natureresearch

Peer Review File

Manuscript Title: A high-resolution protein architecture of the budding yeast genome

Editorial Notes:

Redactions – unpublished data

Parts of this Peer Review File have been redacted as indicated to maintain the confidentiality of unpublished data.

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

While there is a classical regulatory paradigm for transcriptional regulation that involves transcription factors (TFs), cofactors, chromatin, and the pre-initiation complex (PIC) assembling at gene promoters, the inter-relationships between all of these factors at individual promoters remains unclear. In this paper, the authors use high resolution ChIP (ChIP-exo) to assay >800 DNA-binding proteins across the Saccharomyces cerevisiae genome to create a holistic near-single-base-pair resolution map of the genome’s structural organization and infer effects on genome function.

Through clustering and UMAP analysis, 21 “meta-assemblages” were identified, most of which corresponded to known biological complexes. The authors examine meta-assemblages at non-transcribed features and transcribed features independently, concluding much about the architecture at promoters for these individual elements (e.g. ACS, XCE, CEN, rRNA, tRNA, protein-coding genes, and LTRs).
Narrowing their scope to focus on Pol II promoters of coding genes, the authors find 4 classes with distinct architectures: 1) ribosomal protein (RP) promoters, 2) inducible promoters bound by SAGA, Tup1, and/or Mediator (STM), 3) promoters with TF organization (TFO) and without STM, and 4) promoters unbound (UNB) by anything except the PIC. They find that these different promoter architectures are associated with different nucleosome positioning mechanisms and can have different effects on the PICs of nearby tandem genes. Importantly, the authors find that there is no evidence for wide-spread TF regulation globally; rather, there are many constitutive promoters that do not rely on TF binding.

The authors identified a set of 78 TFs bound to promoters, finding that the strongest and most well-defined ChIP-exo signals come from those promoters lacking cofactor interactions. The authors examine the genes regulated by these TFs to uncover the regulatory circuit controlling TF gene transcription regulation and the cascading effects of TF transcription.

This paper concludes by describing the general regulatory mechanisms observed across the yeast genome, suggesting a holistic view of gene regulatory architecture in S. cerevisiae. The strength of this paper lies in its comprehensive survey of the position of regulatory proteins across the genome. Its relative weakness, however, is in some of the strong claims that approach claims of causation, with only evidence of correlation in one condition.

Major Critiques

- It appears that the “four fundamentally distinct architectural themes” (lines 194-5) of promoters were identified in large part through human classification (“Gene Classes” methods, lines 800-826). If unsupervised classification methods were used, would the same classes emerge?
- Quantification and statistical analysis are lacking throughout. With nearly bp resolution and a quantitative assay, the authors should systematically support all conclusions with statistical tests. Metagene analyses are helpful for visualization, but they should not solely be relied on as evidence of the authors’ claims. Here are some examples where quantification and statistics are needed, but the authors should systematically go through all results and bolster their claims with statistical analysis.
  - On lines 155-6, the authors claim that “elongation-associated targets generally matched Pol II occupancy across gene bodies, but were not enriched at promoters.” This would be best illustrated as a correlation between average Pol II occupancy across each gene body with occupancy of each elongation factor in the same gene body. This correlation could then be compared not only across each elongation factor, but also to compare between gene body and promoter.
  - The results displayed in Figure 5 could all be quantified and statistical tests should be performed to support claims.
  - How diverse are the promoters within each class? Are they always archetypical of the class and have the same factors associate or not associated? Perhaps Venn diagrams or another analysis could display how stereotypical loci are genome-wide.
- Overall, causality is frequently assumed without justification or testing. For example:
  - The authors make a strong statement about the role of TFs in PIC assembly on lines 277-80: “Thus, a long-standing paradigm that TFs direct PIC assembly through stable TFIID or TBP interactions was not evident. Instead, we found that TFs engaged a complex mixture of cofactors that may regulate NDR accessibility… and enhance Pol II recruitment.” Is there an orthogonal experimental approach that could be performed to validate this claim, even at one example locus?
  - For the regulatory circuit analysis, it is assumed that if a TF binds a promoter, it controls the expression of that gene. What supports that assumption?
- The final section states much as fact when it is rather the authors’ model. The authors should make clear what parts of their model is supported by their and others’ data, what parts are speculation and what parts remain unclear.
- For most figures, only one subunit of a protein complex is displayed. How similar are the binding patterns for the different subunits of a complex? They cluster together at the global level, which makes sense, but how similar are they at the bp level? If there are differences, what can be made of them?
- For a subset of yeast genes, the location of a protein tag (C or N terminal) can dramatically alter the protein’s subcellular location (PMID: 30550779). How many of the ~800 proteins mapped here are sensitive to tag location?
- It seems that replicate measurements were not made. Can the authors comment on that in the text and justify the decision? How reproducible are ChiP-exo data?

Minor Critiques

- It is not completely clear the thresholds that are used to determine “genome-wide binding patterns that were significantly distinct from background” (lines 69-70). Please clarify and elaborate.
- How stereotypical are the ChIP-exo patterns for the TFs displayed in Extended Data Fig. 7? A composite analysis of a few loci will always produce some shape, but it not clear whether the shape is meaningful. Please comment on whether composite binding patterns are representative of binding at all loci and include heatmaps for some factors.
- Figure 1F is not included in legend.
- It would be nice to include a supplementary figure similar to that of Figure 1F but colored by the meta-assemblages identified by the K=21 k-means clustering (similar, but more complete that Extended Data Figure 3). This would be helpful to visualize both those clusters that correspond to known biological complexes and those that do not.
- For all metagene plots (e.g. Figure 2, Figure 3, etc.), confidence intervals surrounding the mean occupancy traces should be displayed or that information should be displayed another way; Metagene plots averaged over more loci (i.e. ACS, N=253 in Figure 2A) will have more narrow confidence intervals

than those averaged over a few loci (i.e. CEN, N=16 in Figure 2C).

- The lines connecting Figure 3A and Figure 3B are misleading, as Figure 3A is illustrating occupancy around rRNA TSSs, transcribed by Pol I while Figure 3B is illustrating occupancy around tRNA TSSs, transcribed by Pol III.
- Add color legend to Figure 4C to clarify.
- Emphasize that these experiments were conducted in rich media (as stated in line 258); this may partially explain why a relatively small set of genes is found to be regulated by TFs (more would be activated in response to stress conditions, for example).
- How is Figure 5C mode-centered when the RSTM curve peaks right of the center – might be useful to clarify this plot for the readers.
- Font size is very small on many figures, especially Figure 6.
- Figure 5a composite graphs are challenging to understand without reading the text.

Referee #2 (Remarks to the Author):

This work is a tour-de-force that defines the in vivo DNA targets of essentially all DNA and chromatin-associated transcriptional regulatory proteins in yeast, at high resolution using the updated ChIP-exo method. The data look to be of extremely high quality and comprehensive in terms of coverage of proteins. A major strength is that they tested more than 800 factors and present the data from about 370 factors that gave reliable results. This is not likely to be bettered any time soon. It stands to supersede the landmark ChIP-chip datasets of yesteryear generated by the Young lab, and like them, it's likely that this new dataset will sustain fruitful new analyses by scores of labs for a long time to come. For these reasons primarily, it is worth publishing in a high-profile venue.

As with many genomics studies, much of the present study is descriptive and correlative. This is not a knock against it, but at the same time, care must be taken to avoid mechanistic interpretations that may be consistent with the data but don't necessarily follow from it and/or are known based on prior work. For example, some of their conclusions about the function of yeast insulators are unsupported without directed experiments where insulators and their binding proteins are deleted (see below).

The broadest and most significant finding reported here is that about 80% of all yeast protein coding genes (the UNB + TFO classes plus the ribosomal protein RP genes as an additional special class) are dominated by TFIID and are constitutively active (and TATA-less), whereas about 20% (the STM class) are dominated by TBP, SAGA + Mediator, and are inducible (and TATA-containing). However, in its broad outlines this was shown years ago by the Pugh lab's prior studies and verified by independent studies from many other groups. Perhaps the most significant new contribution of this work is that most classical sequence-specific transcription factors (TFs) associate with the minority STM class but not the majority constitutive class, which, if true, may be underappreciated. But the constitutive class includes the TFO group which does show binding by many sequence-specific TFs (which are not clearly detailed, see below), so this distinction is a bit muddled.

In particular, the in vitro nucleosome reconstitution experiments showing that STM promoters can become nucleosome depleted in a TF dependent manner in vivo while UNB promoters are intrinsically nucleosome free and made more so by chromatin remodelers, clarifies nicely the difference between NDRs and NFRs and is insightful.

Other points to consider and address before publication follow below in no particular order. Mainly they have to do with presentation, writing and interpretation which can be improved in many places.

The core of the paper is the classification of promoter architectures into STM, UNB and TFO groups. I could not find anywhere a clear list of which of the 371 factors associate with which promoter architecture. It's not in any of the figures or supplementary tables. Yet they constantly refer to factors belonging to each of the groups throughout the latter half of the manuscript. The Methods states that STM shows binding by at least one of the SAGA-Mediator-TUP1 group of factors (okay), but the TFO group is poorly defined. It does show binding by some TFs, but which ones, other than Abf1 and Reb1? The reader should be able to see in a main figure how well demarcated these 3 groups are and what their constituent factors are. This is key to evaluating their contention that sequence-specific TFs don't associate with the majority constitutive class of promoters in yeast.

Terminology: The acronyms TFO and UNB are un-memorable and confusing. I had a hard time keeping them straight even as I was reading the paper. Why not use descriptive names like "PIC-TF" (PIC + Abf1/Reb1 + ?TFs) and "PIC" respectively? Isn't that what they really are? STM – the inducible class – stands for SAGA-TUP1-Mediator but using TUP1 as the representative member of its group is pretty arbitrary and, in my opinion, misleading. Their Tup1 group combines well characterized activators as well as repressors which are mostly unrelated. It makes sense that they represent an "inducible" group but Tup1 is unlikely to be the unifying functional factor for this assemblage. In general, if you're making up new acronyms it's better to make them descriptive and functional.

Their statement that the budding yeast centromere lacks a nucleosome altogether despite having the histone H3 variant Cse4 (Fig. 2c) is pretty remarkable. This flies in the face of much evidence showing that a single octameric Cse4-containing nucleosome comprises the yeast centromere. This evidence includes a cryo-EM structure of the kinetochore complex formed around such a nucleosome, published in Nature (PMID: 31578520). The lack of signal from other histone proteins in the present study could just reflect a limitation of ChIP for the other histones at this unusual nucleosome.

I found many issues with the analysis of divergent promoters and insulators presented in Fig. 4e. First of all, you have to show the spread of the correlation values plotted, minimally by showing ±SD or ± 95% CI error bars and give the p-values for all differences that are mentioned. Second, provide the number of divergent gene pairs in each class. Third, the TFO class appears to show lower correlations even in the absence of Abf1/Reb1 binding (but see the prior two points), undercutting the argument that these two factors function as insulators. Is TFO + Abf1/Reb1 truly lower than TFO without? It's hard to tell because we don't know the spread or the p-values (first two points above). Fourth, the relevant correlation to measure if you want to talk about gene regulation, is not just TFIIB (Sua7) binding, but RNA expression levels, ideally measured using NET-seq or GRO-seq, which is freely available data. Fifth, the correlation of binding of Sua7 or any PIC component between the two opposites sides of the divergent promoter is fraught with issues of the resolution of the assay and the distance between the two ends of the promoter, which is not considered at all in this analysis. When the two signals are overlapping, they will appear correlated even when they are not. ChIP-exo is claimed to have motif level resolution and it is indeed better than standard ChIP-seq but the many profiles shown here reveal clearly that there is a huge spread of signal over 100-200 bp for many factors including Sua7. This is another reason it's better to use RNA expression to measure correlation which does not have this confounding issue. Finally, to truly say anything about the insulating property of Reb1/Abf1, it is necessary to do the experiment where the site is deleted or Reb1/Abf1 are conditionally turned off and it results in loss of insulation. This type of experiment may be beyond the scope of this study, but making mechanistic statements solely based on occupancy patterns and correlations is an overreach.

