Chromatin-dependent cryptic promoters encode alternative protein isoforms in yeast.

Wu Wei[1,2,3§*], Bianca P. Hennig[4], Jingwen Wang[5], Yujie Zhang[5], Christophe D. Chabbert[6], Sophie H. Adjalley[7], Lars M. Steinmetz[2,3,4§] and Vicent Pelechano[5§*]

[1] CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2] Stanford Genome Technology Center, Stanford University, Palo Alto, CA, USA.
[3] Department of Genetics, School of Medicine, Stanford University, Stanford, CA, USA.
[4] European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany.
[5] SciLifeLab, Department of Microbiology, Tumor and Cell Biology. Karolinska Institutet, Solna, Sweden.
[6] Institute of Molecular Health Sciences, ETH Zurich, 8093 Zürich, Switzerland.
[7] Wellcome Sanger Institute, Hinxton, UK.
* These authors contributed equally to the work.
§ Correspondence may be addressed to: vicente.pelechano.garcia@ki.se, wuwei@picb.ac.cn or larsms@embl.de

Abstract

Cryptic transcription is widespread and generates a heterogeneous group of RNA molecules of unknown function. To improve our understanding of cryptic transcription, we investigated their transcription start site usage, chromatin organization and post-transcriptional consequences in *Saccharomyces cerevisiae*. We show that transcription start sites of chromatin-dependent internal cryptic transcripts resemble those of protein coding genes in terms of DNA sequence, directionality and chromatin accessibility. We define the 5' and 3' boundaries of cryptic transcripts and show that, contrary to RNA degradation-dependent ones, they often overlap with the end of the gene thereby using the canonical polyadenylation site and associate to polyribosomes. In fact, we show that chromatin-dependent cryptic transcripts can be recognized by ribosomes and may produce truncated polypeptides from downstream, in-frame start codons. Our work suggests that a significant fraction of chromatin-dependent internal cryptic promoters are in fact alternative truncated mRNA isoforms. The expression of these chromatin-dependent isoforms is conserved from yeast to human expanding the functional consequences of cryptic transcription and proteome complexity.

Introduction

Genomes are pervasively transcribed, producing a wide diversity of coding and non-coding RNAs (Jensen et al. 2013; Wei et al. 2011; Pelechano and Steinmetz 2013; Kaikkonen and Adelman 2018), raising the question of the biological

significance of such transcriptional activity (Jensen et al. 2013). Some of those transcripts are functionally relevant, such as the well-characterized long non-coding RNAs, antisense transcripts or alternative isoforms (reviewed in (Jensen et al. 2013; Pelechano and Steinmetz 2013; Pelechano 2017; Kaikkonen and Adelman 2018)). However it remains unclear which fraction of those transcripts exert a biological role (direct or regulatory). Such question is particularly difficult to address when these transcriptional units do not have defined or independent genomic boundaries and arise within, or in close proximity to protein coding genes. Among pervasively produced transcripts, the so-called cryptic transcripts constitute a particularly heterogeneous group. Cryptic transcription is typically defined as the production of non-canonical transcripts of unknown function (Wei et al. 2011). The breadth of this definition shows that, despite their abundance and potential relevance for gene expression, our knowledge of this process remains very limited.

Cryptic transcripts can be classified according to the mechanisms by which cells control their abundance: Cryptic transcripts levels may be modulated either by restricting transcription initiation or by selectively degrading them (reviewed in (Jensen et al. 2013)). For simplicity, we will refer to the first class of processes as "chromatin-dependent" and to the second class as "RNA degradation-dependent". A classical example of chromatin-dependent mechanisms is the emergence of cryptic transcripts from within gene bodies when histone deacetylation patterns are disrupted. Specifically, interfering with the activity of the Rpd3S deacetylase complex, which recognizes Histone 3 Lys36 trimethylation (H3K36me3) deposited by the histone methyltransferase Set2 during RNA polymerase II elongation, leads to intragenic cryptic transcription (Carrozza et al. 2005; Lickwar et al. 2009; Churchman and Weissman 2012; Chabbert et al. 2015; Malabat et al. 2015; Kim et al. 2016). Likewise, modulating nucleosome positioning by impairing the function of chromatin remodelers such as Spt6p also leads to the appearance of intragenic cryptic transcripts (Kaplan et al. 2003). In contrast, the second class of cryptic transcripts (RNA degradation-dependent) are constitutively produced and degraded by the cell, and thus become apparent only when RNA degradation is impaired (Jensen et al. 2013). For instance, cryptic unstable transcripts (CUTs) are only detected after depletion of the nuclear RNA exosome (e.g. *rrp6Δ*) (Xu et al. 2009; Neil et al. 2009).

Due to their proximity to or even overlap with protein- coding genes, dissecting the function of cryptic transcription units is especially complicated. In some contexts, cryptic transcription has been assimilated to "opportunistic transcription", whereby RNA polymerase II is recruited to any open chromatin region, generating spurious molecules. However, annotating cryptic transcripts as functional or spurious is not trivial. This has been exemplified in multiple instances where either the RNA product itself or the transcriptional activity per se may have a clear functional impact. For example, the act of transcription itself can regulate the expression of neighboring genes through chromatin modulation (van Werven et al. 2012; Kim et al. 2012; Xu et al. 2011; Brown et al. 2018). On the other hand, previous reports have shown apparent cryptic promoters to drive the expression of alternative isoforms with different post-transcriptional regulation or even to encode alternative protein isoforms (Arribere and Gilbert

2013; Pelechano et al. 2013; Fournier et al. 2012; Lycette et al. 2016; Carlson et al. 1983; Gupta et al. 2014).

To improve the classification of such events and further our understanding of cryptic transcription, we performed a comprehensive characterization of cryptic promoters in *Saccharomyces cerevisiae.* This included the analysis of both the biogenesis of cryptic transcripts and their post-transcriptional life with a focus on those derived from chromatin-dependent mechanisms (*i.e. set2Δ, rco1Δ* and *eaf3Δ*). As a comparison, we also examined the biogenesis of the RNA degradation-dependent CUTs (*rrp6Δ*, RNA degradation dependent). We identified their transcription start sites (TSSs), and demonstrate that their sequence preference and chromatin organization are similar to those of canonical coding-gene promoters. To assess the post-transcriptional life of the cryptic transcripts and better define their boundaries, we examined the association between TSS and polyadenylation site usage. We showed that chromatin-dependent cryptic transcripts encompass the canonical polyadenylation site and thus have the potential to encode N-terminal truncated proteins. To investigate their coding potential, we performed polyribosome fractionation followed by 5' cap sequencing and show that, even in wild-type conditions when only low levels of cryptic transcripts are detected, these often associate with polyribosomes. Finally, examination of the ribosome protection pattern of cryptic transcripts showed that internal methionine codons predicted to act as novel start codons, present a ribosome footprint characteristic of translation initiation. This translation initiation signature is only apparent in conditions where the cryptic transcripts are expressed. All these suggest that a significant fraction of chromatin-dependent cryptic transcripts behave as alternative canonical mRNA isoforms with coding potential.

Results

Chromatin-dependent and nuclear exosome-regulated cryptic transcripts show distinct TSS profiles.

To understand how cryptic transcripts are generated, we performed a genome-wide mapping of their transcription start sites (TSSs) in *Saccharomyces cerevisiae*. 5'cap sequencing (Pelechano et al. 2016), which enables a precise identification of the 5' end of transcripts was conducted in a wild-type strain (BY4741) and in multiple mutants previously associated with cryptic transcription (Fig. 1A). To illustrate chromatin-dependent cryptic transcription, we examined the TSSs profile of cells lacking Set2, the histone methyltransferase responsible for the co-transcriptional deposition of H3K36me3 (Carrozza et al. 2005). We also investigated the TSS profile of strains deficient in Rco1 and Eaf3, components of the Rpd3S histone deacetylase complex acting downstream of Set2. Furthermore we examined the emergence of cryptic TSSs in cells deficient for Set1, the histone methyltransferase responsible for H3K4 methylation and associated with cryptic transcription from promoter-proximal regions (van Werven et al. 2012; Kim et al. 2012). Finally, we conducted a comparative analysis of these TSS profiles with those of the cryptic unstable transcripts (CUTs) that emerge upon depletion of the nuclear RNA exosome subunit Rrp6 (*rrp6Δ*) (Xu et al. 2009; Neil et al. 2009) and exemplify RNA degradation-dependent cryptic transcription.

