Université de Montréal

Développement d'outils bio-informatiques pour l'étude de la transcription cryptique

par Nicole Uwimana

Département de biologie moléculaire
Faculté de Médecine

Mémoire présenté
en vue de l'obtention du grade de maitrise en biologie moléculaire

Août 2016

Résumé

Les expériences de séquençage à haut débit ont permis de démontrer que la transcription ne se limite pas aux régions codantes et qu'une grande partie du génome est transcrite en ARN non-codants (ARNnc). Parmi eux, les transcrits cryptiques sont initiés à l'intérieur des régions codantes. Des études faites chez la levure *Saccharomyces cerevisiae*, ont pu identifier plusieurs facteurs qui répriment la transcription cryptique. Un de ces facteurs est Spt6, une chaperonne d'histones requise pour le maintien d'un bon niveau de nucléosomes le long des gènes transcrits. Lorsque Spt6 est muté, on observe une déplétion des nucléosomes conduisant à l'activation des promoteurs cryptiques. Cependant, le mécanisme par lequel ces transcrits cryptiques sont régulés n'est pas encore clair.

Dans ce mémoire, nous présentons un travail dans lequel nous avons développé une méthode probabiliste dans le but de caractériser les transcrits cryptiques à partir de données de RNA-Seq. Cette méthode est basée sur une cumulation des données et permet de tenir compte des variations dans l'expression et dans la longueur des gènes, grâce à une étape de randomisation des données. Les résultats démontrent que notre méthode est au moins aussi efficace que les méthodes précédemment décrites dans la littérature et offre un bon compromis entre le taux de faux positifs et de faux négatifs. Enfin, le plus important est que cette méthode permet de prédire les régions génomiques où les transcrits cryptiques sont initiés.

Nous avons mis en évidence la présence de transcrits cryptiques sur les brins sens et antisens par rapport au gène. Nous avons également montré que les promoteurs cryptiques sens et antisens sont enrichis en motif TATA et que les transcrits cryptiques sont polyadénylés, ce qui suggère qu'ils peuvent être régulés par les mêmes mécanismes qui régulent les gènes. Alors que les transcrits cryptiques sur le brin sens se terminent à la même position que les gènes dont ils sont issus, les transcrits cryptiques sur le brin antisens terminent préférablement aux extrémités 3' des gènes situés en amont. Nous proposons donc que les terminateurs chez *S. cerevisiae* ont évolué pour terminer la transcription de manière bidirectionnelle afin d'empêcher une transcription aberrante qui pourrait envahir les gènes voisins.

Mots-clés : Transcription cryptique, Terminateurs bidirectionnels, RNA-Seq, Spt6.

Abstract

High throughout sequencing experiments have shown that transcription in not limited to coding regions and that most of the genome is transcribed into non-coding RNA (ncRNA). Among them, cryptic transcripts are aberrantly initiated from within the coding regions. Several studies in *Saccharomyces cerevisiae* have identified many factors that suppress cryptic transcription. One such factor is Spt6, a histone chaperone required for maintaining appropriate nucleosome levels on transcribed genes. In Spt6 mutant cells, nucleosomes are depleted, leading to activation of cryptic promoters. However, the mechanism by which these cryptic transcripts are regulated remains unclear.

In this thesis, we present the development of a probabilistic method for the characterization of cryptic transcripts from RNA-Seq data. The method is used to characterize cryptic transcription in *spt6*-1004 cells. The method is based on a cumulative distribution function, thus taking into account variations in gene expression and gene length thanks to a data randomization step. Results show that our method is at least as good as previously published methods and provides a good compromise between false positives and false negatives. Importantly, this method allows for the prediction of genomic regions where cryptic transcripts are initiated.

We have demonstrated the presence of cryptic transcripts running on the sense and antisense strands relative to genes. We also showed that, both sense and antisense cryptic promoters are enriched for TATA-like sequences and that cryptic transcripts are polyadenylated, suggesting that they may be regulated by the same mechanism that occurs on genes. While the cryptic transcripts on the sense strand terminate at the same position as the genes from which they are derived, cryptic transcripts on the antisense strand preferentially terminate at the 3'-end of upstream genes. We therefore propose that *S. cerevisiae* terminators have evolved to terminate transcription bidirectionally in order to prevent an aberrant transcription that could invade neighboring genes.

Keywords: Cryptic transcription, bidirectional terminators, RNA-Seq, Spt6.