The fact that they don't detect TF-cofactor signatures matching that of GTFs like TFIIB or TBP (Fig. 5b) is not surprising because they are centering this at TF motifs which are likely to be upstream of the site of GTF assembly. At any rate, this observation by itself doesn't challenge the idea that TFs binding to upstream motifs direct PIC assembly at the core promoter. This could well be mediated by the Mediator complex, SAGA and other co-factors and indeed that's how it's thought to occur.

Abstract: "Most Pol II promoters lacked a regulatory region **by design**." And later (p 22), "selectively **designed** into the edge of many NDRs is a TATA box" These are loaded statements and prone to misinterpretation. Designed by who or what?! Perhaps they mean to say that these features have been selected for (by natural selection) to optimize various gene expression outcomes? If so, better to state that rather than use the word design.

Presentation: In Fig. 2 and in many other figures like this, they show selected target binding profiles at the chosen features, which could be misleading with regard to what is actually found there. They should show a version of these figures where not just the selected factors, but ALL target profiles are plotted. Yes, it could get ugly but that's the point. Does the DNA replication origin (ACS) truly have only MCM2-7 and ORC1-6 binding there, or do a whole bunch of other factors also bind there to varying extents? Certainly these profiles should be shown for all factors that show any signal above some numerical threshold, rather than selectively plotting only the factors that we know based on the biology. That really is the value of this work in that it allows one to visualize everything at a given locus.

In Fig. 2b showing the sub-telomeric factors, it's hard to distinguish which profile corresponds to which factor in the schematic. In general, the schematic depicts a much greater level of spatial resolution than warranted by the data. Many of the indicated factors have molecular weights that don't line up with their sizes in the schematic. There's a significant signal for one (or two) red read density traces on the left side of this region, but nothing is shown on top in the schematic. Is this Tup1 and why are there two similar colored traces? Perhaps a better representation would be to split out each of the factors into its own plot and show a better schematic on top, or at least label every trace as they do in panels a and b.

Figure 3a. It's not clear again which profile is which factor. The green TBP profile looks like it's a composite of two profiles, both green. Or it could be one of the other factors superimposed on top of the TBP profile, which looks odd if one is supposed to be behind the other. What is the second purple profile in the gene body that is not labeled? Their figures of this type could be significantly improved if they either plot them independently and stacked vertically or use lower alpha (transparency) settings for composite profiles.

Much of the mechanistic description of "futile cycles" of PIC assembly in the penultimate paragraph are more than what can be learned from the static snapshot data presented here, and sound like a review of our understanding based on many other studies. This should be clearly separated out and is perhaps more appropriate in the introduction.

The TF to TF regulatory circuit shown in Fig. 6 is somewhat fanciful because the data only shows binding, but the implication that there are extended cascades of regulation of TFs by one another is not justified. There is much evidence and analysis suggesting that such elaborate cascades likely don't occur in yeast (most recently see PMID: 32060051). Also, the way it's shown with both a gene and a protein represented for every TF with a dotted line connecting the gene to the protein only serves to clutter the figure. The standard way of showing this is much cleaner, with just one name abstracted out in each node (gene/protein doesn't matter because it is understood that you're only measuring binding of a protein to DNA), and edges representing binding to its target(s).

Fig. 3c shows ribosomal protein (RP) genes. But the font for the word "GENE" is so big the E has fallen out of its box in the schematic on top. What are the two grey downward pointing arrows? I'm guessing nucleosome positions, but this is not stated anywhere. In the TES plot, are the distances and scale the same as for the TSS plot? One could assume that, but why not show the numbers like one would for the X-axis in any plot? Finally, in Extended Fig. 4a they state that the middle sub-panel, showing the top 200 coding genes is identical to Fig. 3c, but that's wrong. It should be the top sub-panel, which shows the same RP gene plot.

Fig. 1f: legend for this panel is completely missing.

Some clarification on numbers. It's stated that 386 targets were replicated and analyzed, while 454 failed threshold, out of the total 840 attempted (Extended Fig. 1). In Fig. 1e and 1f, only 371 targets are shown to cluster, and these 371 are listed in Supplementary Table 3. But even this list of 371 factors includes 78 that are annotated as "Isolated or failed" in either their Meta-assemblage or Sub-assemblage labels. Which of these are failed and why are they not added to the 454 that failed thresholds? What is the difference between meta-assemblage and sub-assemblage? It's very confusing.

Supplementary Table 4. Is the "Miscellaneous" group included in the K-means cluster counts? According to "20Kmeans" it is not included but according to "40Kmeans" it is included.

All misassignments should be called out and commented on. For example, SET2 is assigned to the THO complex though it's not truly part of the THO complex (PMID: 28059701), rather it is expected to be firmly part of the Pol II elongation/SET complex assemblage. Similarly, GCN4 as part of the "cell-cycle regulation" meta assemblage (Supp. Table 3) and RPD3 complex (Ext Data Fig. 1) doesn't really make sense. What is driving these misassignments and how frequent is it?

Fig. 2: I think it should say "mirrored on the y-axis"?

ACS elements: 252 on p 7 but the figure referred to, Fig. 2a, says it's showing 253 elements.

Abstract: to formed -> to form

Referee #3 (Remarks to the Author):

The authors performed a large number of ChIP-exo experiments in yeast under rich medium conditions. The binding patterns were compared between factors, and clustered based on colocalization to suggest 21 meta-assemblies. The authors then classified promoters into 4 distinct groups based on the factor occupancy patterns and perform comparisons. One may argue this work provides a basis for obtaining the first holistic view of a functional eukaryotic genome. A lot of genome mapping data is already available under these conditions, yet the current study provides more mapped factors and higher resolution. The authors made several observations relevant to understanding gene regulation, for example, that SAGA binding is not concomitant with TFIID-independent PIC assembly and that there is no evidence that TFs direct PIC assembly through stable TFIID or TBP interactions. The study provides a resource for yeast and gene expression biologists. Additionally, the authors explored transcription factor regulatory circuits. However I have difficulties to judge the network part of the manuscript and hope another reviewer can do so. I recommend publication after the authors have addressed the following concerns:

1. The description of ChIP-exo to have base-pair resolution seems misleading. ChIP-exo involves formaldehyde, so cross-linking of entire complexes can occur. Also, in case factors do not bind to DNA directly, but via polymerases, the cross-linking sites are not that of the factor to DNA, but that of the polymerase complex. This is also alluded to by the authors in lines 270-272. Please go through text and make sure this is correctly understood. I recommend to also revise the abstract accordingly.
2. The authors should discuss the limitations of ChIP-exo better. For example, some of the factors analyzed bind RNA and a ChIP-exo can then not give a complete picture.
3. It is unclear how the authors determined colocalization. Was the 100 bp window around the peak midpoint from the entire peak region +50 bp on each side? Please clarify how these analyses were done. This relates to concern 1.
4. Why is the strand separated data shifted 50 bp in 3` direction before combining? Please clarify how the data was processed for analyses and visualization.
5. The authors should compare the expression levels of the 4 promoter classes, i.e. are genes with unbound promoters less expressed or are STM and TFO expression levels comparable?
6. In lines 127-129, the authors say that they did not detect histones at centromeres except Cse4. Please offer explanations considering that the Cse4-containing nucleosome structure exists.

7. In line 197, the subset of STM promoters (20% of all promoters), on which the major cofactors are present, is defined. This disagrees with literature (Bonnet et al, PMID 25228644; Baptista et al, PMID 28918903) that detected SAGA on almost all transcribed genes. Please comment on this.
8. Lines 283-298. If TFIID deficient promoters (RSTM subset, lower PIC/TFIID ratio) are associated with the presence of repressive co-factors (Tup1, RPD3-L) does this mean TFIID independent PIC assembly happens at promoters of repressed genes? Are these promoters active? Please add some discussion.
9. I recommend the authors think of the title. It is not obvious why a ‘genome’ should have a ‘protein architecture’, in particular also since protein architecture does not necessarily mean multiple proteins are involved. I suggest: ‘Comprehensive high resolution protein mapping along a eukaryote genome’
10. Abstract and text: Due to the nature of bulk experiments, I do not think that one can deduce from co-occupancy the existence of ‘meta-assemblages’. I think one can only talk about co-localization of proteins and suggest the existence of underlying assemblies, and point to other studies where a corresponding assembly was identified biochemically. The authors mention this in the beginning of the text, but the abstract suggests otherwise and later in the text their warning is largely ignored. Please edit accordingly.
11. There are important references lacking. I understand the list is limited but at least for some key conclusions that have been drawn before the authors should add references. Several important points that the authors make have been made before with the use of ChIP profiling. For example, it is well known from previous yeast factor profiling that Mediator binds to upstream sites and not TSSs as said in lines 166-167 (PMID 27773677), that elongation factors enter at fixed distances from the TSS as said in line 163 (PMID 20818391), and that Pcf11 binds around the TES (the poly-adenylation site for protein-coding genes) as said in line 159 (PMID 28318822).
12. It is unclear to me how the authors decisively conclude that the PIC assembles at the +1 nucleosome only via TFIID. Can it be excluded that other PIC components can bridge to the nucleosome? Please edit accordingly.
13. The authors say in paragraph 7 that intronic genes additionally load splicing factors. I am not sure what the evidence is that this is limited to the few percent of intronic genes. Please clarify.
14. Minor comments:
a. Missing panel description for Fig 1f.
b. Missing descriptions in Fig 1d for “UNB”, “STM”, “TFO” and “RPG”
c. There are very few references for the introduction and sections 1-3.
d. Missing reference to results: “In addition to ChIP-exo validation, the high positional concordance of subunits with each other, along with congruence from orthogonal methods, confirmed that epitope-tagging did not functionally alter the targets.”
e. Fig 3b appears to be an enlarged view of gene in Fig 3a. Please clarify. Also, no color consistency, Fig 3d has different design.
f. In Extend Data Fig. 3b, what is the distinction between the “TFIID” and “TAF” assemblages?

Referee #4 (Remarks to the Author):

This manuscript starts with an experimental tour de force: the profiling of more than 800 (TAP-tagged) proteins that are known or hypothesized to be associated with chromatin in rich media conditions using the ChIP-exo assay in yeast. The resulting dataset is a treasure trove that could be used to uncover new functions and mechanisms.

The first analysis step in the paper is to cluster the proteins – called “targets” here, somewhat confusingly, presumably because in each strain used a different gene has been TAP-tagged, and consequently is targeted by the antibody – based on their genome-wide profile. During this procedure, the profiles are compared after binning over 100bp windows; in other words, the almost single-base-pair resolution that sets the ChIP-exo assay apart is not taken full advantage off. This is a potential missed opportunity in terms of uncovering detailed mechanisms.

What follows is a cluster-by-cluster description of what the average spatial profiles of proteins associated with the cluster look like when centered on a related genomic feature. It all reads more like a review paper than a research article, in that many things that are known or expected are shown to emerge from the data. This is nice, but there is little here that is surprising.

More problematically in my opinion, there is hardly a single statement that is backed up by statistical analysis. In main text, not a single p-value is reported, even though there are many claims about observed differences. This is unacceptable. One example is on line 74: “resulted in ~400 targets displaying significant binding.” By what criterion?

There is also no attempt whatsoever to use the dataset to make testable predictions and then validate these predictions. Again, just as one example, on lines 162-164 the authors “suggest that elongation factors stably enter/exit … gene length”, but do not seem interested in testing this hypothesis. For an organism for which such an excellent genetic toolbox is available, this is disappointing.

With 800 proteins profiled, I would also have expected at least a handful of attributions of specific functions to a particular protein, which could have been tested by genetically perturbing this protein and monitoring a specific aspect of transcriptional regulation using an established reporter system. It is nice to recover many well-characterized protein complexes in unbiased manner, but more is needed to get the paper at the level expected for a top journal in my opinion.

To describe the protein clusters that emerge from the data analysis, a colorful palette of words and terms are used – “architecture”, “meta-assemblage”, “ensemble of assemblages”, “entourage of meta-assemblages” – without clear definition. This feels gratuitous, given that these are just clusters found in the data and there is no attempt to analyze the detailed 3D structure or mechanisms governing these complexes. The coining of new terms for classes of promoters such as “TFO” and “UNB” also feels unnecessary.