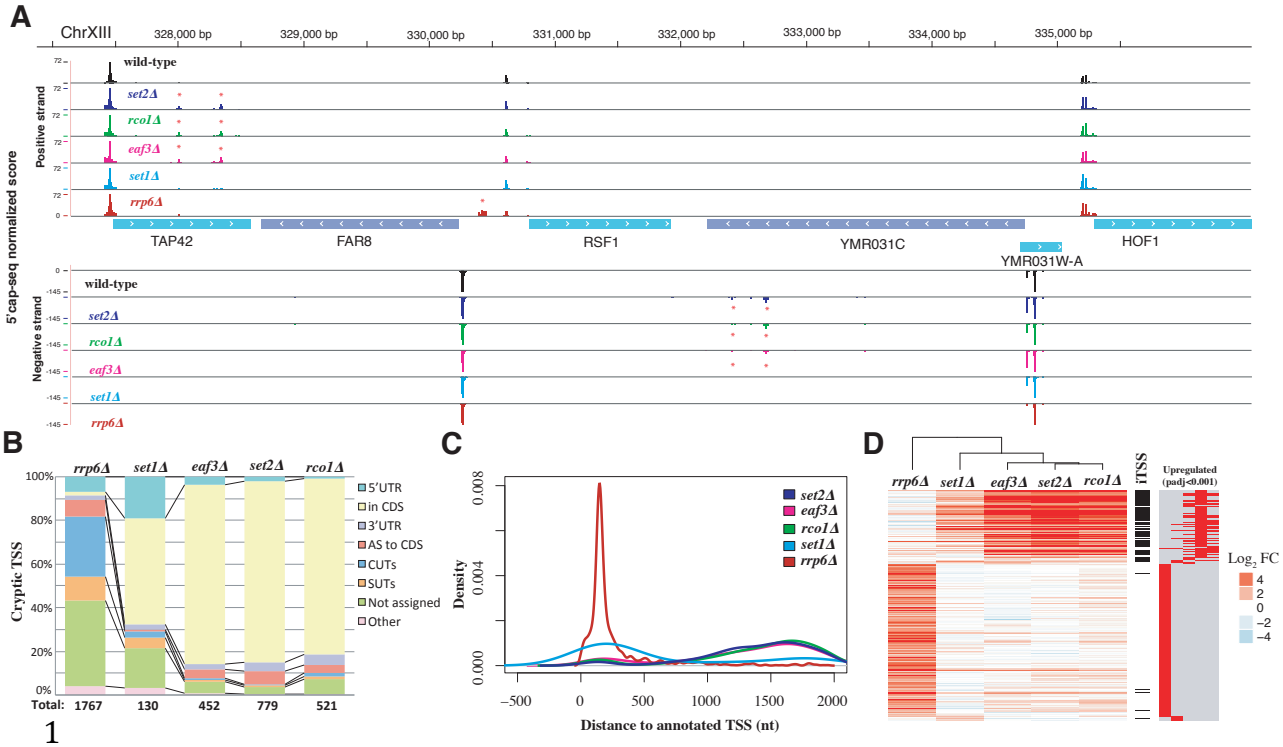


Figure 1. Genome-wide identification of chromatin and RNA-degradation dependent TSSs. (A) Representative 5'cap sequencing track. Score (normalized counts) of collapsed replicates is shown (see methods). Significantly differential expressed TSSs clusters marked by * (p-adj <0.001). (B) Classification of differentially expressed TSSs in respect to annotated features. (C) Distribution of differentially expressed TSSs in respect to annotated ORF-T TSSs. (D) Relationship between TSSs identified in the analyzed strains. Only TSSs identified as differentially expressed with respect to the wild type in at least one condition are shown.

In total, 44963 TSS clusters were identified across all datasets (Supplementary table S1). We used information from our biological replicates and unique molecular identifiers (UMIs) to identify differentially expressed TSSs across strains (adjusted p-val<0.001; see methods) (Fig. 1B and S1A). The disruption of the nuclear exosome (*rrp6Δ*) led to the highest number of up-regulated TSS clusters in comparison to wild-type (1767), while deletion of SET1 had a moderate effect (130 TSS up-regulated clusters). The other mutant strains, *set2Δ, rco1Δ* and *eaf3Δ*, presented an intermediate phenotype (*i.e.* with 779, 521 and 452 up-regulated TSS clusters, respectively). Up-regulated *rrp6Δ*-dependent TSSs were detected in close proximity to the annotated TSSs of coding genes (often in opposite orientation to annotated TSSs (Xu et al. 2009; Neil et al. 2009)) while *set2Δ-, rco1Δ-* and *eaf3Δ*-dependent TSSs occurred preferentially within the body of genes (Fig. 1C and S1B) (Carrozza et al. 2005; Lickwar et al. 2009). Strains with mutations affecting the same pathway (*e.g. set2Δ, rco1Δ* and *eaf3Δ*) shared a high number of up-regulated cryptic TSSs, while cryptic TSSs resulting from disruption of the nuclear exosome (CUTS, *rrp6Δ*) occurred mainly outside of the coding regions (Fig. 1D and S1C). We then characterized the intragenic TSSs (iTSSs) that occur inside the coding region of genes and mostly originate from the Set2-Rco1-Eaf3 pathway (Fig. 1D). Our TSSs detection approach being strand-specific enabled us to determine that most chromatin-dependent cryptic iTSS are expressed in the same orientation as the corresponding ORF. This

contrasts with what is observed for the RNA degradation-dependent ones that arise more often antisense to the CDS than in the same orientation (red vs yellow in Fig. 1B). That may be due to either an increased production or a greater stability of the cryptic transcripts in the same orientation as the main ORF. In fact, chromatin-dependent iTSSs can also be detected, albeit at a much lower level, in wild-type conditions (see below). To gain a better understanding of their regulation, we focused our analysis on the promoters of the chromatin-dependent iTSSs in the same orientation as their overlapping coding gene.

Characterization of cryptic iTSS promoters

After identification of the putative promoter regions with cryptic iTSSs, we compared these with the canonical TSSs of protein-coding genes. iTSSs in all analyzed strains present a similar sequence composition to canonical TSSs, with a pyrimidine enrichment at the -1 and adenine at the 0 position (Pelechano et al. 2013) (Fig. 2A and S2A). We then examined the evolutionary conservation of the iTSSs using the base-by-base conservation scores computed by comparing phylogenetically related yeast species (PhastCons score, UCSC) (Siepel et al. 2005) (Fig. 2B). By definition iTSSs are located within coding regions and therefore expected to be highly conserved. However, chromatin-dependent iTSSs show a higher degree of conservation than RNA degradation–dependent iTSSs , suggesting that the latter are evolutionary more recent. Analysis of total RNA abundance shows a greater fraction of bidirectional iTSSs among the *rrp6Δ*-dependent iTSSs than in those derived from chromatin-dependent processes (Fig. S2B). This is in agreement with recent reports suggesting that the native status of newly evolved promoters is bidirectional (Jin et al. 2017). On the other hand, it is likely that chromatin-dependent iTSSs are under higher evolutionary pressure. It is also important to note that molecules derived from cryptic iTSSs can also be detected in wild-type cells, although at a lower level (Fig. S2B). This suggests that cryptic iTSSs are used by at least a fraction of cells in normal growing conditions.