Table des matières

Liste des figures

Liste des abréviations

3’-Long-SAGE : 3’ long serial analysis of gene expression

ADN : Acide désoxyribonucléique

ADNc ; cDNA : Acide desoxyribonucléique complémentaire

ARN : Acide ribonucléique

ARN Pol II : ARN polymérase II

ARNnc ; ncRNA : Acide ribonucléique non-codant

ChIP : Immunoprécipitation de la chromatine

ChIP-chip : Immunoprécipation de la chromatine sur puces à ADN

cTSS : Sites d’initiation de la transcription cryptique

CUTs : Cryptic Unstable Transcripts

dTTP : Thymidine triphosphate

dUTP : Deoxyuridine triphosphate

FACT : Facilitates chromatin transcription

FPM : Fragments per million base pair

H2Bub : monoubiquitinated histone H2B

NNS : Nrd1/Nab3/Sen1

Pb : Paires de bases

SPT : Suppressor of Ty

SUTs : Stable Unannotated Transcripts

TFIIS : Transcription factor S-II

TSS : Sites d’initiation de la transcription

TTS : Sites de terminaison de la transcription

WT : Cellules sauvage

XUTs : Xrn1- sensitive Unstable Transcripts

Remerciements

Je désire remercier les membres du laboratoire de chromatine et expression du génome à l'institut de recherches cliniques de Montréal. Je leurs remercie d'avoir rendu mes années de maîtrise inoubliables. Un merci particulier à mon directeur de recherche pour m'avoir encouragé à toujours mener plus loin mes recherches et de m'avoir amené à remettre en question mes méthodes. Ceci m'a permis d'être rigoureuse dans mes recherches. Je repars avec un esprit critique, grandement nécessaire pour continuer en recherche.

Je voudrais également remercier toutes les personnes qui ont rendu ma maîtrise plus agréable. Mes amis et famille pour leur support ainsi qu'à tous les membres de l'association étudiante de l'IRCM.

Introduction

La transcription cryptique

Avec le développement des technologies de séquençage à haut débit, il est désormais évident que la transcription ne se limite pas qu'aux régions codantes pour des gènes connus. Chez la levure *Saccaromyce cerevisiae*, dont 75 % du génome correspond à des régions codantes annotées, des expériences de puces à ADN ont démontrées qu'au moins 85% du génome est transcrit (David et al., 2006). En plus de ce 10 % qui correspondent aux ARNs non-codants issus des régions non annotées du génome, certains ARNs sont issus de régions codantes mais empruntent des promoteurs non-canoniques. Ceci a été conclu suite à l'identification de sites d'initiation de la transcription (TSS) situés dans des régions non-promotrices (David et al., 2006; Miura et al., 2006; Zhang and Dietrich, 2005).

Les différentes classes d'ARN non-codants cryptiques

À ce jour, trois classes distinctes d'ARNnc, initiés de manière cryptique, ont été décrites. Les « Cryptic Unstable Transcripts » (CUTs) ont été les premiers à être identifiés par des analyses génomiques de puces à ADN dans des cellules *rrp6Δ* (Davis and Ares, 2006; Wyers et al., 2005). Rrp6 est une enzyme avec une fonction 3'-5' exonucléase impliquée dans la dégradation de l'ARN. Cette découverte, faite dans un contexte où la machinerie de dégradation de l'ARN est altérée, indique que la transcription est bien plus complexe mais qu'il existe des mécanismes permettant de garder l'intégrité du génome en limitant l'expression des transcrits aberrants. Ces transcrits sont rapidement dégradés de sorte qu'ils demeurent difficile à détecter à moins d'inactiver la machinerie de dégradation. Des analyses d'immunoprecipitation de la chromatine couplée aux puces à ADN (ChIP-chip) ont démontré que l'ARN polymérase II (ARN Pol II), est localisée aux sites d'initiation de certains CUTs, suggérant que ceux-ci sont bel et bien transcrits par l'ARN Pol II (Steinmetz et al., 2006). Les CUTs sont en général plus courts que les ARN messagers, mesurant de 200-500 paires de bases (pb) (Wyers et al., 2005). Ces transcrits sont terminés par la voie NNS (Nrd1-Nab3-

Sen1) qui les amène à être polyadénylés par le complexe TRAMP et ensuite rapidement dégradés par l'exosome nucléaire (Arigo et al., 2006; Thiebaut et al., 2006). Deux études ont contribué à identifier la distribution génomique des CUTs en utilisant des méthodes de 3'-Long-SAGE et de puces à ADN (Neil et al., 2009; Xu et al., 2009). Ceci a permis de mettre en évidence que ces transcrits sont initiés dans des régions exemptes de nucléosomes et qu'ils sont en majorité divergents des promoteurs des gènes annotés, grâce à la bidirectionnalité des promoteurs chez la levure.

Contrairement aux CUTs, les « Stable Unannotated Transcripts » (SUTs) sont observables dans des cellules sauvages (Neil et al., 2009; Xu et al., 2009). Similairement aux CUTs, ils sont transcrits à partir des régions exemptes de nucléosomes aux extrémités 5' et 3' des gènes et principalement sur le brin antisens (Xu et al., 2009). Par contre, ils sont en général plus long, avec une longueur médiane de 761 pb et leur mode de terminaison diffère de celui des CUTs (Jacquier, 2009). En effet, les SUTs pourraient être terminés à un site de poly(A) par un complexe de clivage/polyadénylation Pcf11-dépendant (Jacquier, 2009).