The TF network analysis in section 7 in terms of “archetype” motifs also lacks depth. Why is it notable that some TF’s target multiple other TF-encoding genes, and why does this suggest that “they represent major control junctions” (lines 312-313). There is no statistical analysis, no notion of what the expected occurrence would be under some null hypothesis. Autoregulation of TF genes by the proteins they code is well known.

Finally, the blanket statement that since many of the proteins are conserved, the same complexes are expected to form in higher eukaryotes (lines 374-396) is too simplistic. For one thing, the genome sequence is different, and is an equally crucial determinant of chromatin structure.

In summary, while I have great respect for this group’s work in general, and I appreciate the scope and the potential usefulness of the dataset, the present manuscript falls short of expectation by being far too superficial and descriptive.

Minor comments:

- Figure 1b: define “TAP” in caption; antibody binds oddly to TAP tag in schematic
- Figure 1e: typo in “Hierarchical”

Author Rebuttals to Initial Comments:

General response to Referees

We thank the reviewers for a deep and thorough review of the manuscript, and for bringing to our attention many of its shortcomings. We agree with the vast majority of issues and have now fix them. Statistical significance had been a foundational component of this work and is now more prominently indicated. The request to increase the scope of the manuscript by testing key hypotheses via mutations/depletions is very important to us. We have done this. We include some mutational tests in the revised manuscript. However, as the reviewers know well, such studies require a deep dive that can hardly be achieved in a figure subpanel. So, additional mutational tests are now included in this Response, where requested, noting that they are part of other manuscripts under review. An obvious concern is that mutational analyses is very focused, which in our view will take away from the breadth of the study (and its appeal to a general audience).

Referees' comments:

Referee #1 (Remarks to the Author):

While there is a classical regulatory paradigm for transcriptional regulation that involves transcription factors (TFs), cofactors, chromatin, and the pre-initiation complex (PIC) assembling at gene promoters, the inter-relationships between all of these factors at individual promoters remains unclear. In this paper, the authors use high resolution ChIP (ChIP-exo) to assay >800 DNA-binding proteins across the Saccharomyces cerevisiae genome to create a holistic near-single-base-pair resolution map of the genome's structural organization and infer effects on genome function.

Through clustering and UMAP analysis, 21 "meta-assemblages" were identified, most of which corresponded to known biological complexes. The authors examine meta-assemblages at non-transcribed features and transcribed features independently, concluding much about the architecture at promoters for these individual elements (e.g. ACS, XCE, CEN, rRNA, tRNA, protein-coding genes, and LTRs).
Narrowing their scope to focus on Pol II promoters of coding genes, the authors find 4 classes with distinct architectures: 1) ribosomal protein (RP) promoters, 2) inducible promoters bound by SAGA, Tup1, and/or Mediator (STM), 3) promoters with TF organization (TFO) and without STM, and 4) promoters unbound (UNB) by anything except the PIC. They find that these different promoter architectures are associated with different nucleosome positioning mechanisms and can have different effects on the PICs of nearby tandem genes. Importantly, the authors find that there is no evidence for wide-spread TF regulation globally; rather, there are many constitutive promoters that do not rely on TF binding.

The authors identified a set of 78 TFs bound to promoters, finding that the strongest and most well-defined ChIP-exo signals come from those promoters lacking cofactor interactions. The authors examine the genes regulated by these TFs to uncover the regulatory circuit controlling TF gene transcription regulation and the cascading effects of TF transcription.

This paper concludes by describing the general regulatory mechanisms observed across the yeast genome, suggesting a holistic view of gene regulatory architecture in S. cerevisiae. The strength of this paper lies in its comprehensive survey of the position of regulatory proteins across the genome. Its relative weakness, however, is in some of the strong claims that approach claims of causation, with only evidence of correlation in one condition.

Follow-up experiments and analyses are now presented that confirm claims of causation.

Major Critiques

1. It appears that the “four fundamentally distinct architectural themes” (lines 194-5) of promoters were identified in large part through human classification (“Gene Classes” methods, lines 800-826). If unsupervised classification methods were used, would the same classes emerge?

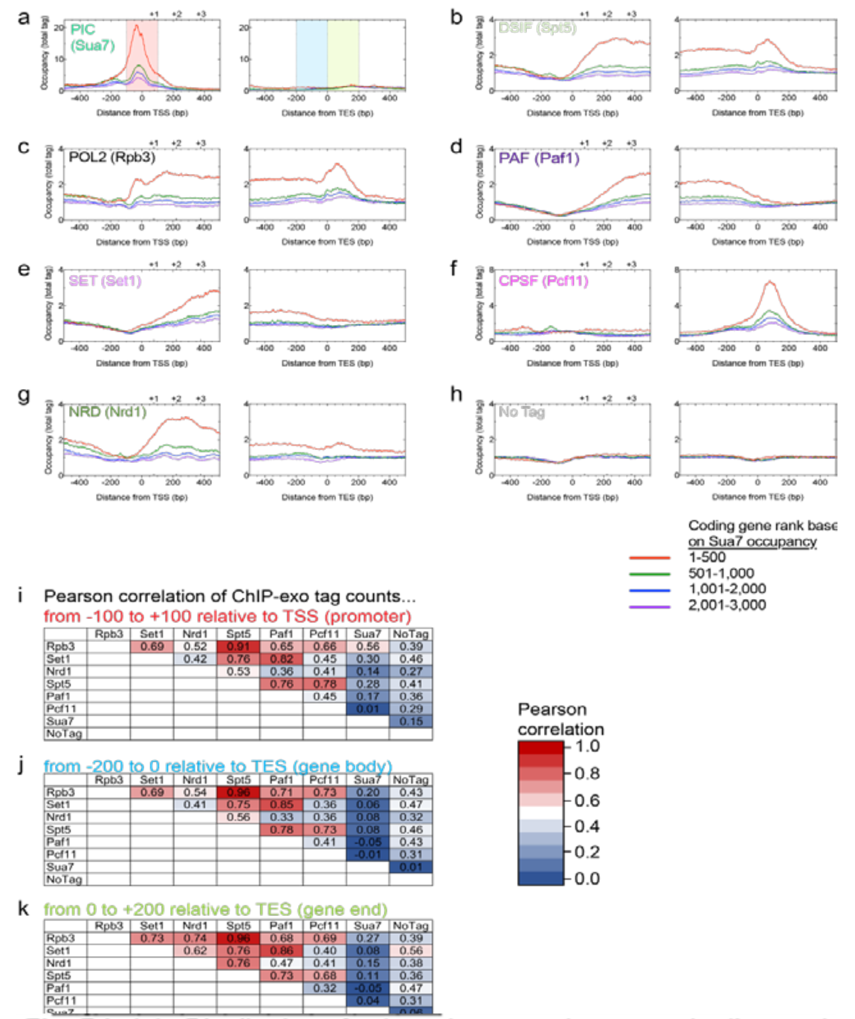
Not likely. Our unsupervised classification is based on target factor binding using 100 bp intervals, where it pulls out known complexes. It treats every target equally. For promoter classification, it does not consider the extent to which each target functionally contributes to distinct architectural themes, based on regulatory importance defined by human interest.

Thus, one of our classifications (RP, ribosomal protein promoters) represents the largest set of known co-regulated genes. Another (TFO), has a single protein at each promoter (primarily Reb1 or Abf1) as the main classification driver. The two are quite separated on the UMAP projection. UNB is driven by the lack of TFs/cofactors but the presence of PIC components, which forms a separate UMAP cluster that is found at all genes. STM is driven by numerous and various TFs and cofactors (which form a broad UMAP cluster). So, aspects of our promoter classification are evident in our unsupervised approach. We now revised the text (Methods) to indicate the classifications are supported by an unsupervised approach, and also explain why unsupervised approaches alone are not sufficient. We also added a main figure (**Fig.4a**) showing all targets and all genes organized by architectural theme, and whether target binding passed threshold or not.

Quantification and statistical analysis are lacking throughout. With nearly bp resolution and a quantitative assay, the authors should systematically support all conclusions with statistical tests. Metagene analyses are helpful for visualization, but they should not solely be relied on as evidence of the authors’ claims. Here are some examples where quantification and statistics are needed, but the authors should systematically go through all results and bolster their claims with statistical analysis.

Every step of our data processing and analysis pipeline was quantitative and statistically rigorous, where appropriate. We used ChExMix for peak calling, which reports p-value significance. Unlike many large genome-wide studies, we require replicates of each dataset that we analyzed, and report conclusions only where the same conclusion is evident from both datasets. We now conduct more statistical tests throughout.

2. On lines 155-6, the authors claim that “elongation-associated targets generally matched Pol II occupancy across gene bodies, but were not enriched at promoters.” This would be best illustrated as a correlation between average Pol II occupancy across each gene body with occupancy of each elongation factor in the same gene body. This correlation could then be compared not only across each elongation factor, but also to compare between gene body and promoter.



i Pearson correlation of ChIP-exo tag counts...

from -100 to +100 relative to TSS (promoter)

	Rpb3	Set1	Nrd1	Spt5	Paf1	Pcf11	Sua7	NoTag
Rpb3		0.69	0.52	0.91	0.65	0.66	0.56	0.39
Set1			0.42	0.76	0.82	0.45	0.30	0.46
Nrd1				0.53	0.36	0.41	0.14	0.27
Spt5					0.76	0.78	0.28	0.41
Paf1						0.45	0.17	0.36
Pcf11							0.01	0.29
Sua7								0.15
NoTag								

j from -200 to 0 relative to TES (gene body)

	Rpb3	Set1	Nrd1	Spt5	Paf1	Pcf11	Sua7	NoTag
Rpb3		0.69	0.54	0.96	0.71	0.73	0.20	0.43
Set1			0.41	0.75	0.85	0.36	0.06	0.47
Nrd1				0.56	0.33	0.36	0.08	0.32
Spt5					0.78	0.73	0.08	0.46
Paf1						0.41	-0.05	0.43
Pcf11							-0.01	0.31
Sua7								0.01
NoTag								

k from 0 to +200 relative to TES (gene end)

	Rpb3	Set1	Nrd1	Spt5	Paf1	Pcf11	Sua7	NoTag
Rpb3		0.73	0.74	0.96	0.68	0.69	0.27	0.39
Set1			0.62	0.76	0.86	0.40	0.08	0.56
Nrd1				0.76	0.47	0.41	0.15	0.38
Spt5					0.73	0.68	0.11	0.36
Paf1						0.32	-0.05	0.47
Pcf11							0.04	0.31
Sua7								0.06

Fig. R1. A-h, Distribution of selected targets along gene bodies, and separated by quartiles of Sua7 occupancy. i-j, correlation values of factors in promoter regions, gene bodies, and gene ends, defined as indicated in each table. Since targets are not always cleanly separated, relative correlation values are most informative.

The requested analysis is largely covered by our UMAP projections, showing that elongation factors cluster together, along with Pol II. What UMAP does not do is offer promoter/gene-body context, which is what Fig. 3c and current ED_Fig. 5a,b offers. Nonetheless, here we provide the requested analysis for the major elongation UMAP clusters (Fig. R1, panels i-k) and will include it in the manuscript if the reviewer still deems it important. Note that we do expect correlations of elongation factors in promoter regions, since there is some bleed-through in the regions analyzed. But the relative depletion of elongation factors at promoters compared to gene bodies is robust to promoter activity, as evidenced in panels a-h.

3. The results displayed in Figure 5 could all be quantified and statistical tests should be performed to support claims.

Each panel in this figure did include quantification of binding. Fig. 5a (Venn) is based on statistically-based ChExMix calls. The null hypothesis is that these classifications occur by random chance among 5378 promoters. The p-value for each venn overlap between SAGA, TUP1 complex, and Mediator is less than 1E-05 (Z > 10) compared to a random distribution of overlapping genes. These are now added to Fig. 5a. The exact patterns in composite plots of Fig. 5a, 5b are observed in replicates and in multiple subunits of the same complex. We do not know what statistical metric would be most appropriate in defining peaks and valleys, since we examine surrounding areas for context, with much of that surrounding area being noise. The reproducible peaks and valleys appear quite evident in the multiple traces. For Fig. 5c, we now provide frequency distributions for six different PIC subunits (Extended Data Fig. 9), all of which produce the same conclusion.

4. How diverse are the promoters within each class? Are they always archetypical of the class and have the same factors associate or not associated? Perhaps Venn diagrams or another analysis could display how stereotypical loci are genome-wide.