Given that chromatin-dependent iTSSs resemble canonical gene-coding TSSs in their base composition, evolutionary conservation and directionality, we assessed whether this also applies to their chromatin organization. We used information on nucleosomal and subnucleosomal fractions from previous high-throughput ChIP-Seq experiments (Chabbert et al. 2015; 2018) (Fig. 2C, S2C and S2D) to analyse the MNase protection pattern around cryptic iTSSs. Cryptic iTSSs present the same MNase protection architecture than canonical TSSs, with an organized nucleosome array downstream of the TSS and a sub-nucleosomal protection site overlapping the region where TFs would typically associate (Henikoff et al. 2011)(Fig. 2D and S2E). This is particularly evident for the Set2-Eaf3-Rco1 dependent iTSSs, as they are further away from the canonical TSSs (Fig. 2D). Surprisingly, a similar, although more discrete pattern (*i.e.* nucleosome array and upstream sub-nucleosomal protection pattern) can also be observed around the same iTSSs in the wild-type strain (Fig. 2D and S2E). The presence of a periodical nucleosome organization in gene bodies suggest that iTSSs tend to occur or contribute to synchronizing, regular nucleosome arrays that are detectable even in mixed cell populations. This, together with the detection of a basal level of cryptic iTSS expression (Fig. S2A), suggests that a small proportion

of cells are expressing these cryptic transcripts even under normal conditions. Additionally, our observation that cryptic iTSSs may be bound by TF al low levels even in normal conditions, are in agreement with recent evidence suggesting that TF such as Gcn4 can also bind and activate internal promoters (Rawal et al. 2018).

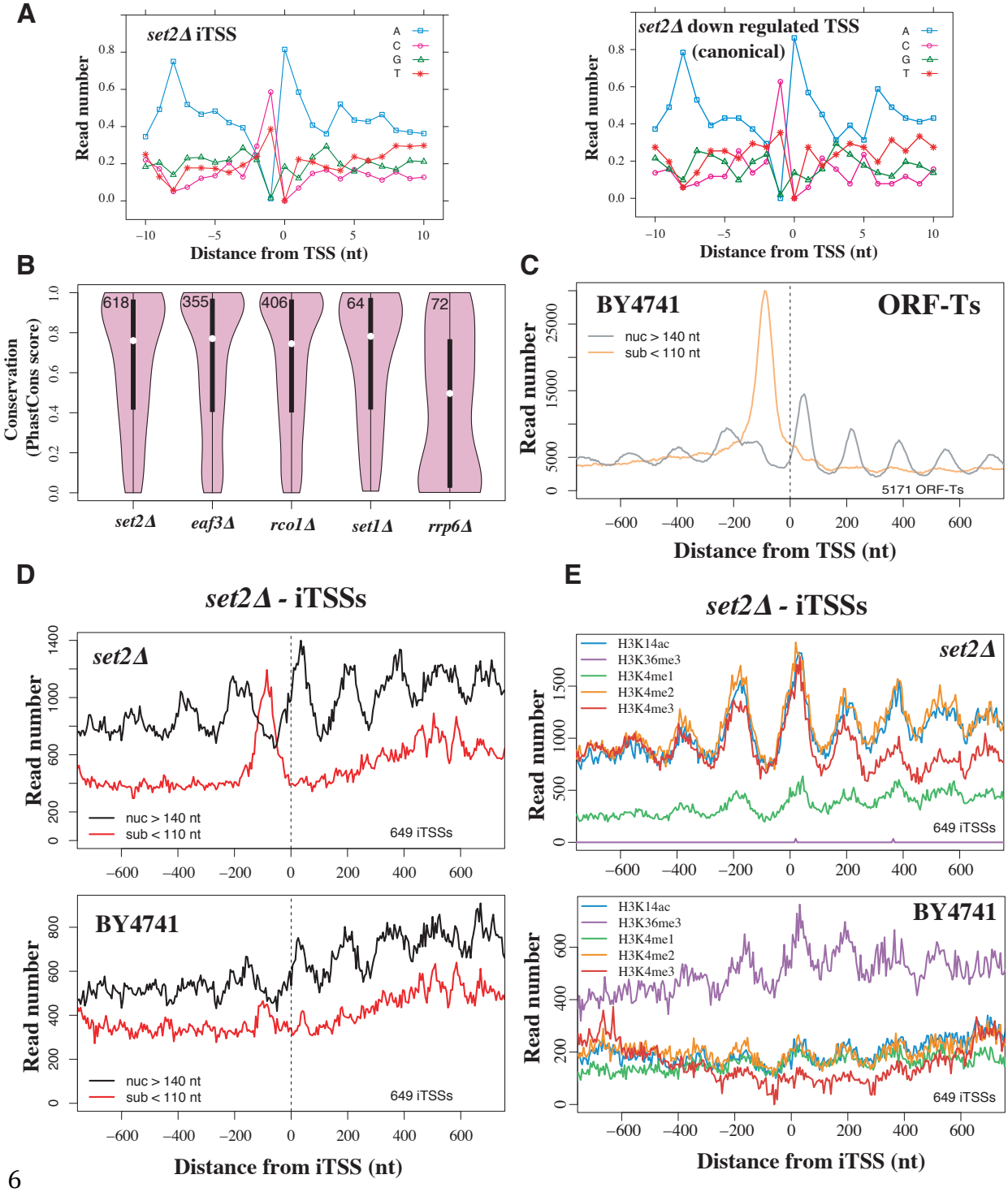


Figure 2. Sequence and chromatin features of iTSSs. (A) Sequence preference of *set2Δ* iTSSs compared to canonical TSSs (*set2Δ* down regulated). (B) Evolutionary conservation of iTSSs. PhastCons conservation score from USCF (Siepel et al. 2005). (C) MNase protection pattern for canonical ORF-T TSSs. MNase fragments are distributed in nucleosome protection fragments (nuc) and sub-nucleosomal ones (sub) according to

their length. (D) MNase protection pattern for *set2Δ* iTSSs. (E) Histone marks distribution profiles around *set2Δ* iTSSs.

To confirm that cryptic iTSSs present the canonical marks associated with promoter activity, we analyzed other chromatin features. We observed an increased signal of H3K4me3 at the first nucleosome (+1 nucleosome) downstream of the iTSSs in *set2Δ* that decreases downstream of the cryptic promoters (Fig. 2E). As expected this is only apparent in this mutant strain as cryptic transcripts are expressed at a sufficient level to be detectable.

Post-transcriptional life of iTSSs derived transcripts

Once confirmed that iTSSs present a canonical promoter structure, we sought to determine the complete length of the transcripts derived from iTSSs in order to gain information on their post-transcriptional life. We applied our previously developed Transcript Isoform Sequencing (TIF-Seq (Pelechano et al. 2013)) approach that allows to jointly and unambiguously determine the start and end sites of each RNA molecule within a sample. By doing so, we identified that most transcripts originating from an iTSS in the *set2Δ* strain use the same polyadenylation site as the canonical mRNAs. This was observed at both individual and genome-wide levels (Fig. 3). Specifically, most transcripts emerging from an iTSS in *et2Δ* originate within the gene body but use the canonical polyadenylation site (Fig. 3A). In contrast, TSSs down-regulated in *set2Δ*, that in the vast majority correspond to canonical mRNA TSSs, generate transcripts that use the previously annotated polyadenylation site (Fig. 3B) (Xu et al. 2009). This suggests that stable chromatin-dependent cryptic transcripts have the potential to encode N-terminal truncated proteins.