Les « Xrn1- sensitive Unstable Transcripts » (XUTs) sont, quant à eux, des transcrits cryptiques exprimés sur le brin antisens suite à l'inactivation de l'exonucléase cytoplasmique Xrn1p. En résumé, les trois classes de transcrits cryptiques sont majoritairement initiées dans des régions exemptes de nucléosomes et sont exprimées le plus souvent sur le brin antisens (Neil et al., 2009; Xu et al., 2009). La distinction entre les trois types de transcrits cryptiques est ambigue puisqu'il existe un fort chevauchement entre eux (van Dijk et al., 2011).

Les transcrits cryptiques intragéniques

Peu de temps avant ces expériences à l'échelle génomique, le laboratoire du Dr. Fred Winston fut le premier à démontrer que des transcrits étaient initiés à l'intérieur des régions codantes pour des gènes (Kaplan et al., 2003a). Cette transcription survient dans un contexte où la structure de la chromatine est perturbée, ce qui permet à la machinerie de transcription de former le complexe d'initiation à des sites normalement inaccessibles. Ces transcrits peuvent être sur les brins sens et antisens par rapport aux gènes. Ces transcrits sont

observables dans des extraits d'ARN totaux (Cheung et al., 2008; DeGennaro et al., 2013) ce qui indique qu'ils sont stables et en général polyadénylés.

Toutefois, ces transcrits cryptiques peuvent également être détectés dans le transcriptome provenant de cellules sauvages. En effet, l'étude du laboratoire de Ito a démontré qu'il existe bel et bien une transcription intra- et inter-génique en utilisant une méthode permettant de capter puis séquencer les régions coiffées en 5' des transcrits. Dans cette étude, des sites d'initiations de la transcription ont été découverts dans la région 3' de 384 régions codantes (Miura et al., 2006). Ce fut la preuve que la transcription peut être initiée à l'intérieur d'un gène dans des souches sauvages.

La transcription cryptique peut donc être définie comme l'initiation de la transcription à partir de sites « cryptiques », c'est-à-dire, des sites qui ne correspondent pas aux promoteurs de gènes annotés. Dans le cadre de ce mémoire, nous définissons un transcrit cryptique sens comme étant toujours initié à l'intérieur d'une région codante (tel que montré dans la Figure 1), alors qu'un transcrit antisens peut être initié à la fois dans des régions intergéniques ou intragéniques.

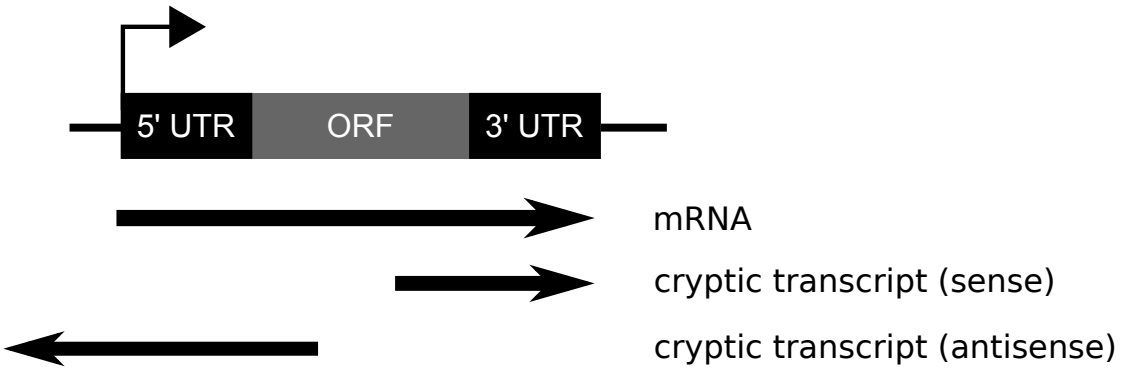

Figure 1. Les transcrits cryptiques.
Schéma simplifié représentant les ARNs transcrits à partir d'une région codante (ARN messager) et les ARNs cryptiques sens et antisens par rapport au gène.

La chromatine et la transcription cryptique

Les transcrits cryptiques étant d'avantage exprimés dans des souches mutantes, plusieurs groupes de recherche ont centré leurs efforts pour identifier les facteurs nécessaires pour réprimer la transcription cryptique chez *Saccharomyce cerevisiae*. Une étude en particulier, celle du Dr. Fred Winston et de son équipe (Cheung et al., 2008) a mené, de façon globale, à l'identification d'une cinquantaine de facteurs qui participent à la répression de la transcription cryptique. La majorité des facteurs identifiés sont des mutants d'histones, des gènes régulateurs d'histones ou des remodeleurs de la chromatine (Cheung et al., 2008), ce qui indique que la structure de la chromatine joue un rôle important dans le contrôle de la transcription cryptique.