RP genes are not diverse, as they were chosen for their similarity and coregulation. UNB genes are not diverse, having only a PIC and the transcription/elongation machinery. TFO are like UNB, except that >75% of them have one of 10 insulator-like TFs. STM promoters are quite diverse, having a wide variety of TFs. Promoter diversity is now illustrated in Fig. 4a, and analyzed in Supplementary Table 3.

• Overall, causality is frequently assumed without justification or testing. For example:

5. The authors make a strong statement about the role of TFs in PIC assembly on lines 277-80: "*Thus, a long-standing paradigm that TFs direct PIC assembly through stable TFIID or TBP interactions was not evident. Instead, we found that TFs engaged a complex mixture of cofactors that may regulate NDR accessibility… and enhance Pol II recruitment.*" Is there an orthogonal experimental approach that could be performed to validate this claim, even at one example locus?

In the first quoted sentence, we question the causality reported in the literature. There may not be a simple orthologous experimental approach to prove a lack of causality. While there is historical published evidence for stable TF interactions with GTF/PIC components **in vitro**, we find no definitive evidence (including in the literature) for equivalent interactions **in vivo** in a natural setting. Our results demonstrate stable/measurable TF interactions with SAGA, Mediator, and TUP, but the same is not observed with the GTFs including TFIID TAFs. Without all of these positive control interactions, we are quite confident that if stable interactions occur between TFs and TFIID/GTFs, we would have detect it as a TFIID/GTF crosslinking pattern around the TF. This is why our combined study of many factors is so very powerful.

The second quoted sentence (that is, TF-cofactor interactions) was not intended to claim causality, but was stating it as a generally accepted concept. We now break that sentence into two parts, with the latter being referenced to the published literature.

6. For the regulatory circuit analysis, it is assumed that if a TF binds a promoter, it controls the expression of that gene. What supports that assumption?

The general assumption has ample support in the literature, that TF-bound sites within promoters contribute to the expression of that gene. We now reference examples of such evidence (PMID: 7739554, 11125145). Nonetheless, this may not be true in every case, particularly when there is built-in redundancy such as multiple TFs bound to a promoter. It was not our intent to establish the assumptions de novo, but rather just to point out where our data is supportive and where it is not supportive. Where it is not supportive, we assess the prior evidence on which the assumption is based. In such cases (e.g. centromeric nucleosome, and TF-PIC interactions) the prior evidence turns out to be weak.

7. The final section states much as fact when it is rather the authors' model. The authors should make clear what parts of their model is supported by their and others' data, what parts are speculation and what parts remain unclear.
We have now redoubled our efforts to do so. We now more clearly distinguish what we are concluding based on our work, the published literature (by referencing), and what we are speculating on.

8. For most figures, only one subunit of a protein complex is displayed. How similar are the binding patterns for the different subunits of a complex? They cluster together at the global level, which makes sense, but how similar are they at the bp level? If there are differences, what can be made of them?

The ChIP-exo patterns of different subunits of a complex are typically quite similar (within experimental error). This is now shown in Extended Data Fig. 8b for Mediator bound to the Yrr1 TF (but also for TFIIIC in Fig. 3b, and the Kar4/Dig1/Ste12 complex in Fig. 3d). This is to be expected because many subunits crosslink to each other and to DNA, and so they show essentially the same chip-exo pattern. This was noted in the original draft. The local signal intensity may vary somewhat between subunits (whether in same or different complexes). We expect subunits that do not directly interact with the TF to have weaker intensity because their signal is dependent on a greater number of indirect protein-protein crosslinks. We now note the consistency of crosslinking points among subunits of the same complex (for TFIIIC, Kar4/Dig1/Ste12, and Mediator). Related to the reviewers point, histone subunits which are spread across 150 bp of nucleosomal DNA do not have ChIP-exo patterns that precisely coincide with each other, but instead are positioned next to each other in accord with the wrap of DNA around the histone core and atomic structures.

9. For a subset of yeast genes, the location of a protein tag (C or N terminal) can dramatically alter the protein's subcellular location (PMID: 30550779). How many of the ~800 proteins mapped here are sensitive to tag location?

Since we were canvasing so widely, we did not dig deeper on the factors that did not work. Of the 400 proteins we tested that did not produce a ChIP signal, over half were not expected to bind DNA. Another 100 were not expected to bind under nonstress conditions. Thus, we estimate that no more than 25% of the tagged proteins that failed to produce a signal in ChIP were rendered nonfunctional by tagging, which is in line with PMID: 30550779.

Our high success rate among proteins known to bind a DNA motif (and thus orthogonally validated) led us to conclude that for the vast majority of proteins the tag was not dramatically altering protein activity. Also, the genome-wide colocalization of targets that are known to bind each other provide additional validation.

At the beginning of this project we tested about a half dozen targets with a target-specific antibody. We had done the same with our earlier low resolution work and found no significant differences between the binding locations or growth pattern in the tagged strain versus our negative control (BY4741) strain. One exception was Rap1; binding specificity of the tagged version was unchanged, but the strain had a doubling time ~75% longer than WT.

10. It seems that replicate measurements were not made. Can the authors comment on that in the text and justify the decision? How reproducible are ChiP-exo data?

This is not true. Replicate measurements were in fact made for all analyzed data, and stated as such in the main text, in the sentence after Fig. 1c was called out. Individual replicates can be viewed at yeastepigenome.org. Additionally, all analyses in the paper were performed using ChExMix bound locations that were reproducible as described in the Methods (ChExMix Locations). Please note that targets that did not work the first time were attempted a second time, but were not analyzed further following a second failure.

Minor Critiques

11. It is not completely clear the thresholds that are used to determine “genome-wide binding patterns that were significantly distinct from background” (lines 69-70). Please clarify and elaborate.

This is now stated in the main text (“Benjamini-Hochberg corrected p-values <0.01, and 1.5 fold enrichment over untagged controls”) and detailed in the Methods (ChExMix Locations).

12. How stereotypical are the ChIP-exo patterns for the TFs displayed in Extended Data Fig. 7? A composite analysis of a few loci will always produce some shape, but it not clear whether the shape is meaningful. Please comment on whether composite binding patterns are representative of binding at all loci and include heatmaps for some factors.

These patterns are very stereotypical for each TF, and robust across all sites. See for example ED_Fig. 2 (lower right). Heatmaps for each and every TF can be found at yeastepigenome.org. We also demonstrate 95% and 5% confidence interval for the first instance of a composite (Extended Data Fig. 4). Peak shapes are meaningful because they are highly reproducible and define the exact position of protein crosslinks to DNA (directly or indirectly). This is now stated in section 1. Moreover, since each peak on one strand is accompanied by its pair on the opposite strand, offset in the 3’ direction, the shape is internally quite robust and interpretable. We also see the same patterning for other subunits of the same complex (Extended Data Fig. 8b).

13. Figure 1F is not included in legend.

It was there, but it got cut off when the figure was imbedded into Word and converted to PDF.

14. It would be nice to include a supplementary figure similar to that of Figure 1F but colored by the meta-assemblages identified by the K=21 k-means clustering (similar, but more complete that Extended Data Figure 3). This would be helpful to visualize both those clusters that correspond to known biological complexes and those that do not.

We believe the reviewer's request is in fact Fig. 1f. However, the reviewer may be asking the reciprocal question, where Fig. 1f is colored by known biochemical complexes instead of by clustering. When we do this it looks virtually identical to Fig. 1f, but then we lose the connectivity to other complexes (e.g., Mediator and SWI/SNF cluster together). Essentially all clusters are associated with at least one particular biochemical complex, and one can look this up for individual targets in Supplementary Table 3. This point was noted in the previous draft, and we now taken steps to ensure this point is clearer in the revised draft.

15. For all metagene plots (e.g. Figure 2, Figure 3, etc.), confidence intervals surrounding the mean occupancy traces should be displayed or that information should be displayed another way; Metagene plots averaged over more loci (i.e. ACS, N=253 in Figure 2A) will have more narrow confidence intervals than those averaged over a few loci (i.e. CEN, N=16 in Figure 2C).

Our interpretation of this comment is that the 95% and 5% “confidence intervals” for metagene plots reflect the variance across different instances of features that belong to the same class. So, they are not an independent measure of the same instance, and thus confidence intervals are not reporting on statistical variation.

We now present examples of the observed “meta variance” in Extended Data Fig. 4, by plotting the mean with the 95 and 5 percent confidence intervals. We find that the heatmaps presented alongside provide a more comprehensive view of the data spread. Heatmaps of all datasets are available at yeastepigenome.org and we have updated the manuscript to re-emphasize the availability of this information.

16. The lines connecting Figure 3A and Figure 3B are misleading, as Figure 3A is illustrating occupancy around rRNA TSSs, transcribed by Pol I while Figure 3B is illustrating occupancy around tRNA TSSs, transcribed by Pol III.

We now remove the lines.

17. Add color legend to Figure 4C to clarify.

Done.

18. Emphasize that these experiments were conducted in rich media (as stated in line 258); this may partially explain why a relatively small set of genes is found to be regulated by TFs (more would be activated in response to stress conditions, for example).

We have now emphasized this more. However, our evidence suggests that only a relatively small subset of all genes have any potential to be regulated by TF/cofactors under any condition. Most genes are not designed to be regulated under any condition. This is explained extensively in the Methods. This is a very important point.

19. How is Figure 5C mode-centered when the RSTM curve peaks right of the center – might be useful to clarify this plot for the readers.

All traces are linked to each other, but then centered on the most frequent mode. Since mode-centering just shifts the x-axis, for simplicity we now do not mode-center.

20. Font size is very small on many figures, especially Figure 6.

We have now gone through the main figures to ensure that they are compliant with publishing guidelines. Fig. 6 is particularly dense and complex, that may be best visualized by magnifying online, and may not conform to best practices when shown in print.

21. Figure 5a composite graphs are challenging to understand without reading the text.

We have now redrawn and re-annotated the figure to be more consistent with prior figure layouts.

Referee #2 (Remarks to the Author):

This work is a tour-de-force that defines the in vivo DNA targets of essentially all DNA and chromatin-associated transcriptional regulatory proteins in yeast, at high resolution using the updated ChIP-exo method. The data look to be of extremely high quality and comprehensive in terms of coverage of proteins. A major strength is that they tested more than 800 factors and present the data from about 370 factors that gave reliable results. This is not likely to be bettered any time soon. It stands to supersede the landmark ChIP-chip datasets of yesteryear generated by the Young lab, and like them, it's likely that this new dataset will sustain fruitful new analyses by scores of labs for a long time to come. For these reasons primarily, it is worth publishing in a high-profile venue.

As with many genomics studies, much of the present study is descriptive and correlative. This is not a knock against it, but at the same time, care must be taken to avoid mechanistic interpretations that may be consistent with the data but don't necessarily follow from it and/or are known based on prior work. For example, some of their conclusions about the function of yeast insulators are unsupported without directed experiments where insulators and their binding proteins are deleted (see below).

1. The broadest and most significant finding reported here is that about 80% of all yeast protein coding genes (the UNB + TFO classes plus the ribosomal protein RP genes as an additional special class) are dominated by TFIID and are constitutively active (and TATA-less), whereas about 20% (the STM class) are dominated by TBP, SAGA + Mediator, and are inducible (and TATA-containing). However, in its broad outlines this was shown years ago by the Pugh lab's prior studies and verified by independent studies from many other groups. Perhaps the most significant new contribution of this work is that most classical sequence-specific transcription factors (TFs) associate with the minority STM class but not the majority constitutive class, which, if true, may be

underappreciated. But the constitutive class includes the TFO group which does show binding by many sequence-specific TFs (which are not clearly detailed, see below), so this distinction is a bit muddled.

