indicated in Fig. 1a, along with their relationship to the TFIIIDomin and SAGA_snn gene classes. Relative PIC occupancy (green-dot count) is based on average TFIIIB (Sua7) occupancy (Supplementary Data 1 (1AJ)) but confirmed with nascent and steady-state transcription.

**Stringent Pol II promoter classes**

These classifications were more stringent than those above and relate to Fig. 5b, c, and Extended Data Fig. 9b, c. The ‘SAGA-bound’ classification required a promoter to have a ChExMix peak call (‘1’ in Supplementary Data 2 (3)) for two or more of the following targets: Sp7, Ada2, Sgf11 or Sgf73. The ‘STM-bound’ classification required a promoter to have all three of the following labels: SAGA-bound, TUP-bound and Mediator/SWI–SNF-bound, as follows. The ‘TUP-bound’ classification required a promoter to have a ChExMix call (‘1’) for two or more of the following targets: Tup1, Cyc8, Sok2 and Cio5. The ‘Mediator/SWI–SNF-bound’ classification required a promoter to have a ChExMix call (‘1’) for two or more of the following targets: Swi1, Med2, Snf6 and Swi3. The ‘RSTM-bound’ classification required a promoter to have both of the following labels: STM-bound and RPD-bound. The RPD-bound classification required a promoter to have a ChExMix call (‘1’) for two or more of the following targets: Rpd3, Rxt1/Cit6, Rxt2, Rxt3, Nrm1 and Ume6.
Heatmaps and composite plots

Analysis was performed using the GUI ScriptManager version 012, which is available for download at https://github.com/CEGRcode/scriptmanager. ScriptManager provides a simple user-friendly interface for ChIP–exo analysis, and includes simple installation instructions. Heatmaps and composite plots were generated using Tag Pileup script. For ChIP–exo data, the following settings were used: Read_1 5′ end; separate strands; 0 bp tag shift; 1 bp bin size, sliding window (moving average) 11. For MNase ChIP–seq data the following settings were used: (paired-end) read midpoint; combined strands, 0 bp tag shift, 1 bp bin size, sliding window 21. All data are oriented by TSS or reference point strand.

For graphical display of composite plots, output data consisted of frequency counts of Read_1 5′ ends for ChIP–exo or Read_1/Read_2 midpoint for MNase H3/H2B ChIP–seq dyads (BAM files) that were at x-axis base-pair distances from sets of genomic reference points (BED files). Underlying patterns and datapoints are available at yeastepigenome.org and as Excel_Composite_Data_Processed.xlsx in 01_Composite_Files at https://github.com/CEGRcode/2021_Rossi_Nature. An additional moving average of 20 bp (30 bp for Pol II elongation and Yrr1 composites) was performed for the purpose of improving visual clarity. Without this, the high-bp resolution of ChIP–exo resulted in peaks that were quite narrow in the 1-kb visualization window, such that their fill patterns were less visually obvious. For gene-body targets (Fig. 3c and Extended Data Fig. 5), smoothed strand-separated data were shifted 50 bp in the 3′ direction before combining strands. The rationale for this is that when we examined each strand separately, we noticed that patterns on the transcribed strand showed some mirroring on the nontranscribed strand. But this pattern was shifted in the 3′ direction relative to transcribed strand (that is, down more of the TSS). We surmise that this ‘double-vision’ effect was caused by efficient crosslinking such that the 5′–3′ exonuclease is generally stopped at the backend of the Pol II entourage on the transcribed strand and stopped at the front-end of the entourage on the nontranscribed strand. Shifting data on both strands by 50 bp in their respective 3′ directions partially corrected this double vision and reflects the middle of the complex.

In the absence of a strand-specific 3′ shift for gene-body targets, patterns near the TSS reflect the backbone of the Pol II entourage, and patterns near the TES represents its front end. The data in Fig. 3b and Extended Data Fig. 9b were not strand-shifted before removing strand information.