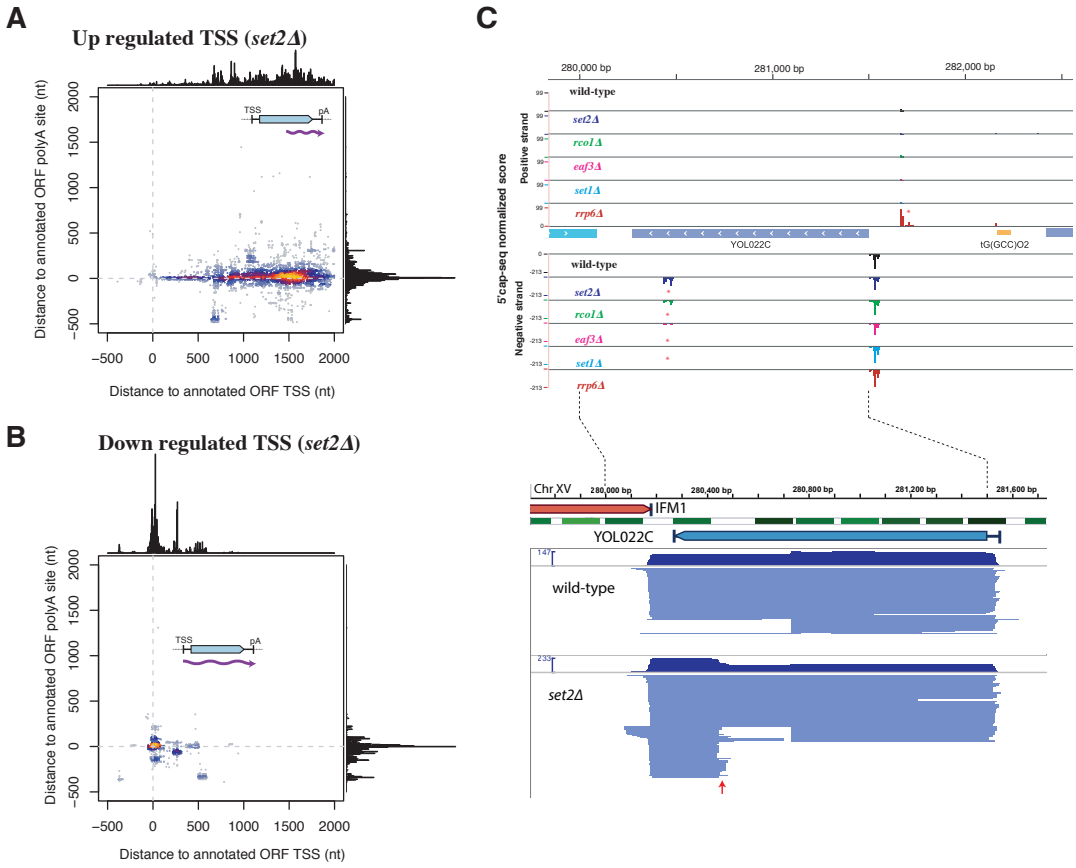


Figure 3. Full length of *set2Δ* iTSSs derived transcripts. (A) TIFSeq reads for *set2Δ* iTSSs derived transcripts with respect to annotated ORF-T boundaries. *set2Δ* iTSSs derived transcripts originate within the body of the gene (internal 5') but use canonical 3' polyadenylations sites. (B) Down-regulated TSSs in *set2Δ* use canonical 5' and 3' sites. (C) Example of TIFSeq coverage for a selected gene.

As most chromatin-dependent cryptic iTSSs produce 5' truncated mRNAs we investigated their ability to associate to ribosomes. This is particularly interesting as those molecules are present at low levels even in wild-type conditions, as would be expected for alternative mRNA isoforms. To measure association with ribosomes, we combined isolation of polyribosomes by sucrose fractionation with 5' cap sequencing (Supplementary Table S2). As expected ORF-Ts TSSs are associated with polyribosome fractions, while non-coding RNAs such as SUTs (Stable Unannotated Transcripts) or CUTs are not (Xu et al. 2009) (Figure 4A). Interestingly, mRNA molecules originating from chromatin-dependent cryptic iTSSs are also enriched in the heavy polyribosome fractions that are associated with active translation. By contrast, RNA degradation-dependent cryptic iTSSs behave as the rest of the CUTs and are not preferentially associated with the polyribosome fraction. This suggests that cryptic transcripts originating from chromatin-depending cryptic promoters associated with ribosomes and may possibly produce truncated proteins.

To assess whether ribosome-bound cryptic transcripts also are engaged in active translation, we assayed the ability of ribosomes to recognize such cryptic transcripts. To this aim, we used our previously developed 5PSeq approach, that measures ribosome dynamics by sequencing co-translational mRNA degradation intermediates (Pelechano et al. 2015; 2016). To distinguish the translation of the canonical full length mRNAs from the shorter overlapping transcripts derived from iTSSs we combined 5PSeq with glucose starvation (Zid and O'Shea 2014). We have previously shown that yeast cells in slow growth conditions such as growth in minimal media or stationary phase present a characteristic ribosome protection pattern at the translation start codon consistent with inhibition of translation initiation (Pelechano et al. 2015; Pelechano and Alepuz 2017). To further enhance this start codon signature, we exposed cells to a glucose-free media for 5 minutes. By limiting translation initiation we increased the start codon signature and allowed the ribosomes engaged in translation *to run-off* the mRNA (Yujie Zhang *et al., in preparation*), an effect that can be readily observed at the canonical start codons of annotated protein-coding genes (Fig. 4B). We then analyzed the ribosome pattern associated with internal methionines and focused on those in-frame that could potentially be recognized as new start codons in transcripts derived from cryptic iTSSs but not in full-length mRNAs. Interestingly, we observed the start-codon signature in the *set2Δ* strain but not in the wild type strain (Fig. 4C). This is particularly striking as even in the *set2Δ* strain, the ribosome protection pattern is a composite of the translation signature of both canonical and iTSSs-derived transcripts. This result suggests that cryptic transcripts are not only associated to polyribosomes, but that ribosomes identify new start codons as canonical ones.

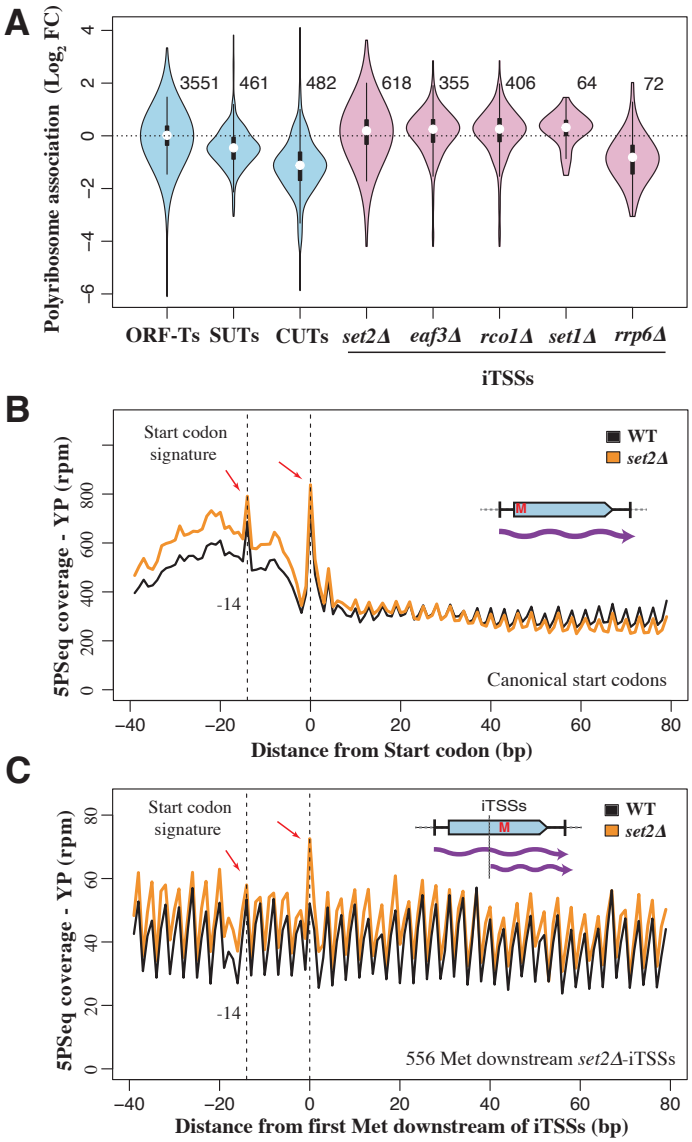


Figure 4. Post-transcriptional life of iTSSs derived transcripts. (A) Relative association with polyribosome fraction after sucrose fractionation versus total extract. Analyzed events (present at a sufficient level in the wild-type strain) are indicated to the right of each plot. (B) Example of 5PSeq start-codon associated signature after glucose depletion for coding genes. (C) Start-codon associated signature after glucose depletion for predicted novel start codons in *set2Δ* iTSSs derived transcripts. Those positions are expected to behave as internal methionines in a wild-type strain.