We agree that TFIID vs SAGA/Mediator is an important finding, which continues to confuse the field. Our prior description of SAGA-dominated promoters being highly selective for SAGA and TBP has been challenged in recent publications (incorrectly, in our opinion, but this was recently dialed back by Donczew et al 2020). We touch on this because it is part of the whole picture and has been incorrectly portrayed. Thus, it continues to be confusing. Nonetheless, we are led to believe that the broadest and most significant findings include the following: 1) The ORC replication complex engulfs an adjacent nucleosome; 2) Centromeres lack nucleosomes. 3) Subtelomeric repressive domains have a highly focused but well defined architecture of proteins encompassing three nucleosomes. 4) That most promoters evolved to not have an inducible architecture. The inducible architecture includes TFs/cofactors/chromatin/TBP, whereas the constitutive architecture lacks this, and instead involves TFIID without TFs and cofactors (acknowledged by the reviewer). It was very surprising to find that TFs do not interact with TFIID, but this now makes sense in light of our new model of gene induction. 5) A comprehensive TF regulatory network is described, where none has previously existed. We identify the cognate cofactor(s) for each TF – this has not previously existed. These observations are novel, and in our opinion create a highly enlightened view of chromatin.

2. In particular, the in vitro nucleosome reconstitution experiments showing that STM promoters can become nucleosome depleted in a TF dependent manner in vivo while UNB promoters are intrinsically nucleosome free and made more so by chromatin remodelers, clarifies nicely the difference between NDRs and NFRs and is insightful.

We agree. This figure contributes one of a number of experimental tests of hypotheses in the manuscript.

Other points to consider and address before publication follow below in no particular order. Mainly they have to do with presentation, writing and interpretation which can be improved in many places.

3. The core of the paper is the classification of promoter architectures into STM, UNB and TFO groups. I could not find anywhere a clear list of which of the 371 factors associate with which promoter architecture. It's not in any of the figures or supplementary tables. Yet they constantly refer to factors belonging to each of the groups throughout the latter half of the manuscript. The Methods states that STM shows binding by at least one of the SAGA-Mediator-TUP1 group of factors (okay), but the TFO group is poorly defined. It does show binding by some TFs, but which ones, other than Abf1 and Reb1? The reader should be able to see in a main figure how well demarcated these 3 groups are and what their constituent factors are. This is key to evaluating their contention that sequence-specific TFs don't associate with the majority constitutive class of promoters in yeast.

We apologized for this oversight. It is now included in Supplementary Table 3. This should answer the question about TFO, in detail, but the major TFs that constitute the TFO class are now listed in ED_Fig. 7. We now present a main Fig. 4a showing a graphical matrix of geneID (rows) sorted by group, and bound by each of the 371 targets sorted by their role in defining the four gene groups. Fig.4a row/column names is presented in Supplementary Table 3$_3$. (Subscripts refer to embedded worksheet order)

4. Terminology: The acronyms TFO and UNB are un-memorable and confusing. I had a hard time keeping them straight even as I was reading the paper. Why not use descriptive names like "PIC-TF" (PIC + Abf1/Reb1 + ?TFs) and "PIC" respectively? Isn't that what they really are? STM – the inducible class – stands for SAGA-TUP1-Mediator but using TUP1 as the representative member of its group is pretty arbitrary and, in my opinion, misleading. Their Tup1 group combines well characterized activators as well as repressors which are mostly unrelated. It makes sense that they represent an "inducible" group but Tup1 is unlikely to be the unifying functional factor for this assemblage. In general, if you're making up new acronyms it's better to make them descriptive and functional.

We agree. The acronyms are indeed unmemorable, and probably should be for the reasons discussed below. The reviewer's suggested names seem cumbersome when verbally expressed, and may also be confusing (PIC is a term for a complex, rather than a gene class). The challenge lies in the realization that there is no classification schema (not to mention labels) that is best in an absolute sense, since all classifications are based on a limited set of criteria, which are experiment-based (and thus subject to variability). Indeed, there are many other ways to carve up the data to achieve different (but largely overlapping) classifications (e.g., STM=SAGA$_{dom}$=TAF$_{depleted}$), which would be appropriate when extracting certain aspects of promoters. So, our reasoning is that we are creating general classes based on what is measurably bound (i.e., the underlying experimental basis), but whose membership is locked in this snapshot of criteria. In accord with the reviewer's point, we do use terms like "inducible", "insulated", and "constitutive" as general descriptors for sets of equivalent gene classes (e.g., inducible=STM=SAGA$_{dom}$ =TAF$_{depleted}$).

The reviewer's comment that Tup1 is not likely a unifying functional factor may indeed be correct, but the naming scheme is not a product of functional studies. From a ChIP perspective, Tup1 is a major factor. As shown in Fig 5, the Tup1 complex binds many of the same genes as activator TFs. At the TF sites, the ChIP-exo patterns of Tup1, SAGA, and Mediator are essentially equivalent (Fig. 5a, bottom). That is why we think they warrant being grouped together.

5. Their statement that the budding yeast centromere lacks a nucleosome altogether despite having the histone H3 variant Cse4 (Fig. 2c) is pretty remarkable. This flies in the face of much evidence showing that a single octameric Cse4-containing nucleosome comprises the yeast centromere. This evidence includes a cryo-EM structure of the kinetochore complex formed around such a nucleosome, published in Nature (PMID: 31578520). The lack of signal from other histone proteins in the present study could just reflect a limitation of ChIP for the other histones at this unusual nucleosome.

Cse4, Mcm16, Nkp2, etc… map to the centromere, and are among our most robust data of the 800+ targets assayed. Thus, we do not think this result is due to a limitation of ChIP or a bioinformatic artifact. All histone components show strong signal adjacent to the centromere, but a depletion of signal at the centromere. It is hard to imagine being able to ChIP centromeric components, and Cse4, but not two copies each of the other three core histones, particularly when they chip well throughout the rest of the genome. We also get the same result with histone and histone modification antibodies in WT strains. We note that all of the in vivo studies that concluded a nucleosome was present was based on MNase resistance (whithout ChIP), which could have also been produced by the kinetochore. Where histone ChIP was done, it was such low resolution that it likely picked up surrounding histones.

PMID: 31578520 cryoEM structure of the kinetochore complex (CCAN) with a Cse4-containing nucleosome, was achieved at 4.15Å resolution using non-centromeric "601" sequence. The published structure included modeling a standard nucleosome rather than a Cse4 nucleosome (whose independent structure has not been determined). The electron density did not exactly fit a canonical nucleosome. Is it possible that the super-stable unnatural 601 DNA sequence forced Cse4 into a nucleosome-like structure, which then bound CCAN primarily because Cse4 has strong interactions with CCAN (which our data supports)? How certain can one be at 4.15 Å resolution? Note, there is no definitive in vivo experiments on this. Since no method has perfect "vision", we now note the discrepancy, and leave it as an open question.

6. I found many issues with the analysis of divergent promoters and insulators presented in Fig. 4e. First of all, you have to show the spread of the correlation values plotted, minimally by showing ±SD or ± 95% CI error bars and give the p-values for all differences that are mentioned.

We now add ±SD (6 replicates) and have added the analysis suggested in comment #9 below. This new figure panel is based on published CRAC data of nascent transcription, which did not present enough biological replicates to provide SD values on the correlations. However, there are a sufficient number of negative controls presented, that provide a sense of the variance in the data.

7. Second, provide the number of divergent gene pairs in each class.

This is now provided.

8. Third, the TFO class appears to show lower correlations even in the absence of Abf1/Reb1 binding (but see the prior two points), undercutting the argument that these two factors function as insulators. Is TFO + Abf1/Reb1 truly lower than TFO without? It's hard to tell because we don't know the spread or the p-values (first two points above).

Abf1/Reb1 are not the only two insulator factors. They were only used as predominant examples. There are other insulator TFs in the TFO group that may also be contributing to the low correlation.

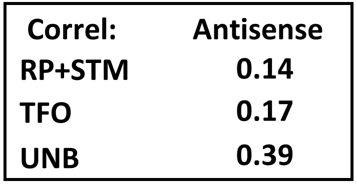
9. Fourth, the relevant correlation to measure if you want to talk about gene regulation, is not just TFIIB (Sua7) binding, but RNA expression levels, ideally measured using NET-seq or GRO-seq, which is freely available data.

We have now performed this analysis with existing nascent RNA data (CRAC) from the Libri lab, and added it to Fig. 4e. We focus on this data because it includes

Reb1 and Rap1 depletion. Their analysis was directed at addressing insulator ("roadblock") function of these factors at terminators. The results confirm and provide additional evidence that Reb1 and Rap1 insulate divergent transcription (as well as tandem transcription).

10. Fifth, the correlation of binding of Sua7 or any PIC component between the two opposites sides of the divergent promoter is fraught with issues of the resolution of the assay and the distance between the two ends of the promoter, which is not considered at all in this analysis. When the two signals are overlapping, they will appear correlated even when they are not. ChIP-exo is claimed to have motif level resolution and it is indeed better than standard ChIP-seq but the many profiles shown here reveal clearly that there is a huge spread of signal over 100-200 bp for many factors including Sua7. This is another reason it's better to use RNA expression to measure correlation which does not have this confounding issue.

As the reviewer points out, ChIP-exo provides near bp resolution. The overlap spread described by the reviewer, and shown as composite plots, results from gene averaging. Since different genes have different length promoter/NFs, there will be overlap. However, our data points are based on individual promoters, not composites, and so the overlap described by the reviewer is not part of this correlation analysis. Nonetheless, we now take a smaller window (100 bp centered on the TSS), which we estimate provides about 50 bp gap between the closest ends of the divergent PICs. We have also examined only the gene-body antisense reads, which originate on the distal sides of the two PICs, and thus are not subject to the constraints posed by the reviewer. We obtained essentially the same results (see box below). Finally, we now use nascent RNA data (published CRAC data from Libri lab) as the reviewer requests. These not only give confirmatory results, but also demonstrate that the correlation increases when at least two insulator TFs are depleted (Rap1 and Reb1). The correlation is specific to the gene class to which these factors bind.



Correl:	Antisense
RP+STM	0.14
TFO	0.17
UNB	0.39

11. Finally, to truly say anything about the insulating property of Reb1/Abf1, it is necessary to do the experiment where the site is deleted or Reb1/Abf1 are conditionally turned off and it results in loss of insulation. This type of experiment may be beyond the scope of this study, but making mechanistic statements solely based on occupancy patterns and correlations is an overreach.

This has now been done in response to comment #9.

12. The fact that they don't detect TF-cofactor signatures matching that of GTFs like TFIIB or TBP (Fig. 5b) is not surprising because they are centering this at TF motifs which are likely to be upstream of the site of GTF assembly. At any rate, this observation by itself doesn't challenge the idea that TFs binding to upstream motifs direct PIC assembly at the core promoter. This could well be mediated by the Mediator complex, SAGA and other co-factors and indeed that's how it's thought to occur.

There is a misunderstanding here. The analysis was performed by treating TFIID, SAGA, and Mediator equivalently, and designed to detect TF interactions, not PIC-core promoter interactions. We agree with the second point (and this was our point in the manuscript). The point was that we see no evidence for stable TF-TAF/TFIID interactions like we see for SAGA and Mediator. We are quite supportive of the general view that TFs recruit TBP indirectly through cofactors like SAGA.

13. Abstract: "Most Pol II promoters lacked a regulatory region **by design**." And later (p 22), "selectively **designed** into the edge of many NDRs is a TATA box" These are loaded statements and prone to misinterpretation. Designed by who or what?! Perhaps they mean to say that these features have been selected for (by natural selection) to optimize various gene expression outcomes? If so, better to state that rather than use the word design.

Yes, poor word choice on our part. We have now changed this.

14. Presentation: In Fig. 2 and in many other figures like this, they show selected target binding profiles at the chosen features, which could be misleading with regard to what is actually found there. They should show a version of these figures where not just the selected factors, but ALL target profiles are plotted. Yes, it could get ugly but that's the point. Does the DNA replication origin (ACS) truly have only MCM2-7 and ORC1-6 binding there, or do a whole bunch of other factors also bind there to varying extents? Certainly these profiles should be shown for all factors that show any signal above some numerical threshold, rather than selectively plotting only the factors that we know based on the biology. That really is the value of this work in that it allows one to visualize everything at a given locus.

In principle, we agree. However, 400+ traces on a single plot is not practical or helpful to readers. For this reason, we had set up a website (yeastepigenome.org) in which every dataset could be visualized around every individual feature or group of features, as the reviewer indicates. What is exciting about this, is that it allows the reader to explore very easily. The UMAP clusters will allow readers to focus. We also now provide Supplementary Table 3 for each feature and class, along with the enrichment of each target. Sorting by the targets that are most enriched at each feature class provides the objective basis for the selected profiles shown as figures in the manuscript. Specifically for ACS, we did not find anything more than what was indicated. However, we did not canvas replication proteins broadly.