In composite plots, the y-axis is labelled ‘Occupancy (a.u.)’ (arbitrary units), reflecting y-axis scaling that was adjusted to highlight the pattern of the data. Within a single figure (including any Extended Data figure counterparts), occupancy levels can be compared across multiple panels only for the same dataset. Occupancy levels of different datasets in the same or different panels cannot be compared directly. Only the peak positions are comparable. For Fig. 2, the MEME motif obtained and shown for Orc6 starts at position 2 of the ACS. For Cbf1, the MEME motif starts at position 1 of CEN. Schematics reflect subjective interpretations of peak locations, are nonlinear with respect to the diagrammed DNA linearity, and do not reflect protein molecular weights.

Circuitry involving ssTFs

This analysis relates to Extended Data Fig. 10. We analysed the set of 78 genes encoding ssTFs (defined in YPD) along with the ssTFs that bound their promoter regions site specifically (Supplementary Data 2 (1K)). A circuit-like diagram was then constructed by connecting ssTFs to the ssTF-binding genes in YPD at 25°C (that is, they had bound locations, but were not bound site-specifically). Their site specificity could be condition specific. For example, Nrg1 and Nrg2 bind the same motif, although JASPAR assigns this motif to Nrg1. We labelled both as ssTFs. Another equivalent example involves Met4, Met31 and Met32. Both Yox1 and Mcm1 have distinct motifs reported in JASPAR, and both biochemically interact. However, ChIP–exo reported the Mcm1 motif for both, with Mcm1 being much stronger. We therefore classified Yox1 as a cofactor in YPD at 25°C, instead of a ssTF. Eight targets had GO annotations indicative of a ssTF and yielded robust motifs by ChIP–exo with a robust ChIP–exo pattern, but five of them had no motif in JASPAR (Nrg2, Hms2, Hmo1, War1 and Pip2), and three had a different motif in JASPAR (Tea1, Rds2 and Sum1). These eight were also labelled as ssTFs. This resulted in 78 ssTFs that ChIP–exo/ChExMix detected as bound to a motif in YPD at 25°C. The remaining candidate targets that had JASPAR motifs were not labelled as ssTFs for the following reasons. First, one (Yox1) appeared site specific but was classified as a cofactor. Second, one is a GTF (TBP/Spt15). Third, 21 produced ChExMix binding locations but were deemed to be cofactors in YPD at 25°C (that is, they had bound locations, but were not bound site-specifically). Their site specificity could be condition specific. Fourth, 37 were not tested or not epitope-tagged (possibly because of lethality or technical difficulty in tagging). The remaining targets did not pass our detection thresholds. See Supplementary Data 2 (1, 9, 11) for the complete list of candidate factors, JASPAR/cis-bp motifs, and matches to ssTF-bound location in ref. 33.

The yeastepigenome.org website

Design. The backend of yeastepigenome.org is composed of two internal modules: a nodejs REST application and MongoDB database (version 3.4.5)
To generate composites, the ‘TagPileUpFrequency’ tool was used with no tagshifts, single-base-pair bins and tags set to equal with combined strands. One of the inputs to this tool is a BED file containing regions that have at least one overlapping ChExMix peak; the other is a BAM file. The tool was run on the target and master NoTag_20180928 control BAM files individually, to generate two data files that were fed into a composite generation script. The script uses matplotlib, a python plotting library, to generate a combined composite plot.

**Data availability**

See Supplementary Data 4 for a list of where to find available data and code online. In essence, all raw sequencing data and peak files from this study are available at the NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE147927. Processed data are available at https://doi.org/10.26208/rykf-6050. Additional analyses and data are at yeastepigenome.org. We warn that single-replicate data files are not likely to have meaningful data and should not be used without further replication. All underlying data used to generate composite plots, coordinate files and script parameters for Figs. 2–5, Extended Data Figs. 4, 5, 7, 8b and Supplementary Fig. 1 can be downloaded from https://github.com/CEGRcode/2021-Rossi_Nature. Final composite plot values can be found in Supplementary Data 5.

**Code availability**

code is available at https://github.com/CEGRcode/scriptmanager.

**Acknowledgements**

This work was supported by National Institutes of Health (NIH) grants ES013768, GM059055 and HG004160 to B.F.P.; National Science Foundation (NSF) ABI INNOVATION grant 1564466 to S.M.; grants from the Pennsylvania State University Institute for Advanced Research Computing 285–292 (Association for Computing Machinery, Portland, OR, 2020).