Discussion

Here, we have shown that chromatin-dependent cryptic promoters present a structure very similar to canonical gene-coding promoters. We focused on *set2Δ*-, *rco1Δ*- and *eaf3Δ*-dependent internal cryptic TSSs, and demonstrated that their DNA sequence, evolutionary conservation, transcription directionality and chromatin organization are similar to those of canonical promoters (Fig. 1 and 2). Our MNase footprint analysis showed that those promoters present a canonical nucleosome array organization and suggested that canonical TFs bind upstream of the iTSSs in the body of genes (Fig. 2). Our observations are in agreement with recent reports that demonstrate how Gcn4 binds frequently in coding regions and can activate transcription from internal promoters (Rawal et

al. 2018; Mittal et al. 2017). This suggests that a significant fraction of the cryptic promoters are in fact alternative promoters, whose expression under standard conditions is restricted by the chromatin organization. Interestingly, previous studies have shown that a significant number of Set2-repressed cryptic promoters can be regulated by carbon sources (Kim et al. 2016). Altogether, this suggests that our classification of cryptic and canonical promoters may be influenced by the environmental conditions under which cells are profiled.

To assess to what degree these cryptic iTSSs could represent *bona fide* alternative transcript isoforms, we investigated their full boundaries (Fig 3). Using our previously developed TIF-Seq approach, we identified that most of them employ the canonical polyadenylation sites used by full-length isoforms. It has been previously described that a significant proportion of iTSSs are targets of Non-sense Mediated Decay (NMD) (Malabat et al. 2015). Here we focused on the molecules that are and thus more likely to have a post-transcriptional effect. We found that, even in a wild type strain, chromatin-dependent iTSSs are typically associated to polyribosome fractions. To further dissect if these short isoforms are not only bound to polyribosomes, but actually translated, we applied an optimized version of our 5PSeq approach (Yujie Zhang *et al., in preparation)* detected even in the presence of an active NMD pathway (Malabat et al. 2015). We identified that the first methionine in the truncated transcripts presents a ribosome protection signature characteristic of translation start sites. In contrast, , this signal is not detected in the wild type strain, in which truncated isoforms are expressed at low levels. Our observation using transcription evidence and ribosome footprint analysis extends on previous observations from our group and others that variations in the transcripts' 5' boundaries potentially leading to truncated proteins are common in yeast. These variations may be environmentally regulated or occur simultaneously in a apparently homogenous population of cells (Pelechano et al. 2013; Fournier et al. 2012; Lycette et al. 2016; Carlson et al. 1983).

N-terminal proteomics approaches showed that downstream in-frame methionines often define alternative amino termini in the budding yeast proteome (Fournier et al. 2012; Lycette et al. 2016). These alternative proteoforms can be detected even in standard laboratory conditions suggesting that their expression coexist with the full-length proteoforms. However, these studies focused their analysis on the transcripts' first 100 nucleotides, and thus did not investigate the downstream truncations that were commonly disregarded as cryptic transcripts. A similar phenomenon has been described in human cells, where alternative N-terminal proteoforms can lead to different protein stability (Gawron et al. 2016; Na et al. 2018). Regardless of their origin, it is clear that truncated proteins can have significant phenotypical impacts such as changes in protein localization (Carlson et al. 1983) or may even act as dominant-negative factors opposing the function of the full-length protein (Ungewitter and Scrable 2010). In the future, further characterizing the abundance and functionality of alternative proteoforms derived from previously considered "cryptic" transcripts will be extremely valuable.

Although we focused our study on budding yeast, our conclusion that chromatin-dependent cryptic iTSSs may act as alternative canonical TSSs have further implications. In mammals, alternative transcription start and termination sites, rather than alternative splicing, accounts for the majority of isoform differences across tissues (Reyes and Huber 2017). This highlights the importance of TSS selection in the definition of the transcriptome. Interestingly, it has been recently reported that the treatment of human cancer cell lines with DNA methyltransferase and histone deacetylase inhibitors (DNMTi and HDACi, respectively) results in the appearance of thousands of unannotated TSSs (TINATs) (Brocks et al. 2017). TINATs frequently splice into coding-protein exons and, in some cases, are associated with polyribosomes. Thus disruption of the epigenome by the DNMTi and HDACitreatments leads to the expression of cryptic TSSs similar to the chromatin-dependent iTSSs defined here, both in terms of biogenesis and potential post-transcriptional consequences. This suggests that the expression of cryptic TSSs is likely to be evolutionary conserved and a source of alternative (functional or aberrant) proteoforms that should be further investigated. The study of chromatin-dependent cryptic promoter regulation will help to better distinguish spurious transcripts from those functionally relevant although only expressed in a subpopulation of cells or under specific environmental conditions.

Methods

Cell growth

All *Saccharomyces cerevisiae* strains used in this study were derived from BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*). BY4741, *rrp6Δ* (*rrp6*::kanMX4), *set2Δ* (*set2*::kanMX4), *rco1Δ* (*rco1*::kanMX4) and *eaf3Δ* (*eaf3*::kanMX4) were obtained from Euroscarf. *set1Δ* (*set1*::kanMX4) was generated using standard yeast chemical transformation as previously described (Chabbert et al. 2015). Cells were grown in YPD (1% yeast extract, 2% peptone, 2% glucose and 40 mg/L adenine) and harvested at OD600 ~1. For 5PSeq start codon identification, cells were shifted for 5 minutes to YP media without glucose (1% yeast extract, 2% peptone) prior to harvesting. Total RNA was phenol extracted using standard methods and contaminant DNA was removed by DNase treatment (Turbo DNA-free kit, Ambion).

5'cap library preparation

Identification of 5'capped mRNAs was performed as previously described (Pelechano et al. 2016). In brief, 10µg total RNA was treated with Calf intestinal alkaline phosphatase (NEB) to remove 5′P from fragmented and non-capped molecules. After purification, mRNA caps were removed using 3.75 units of Cap-Clip (Biozyme) exposing a 5'P in those molecules previously capped. Samples were ligated overnight at 16ºC with a DNA/RNA oligo (rP5_RND: TTTCCCTACACGACGCTCTTCCGATrCrUrNrNrNrNrNrNrNrN) using T4 RNA ligase 1 (New England Biolabs). RNA integrity after ligation was assayed by agarose gel electrophoresis and polyA RNA was purified using oligo dT magnetic beads. After this, ligated mRNA was fragmented at 80ºC for 5 min in the presence of RNA fragmentation buffer (40 mM Tris-acetate, pH 8.1, 100 mM KOAc, 30 mM MgOAc). Ligated RNA was subjected to reverse transcription using random hexamers with Superscript II (Life Technologies) with the following program: 10

min at 25ºC, 50 min at 42ºC and heat inactivated for 15 min at 72ºC. Second strand cDNA synthesis was performed by a single PCR cycle (1 min at 98ºC; 2 min at 50ºC and 15 min at 72ºC) using Phusion High-Fidelity PCR Master Mix (New England Biolabs). A biotinilated oligo (BioNotI-P5-PET: [Btn]TATAGCGGCCGCAATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC ACGACGCTCTTCCGATCT) was added during the generation of the second cDNA strand. Double stranded cDNA was purified using Ampure XP (Beckman Coulter) or HighPrep (Magbio) beads. After the samples were bound to streptavidin coated magnetic beads (M-280 Dynabeads, Life Technologies) and subjected to standard Illumina end-repair, dA addition and adapter ligation was performed as previously described (Pelechano et al. 2016). Libraries were enriched by PCR and sequenced in an Illumina HiSeq 2000 instrument.

TIF-Seq sequencing.