Structure de la chromatine

Dans le noyau, l'ADN est présent sous forme de chromatine. Cette structure consiste en l'enroulement de l'ADN autours d'octamères d'histones pour former des nucléosomes. Ce premier niveau de compaction est composé de 146 paires de bases d'ADN entourées autour d'un octamère d'histones (Luger et al., 1997). Cet octamère d'histones est composé d'un tétramère d'histones H3-H4 et de deux dimères d'histones H2A-H2B. Pour former le deuxième niveau de compaction, l'histone H1 est nécessaire afin de lier les nucléosomes adjacents pour former une fibre de 30 nm. Cette fibre peut être plus ou moins condensée selon les différentes étapes du cycle cellulaire.

Rôle de la chromatine dans le contrôle de l'expression génique

La chromatine a pour but de compacter et de protéger le génome. Cette forme condensée permet surtout de limiter l'accès à l'ADN, puisque lorsqu'enroulé autour d'histones, l'ADN est inaccessible. Ceci donne un rôle important à la chromatine dans l'initiation et l'élongation de la transcription. En effet, les nucléosomes peuvent agir comme des barrières physiques en empêchant les facteurs d'initiation de la transcription de se lier à l'ADN (Knezetic and Luse, 1986; Lorch et al., 1987). Des études faites avec le gène *PHO5* chez la levure ont permis de démonter un lien direct entre la présence de l'ADN sous forme

chromatinienne et l'initiation de la transcription. L'activation de la transcription du gène est accompagnée d'une perte des nucléosomes se trouvant dans la région promotrice. Par contre, lorsque ces nucléosomes localisés dans le promoteur sont hyper-stabilisés, la transcription du gène n'est plus possible (Straka and Horz, 1991). Ceci indique que la présence de nucléosomes réprime l'expression des gènes. La chromatine participe également à l'élongation de la transcription puisque les nucléosomes sont capables *in vitro* de bloquer l'élongation de l'ARN Pol II (Chang and Luse, 1997; Izban and Luse, 1991, 1992). Un remodelage de la chromatine est donc requis pour l'initiation et l'élongation de la transcription, ainsi, on observe une déplétion des nucléosomes sur les régions codantes des gènes activement transcrits (Schwabish and Struhl, 2006).

Rôle de la chromatine dans la transcription cryptique

Étant donné le rôle de la chromatine dans la régulation de l'expression des gènes, plusieurs mécanismes par lesquels la transcription cryptique pourrait être régulée par la chromatine sont envisageables. Tout mécanisme pouvant mener à une déstabilisation, une hyper-acétylation ou une diminution du niveau de nucléosomes et résultant en une augmentation de l'accessibilité à l'ADN pourrait permettre l'accessibilité aux promoteurs cryptiques qui ne sont normalement pas accessibles par la machinerie de transcription (Hennig and Fischer, 2013; Smolle and Workman, 2013).

Diminution du niveau de nucléosomes.

Pour permettre la transcription, les nucléosomes doivent être expulsés à l'aide de plusieurs remodeleurs de la chromatine. Toutefois après cette expulsion les nucléosomes doivent être rapidement remis en place afin de préserver l'intégrité du génome (Schwabish and Struhl, 2006). Les chaperonnes d'histones sont des facteurs qui permettent de désassembler et de réassembler les nucléosomes durant la transcription. L'incapacité de reformer les nucléosomes va diminuer leur niveau le long des régions codantes, détruisant ainsi la barrière physique qui empêche normalement l'activation des mécanismes de transcription. Ce phénomène a été décrit pour la première fois en 2003, lorsque des transcrits cryptiques ont été

identifiés dans des souches mutantes pour les gènes *Spt6* (Kaplan et al., 2003a) et *FACT (Mason and Struhl, 2003)*, deux chaperonnes d'histones. Par la suite, d'autres chaperonnes d'histones impliquées dans la répression de la transcription cryptique ont été identifiées, tels que Rtt106 *(Imbeault et al., 2008; Silva et al., 2012)*, Asf1 (Schwabish and Struhl, 2006; Silva et al., 2012) et *HIR (Silva et al., 2012)*.

Augmentation du niveau d'acétylation des histones.