**References**

36. Vinayachandran, V. et al. Widespread and precise reprogramming of yeast protein–DNA interactions in response to heat shock. Genome Res. 28, 357–366 (2018).

37. Wang, M. & Pugh, B. F. Genome-wide mapping of nucleosome positions in yeast using high-resolution MChase ChIP-Seq. Methods Enzymol. 513, 233–250 (2012).

38. Shao, D., Kellogg, G. D., Lai, W. K. M., Mahony, S. & Pugh, B. F. In Practice and Experience in Advanced Research Computing 285–292 (Association for Computing Machinery, Portland, OR, 2020).

39. Picard Toolkit. http://broadinstitute.github.io/picard (2019).

40. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

41. Yamada, N., Lai, W. K. M., Farrell, N., Pugh, B. F. & Mahony, S. Characterizing protein–DNA binding event subtypes in ChIP–exo data. Bioinformatics 35, 903–913 (2019).

42. Benjaminii, V. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 57, 289–300 (1995).

43. de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. Bioinformatics 20, 1453–1454 (2004).

44. Becht, E. et al. Dimensionality reduction for single-cell data using UMAP. Nat. Biotechnol. 37, 38–44 (2019).

45. Xu, Z. et al. Bidirectional promoters generate pervasive transcription in yeast. Nature 457, 1033–1037 (2009).

46. Rape, H. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature 463, 295–301 (2012).

47. van Dijk, E. L. et al. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature 475, 114–117 (2011).

48. Albert, I., Wachi, S., Jiang, C. & Pugh, B. F. GeneTrack—a genomic data processing and visualization framework. Bioinformatics 25, 1305–1306 (2008).

49. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).

50. Jiang, C. & Pugh, B. F. A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome. Genome Biol. 10, R109 (2009).

51. Yen, Y., Vinayachandran, V., Batta, K., Koerber, R. T. & Pugh, B. F. Genome-wide nucleosome specificity and directionality of chromatin remodelers. Cell 149, 1461–1473 (2012).

52. Badis, G. et al. A library of yeast transcription factor motif reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. Mol. Cell 32, 878–887 (2008).

53. MacIsaac, K. D. et al. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 7, 113 (2006).

54. Liang, K. & Keles, S. Normalization of ChIP-seq data with control. BMC Bioinformatics 13, 199 (2012).

55. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2, 28–36 (1994).

**Innovation grant 1564466 to S.M.; grants from the Pennsylvania State University Institute for
Computational and Data Sciences to B.F.P. and W.K.M.L.; and computation from Advanced Cyberinfrastructure (ROAR) at the Pennsylvania State University. We thank D. Shao for her role as lead software engineer for the PEGR platform and for support through the Penn State Institute and Computational Data Sciences (ICDS) Research Innovations with Scientists and Engineers (RISE) team. We thank O. Lang for operating EpitopeID.

Author contributions M.J.R. designed and conducted experiments; performed library sequencing and data analysis; designed and tested the quality-control pipelines and web page; trained and managed lab personnel to produce data; supervised the project and co-wrote the manuscript. P.K.K. designed, developed and implemented the quality-control pipeline, analysis pipeline and website; organized and maintained data files; and provided bioinformatic support. W.K.M.L. performed high-throughput data processing and analysis, and provided bioinformatic support and scientific discussion. N.Y. provided bioinformatic support and developed the initial quality-control pipeline. N.B. and C.M. performed ChIP–exo and MNase ChIP–seq experiments and provided scientific discussion. G.K. provided bioinformatic support. K.B. and N.P.F. conducted ChIP–exo experiments and performed library sequencing. T.R.B., J.D.M., A.V.B., K.S.M., D.J.R. and E.S.P. conducted ChIP–exo experiments. G.D.K. provided high-performance infrastructure architecture and development, and edge-computing infrastructure design and support. S.M. provided bioinformatic guidance and support. B.F.P. conceptualized the project and conclusions, designed experiments, analysed the data, wrote the main text of the manuscript and co-wrote the remaining parts.

Competing interests B.F.P. has a financial interest in Peconic, LLC, which offers the ChIP–exo technology (US Patent 20100323361A1) implemented herein as a commercial service and could potentially benefit from the outcomes of this research. The remaining authors declare no competing interests.