15. In Fig. 2b showing the sub-telomeric factors, it's hard to distinguish which profile corresponds to which factor in the schematic. In general, the schematic depicts a much greater level of spatial resolution than warranted by the data. Many of the indicated factors have molecular weights that don't line up with their sizes in the schematic. There's a significant signal for one (or two) red read density traces on the left side of this region, but nothing is shown on top in the schematic. Is this Tup1 and why are there two similar colored traces? Perhaps a better representation would be to split out each of the factors into its own plot and show a better schematic on top, or at least label every trace as they do in panels a and b.

Underlying patterns for each figure can be found at yeastepigenome.org and at https://github.com/CEGRcode/2020-Rossi_YEP. Schematics were not intended to reflect MW, but of genomic regions of crosslink-ability. We now indicate this in the figure legend. Schematic representation is also challenging in that the linear x-axis of a figure cannot readily accommodate the 3D architecture of the DNA wrap on nucleosomes. However, this can be accommodated in our schematic, showing the wrap of the DNA on the nucleosome and the corresponding Tup1 crosslinks, ~150 away. So there is consistency with where factors are placed in the context of nucleosomes and what is observed by the traces (albeit not linear).

16. Figure 3a. It's not clear again which profile is which factor. The green TBP profile looks like it's a composite of two profiles, both green. Or it could be one of the other factors superimposed on top of the TBP profile, which looks odd if one is supposed to be behind the other. What is the second purple profile in the gene body that is not labeled? Their figures of this type could be significantly improved if they either plot them independently and stacked vertically or use lower alpha (transparency) settings for composite profiles.

We now apply the reviewer's recommendation and revised the panel

17. Much of the mechanistic description of "futile cycles" of PIC assembly in the penultimate paragraph are more than what can be learned from the static snapshot data presented here, and sound like a review of our understanding based on many other studies. This should be clearly separated out and is perhaps more appropriate in the introduction.

We now make this separation. However, it might be more appropriate as a discussion item, as it is meant to be thought provoking.

18. The TF to TF regulatory circuit shown in Fig. 6 is somewhat fanciful because the data only shows binding, but the implication that there are extended cascades of regulation of TFs by one another is not justified. There is much evidence and analysis suggesting that such elaborate cascades likely don't occur in yeast (most recently see PMID: 32060051). Also, the way it's shown with both a gene and a protein represented for every TF with a dotted line connecting the gene to the protein only serves to clutter the figure. The standard way of showing this is much cleaner, with just one name abstracted out in each node (gene/protein doesn't matter because it is understood that you're only measuring binding of a protein to DNA), and edges representing binding to its target(s).

PMID: 32060051 is a great attempt to reconcile binding and regulation. It largely relies on the old Harbison chip-chip data, which the authors of that paper show is inferior to ChIP-exo. TF binding is then compared in the TFKO dataset. Since these are knockout strains that are grown without the TF for generations, there will be indirect effects, and altered cell cycles. Studies (e.g. Holstege) have demonstrated that these indirect effects often drive discordance. Finally, PMID: 32060051 does not take into account multiple TF binding events at promoters, which produces redundancy that masks TFKO effects.

In regards to the display of the diagram, we followed the lead of Harbison et al, and used their symbols of nodes and edges. We agree that the reviewer's method would make it appear cleaner. However, we are concerned that a general audience may see it differently/incorrectly. A simpler node icon might be confusingly interpreted as a protein-protein interaction network, or a transcriptomic interaction network, rather than protein-DNA interactions, which in our case may be unprecedented as of late, but fully in line with Harbison et al.

19. Fig. 3c shows ribosomal protein (RP) genes. But the font for the word "GENE" is so big the E has fallen out of its box in the schematic on top. What are the two grey downward pointing arrows? I'm guessing nucleosome positions, but this is not stated anywhere. In the TES plot, are the distances and scale the same as for the TSS plot? One could assume that, but why not show the numbers like one would for the X-axis in any plot? Finally, in Extended Fig. 4a they state that the middle sub-panel, showing the top 200 coding genes is identical to Fig. 3c, but that's wrong. It should be the top sub-panel, which shows the same RP gene plot.

These issues are now fixed. Distance and scale are the same in the TES plot. This is now stated.

20. Fig. 1f: legend for this panel is completely missing.

It is there, but it got cut off when the figure was imbedded into Word and converted to PDF.

21. Some clarification on numbers. It's stated that 386 targets were replicated and analyzed, while 454 failed threshold, out of the total 840 attempted (Extended Fig. 1). In Fig. 1e and 1f, only 371 targets are shown to cluster, and these 371 are listed in Supplementary Table 3. But even this list of 371 factors includes 78 that are annotated as "Isolated or failed" in either their Meta-assemblage or Sub-assemblage labels. Which of these are failed and why are they not added to the 454 that failed thresholds?

While 386 targets were successful, some had so few binding events that they were not confidently handled by our scripts. Thus, we required at least five peaks to overlap the 8,795 mappable features in the genome we considered. 15 targets (386 – 371) failed to meet this threshold and were excluded from further analysis. Some of these include targets that bind to the repetitive rDNA locus, which for the yeast genome is annotated as having only two copies, and thus cannot reach five peak threshold. Plus, rDNA is inherently noisy making peak calling there untenable, except for the most robust factors. The "Isolated or failed" label was a misnomer. They simply did not cluster with other factors, despite having valid binding. We now correct this labeling in Supplementary Table 3 to "Not clustered".

22. What is the difference between meta-assemblage and sub-assemblage? It's very confusing.

Meta-assemblage is defined by all data via UMAP clustering. As such those assemblages may not exist at any particular gene (equivalent to metagene analysis). We now make this clearer, where first mentioned. "Sub-assemblage" was misnomer, which we now remove.

23. Supplementary Table 4. Is the "Miscellaneous" group included in the K-means cluster counts? According to "20Kmeans" it is not included but according to "40Kmeans" it is included.

Yes, the miscellaneous group was counted. These groups represent the few datasets where the UMAP and K-means clustering lacked sufficient granularity to provide separation. We do not find that they co-localize to the same genomic regions.

24. All misassignments should be called out and commented on. For example, SET2 is assigned to the THO complex though it's not truly part of the THO complex (PMID: 28059701), rather it is expected to be firmly part of the Pol II elongation/SET complex assemblage. Similarly, GCN4 as part of the "cell-cycle regulation" meta assemblage (Supp. Table 3) and RPD3 complex (Ext Data Fig. 1) doesn't really make sense. What is driving these misassignments and how frequent is it?

We don’t think it is appropriate to call out each “mis-assignment”, since these are the fuzzy regions of an unsupervised clustering. The “mis-assignments” are few, but we are unsure what criteria we would use to define them as “mis-assignment”, since it is possible that some a priori knowledge may be missing (do we really know the entirety of what Set2 and Gcn4 do?). Also, binding patterns need not be absolutely linked to function. It would seem possible that two distinct complexes might bind to the same region of the genome but do different things. We do not feel it appropriate to over-ride unsupervised learning by what might be a narrow slice of biology that exists in the literature.

25. Fig. 2: I think it should say "mirrored on the y-axis"?

We now reword this.

26. ACS elements: 252 on p 7 but the figure referred to, Fig. 2a, says it's showing 253 elements.

Fixed. 254 is correct.

27. Abstract: to formed -> to form

Fixed.

Referee #3 (Remarks to the Author):

The authors performed a large number of ChIP-exo experiments in yeast under rich medium conditions. The binding patterns were compared between factors, and clustered based on colocalization to suggest 21 meta-assemblies. The authors then classified promoters into 4 distinct groups based on the factor occupancy patterns and perform comparisons. One may argue this work provides a basis for obtaining the first holistic view of a functional eukaryotic genome. A lot of genome mapping data is already available under these conditions, yet the current study provides more mapped factors and higher resolution. The authors made several observations relevant to understanding gene regulation, for example, that SAGA binding is not concomitant with TFIID-independent PIC assembly and that there is no evidence that TFs direct PIC assembly through stable TFIID or TBP interactions. The study provides a resource for yeast and gene expression biologists. Additionally, the authors explored transcription factor regulatory circuits. However I have difficulties to judge the network part of the manuscript and hope another reviewer can do so. I recommend publication after the authors have addressed the following concerns:

1. The description of ChIP-exo to have base-pair resolution seems misleading. ChIP-exo involves formaldehyde, so cross-linking of entire complexes can occur. Also, in case factors do not bind to DNA directly, but via polymerases, the cross-linking sites are not that of the factor to DNA, but that of the polymerase complex. This is also alluded to by the authors in lines 270-272. Please go through text and make sure this is correctly understood. I recommend to also revise the abstract accordingly.

ChIP-exo “can” provide near-bp resolution. But as the reviewer points out, the nature of large complexes and translocating complexes are intrinsically low resolution and crosslinking can be indirect (a feature we exploit to define interactions). We now have tried to provide clarity so as to not be misleading.

2. The authors should discuss the limitations of ChIP-exo better. For example, some of the factors analyzed bind RNA and a ChIP-exo can then not give a complete picture.

We now do this, emphasizing a DNA-centric perspective.

3. It is unclear how the authors determined colocalization. Was the 100 bp window around the peak midpoint from the entire peak region +50 bp on each side? Please clarify how these analyses were done. This relates to concern 1.

We now provide more explanation in the main section and Methods, that the genome was split up into nonoverlapping 100 bp windows, but the peak was a single coordinate defined by ChExMix.

4. Why is the strand separated data shifted 50 bp in 3` direction before combining? Please clarify how the data was processed for analyses and visualization.

This is relevant to Fig. 3c and ED_Fig. 5, performed for targets located in gene bodies. 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 (i.e., more downstream of the TSS). We surmise that this “double-vision” effect is caused by efficient crosslinking such that the 5’-3’ lambda 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 corrects this double vision and reflects the middle of the complex. The 100 bp size is also consistent with the expected footprint of Pol II (plus the exonuclease headroom on both sides). If we do not perform the shift, then the pattern near the TSS reflects the backend of the Pol II entourage, and the pattern near the TES represents the front end.

5. The authors should compare the expression levels of the 4 promoter classes, i.e. are genes with unbound promoters less expressed or are STM and TFO expression levels comparable?

Expression level follows the following order: RP > STM > TFO > UNB. This was shown in Fig. 1d, based on PIC (Sua7) occupancy. Essentially the same results are obtained when examining nascent transcription or steady-state mRNA.

6. In lines 127-129, the authors say that they did not detect histones at centromeres except Cse4. Please offer explanations considering that the Cse4-containing nucleosome structure exists.

We now provide more explanation, but leave it equivocal. To our knowledge a Cse4-containing nucleosome structure on centromeric DNA (as opposed to 601 DNA) at atomic level resolution remains unresolved, and in our read of the literature, has not been unequivocally demonstrated to exist in vivo.

7. In line 197, the subset of STM promoters (20% of all promoters), on which the major cofactors are present, is defined. This disagrees with literature (Bonnet et al, PMID 25228644; Baptista et al, PMID 28918903) that detected SAGA on almost all transcribed genes. Please comment on this.

We now comment on this. Those authors used the ChEC-seq assay, which has major problems (see *Rossi, M., Lai, W. & Pugh, B. Correspondence: DNA shape is insufficient to explain binding. Nat Commun* ***8,*** *15643 (2017). https://doi.org/10.1038/ncomms15643*). We have also challenged the validity of parts of the Bonnet and Baptista papers, in work that was accepted for publication by Molecular Cell more than one year ago, but currently awaits action by the editor. We now add this as an appendix to this Reviewer Response document. The “signal” they see at the vast majority of promoters is equal to or less than the signal in their negative control, meaning that the ChEC-seq signal at most transcribed genes is just noise. The signal is only significantly above background at about 25% of genes, which agrees with our study.

8. Lines 283-298. If TFIID deficient promoters (RSTM subset, lower PIC/TFIID ratio) are associated with the presence of repressive co-factors (Tup1, RPD3-L) does this mean TFIID independent PIC assembly happens at promoters of repressed genes? Are these promoters active? Please add some discussion.