Les molécules d'histones étant chargées positivement, elles interagissent fortement avec l'ADN qui est chargé négativement, ce qui contribue à la compaction de la chromatine. Dans sa forme condensée, les queues N-terminal et C-terminal des histones sont accessibles. Les histones sont des molécules basiques et leurs extrémités N-terminal sont riches en résidus lysines qui peuvent subir des modifications post-traductionnelles. En effet, les queues d'histones peuvent être acétylées, méthylées, phosphorylées et ubiquitinylées. Lorsque les nucléosomes sont acétylés, le groupement acétyle annule leur charge positive, ce qui diminue la force d'interaction avec l'ADN (Kuo and Allis, 1998; Luger and Richmond, 1998). Ces études ont ainsi démontré que la chromatine a une structure plus relâchée lorsque les histones sont acétylées. Bien que l'acétylation soit importante pour activer la transcription, la déacétylation des nucléosomes suite au passage de l'ARN Pol II est tout aussi importante pour préserver l'intégrité du génome. Alors que les histones-acétyltransférases ajoutent un groupement acétyle aux lysines, les histones-déacétylases clivent le groupement N-acétyle et permettent ainsi à la chromatine de retrouver sa structure compacte. Ces deux groupes d'enzymes permettent ainsi de garder un bon équilibre entre l'ouverture et la fermeture de la chromatine. Plusieurs mutations qui causerait une hyper-acétylation dans les régions codantes ont été liées à l'activation de promoteurs cryptiques (Carrozza et al., 2005; Joshi and Struhl, 2005a; Keogh et al., 2005; Lickwar et al., 2009; Pattenden et al., 2010; Venkatesh et al., 2012).

Les facteurs permissifs pour la transcription cryptique

L'histone chaperonne Spt6

Spt6 est un gène essentiel et hautement conservé chez les eucaryotes. Spt6 a été identifié pour la première fois comme étant un suppresseur d'insertions des transposons Ty et de mutations δ dans le promoteur du gène *HIS4* (Clark-Adams and Winston, 1987; Winston et al., 1984). Les changements dans la transcription causés par les insertions Ty et δ peuvent être réprimées lorsque les gènes SPT sont mutés, ce qui indique que ces gènes jouent un rôle important dans le contrôle de la transcription. De fait, plusieurs évidences démontrent que Spt6 joue un rôle important dans le contrôle de la transcription chez la levure et ce, à l'échelle génomique.

Premièrement, Spt6 joue un rôle essentiel dans **l'élongation** de la transcription. Plusieurs études décrivent Spt6 comme étant un facteur associé à l'ARN Pol II en élongation (Hartzog et al., 1998; Krogan et al., 2002). Spt6 peut interagir physiquement avec le complexe d'élongation Spt4/Spt5 et la mutation d'un de ces trois gènes cause un défaut d'élongation (Hartzog et al., 1998; Lindstrom et al., 2003; Swanson and Winston, 1992). En plus d'interagir avec le complexe Spt4/Spt5, Spt6 interagit également avec FACT et Iws1, deux facteurs d'élongation (Krogan et al., 2002). De plus, Spt6 est localisé dans les régions codantes pour des gènes activement transcrits, ce qui vient appuyer son rôle dans l'élongation de la transcription (Kaplan et al., 2005; Kaplan et al., 2000; Krogan et al., 2002).

Deuxièmement, Spt6 joue un rôle dans le **remodelage** de la chromatine. Spt6 est connu pour interagir génétiquement et physiquement avec les histones, ce qui suggère un rôle dans le remodelage de la chromatine. Des expériences faites *in vitro* ont démontré que Spt6 interagit directement avec l'histone H3 et est capable d'assembler des nucléosomes (Bortvin and Winston, 1996). *In vivo*, Spt6 est nécessaire pour une bonne organisation des nucléosomes sur les régions codantes, puisque dans les souche *spt6-1004*, les cellules sont hypersensibles au traitement micrococcal, une endo-exonucléase, à une température non-permissive (Kaplan et al., 2003a). D'autres études ont également montré l'importance de Spt6 dans l'organisation de

la chromatine sur les régions codantes (Bortvin and Winston, 1996; DeGennaro et al., 2013; Jeronimo et al., 2015).

Spt6 et la transcription cryptique

Le mécanisme par lequel Spt6 réprime la transcription cryptique est assez bien caractérisé. Les premières études dans le domaine ont démontré, en utilisant un allèle sensible à la température dans lequel un domaine de liaison à l'ADN est muté (*spt6-1004*), l'apparition de plusieurs transcrits intragéniques (Cheung et al., 2008; Kaplan et al., 2003a). Ces transcrits ont été confirmés par Northern Blot pour une douzaine de gènes (Cheung et al., 2008; Kaplan et al., 2003a). À l'échelle génomique, Spt6 est impliqué dans la répression des transcrits cryptiques sur le brin sens (Cheung et al., 2008; Kaplan et al., 2003a) et sur le brin antisens (van Bakel et al., 2013).