Additional information Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03314-8. Correspondence and requests for materials should be addressed to B.F.P.

Peer review information Nature thanks Vishwanath Iyer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | ChIP–exo targets within meta-assemblages.

a, Simplified view of transcriptional regulation. A ssTF (TF; for example, Gal4) binds to its cognate motif (a UAS) within promoters in competition with chromatin/nucleosomes (red bar). The ssTF recruits (pink/green arrow) cofactors (for example, SAGA and Mediator) that assist in the assembly of a PIC (comprising TBP, TFIIB, and so on) and of Pol II at the transcript start site (TSS) of genes. Pol II then traverses the gene to the transcription end site (TES).

b, Diagram showing the ChIP–exo assay. Protein targets are crosslinked to DNA, which is then fragmented. Specific proteins are captured through an engineered TAP tag that binds the common Fc region of any immobilized IgG.

Near base-pair resolution is achieved using a strand-specific λ exonuclease that digests each strand of DNA in the 5′–3′ direction up to the point of crosslinking. c, Pie chart showing assayed targets separated by broad GO-based classifications (inner ring), or by UMAP-based clustering of genome-wide binding locations (outer ring). Listed are the common names of ChIP–exo targets that generated significantly enriched locations (with ‘significance’ defined in the Methods section ‘ChExMix locations’), grouped by their UMAP/K-means-derived meta-assemblage abbreviations (along with membership count), which are further grouped by the simplified GO-related categories. See also Supplementary Data 2 (2H).
Extended Data Fig. 2 | Data visualization and discovery in yeasteptiome.org. Shown is an example web browser view at yeasteptiome.org of ChIP-exo occupancy patterns for all targets (for example, Reb1) around predefined genomic features. Rows are sorted by gene or promoter (NFR/NDR) length, or by distance from the indicated reference feature (where $x = 0$). Promoter classes include (from top to bottom) RP, STM, TFO, UNB and others. See Supplementary Data 1 (1G, 1J, 1C) for the identification numbers and coordinates of respective row features, and for the sort order of features that are constant in all target display windows. The lower right box (when present) provides strand-separated tag 5′ ends distributed around the target’s cognate DNA motif, with the motif’s opposite strand (red) inverted in the composite plot. Corresponding colour-coded nucleotide sequences are shown. All images, underlying data values and datasets can be downloaded through embedded ‘META DATA’ target-specific links at yeasteptiome.org. Each dataset download includes a ReadMe file describing the contents of the download. We warn that targets with only a single replicate did not pass our significance threshold. See Supplementary Data 1 (1C) for sort orders that are not provided in the download.
**Extended Data Fig. 3 | UMAP granularity.** UMAP projection from Fig. 1c, with zoomed-in inserts. Labels are 40 K-means-based abbreviations (Supplementary Data 2 (Ij)). For coordinate values for individual targets, see Supplementary Data 2 (IC, ID).
Extended Data Fig. 4 | Confidence intervals for two example ChIP−exo datasets. Left, plots showing the ChIP−exo patterns for Orc6 and Mcm5. Bold lines represent means; dotted lines represent the 5–95% confidence interval (CI). The CI was calculated for each base pair in the 1 kb window across all ACSs (n = 253). Right, heatmaps showing ACS occupancy by Orc6 and Mcm5. Blue represents ChIP−exo data on the ACS motif strand; red represents data on the opposite strand.
Extended Data Fig. 5 | Protein architecture at regions transcribed by Pol II. Shown is gene-body occupancy (strands combined) of selected Pol II elongation-associated targets (a and b have different datasets, as annotated). In each panel, data were aligned and orientated by TSS (left) and TES (right). Shown are the top 200 coding genes (middle) and the top 200 noncoding genes (bottom) (based on Sua7 occupancy). See also Fig. 3c and note that the RP panels are identical to Fig. 3c. The y-axis values represent arbitrary linear units (a.u.), and are not comparable across different datasets, but are scaled equivalently across each of the six subpanels for the same dataset (shown with the same fill colour). Individual plotted values can be found in Supplementary Data 5.
Extended Data Fig. 6 | Architecture at LTRs. Heatmaps show the occupancy of the TFIIIB (Bdp1) (also TFIIIC & Pol III) and TBP (Spt15) components of the Pol III PIC and the TFIIIB (Sua7) components of the Pol II PIC at the five Ty LTR classes, along with the nucleotide composition (±100 bp from the LTR start; from yeastepignome.org; Gs, As, Ts and Cs are in yellow, red, green and blue, respectively). All rows are linked and sorted by LTR class, then length.
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | Properties of inducible (STM), insulated (TFO) and constitutive (UNB) Pol II promoters. a, NDRs are nucleosomal in vitro, while NFRs are mostly nucleosome-free. Heatmap of in vitro reconstituted MNase H3–ChIP nucleosomes (right), aligned by in vivo +1 nucleosome dyads and sorted by distance between the in vivo +1 nucleosome dyad and the first upstream stable nucleosome (USN) dyad (in vivo, left). b, Insulator ssTFs uncouple divergent PIC assembly. Shown are correlation coefficients between PIC (TFIIB/Sua7) occupancy (100-bp window centred on the TSS) at divergent TSSs that share the same promoter region. Data are presented as means ± s.d., from N = 6 biologically independent experiments, using two-tailed t-tests with no multiple comparisons. RP and STM promoters are merged here. c, Insulation at tandem genes. Shown are composite plots of PIC occupancy (green, TFIIB/Sua7) for promoter regions that share an upstream termination region (that is, tandem genes). Pcf11 (a representative termination factor) is in light brown, along with all 78 ssTFs (cyan), either collectively (‘TF’, top two panels) or individually, as indicated. STM, TFO and UNB composites are shown. The top ten insulator-associated ssTFs are based on the number of TFO promoters bound.
Extended Data Fig. 8 | ChIP–exo patterning reveals distinct local ssTF environments. a, Shown are strand-separated composite plots of 78 ssTFs bound at their cognate sites, and grouped by their meta-assemblage label (coloured borders, as defined in Extended Data Fig. 1c). Plots are oriented and centred by motif, and extend from −100 bp to +100 bp. Patterns were highly penetrant across individual sites for each ssTF (for example, see the lower right panel in Extended Data Fig. 2 for Rebl and yeastepigenome.org for other ssTFs). b, ChIP–exo composite profiles of individual subunits of the Mediator complex at motifs that are bound by the ssTF Yrr1 (from −500 bp to +500 bp), showing consistency of patterning across Mediator subunits.
Extended Data Fig. 9 | Properties of promoters bound by STM cofactors. 