TIF-Seq libraries were performed as described in (Pelechano et al. 2013) using 60 µg of DNA-free total RNA as input. In brief, 5' non-capped molecules were dephosphorylated using 6 units of Shrimp Alkaline Phosphatase (Fermentas). RNA was phenol purified, and the 5'P of capped molecules was exposed by treatment with 5 units of Tobacco Acid Pyrophosphatase (Epicentre). RNA samples were ligated with the TIF-Seq DNA/RNA 5oligo cap using T4 RNA ligase 1 (NEB). Full-length cDNA (FlcDNA) was generated with Superscript III reverse transcriptase and amplified by PCR with HF Phusion MasterMix (Finnzymes). FlcDNA was digested with NotI (NEB) to generate cohesive ends. Samples were subjected to intramolecular ligation using T4DNA ligase. TIF-Seq chimeras were controlled mixing 2 aliquots of differentially barcoded FlcDNA during the ligation, as described in the original TIF-Seq manuscript. Non-circularized molecules were degraded using Exonuclease III and Exonuclease I (NEB). Circularized FlcDNAs was fragmented by sonication using a Covaris S220 (4 min, 20% Duty Cycle, Intensity 5, 200 cycles/burst). Fragmented DNA was purified, and biotin-containing fragments were captured with Streptavidin-conjugated Dynabeads M-280 (Invitrogen). Forked barcoded adapters were added using the standard Illumina DNA-Seq library generation protocols. Libraries were enriched by 20 cycles of PCR Phusion polymerase (Finnzymes). 300 bp libraries were isolated using e-Gel 2% SizeSelect (Invitrogen) and sequenced in an Illumina HiSeq 2000 instrument (105 paired-end sequencing).

Polyribosome fractionation

100 mL of *S. cerevisiae* cells at OD_{600} ~1 were treated with cycloheximide for 5 minutes (100 µg/mL, final concentration), harvested by centrifugation and transferred to ice. Pellets were washed with ice-cold lysis buffer and resuspended in 700µL lysis buffer. Lysis buffer contains 20 mM Tris-HCl, pH 8, 140 mM KCl, 5 mM $MgCl_2$, 0.5 mM DTT, 1% Triton X-100, 100 µg/mL cycloheximide, 500 µg/mL heparin and complete EDTA-free protease inhibitor (1 tablet per 10 mL, Sigma Aldrich). For cell lysis, samples were transferred to pre-cooled 1.5mL screw-tubes with 300 µL glass beads and supplemented with 100 units of RNAse inhibitor (RNAsin plus, Promega). Cells were lysed using a FastPrep-24 shaker (6.0m/s for 15 seconds, MP biomedicals). Supernatant was recovered after 5 minutes centrifugation at 2300g, and cleared with an additional centrifugation at 5900g. Extracts were supplemented with glycerol

(5% final v/v) and stored at -70C. 10-50% sucrose gradients were prepared with a Gradient Master BIOCOMP (Nycomed Pharma). Sucrose solution contains 20 mM Tris-HCl, pH 8, 140 mM KCl, 5 mM $MgCl_2$, 0.5 mM DTT, 100 µg/mL cycloheximide and sucrose (from 10 to 50 %). Cleared cell extracts were ultracentifuged at 34400 rpm for 2 hours 40 minutes at 4C using a C-1000 XP centrifuge with SW40 rotor (Beckman Coulter). Gradient UV absorption at 254 nm was measured and selected fractions were selected for 5'cap library preparation (5µg purified RNA per sample). Polyribosome fraction (*i.e.,* 2n+) was compared with the total extract prior to fractionation).

5PSeq

5PSeq libraries were prepared as previously described (Pelechano et al. 2015; 2016). 5PSeq protocol is the same as the one described for 5'cap sequencing (see above) with variations only for the RNA ligation and rRNA depletion. Specifically, 6 µg of total RNA were directly ligated with a DNA/RNA oligo (rP5_RND). In that way only molecules with a 5'P in the original sample are ligated. Ribosomal RNAs were depleted using Ribo-Zero Magnetic Gold Kit (Illumina) . Samples were sequenced in an Illumina NextSeq 500 instrument.

Bioinformatic analysis.

For 5' cap sequencing reads, random barcodes were first extracted and added to the reads name. The reads were aligned to yeast genome (S. cerevisiae genome (SGD R64- 1-1; sacCer3) with Novoalign (http://www.novocraft.com) using default setting. A customized script adapted from UMI-tools (Smith et al. 2017) was used for removing PCR duplicates. Specifically we allowed 1 bp shifting at the beginning of 5' ends. CAGEr was employed for clustering the 5' cap TSSs of BY4741 wild-type strain and the mutants (Haberle et al. 2015). TSS counts in different samples were normalized to match a common reference power-law distribution. Every sample was normalized to a referent power-law distribution. Low-fidelity tags supported by less than 2 normalized counts in all samples were filtered out before clustering. In each sample, neighboring tags within 20 bp were spatially clustered into larger tag clusters. If the tag clusters were within 10 bp apart, they were aggregated together into non-overlapping consensus clusters across all samples. The raw expression counts of the consensus clusters were further exported to the DESeq2 (Love et al. 2014) for differential expression analysis, comparing between mutants and wild-type strain. Polyribosome derived 5' cap sequencing reads were assigned to the consensus clusters by featureCounts (Liao et al. 2014), with read counting based on the 5' most base. Differential expression analysis of polyribosome fractionation against total extract was performed using DESeq2.

Bar-ChIP sequencing data were processed as described previously (Chabbert et al. 2015).

TIF-Seq sequencing data were processed as described previously (Pelechano et al. 2013). In general, all reads were first de-multiplexed and random barcodes were extracted. Pairs of transcript 5' and 3' end reads were mapped to yeast genome (S. cerevisiae genome (SGD R64- 1-1; sacCer3) with Novoalign (http://www.novocraft.com) using default setting separately. Only transcripts

with both ends mapped in same chromosome a length ranging from 40 to 5,000bp were used for further analysis.

5PSeq reads were mapped to the *S. cerevisiae* (genome R64-1-1) using STAR 2.5.3a (Dobin et al. 2013) with default parameters except AlignIntronMax (2500). PCR duplicates were removed as described for 5' cap sequencing. Reads were aligned to either the start codon, or the first in-frame methionine downstream of *set2Δ*-specific iTSS.

Data availability

All newly generated data are accessible at GEO with the following access codes. 5'cap data: GSE119114; *set2Δ* TIF-Seq: GSE119160; and 5PSeq data: GSE118758. Our previously produced Bar-ChIP data (Chabbert et al. 2015) are accessible at ENA (accession number ERP007035).

Acknowledgments

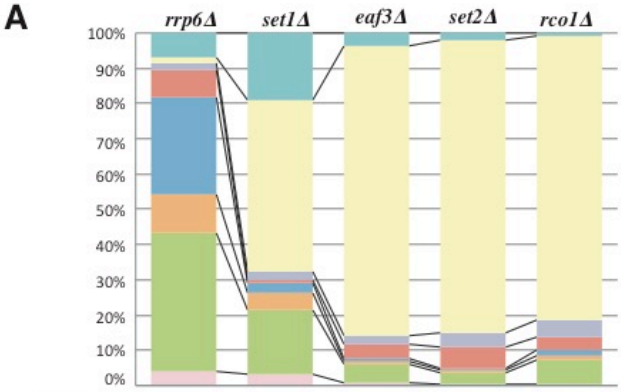
We want to thank all members of Steinmetz, Pelechano and Wei laboratories for discussion. We thank Petra Jakob, Manu Tekkedil, Sandra Clauder-Münster and Yerma Pareja for technical assistance. We thank Bruno Galy for his help with polyribosome fractionation. This study was technically supported by the European Molecular Biology Laboratory (EMBL) Genomics Core Facility and the Science for Life Laboratory (Sweden). This study was financially supported the National Key R&D Program of China (2017YFC0908405) to W.W.; the US National Institutes of Health (NIH grant P01 HG000205), Deutsche Forschungsgemeinschaft (1422/4-1) and a European Research Council Advanced Investigator Grant to L.M.S.; and by the Swedish Research Council (VR 2016-01842), a Wallenberg Academy Fellowship (KAW 2016.0123), the Swedish Foundations' Starting Grant (Ragnar Söderberg Foundation) and Karolinska Institutet (SciLifeLab Fellowship, SFO) to V.P.; B.P.H. and S.H.A were supported by a fellowship from the EMBL Interdisciplinary Postdoctoral (EIPOD) programme under Marie Sklodowska-Curie Actions COFUND (grant number 291772). Y.Z. is funded by a fellowship from the China Scholarship Council. C.D.C was supported by a PhD fellowship from the Boehringer Ingelheim Fonds.