Le niveau de nucléosomes dans les cellules *spt6-1004* dans les régions codantes des gènes actifs est anormalement bas. Ceci suggère que Spt6 réprime la transcription cryptique en assurant un bon niveau de nucléosomes dans les régions codantes ce qui empêche la liaison du complexe de pré-initiation à des promoteurs cryptiques. En effet, des analyses d'immunoprécipitation de la chromatine (ChIP) ont démontré une perte de nucléosomes et des analyses de digestion à la nucléase micrococcale ont démontré que la chromatine est plus ouverte le long du gène *FLO8* (Bortvin and Winston, 1996; Kaplan et al., 2003a). À l'échelle génomique, on observe une déplétion globale du niveau de nucléosomes sur les régions codantes lorsque Spt6 est inactivé (Jeronimo et al., 2015), ce qui pourrait expliquer que l'on observe une activation des promoteurs cryptiques dans au moins 1/6 des gènes chez la levure (Cheung et al., 2008). La répression de la transcription cryptique médiée par Spt6 implique également la restriction de H2A.Z en dehors des régions codantes. H2A.Z est une variante de l'histone H2A qui est conservée chez les eucaryotes. Cette variante d'histone est localisée aux nucléosomes +1 et -1 autour du site d'initiation de la transcription des gènes chez la levure (Guillemette et al., 2005; Raisner et al., 2005). Cette variante n'est normalement pas retrouvée dans les régions codantes des gènes activement transcrits (Hardy et al., 2009). Or, en absence de Spt6 (et FACT), la perte des nucléosomes sur les régions codantes permet au complexe

SWR-C d'incorporer H2A.Z ce qui contribue à l'activation des promoteurs cryptiques (Jeronimo et al., 2015).

Les autres chaperonnes d'histones

FACT

FACT (FAcilitates Chromatin Transcription) est composé de deux sous-unités essentielles et hautement conservées à travers les eucaryotes, Spt16 et Pob3 (Brewster et al., 1998). Spt16 est connue pour interagir physiquement avec deux complexes d'élongation Spt4/Spt5 et PAF (Krogan et al., 2002; Squazzo et al., 2002) ainsi qu'avec l'ARN Pol II en élongation (Mason and Struhl, 2003), ce qui suggère un rôle dans l'élongation de la transcription. De plus, Spt16 désassemble les nucléosomes en interagissant avec les dimères H2A-H2B (Belotserkovskaya et al., 2003) ce qui indique que cette protéine participe au remodelage de la chromatine durant l'élongation. Son rôle dans la répression des promoteurs cryptiques a été découvert en 2003, lorsqu'un transcrit cryptique a été détecté dans le gène *FLO8*, dans des cellules mutantes pour Spt16 (Kaplan et al., 2003a; Mason and Struhl, 2003). À l'échelle génomique, l'inactivation de Spt16 est impliquée dans la répression des transcrits cryptiques sur le brin sens (Cheung et al., 2008). Spt16 possède un mécanisme similaire à celui de Spt6 dans la répression de la transcription cryptique, puisque la mutation de *SPT16* ou *SPT6* active des promoteurs cryptiques dans un même ensemble de gènes (Cheung et al., 2008; van Bakel et al., 2013) et ce, dû à l'incapacité de reformer les nucléosomes durant l'élongation.

Rtt106

Rtt106 est une chaperonne d'histones, impliquée dans la régulation de la structure de la chromatine dans les régions codantes. Dans le but d'identifier le rôle de Rtt106, des expériences de ChIP ont pu associer cette protéine aux régions codantes des gènes actifs (Imbeault et al., 2008). De plus, Rtt106 a des interactions génétiques avec plusieurs facteurs d'élongation tels que Spt6, TFIIS et des membres du complexe PAF, suggérant qu'il joue un rôle dans l'élongation de la transcription (Imbeault et al., 2008). Plus précisément, Rtt106

possède une activité de chaperonne d'histone et interagit physiquement et fonctionnellement avec les histones H3/H4 (Huang et al., 2005) lui permettant ainsi de réguler le niveau de déposition de l'histone H3 sur les régions transcrites durant l'élongation (Imbeault et al., 2008). Étant donné le rôle de Rtt106 dans la régulation de la structure chromatinienne, l'activation de transcrits cryptiques a été observée dans des souches *rtt106Δ* (Imbeault et al., 2008; Silva et al., 2012). Finalement, un double mutant *rttt106Δ spt6-1004* augmente le niveau de transcription cryptique suggérant que les deux chaperonnes coopèrent dans la répression de la transcription cryptique (Imbeault et al., 2008).

Asf1 et le complexe HIR

Durant la transcription, des facteurs sont recrutés afin de combler la perte de nucléosomes. C'est le cas du complexe HIR qui permet le replacement des nucléosomes sur les régions en transcription. En plus de s'associer à Rtt106, la chaperonne d'histone Asf1 s'associe au complexe HIR afin de stimuler la déposition des histones (Lambert et al., 2010). Asf1 s'associe aux régions codantes des gènes actifs et permet l'éviction et la déposition de l'histone H3 durant l'élongation de la transcription (Schwabish and Struhl, 2006). Des mutations dans le gène Asf1 ainsi que dans toutes les sous-unités du complexe HIR mènent à transcription cryptique (Schwabish and Struhl, 2006; Silva et al., 2012) suggérant que ces deux histones chaperonnes participent au maintien de la chromatine dans son état réprimé suite au passage de l'ARN Pol II.