No, this does not mean that TFIID-independent PIC assembly occurs at repressed genes. Just the opposite. These promoters are in a dynamic state along the induction-repression continuum. So we see both inducing and repressing factors in this bulk population ensemble view. We now clarify this in the discussion, that TBP would act during the promoter-accessible portion of a “futile” cycle of induction/repression.

9. I recommend the authors think of the title. It is not obvious why a ‘genome’ should have a ‘protein architecture’, in particular also since protein architecture does not necessarily mean multiple proteins are involved. I suggest: ‘Comprehensive high resolution protein mapping along a eukaryote genome’

We thank the reviewer for the suggested title changed. We have thought about the title quite a bit. A succinct title can mean different things to different people. For example, the reviewer's use of "protein mapping" might be misunderstood by some readers as annotation of coding sequences. It also does not necessarily mean multiple proteins are involved. "Architecture", as in our macro-world, does imply the use of a range of materials, and thus we thought it most appropriate. Another term we favor is "epigenome", although in our opinion it has been usurped to improperly reflect only histones and DNA modifications, when in fact the genome is regulated by many proteins. As such we can suggest the alternative title: "*A comprehensive high-resolution protein architecture of a Saccharomyces epigenome*", but will leave it to the reviewers and editor to weigh in.

10. Abstract and text: Due to the nature of bulk experiments, I do not think that one can deduce from co-occupancy the existence of 'meta-assemblages'. I think one can only talk about co-localization of proteins and suggest the existence of underlying assemblies, and point to other studies where a corresponding assembly was identified biochemically. The authors mention this in the beginning of the text, but the abstract suggests otherwise and later in the text their warning is largely ignored. Please edit accordingly.

We have now edited this in four places. However, what the reviewer writes is indeed within our definition of meta-assemblages. Due to them being bulk experiments, we cannot call them assemblages (which are biochemically defined), and so call them "meta". "Co-localization" suffers from the same limits as bulk studies. We have now redoubled our efforts not to ignore our stated caveats.

11. There are important references lacking. I understand the list is limited but at least for some key conclusions that have been drawn before the authors should add references. Several important points that the authors make have been made before with the use of ChIP profiling. For example, it is well known from previous yeast factor profiling that Mediator binds to upstream sites and not TSSs as said in lines 166-167 (PMID 27773677), that elongation factors enter at fixed distances from the TSS as said in line 163 (PMID 20818391), and that Pcf11 binds around the TES (the poly-adenylation site for protein-coding genes) as said in line 159 (PMID 28318822).

Additional references are now added.

12. It is unclear to me how the authors decisively conclude that the PIC assembles at the +1 nucleosome only via TFIID. Can it be excluded that other PIC components can bridge to the nucleosome? Please edit accordingly.

We agree. Other interactions are not excluded, although we do not know of other PIC interactions with nucleosomes, in the TFIID-anchored PIC. We have edited the manuscript to address this point.

13. The authors say in paragraph 7 that intronic genes additionally load splicing factors. I am not sure what the evidence is that this is limited to the few percent of intronic genes. Please clarify.

While not entirely visible in ED_Fig. 4, we see representative splicing factors being limited to RP gene bodies. We now clarify this conclusion.

14. Minor comments:
a. Missing panel description for Fig 1f.
This was inadvertently cut off in the pdf.

b. Missing descriptions in Fig 1d for “UNB”, “STM”, “TFO” and “RPG”
Now added.

c. There are very few references for the introduction and sections 1-3.
Six more have now been added to these sections, but the journal has limits on the number of references.

d. Missing reference to results: “In addition to ChIP-exo validation, the high positional concordance of subunits with each other, along with congruence from orthogonal methods, confirmed that epitope-tagging did not functionally alter the targets.”
Now added, although the current manuscript provides sufficient evidence.

e. Fig 3b appears to be an enlarged view of gene in Fig 3a. Please clarify. Also, no color consistency, Fig 3d has different design.
This was confusing to another reviewer as well. One is not an enlarged view of the other, but it created confusion by drawing dashed lines between the two panels. This has now been corrected, along with color consistency

f. In Extend Data Fig. 3b, what is the distinction between the “TFIID” and “TAF” assemblages?
These labels are further described in Supp Table 4. The “TAF” group predominantly contained the scaffold Taf proteins that are found in both the SAGA and TFIID complexes (i.e. Taf12), whereas the “TFIID” group contained the Taf proteins that are specific to the TFIID complex (i.e. Taf1).

Referee #4 (Remarks to the Author):

This manuscript starts with an experimental tour de force: the profiling of more than 800 (TAP-tagged) proteins that are known or hypothesized to be associated with chromatin in rich media conditions using the ChIP-exo assay in yeast. The resulting dataset is a treasure trove that could be used to uncover new functions and mechanisms.

1. The first analysis step in the paper is to cluster the proteins – called “targets” here, somewhat confusingly, presumably because in each strain used a different gene has been TAP-tagged, and consequently is targeted by the antibody – based on their genome-wide profile. During this procedure, the profiles are compared after binning over 100bp windows; in other words, the almost single-base-pair resolution that sets the

ChIP-exo assay apart is not taken full advantage off. This is a potential missed opportunity in terms of uncovering detailed mechanisms.

We looked at this at higher resolution, and found no major changes. This was not a missed opportunity because proteins do not just bind a single or a few bp. They typically engage with 20-100 bp of DNA. Binning in smaller bins just makes the process computationally more demanding, with incremental gain of new information. Certainly, follow-up studies may be able to extract deeper nuances.

2. What follows is a cluster-by-cluster description of what the average spatial profiles of proteins associated with the cluster look like when centered on a related genomic feature. It all reads more like a review paper than a research article, in that many things that are known or expected are shown to emerge from the data. This is nice, but there is little here that is surprising.

Here are some novel and/or surprising observations that we report: 1) The ORC replication complex engulfs an adjacent nucleosome. 2) Centromeres lack nucleosomes. 3) Subtelomeric repressive domains have a highly focused but well defined architecture of proteins encompassing three nucleosomes. 4) Most promoters have not evolved a TF/cofactor promoter architecture and thus are not directly regulated by TFs. 5) A defined TF/cofactor/chromatin/TBP promoter architecture of induced promoters that is quite different from constitutive TFIID-based promoters. It was very surprising to find that TFs do not interact with TFIID, but this now makes sense in light of our new model of gene induction. Finally, a comprehensive TF regulatory network is described, where none has previously existed. We identify the cognate cofactor(s) for each TF – this a remarkable achievement. These observations are novel, and in our opinion create a highly enlightened view of chromatin.

We would hope that the reviewer finds it reassuring that many prior discoveries are confirmed in our study. This places our study on solid ground, particularly when we challenge certain existing paradigms.

3. More problematically in my opinion, there is hardly a single statement that is backed up by statistical analysis. In main text, not a single p-value is reported, even though there are many claims about observed differences. This is unacceptable. One example is on line 74: “resulted in ~400 targets displaying significant binding.” By what criterion?

The 400 targets displayed significant binding based on ChExMix, which was designed for ChIP-exo peak calling. This published software includes robust statistical testing using a large number of negative control datasets (untagged strain), which we now emphasize. The patterns displayed around bound genomic features are all based on robust statistical thresholding by ChExMix, and so it is redundant to apply a second statistical test. Moreover, we now show that the ChIP-exo patterns are quite reproducible among replicates, among subunits of the same complex, and even complexes that interact. But only a low flatline signal is observed for negative controls and non-interacting factors. Finally, we now add more statistical tests throughout.

4. There is also no attempt whatsoever to use the dataset to make testable predictions and then validate these predictions. Again, just as one example, on lines 162-164 the authors “suggest that elongation factors stably enter/exit … gene length”, but do not

seem interested in testing this hypothesis. For an organism for which such an excellent genetic toolbox is available, this is disappointing.

We are quite interested in testing these and other hypothesis suggested by the data. We now offer 4 tests. **1)** We test and validate the concept of NDR vs NFR using in vitro nucleosome reconstitution data (Fig. 4c). **2)** We now directly test and validate the model that insulator TFs (like Reb1, Abf1, and Rap1) uncouple divergent transcription, by examining divergent transcription when Reb1 or Rap1 is depleted (Fig. 4e**). 3)** We tested the hypothesis on elongation factor entry, as mentioned by the reviewer. As shown below, depletion of elongation factors Spt4 or Spt5 results in Pol II stalling exactly in the vicinity where these factors were measured to enter (Fig. R2, below). We have not included this in the manuscript due to it being part of a larger study that is currently under review. **4)** We test the concept that STM-bound or RSTM-bound genes are particularly dependent on TFIID-independent SAGA-regulated TBP delivery (Fig. 5c). First, these genes show very high PIC/TFIID (Sua7/Taf2) ratios with 6 different GTF subunits (ED_Fig.9), demonstrating the robustness of the effect. To test this further, SAGA vs TFIID was depleted, then assayed for selective loss of TBP. Indeed, we found that TBP is preferentially lost at promoters having a high ratio PIC/TFIID (Sua7/Taf2), compared to UNB/TFO promoters, upon SAGA depletion (Fig. R3, below). TBP was lost at both types of promoters upon depletion of Taf1.

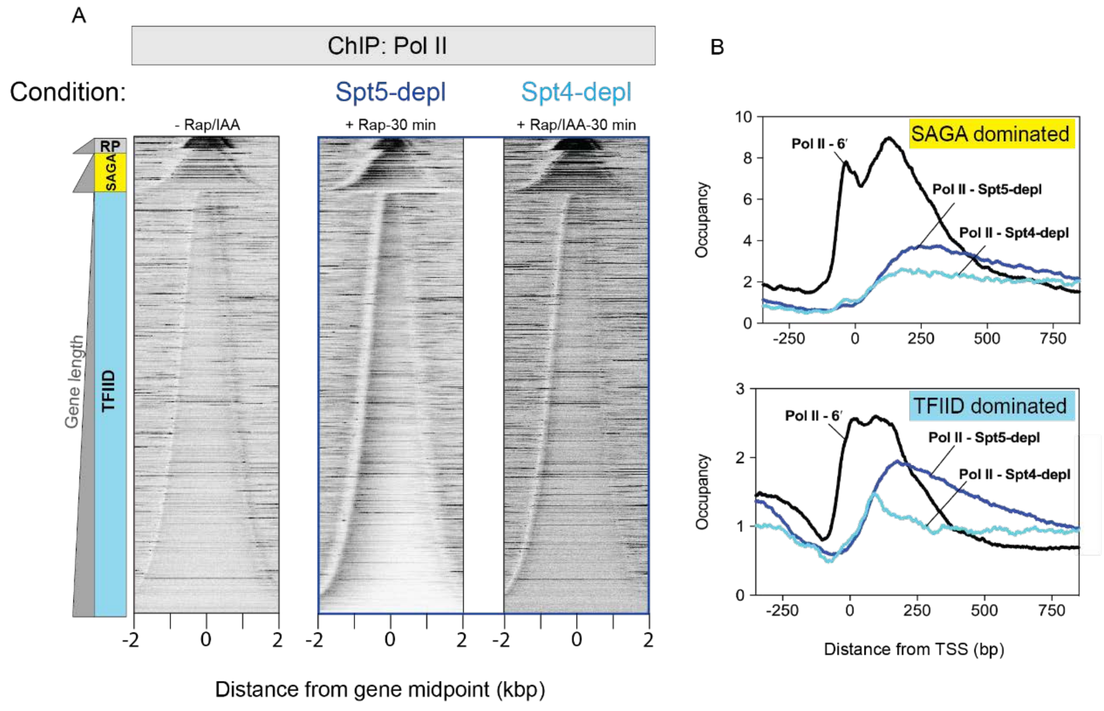


Fig. R2. Demonstration that depleting elongation factors Spt4 and Spt5 results in Pol II stalling downstream of the TSS. Left panels show heat maps centered by gene midpoint (TSS to TES, from left to right). Pol II is distributed across gene bodies in a mock depletion (left-most heatmap). Upon Spt4 or Spt5 depletion, Pol II stalls near the 5' end of genes (shadow on the left side of the "bell"). Right panels provide quantitation, plus a demonstration that Pol II stalls in the same vicinity in response to stress.