Additional information

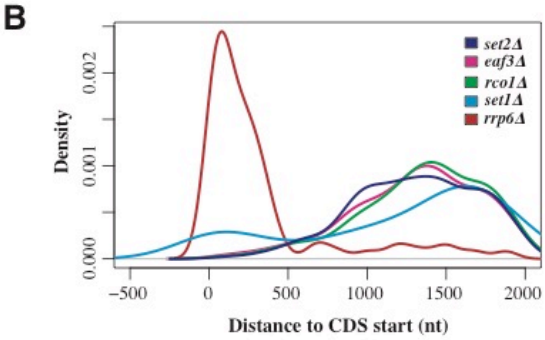
Author contribution

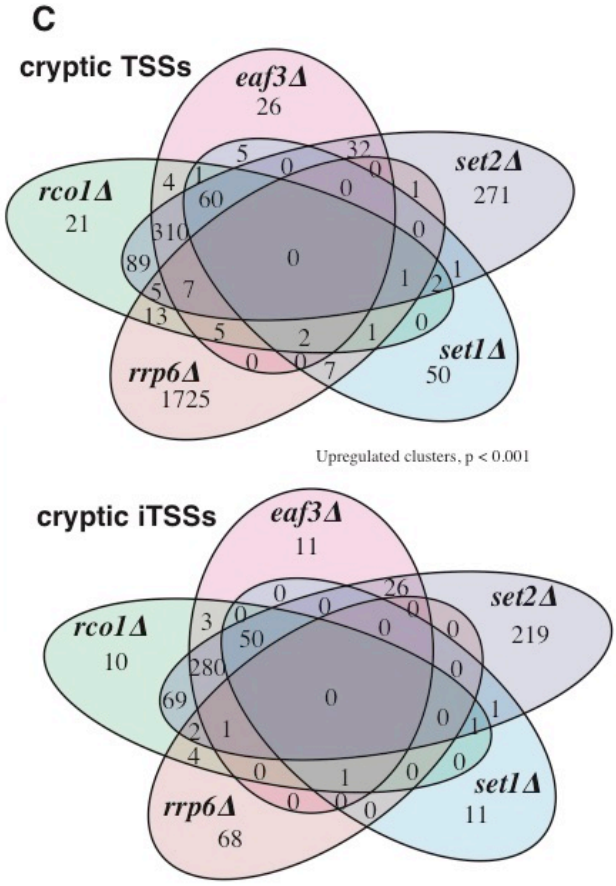
Conceptualization: VP, WW & LSM; Acquisition of data and interpretation: VP, BPH, YZ, CDC & SHA; Computational analysis and interpretation: WW, VP, JW & CDC; Writing – original draft preparation: VP & WW; Writing – review & editing: VP, WW, BPH, JW, YZ, CDC, SHA & LSM; Supervision: VP & LSM; Funding acquisition: LSM, VP & WW.

Supplementary information

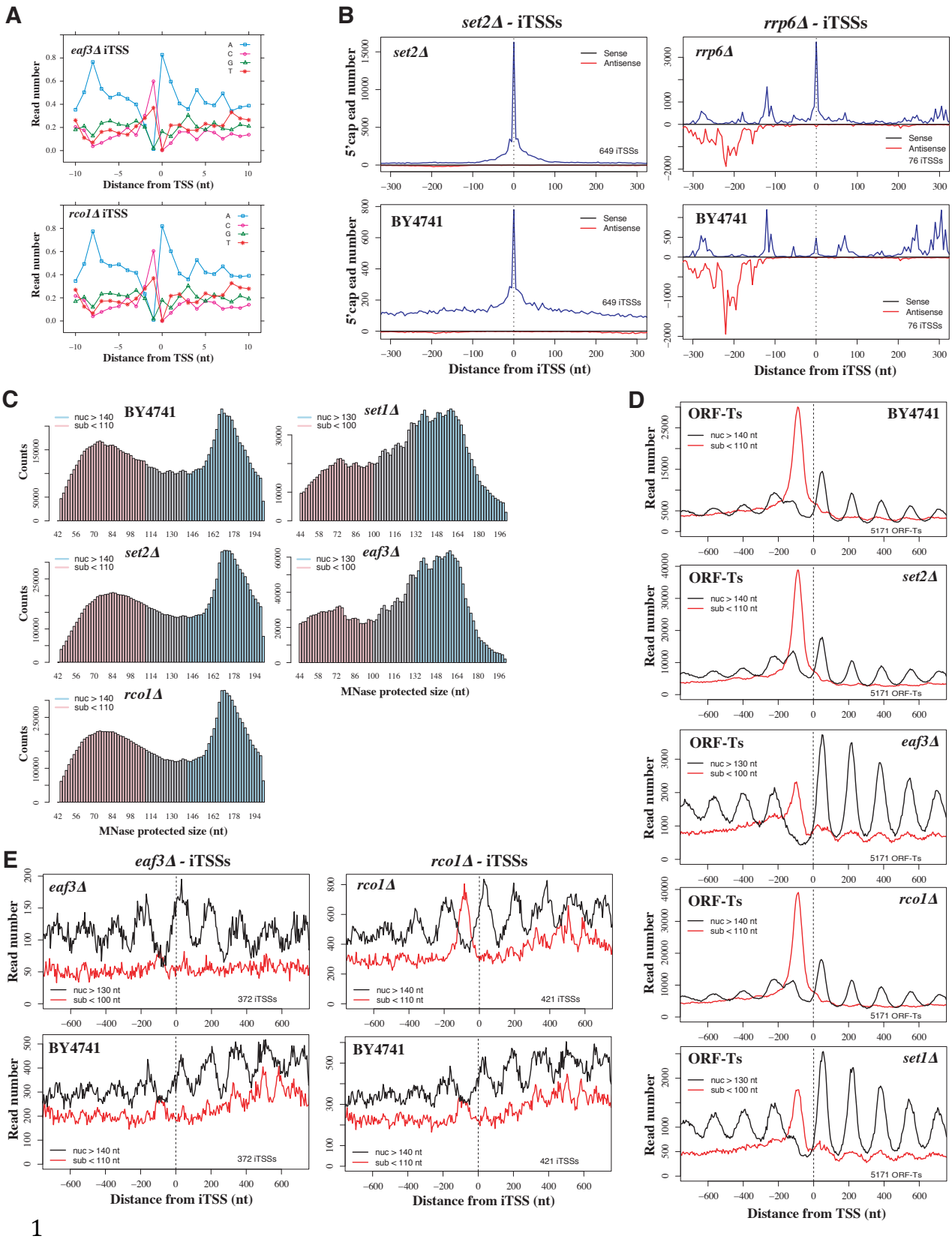


	rrp6Δ	set1Δ	eaf3Δ	set2Δ	rco1Δ
Total:	1767	130	452	779	521
5'UTR	120	25	17	15	5
in CDS	29	63	371	647	419
3'UTR	37	3	11	33	25
AS to CDS	135	1	18	46	20
CUTs	490	4	4	4	7
SUTs	192	6	4	6	6
Not assigned	689	24	23	25	36
Other	75	4	4	3	3





Supplementary Figure S1. Relationship between TSSs across studied strains. (A) Classification of differentially expressed TSSs in respect to annotated features. (B) Distribution of differentially expressed TSSs with respect to annotated start codons. (C) Relationship between TSSs identified across strains. Analysis is shown for all TSSs (upper part) or those occurring inside coding regions in the same orientation (iTSSs, bottom part).