La voie set2/Rpd3S

Set2

Set2 est une histone méthyltransférase, requise pour la méthylation de l'histone H3 sur la lysine 36 (H3K36), une modification post-transcriptionnelle connue pour son rôle dans la répression de la transcription (Strahl et al., 2002). L'inactivation de Set2 rend la cellule sensible au 6-azauracil indiquant qu'en son absence, l'élongation de la transcription est altérée (Li et al., 2002). Des expériences de ChIP ont démontré que Set2 est localisée dans les régions codantes de gènes activement transcrits et que Set2 se lie à plusieurs facteurs d'élongation tels

que le complexe PAF et le remodeleur de chromatine Chd1 (Krogan et al., 2003), supportant le rôle de la protéine dans l'élongation de la transcription. En absence de Set2, on observe une augmentation d'acétylation dans les régions codantes (voir plus bas), ce qui rend l'ADN plus accessible. Ainsi, plusieurs études ont pu démontrer l'apparition de transcrits cryptiques dans les cellules mutantes pour Set2 (Carrozza et al., 2005; Kaplan et al., 2003a; Lickwar et al., 2009). Un autre mécanisme qui réprime la transcription cryptique de manière Set2 dépendante, implique l'histone chaperonne Asf1. La tri-méthylation des histones H3 à la lysine 36 par Set2 empêche la liaison de Asf1 aux histones (Venkatesh et al., 2012). Ceci permet de retenir les nucléosomes tri-méthylés le long des régions codantes et empêche leur hyperacétylation pour ainsi garder une structure non-permissive à la transcription cryptique.

Rpd3S

La méthylation de H3K36 le long des gènes activement transcrits est très important pour la répression de la transcription. Cette marque est spécifiquement reconnue par le complexe déacétylase Rpd3S, qui enlève les groupements acétyles des histones. Ce complexe est recruté aux gènes actifs via la phosphorylation du domaine C-terminal de l'ARN pol II (Drouin et al., 2010). Toutefois, pour être actif aux histones H3K36 méthylés, le complexe Rpd3S requiert ses sous-unités Eaf3 et Rco1. Eaf3 se lie aux histones H3K36 méthylées grâce à son chromodomaine (Joshi and Struhl, 2005a) alors que Rco1 lie les nucléosomes grâce à son domaine PHD (Li et al., 2007a). La déacétylation des histones médiées par Rpd3S durant l'élongation redonne une charge positive aux histones, ce qui permet de recompacter efficacement l'ADN suite au passage de l'ARN Pol II. En absence du complexe Rpd3S, les régions codantes ont un taux d'acétylation plus grand qu'à l'habitude (Carrozza et al., 2005) et donc une plus grande accessibilité à l'ADN. Ainsi, des transcrits cryptiques sont initiés lorsque les sous-unités Eaf3 ou Rco1 sont mutées (Carrozza et al., 2005; Joshi and Struhl, 2005a; Keogh et al., 2005; Lickwar et al., 2009; Pattenden et al., 2010)

Ctk1

Ctk1 est la kinase qui permet la phosphorylation des sérines 2 du domaine C-terminal de Rpb1, la plus grande sous-unité de l'ARN Pol II. Cette phosphorylation est nécessaire pour le recrutement de l'histone methyltransférase Set2 sur l'ARN Pol II via son domaine

d'interaction Set2-Rpb1 (Li et al., 2003). La perte de méthylation dans des cellules mutantes pour Ctk1 provoque l'initiation de transcrits cryptiques (Cheung et al., 2008; Youdell et al., 2008) puisque ceci empêche le recrutement de Set2 (voir le chapitre sur Set2).

Chd1 et Isw1b

L'échange d'histones arrive rarement dans les régions codantes pour la majorité des gènes chez la levure (Dion et al., 2007; Katan-Khaykovich and Struhl, 2011; Rufiange et al., 2007). En effet, il a été démontré que suite au passage de l'ARN Pol II, Set2 méthyle les histones le long des régions codantes empêchant ainsi la liaison de Asf1 et donc l'échange des nucléosomes liés à l'ADN avec ceux du milieu soluble (Venkatesh et al., 2012). Un autre mécanisme qui permet la rétention des histones H3K36 méthylées implique les remodeleurs de la chromatine Iws1b et Chd1. Ensemble, Chd1 et Isw1b empêchent l'échange des nucléosomes sur les régions codantes (Radman-Livaja et al., 2012; Smolle et al., 2012) permettant ainsi la rétention des histones H3K36 méthylées. En absence de Isw1b ou Chd1, les histones sont tout de même méthylées, mais elles ne sont pas retenues sur la région codante (Radman-Livaja et al., 2012; Smolle et al., 2012). Ceci conduit à une augmentation d'acétylation dans les régions codantes et une plus grande accessibilité de l'ADN. Ainsi, des transcrits peuvent être initiés à partir de promoteurs cryptiques en absence de Chd1 ou Isw1 (Cheung et al., 2008; Hennig et al., 2012; Quan and Hartzog, 2010; Smolle et al., 2012; Tirosh et al., 2010).