**a**, Venn diagram showing promoters that have overlapping locations of STM cofactors (greater than 0 ChExMix calls in the dataset ‘SampleIDs’ in Supplementary Data 2 (10A)), as well as $Z$ scores for pair-wise overlaps.

**b**, Representative architecture of STM cofactors or PIC components at a consolidated set of ssTF motifs at 984 STM promoters (not strand-separated; see Methods and Supplementary Data 1 (1A)), oriented by TSS. 

**c**, Frequency distribution of promoters having the indicated PIC/TFIID (GTF/Taf2) occupancy ratios. The GTF that was used to measure occupancy is indicated at the top of each panel. These six GTFs were averaged in Fig. 5c. ‘MA’, three-bin moving average.
Extended Data Fig. 10 | Circuits involving interactions between ssTFs and cofactors. Genes encoding ssTFs are capitalized and are connected to their encoded proteins (ovals). Arrows connect ssTFs to other ssTF-encoding genes, to which they are bound via their cognate motif. ssTFs that bind to their own genes or create a loop are indicated with blue arrows (light blue where a motif was not detected for ssTFs binding their own gene). ssTFs are colour-coded on the basis of their meta-assemblage membership (see key). Promoter-bound ssTFs that are also particularly enriched with bound cofactors have coloured halos. Short diagonal black arrows point to the total number of all 5,378 coding genes that are bound by that ssTF to its cognate motif (first number given) or where no motif was detected (second number). The average relative PIC (TFIIB/Sua7) occupancy levels for those sets of genes is indicated by the number of green dots.
Extended Data Fig. 11 | Isolated ‘circuits’. a, Colours demarcate series paths. b, Colours emphasize different meta-assemblages, as defined in Extended Data Fig. 10.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- n/a Confirmed
- [ ] The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- [ ] A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- [ ] The statistical test(s) used AND whether they are one- or two-sided
  Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- [ ] A description of all covariates tested
- [ ] A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- [ ] A full description of the statistical parameters including central tendency (e.g. mean) or other basic estimates (e.g. regression coefficient) and variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- [ ] For null hypothesis testing, the test statistic [e.g. F, t, r] with confidence intervals, effect sizes, degrees of freedom and P value noted
  Give P values as exact values wherever possible.
- [ ] For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- [ ] For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- [ ] Estimates of effect sizes [e.g. Cohen’s d, Pearson’s r], indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