Supplementary Figure S2. Sequence and chromatin features of iTSSs. (A) Sequence preference of *eaf3Δ* and *rco1Δ* iTSSs. (B) 5'cap sequencing reads with respect to the identified iTSSs. (C) Size distribution of analyzed MNase protected fragments. Selected nucleosome protection fragments (nuc) and sub-nucleosomal fragments (sub) are depicted in blue and pink respectively. (D) MNase protection pattern for canonical ORF-T TSSs in different strains. (E) MNase protection pattern for *eaf3Δ* and *rco1Δ* iTSSs.

Supplementary Table S1. Text file with all identified TSSs and their fold change expression versus wild-type as measured by 5'cap sequencing.

Supplementary Table S2. Text file with differential expression for the presence of 5' cap molecules in polyribosome fractions with respect the total extract. Experiment were performed in the wild-type strain (BY4741).

Bibliography

Arribere JA, Gilbert WV. 2013. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res* **23**: 977–987.

Brocks D, Schmidt CR, Daskalakis M, Jang H-S, Shah NM, Li D, Li J, Zhang B, Hou Y, Laudato S, et al. 2017. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat Genet* **49**: 1052–1060.

Brown T, Howe FS, Murray SC, Wouters M, Lorenz P, Seward E, Rata S, Angel A, Mellor J. 2018. Antisense transcription-dependent chromatin signature modulates sense transcript dynamics. *Mol Syst Biol* **14**: e8007.

Carlson M, Taussig R, Kustu S, Botstein D. 1983. The secreted form of invertase in Saccharomyces cerevisiae is synthesized from mRNA encoding a signal sequence. *Mol Cell Biol* **3**: 439–447.

Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, Shia W-J, Anderson S, Yates J, Washburn MP, et al. 2005. Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123**: 581–592.

Chabbert CD, Adjalley SH, Klaus B, Fritsch ES, Gupta I, Pelechano V, Steinmetz LM. 2015. A high-throughput ChIP-Seq for large-scale chromatin studies. *Mol Syst Biol* **11**: 777.

Chabbert CD, Adjalley SH, Steinmetz LM, Pelechano V. 2018. Multiplexed ChIP-Seq Using Direct Nucleosome Barcoding: A Tool for High-Throughput Chromatin Analysis. *Methods Mol Biol* **1689**: 177–194.

Churchman LS, Weissman JS. 2012. Native Elongating Transcript Sequencing (NET-seq). *Curr Protoc Mol Biol* **Chapter 4**: Unit4.14–17.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Fournier CT, Cherny JJ, Truncali K, Robbins Pianka A, Lin MS, Krizanc D, Weir M. 2012. Amino Termini of Many Yeast Proteins Map to Downstream Start Codons. *J Proteome Res* **11**: –5719.

Gawron D, Ndah E, Gevaert K, Van Damme P. 2016. Positional proteomics reveals differences in N-terminal proteoform stability. *Mol Syst Biol* **12**: 858.

Gupta I, Clauder-Münster S, Klaus B, Järvelin AI, Aiyar RS, Benes V, Wilkening S, Huber W, Pelechano V, Steinmetz LM. 2014. Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. *Mol Syst Biol* **10**: 719–719.

Haberle V, Forrest ARR, Hayashizaki Y, Carninci P, Lenhard B. 2015. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res* 1–11.

Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S. 2011. Epigenome characterization at single base-pair resolution. *Proc Natl Acad Sci USA* **108**: –18323.

Jensen TH, Jacquier A, Libri D. 2013. Dealing with pervasive transcription. *Mol Cell* **52**: 473–484.

Jin Y, Eser U, Struhl K, Churchman LS. 2017. The Ground State and Evolution of Promoter Region Directionality. *Cell* **170**: 1–21.

Kaikkonen MU, Adelman K. 2018. Emerging Roles of Non-Coding RNA Transcription. *Trends Biochem Sci*.

Kaplan CD, Laprade L, Winston F. 2003. Transcription elongation factors repress transcription initiation from cryptic sites. *Science* **301**: 1096–1099.

Kim JH, Lee BB, Oh YM, Zhu C, Steinmetz LM, Lee Y, Kim WK, Lee SB, Buratowski S, Kim T. 2016. Modulation of mRNA and lncRNA expression dynamics by the Set2-Rpd3S pathway. *Nat Commun* **7**: 13534.

Kim T, Xu Z, Clauder-Münster S, Steinmetz LM, Buratowski S. 2012. Set3 HDAC Mediates Effects of Overlapping Noncoding Transcription on Gene Induction Kinetics. *Cell* **150**: 1158–1169.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.

Lickwar CR, Rao B, Shabalin AA, Nobel AB, Strahl BD, Lieb JD. 2009. The Set2/Rpd3S pathway suppresses cryptic transcription without regard to gene length or transcription frequency ed. L. Rusche. *PLoS ONE* **4**: e4886.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lycette BE, Glickman JW, Roth SJ, Cram AE, Kim TH, Krizanc D, Weir MP. 2016. N-Terminal Peptide Detection with Optimized Peptide-Spectrum Matching and Streamlined Sequence Libraries. *J Proteome Res* **15**: 2891–2899.

Malabat C, Feuerbach F, Ma L, Saveanu C, Jacquier A. 2015. Quality control of transcription start site selection by nonsense-mediated-mRNA decay. *Elife* **4**.

Mittal N, Guimaraes JC, Gross T, Schmidt A, Vina-Vilaseca A, Nedialkova DD, Aeschimann F, Leidel SA, Spang A, Zavolan M. 2017. The Gcn4 transcription factor reduces protein synthesis capacity and extends yeast lifespan. *Nat Commun* **8**: 457.

Na CH, Barbhuiya MA, Kim M-S, Verbruggen S, Eacker SM, Pletnikova O, Troncoso JC, Halushka MK, Menschaert G, Overall CM, et al. 2018. Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. *Genome Res* **28**: 25–36.

Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A. 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457**: 1038–1042.

Pelechano V. 2017. From transcriptional complexity to cellular phenotypes: Lessons from yeast. *Yeast* **34**: 475–482.

Pelechano V, Alepuz P. 2017. eIF5A facilitates translation termination globally and promotes the elongation of many non polyproline-specific tripeptide sequences. *Nucleic Acids Res*.

Pelechano V, Steinmetz LM. 2013. Gene regulation by antisense transcription. *Nat Rev Genet* **14**: 880–893.

Pelechano V, Wei W, Steinmetz LM. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**: 127–131.

Pelechano V, Wei W, Steinmetz LM. 2016. Genome-wide quantification of 5'-phosphorylated mRNA degradation intermediates for analysis of ribosome dynamics. *Nat Protoc* **11**: 359–376.

Pelechano V, Wei W, Steinmetz LM. 2015. Widespread Co-translational RNA Decay Reveals Ribosome Dynamics. *Cell* **161**: 1400–1412.

Rawal Y, Chereji RV, Valabhoju V, Qiu H, Ocampo J, Clark DJ, Hinnebusch AG. 2018. Gcn4 Binding in Coding Regions Can Activate Internal and Canonical 5' Promoters in Yeast. *Mol Cell* **70**: 297–311.e4.

Reyes A, Huber W. 2017. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res* **46**: 582–592.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. **15**: 1034–1050.

Smith T, Heger A, Sudbery I. 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. **27**: 491–

499.

Ungewitter E, Scrable H. 2010. Delta40p53 controls the switch from pluripotency to differentiation by regulating IGF signaling in ESCs. *Genes Dev* **24**: 2408–2419.

van Werven FJ, Neuert G, Hendrick N, Lardenois A, Buratowski S, van Oudenaarden A, Primig M, Amon A. 2012. Transcription of Two Long Noncoding RNAs Mediates Mating-Type Control of Gametogenesis in Budding Yeast. *Cell* **150**: –1181.

Wei W, Pelechano V, Järvelin AI, Steinmetz LM. 2011. Functional consequences of bidirectional promoters. *Trends Genet* **27**: 267–276.

Xu Z, Wei W, Gagneur J, Clauder-Münster S, Smolik M, Huber W, Steinmetz LM. 2011. Antisense expression increases gene expression variability and locus interdependency. *Mol Syst Biol* **7**: 468–468.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**: 1033–1037.

Zid BM, O'Shea EK. 2014. Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast. *Nature* **514**: –121.