[Fig. R3 redacted]

5. With 800 proteins profiled, I would also have expected at least a handful of attributions of specific functions to a particular protein, which could have been tested by genetically perturbing this protein and monitoring a specific aspect of transcriptional regulation using an established reporter system. It is nice to recover many well-characterized protein complexes in unbiased manner, but more is needed to get the paper at the level expected for a top journal in my opinion.

This question is the same as the prior question. We have performed the genetic perturbations suggested (see prior answer).

6. To describe the protein clusters that emerge from the data analysis, a colorful palette of words and terms are used – “architecture”, “meta-assemblage”, “ensemble of assemblages”, “entourage of meta-assemblages” – without clear definition. This feels gratuitous, given that these are just clusters found in the data and there is no attempt to analyze the detailed 3D structure or mechanisms governing these complexes. The coining of new terms for classes of promoters such as “TFO” and “UNB” also feels unnecessary.

We agree in principle. When we started using other existing terms, they ran afoul with their own specific meanings that were not entirely accurate to the specific use cases. We use terms as a succinct catch-all for the outputs from complicated analyses. We have to refer to them in some way. In our assessment, the reading would be become tedious and ambiguous if lengthy and imprecise terms were used, such as simply referring to everything as clusters. We now redoubled our efforts to minimize and justify any new terms. The use of “TFO” and “UNB” is fundamentally no different than calling them “Group 1” and “Group 2”, except that latter are difficult to remember in light of their properties that we report on. The purpose of gene groups and names is that not all genes can be described by a single architectural principle. Instead, the minimal architectural themes we find is four. It seems preferable to use the historically useful names like SAGA$_{dom}$ and TFIID$_{dom}$ genes instead of STM and TFO/UNB, respectively. However, while there is substantial overlap in gene membership, they are not identical. Moreover, the latter memberships more cleanly represent the architectural themes than the former.

Given text limits, we are unable to delve deeply into how known 3D structures relate to ChIP-exo patterning. This had been done previously for the Pol II PIC (Rhee et al. Nature volume 483, pp. 295–301(2012), and for individual insulator TFs (Rossi et al PMID: 29563167). Nevertheless, we now make note of concordance with atomic structures of the ORC complex and the Pol III PIC.

7. The TF network analysis in section 7 in terms of “archetype” motifs also lacks depth. Why is it notable that some TF’s target multiple other TF-encoding genes, and why does this suggest that “they represent major control junctions” (lines 312-

313). There is no statistical analysis, no notion of what the expected occurrence would be under some null hypothesis. Autoregulation of TF genes by the proteins they code is well known.

It is notable because it follows from a highly referenced study (Harbison et al) on TF network analysis. That paper has over 2400 citations, and has had a very large impact on the field. We were in a position to follow up on it in a much more comprehensive manner with very high-resolution and highly comprehensive data.

TFs that target many other TFs would be candidates for major control junctions because of the potential to regulate a large number of key genes. We have now reworded this to state "*have the potential to amplify their control through other TFs*" To our knowledge, this has been axiomatic in the field of gene regulatory networks (and all networks).

We are not clear on the reviewer's notion of a null hypothesis for this network.

Bound promoters were defined by rigorous statistical criteria, and we are simply reporting their interconnectivity. It makes little sense to us to now create a null hypothesis that supposes random binding, which we have already demonstrated is not the case. We are willing to consider applying a statistical test to a null hypothesis, if the reviewer can suggest both. Here is one attempt: The probability that the 22 TFs would randomly bind their own gene (given the specific number of genes they bind) is 1 in $7.2x10^{-40}$. It is not clear to us that such a statistic contributes anything to our understanding.

8. Finally, the blanket statement that since many of the proteins are conserved, the same complexes are expected to form in higher eukaryotes (lines 374-396) is too simplistic. For one thing, the genome sequence is different, and is an equally crucial determinant of chromatin structure.

Agreed. Yet, many of the same fundamentals remain: nucleosome-free promoter regions, positioned nucleosomes, Hsf, SAGA, Mediator, Tup, GTFs, etc. exist in yeast and humans. It would seem remiss not to make such an evident connection.

Nevertheless, we now qualify the statement.

9. In summary, while I have great respect for this group's work in general, and I appreciate the scope and the potential usefulness of the dataset, the present manuscript falls short of expectation by being far too superficial and descriptive.

We respectfully disagree with the study being too superficial. There is enormous depth to be appreciated when working with the Supplementary Data and the thousands of graphs posted on yeastepigenome.org. Only a small fraction of knowledge can be conveyed within a single scientific article. A description of an entire epigenome in such limited space would indeed appear to be superficial. However, for every conclusion, albeit brief, we have conducted deep and thorough analyses to ensure its validity.

While deep mechanistic dives are most commonly done, the intent of this study was to create as broad of a coverage as feasible, with the intent of tying together the many deep studies in the literature. Such deep studies often are limited in their ability to link to other studies because one aspect or another of the experimental methods differ. Here, our goal was to create a resource and unifying concepts that allows linkages across most types of genomic interactions. The results support, but also challenge existing paradigms.

While this study is not the final word on many of the models, it does present the first and most robust holistic context of entire genome assemblages, at a resolution that Nature-published studies like ENCODE have yet to achieve. This is likely to be lost in deeper and isolated analyses that would not be suitable for Nature. We see our work as equivalent to the original draft of the yeast genome DNA sequence. That had enormous impact on the field despite it being incomplete, and lacking mutational analyses that the reviewer is expecting.

Minor comments:

10. - Figure 1b: define “TAP” in caption; antibody binds oddly to TAP tag in schematic Done. The Protein A portion of the TAP tag recognizes the common portion (Fc) of the antibody. The antibody was drawn to show that it is not binding through the antigen recognition portion.

11. - Figure 1e: typo in “Hierarchical” Fixed.

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

Concerns have been addressed.

Referee #2 (Remarks to the Author):

They have done a very job of addressing most of my comments and those of the other reviewers. I'll note that one point that was raised by me and another reviewer was the apparent lack of nucleosome/histone signals at the centromere. The original abstract did not mention it, but the revised abstract states it as a key finding, even though they have presented no new data on this point in the revised manuscript. They may well be correct that the centromere lacks a nucleosome in vivo, but if they want to mention this in the abstract, it's probably worth clarifying that they do observe the histone H3 variant Cse4 but not the other histones at the centromere.

Regarding the title, I had no issues with the original title and much prefer it to the alternative in the revised manuscript.

Referee #3 (Remarks to the Author):

The authors have well revised the manuscript and greatly improved the clarity of the text and figures and have addressed most issues, but there are some discrepancies and inconsistencies that should be clarified before publication (at the discretion of the editor):

Major points:

- Ext. Data Figure 1: The initial version reported 840 attempted ChIP-exo experiments, 454 failing threshold and 386 replicated and analyzed. The revised version reports fever attempted experiments (807), 407 failed threshold and 400 experiments and replicated and analyzed. Why are there more datasets included and how can there now be less experiments attempted?
- Figure 3a: the new figure shows now different, smoother profiles for the factors. Was the plot

produced in a different way than previously? If so, why was this plot generated with different parameters? Were all plots regenerated with different parameters? Please include description.

Minor points:

- Line 234: “The ~1,300 noncoding promoters were similarly classified (Supplementary Data 1E), indicating that they are governed by the same regulatory mechanisms.“. Previous version stated: “All ~2,000 noncoding Pol II PIC/TSS were similarly classified (Supplementary Table 1, Column D), indicating that they are governed by the same regulatory mechanisms as coding genes.”. Please clarify.
- Figure 4d was marked as H3 MNase ChIP-seq, this is now labeled Nucleosomes (MNase). What experiment was actually performed: MNase-seq or ChIP-seq?

Referee #4 (Remarks to the Author):

The authors have provided a highly detailed and thoughtful response to the comments raised by all four reviewers. The revised manuscript does a better job of describing the novelty of the findings. The authors have also made a compelling case that they are following up on this initial study with more in-depth experimental validation, and that it would be unreasonable to expect them to include these analyses in the current manuscript. I am fully satisfied with the revised manuscript.

Author Rebuttals to First Revision:

Dec 11, 2020 Author responses to reviewer comments (Author comments in blue)(Updated comments in red)

Referee #1 (Remarks to the Author)

Concerns have been addressed.

Referee #2 (Remarks to the Author)

They have done a very job of addressing most of my comments and those of the other reviewers. I'll note that one point that was raised by me and another reviewer was the apparent lack of nucleosome/histone signals at the centromere. The original abstract did not mention it, but the revised abstract states it as a key finding, even though they have presented no new data on this point in the revised manuscript. They may well be correct that the centromere lacks a nucleosome in vivo, but if they want to mention this in the abstract, it's probably worth clarifying that they do observe the histone H3 variant Cse4 but not the other histones at the centromere.

Dec 11, 2020: We will add the following to the abstract: “(but contain other centromeric components including histone H3 variant Cse4)”.

Jan 10, 2021: The abstract was shortened, and this new text was removed as a result, as it was deemed unnecessary in the context of the revised abstract.

Regarding the title, I had no issues with the original title and much prefer it to the alternative in the revised manuscript.

We now use the title recommended by the editor: “A high-resolution protein architecture of the budding yeast genome”

Referee #3 (Remarks to the Author)

The authors have well revised the manuscript and greatly improved the clarity of the text and figures and have addressed most issues, but there are some discrepancies and inconsistencies that should be clarified before publication (at the discretion of the editor:

Major points:

• Ext. Data Figure 1: The initial version reported 840 attempted ChIP-exo experiments, 454 failing threshold and 386 replicated and analyzed. The revised version reports fever attempted experiments (807), 407 failed threshold and 400 experiments and replicated and analyzed. Why are there more datasets included and how can there now be less experiments attempted?

The original list of 840 contained targets that failed to produce sufficiently complex libraries (at least 200,000 uniquely mappable tags), which was our threshold for further analysis. In principle, deeper sequencing with additional complexity might have allowed them to be successful. However, based on their GO terms and our experience we did not think this likely. These datasets were not included in our supporting website or in the GEO submission at any time, but had been counted in the tally in the original Ext. Data Fig. 1. In the revised version, we now bring everything into alignment, meaning that we no longer consider low-tag count datasets as being attempted. There should be at least 1207 datasets available to the public: two each of 400 successful targets, and at least one dataset for each of the 407 targets that did not meet our threshold. Some of these 407 may have worked in ways that are unknown to us, and thus we feel they should be made available.

• Figure 3a: the new figure shows now different, smoother profiles for the factors. Was the plot produced in a different way than previously? If so, why was this plot generated with different parameters? Were all plots regenerated with different parameters? Please include description.

Figure 3a, was originally plotted using a different graphing program with less smoothing than other figures. The revised figure replotted all data using with the same program and the same parameters, which was indicated in the figure legend. No conclusions were affected.

Minor points:

• Line 234: “The ~1,300 noncoding promoters were similarly classified (Supplementary Data 1E), indicating that they are governed by the same regulatory mechanisms.“. Previous version stated: “All ~2,000 noncoding Pol II PIC/TSS were similarly classified (Supplementary Table 1, Column D), indicating that they are governed by the same regulatory mechanisms as coding genes.”. Please clarify.

The 2,000 number for noncoding features was an estimate from an earlier version of the manuscript that was not caught before initial submission. As was described in the Methods section under “Excluded ncRNA,” we determined that there was no evidence of the transcription machinery at many of these reported noncoding features, and so deemed them to be false positives. The final list we settled on includes 1,346 ncRNA that have a PIC, and can be found in Supplementary Data 1E.

• Figure 4d was marked as H3 MNase ChIP-seq, this is now labeled Nucleosomes (MNase). What experiment was actually performed: MNase-seq or ChIP-seq?

This is MNase H3 ChIP-seq, and was changed to simplify the annotation. It is now indicated in the figure legend, although the same dataset was used in other panels and was appropriately described as MNase H3 ChIP-seq.

Referee #4 (Remarks to the Author)

The authors have provided a highly detailed and thoughtful response to the comments raised by all four reviewers. The revised manuscript does a better job of describing the novelty of the findings. The authors have also made a compelling case that they are following up on this initial study with more in-depth experimental validation, and that it would be unreasonable to expect them to include these analyses in the current manuscript. I am fully satisfied with the revised manuscript.

Reviewer Reports on the Second Revision:

Reviewer 3 was consulted about the authors’ rebuttal and was satisfied with the responses.