| Data collection | None used |
|-----------------|-----------|
| Data analysis   | Read mapping: bwa-mem [v0.7.17]; Peakh calling: ChExMix version 0.31; Analysis: ScriptManager v.0.12 at: https://github.com/CEGIRcode/scriptmanager; Deduplication: Picard; Other: samtools |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing files and peak files from this study are available at NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE147927. See Supplementary Data 4, for listings. All underlying data to generate composite plots, and coordinate files used to generate the figures for this paper can be downloaded from: https://github.com/CEGIRcode/2020-Rossi_YEP or at yeastepigenome.org.
### Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [x] Life sciences
- [ ] Behavioural & social sciences
- [ ] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reportsummary-flat.pdf

### Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | No statistical methods were used to predetermine sample size. Measurements were taken at all genomic locations. Sample size equals the number of annotated features reported by SGD or as calculated in the Methods. |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions | No data was excluded. |
| Replication | Conclusions that are based on data generated in this study are from a merge of at least two independent replicates, where each replicate supports the conclusion. Stringent criteria for sequencing depth and library complexity are reported in the Methods. |
| Randomization | Gene classifications were not random, but based on data clustering and prior knowledge - as described. Negative controls reflecting natural random background are used throughout. |
| Blinding | The investigators were not blinded to allocation during experiments and outcome assessment. |

### Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

#### Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| x   | Antibodies            |
| x   | Eukaryotic cell lines |
| x   | Palaeontology and archaeology |
| x   | Animals and other organisms |
| x   | Human research participants |
| x   | Clinical data         |
| x   | Dual use research of concern |

#### Methods

| n/a | Involved in the study |
|-----|-----------------------|
| x   | ChIP-seq              |
| x   | Flow cytometry        |
| x   | MRI-based neuroimaging |

### Antibodies

Antibodies used

- TAP-reactive Rabbit IgG [Sigma, i5005, various lot fl] 8.3 µg per 630 million cells;
- Anti-HA [Santa Cruz Biotechnology, sc-7392] 8.3 µg per 630 million cells;
- Serine 7, 2, or 5 phosphorylated forms of the C-terminal domain of RNA polymerase II [Millipore, 04-1570-L, 04-1571-L, or 04-1572-L, respectively] 3 µg per 630 million cells;
- H3K9ac [Millipore, 07-352] 3 µg per 630 million cells;
- H2BK123ub [Cell Signaling, s5456] 3 µg per 630 million cells;
- Cse4 [Car Wu, Johns Hopkins University] 3 µg per 630 million cells;
- Hsf1 [David Gross, Louisiana State University] 3 µg per 630 million cells;
- H3 [ab1971], H3K27ac [ab4729], H3K36me3 [ab9050], H3K4me3 [ab8580], H3K79me3 [ab2621], H3K12ac [ab49983], and H2B [Active Motif 39271] 3 µg per 630 million cells.

### Validation

TAP-tagged and HA-tagged strains were used which allows for use of non-specific IgG, and validation based on sequencing the epitope gene junction.

- 04-1570-L, https://www.sigmaaldrich.com/catalog/product:mm/041570?lang=en&region=US
- 04-1571-L, https://www.sigmaaldrich.com/catalog/product/mm/041571?lang=en&region=US
- 04-1572-L, https://www.sigmaaldrich.com/catalog/product/mm/041572?lang=en&region=US
- 07-352, https://www.sigmaaldrich.com/catalog/product/mm/07352?lang=en&region=US
- 55465, https://www.cellsignal.com/products/primary-antibodies/ubiquityl-histone-h2b-lys120-d11-xp-rabbit-mab/55465
- Cse4, https://doi.org/10.1016/j.cell.2007.04.026
- Hsf1, https://doi.org/10.1016/j.cell.2018.12.034
Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)
Dharmacon

Authentication
Validation was based on sequencing the epitope gene junction.

Mycoplasma contamination
N/A or yeast cells

Commonly misidentified lines
N/A; every cell culture is genotyped during sequencing

ChIP-seq

Data deposition

- Confirm that both raw and final processed data have been deposited in a public database such as GEO.
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
GEO submission: GSE147927; https://github.com/CEGcode/2020-Rossi_YEP

Files in database submission
See Supplementary Data 4

Genome browser session
yeastepigenome.org

Methodology

Replicates
All experiments for a given protein target were analyzed by ChxMx individually. The resulting peak calls for each individual replicate experiment can be found at yeastepigenome.org or GEO. In addition, the --lenientplus option enables a multi-replicate reproducibility assessment mode in ChxMx. Using this feature, replicate experiments passing Quality Control were analyzed simultaneously, and the resulting joint peak calls were used to classify Pol II features (see Promoter Classes, below). Locations are defined as ChxMx peaks if their tag counts pass the thresholds in the combined meta-experiment (essentially merging tag counts across replicates), or in one or more individual replicate experiments. However, binding events are only reported if the NCIS-scaled tag counts did not vary significantly across replicates [Binomial; 1.5 fold, p<0.01].

Sequencing depth
Typical read depth: 3 million paired-end reads (40 bp Read_1 and a 36 bp Read_2; entirety of both used for mapping). We empirically determined a minimum of 200,000 deduplicated tags were required to assess the quality of an individual dataset. If a dataset received less than 200,000 tags, then we required the tag duplication level (# of reads discarded by PICARD / # of input reads) of the sample to be less than 70% before we sequenced it deeper. For example: if a dataset had 100,000 mappable deduplicated tags [unique Read1/Read2 combination], but a total of 1 million mappable tags before filtering, then the duplication level was 90% and it was assumed that the library was insufficiently complex to warrant additional sequencing. If the library was insufficiently complex, we performed a technical replicate with the remainder of the chromatin preparation. Following this procedure, we produced a sufficiently complex library for over 95% of targets tested from a single yeast culture.

Antibodies
See Antibody section above.

Peak calling parameters
Read mapping was with BWA [see above], ChxMx12 version 0.31 was run with the following non-default parameters: --noend2 --scalewith1000 --minmodelupdateevents 30 --fixeddepth 0 --memminw 8 --memmaxw 21 --minmodelupdaterafs 25 --lenientplus. We also used the --excludedbed option to exclude analysis of a custom set of blacklist regions that included the rDNA locus, trRNA genes, and telomere regions. This list is available on Github: https://github.com/CEGcode/2020-Rossi_YEP/tree/master/02_references_and_features_files/ChxMx/Peak_Filter_List_190512.bed.

Data quality
By default, ChxMx requires the tag count at binding events to achieve at least 1.5 fold enrichment and a minimum Benjamin-Hochberg II corrected p-value of 0.01 [Binomial], compared with the scaled “Master No Tag” negative control (see Master No Tag negative control). All experiments for a given protein target were analyzed by ChxMx individually. The resulting peak calls for each individual replicate experiment can be found at yeastepigenome.org or GEO. In addition, the --lenientplus option enables a multi-replicate reproducibility assessment mode in ChxMx. Using this feature, replicate experiments passing Quality Control were analyzed simultaneously, and the resulting joint peak calls were used to classify Pol II features (see Promoter Classes, below). Locations are defined as ChxMx peaks if their tag counts pass the thresholds in the combined meta-experiment (essentially merging tag counts across replicates), or in one or more individual replicate experiments. However, binding events are only reported if the NCIS-scaled tag counts did not vary significantly across replicates [Binomial; 1.5 fold, p<0.01]. This latter condition has the effect of...
| Software | Bound locations were obtained using ChExMix software [Ref 12], which takes advantage of ChIP-exo resolution, statistical thresholding, and NCIS normalization [Ref 13] using untagged negative controls. |