BUR2 et le complexe PAF

Une étude a démontré que la délétion des protéines Bur2 ou Paf1 accentue l'effet de la transcription cryptique dans des cellules *set2*Δ, suggérant un autre mécanisme responsable de la transcription cryptique (Chu et al., 2007a). Paf1 et Bur2 sont deux facteurs d'élongation impliqués dans la monoubiquitination de l'histone H2B (H2Bub). Lorsque les deux protéines du complexe Bur1/Bur2 sont mutées, on observe un défaut dans la monoubiquitination de H2B puisque le recrutement du complexe PAF dépend de ces deux protéines (Laribee et al., 2005; Wood et al., 2005). La monoubiquitination de l'histone H2B est nécessaire pour

l'initiation et l'élongation de la transcription. L'ubiquitination sur la lysine 123 est effectuée par le complexe Rad6/Bre1/Lge1 qui voyage avec l'ARN Pol II en élongation via une interaction avec le complexe PAF (Weake and Workman, 2008). En coopération avec l'histone chaperonne FACT, H2Bub joue un rôle important dans l'élongation de la transcription et dans le repliement de la chromatine suite au passage de l'ARN Pol II (Fleming et al., 2008; Pavri et al., 2006; Tanny et al., 2007). H2Bub est aussi impliquée dans la stabilité des nucléosomes puisque des mutations dans l'ubiquitine ligase Rad6 causent une réduction du niveau de l'histone H3 dans les régions codantes (Batta et al., 2011). De plus, H2Bub favorise le réassemblage des nucléosomes durant l'élongation de la transcription (Batta et al., 2011). Tout ceci indique que H2Bub est nécessaire pour la stabilité des nucléosomes et qu'en absence de cette modification post-traductionnelle, la machinerie de transcription peut se lier à des régions non-promotrices pour engendrer des transcrits cryptiques. En effet, des mutations dans la voie H2Bub ont été liées à la transcription cryptique (Fleming et al., 2008; Silva et al., 2012).

Les promoteurs cryptiques

Les differents promoteurs cryptiques qui existent ne sont pas régulés de la même manière par les facteurs qui répriment la transcription cryptique. Dans les cellules mutantes pour la chaperonne d'histone Spt6, l'activation du promoteur cryptique situé dans le gène *FLO8* dépend d'une séquence consensus TATA mais ce phénomène n'est pas observé pour les promoteurs cryptiques issus des gènes *STE11* et *SPB4* (Kaplan et al., 2003a). De même, deux facteurs peuvent réprimer l'activation des promoteurs cryptiques dans les mêmes gènes mais l'expression des transcrits cryptiques dans ces gènes peut varier considérablement (Cheung et al., 2008; Pattenden et al., 2010). Aussi, la transcription cryptique peut être régulée de manière spécifique à certains gènes puisque deux facteurs peuvent activer la transcription cryptique dans deux groupes de gènes distincts (Cheung et al., 2008). Les mécanismes par lesquels les promoteurs cryptiques sont activés ne sont pas encore complètement élucidés. Des évidences suggèrent que les promoteurs cryptiques sont régulés par les même mécanismes qui régulent les promoteurs canoniques (Koster et al., 2014; Pattenden et al., 2010). Cependant, la

régulation du promoteur cryptique se ferait d'une manière indépendante de la régulation du promoteur canonique du gène auquel il correspond (Pattenden et al., 2010). Alors que la transcription cryptique survient en général lorsqu'il y a une grande déplétion en nucléosomes ou en cas d'hyperacétylation des nucléosomes le long des régions codantes, ces modifications ne sont pas suffisantes pour dicter la localisation des promoteurs cryptiques. Une hyperacétylation causée par l'inactivation du complexe Rpd3S dans l'ORF ne conduit pas à une initiation cryptique aléatoire dans l'ORF mais à des sites bien précis (Pattenden et al., 2010).

Les méthodes d'identification des transcrits cryptiques

Il existe plusieurs méthodes permettant l'identification des transcrits cryptiques. Techniquement, presque toutes les méthodes qui permettent l'analyse d'ARNs devraient permettre l'identification de transcrits cryptiques. Dans cette section, nous allons voir les différentes méthodes utilisées au fil des ans pour étudier les transcrits cryptiques.

Northern blot

La méthode de northern blot, également appelée transfert d'ARN ou buvardage de type northern, permet d'étudier les ARNs. L'électrophorèse, technique utilisée pour séparer et caractériser des molécules selon leur taille moléculaire, est à la base du northern blot. Les ARNs sont par la suite transférés sur une membrane puis détectés grâce à une sonde d'ARN ou d'ADN simple ou double brin, marquée de manière radioactive ou fluorescente. Pour l'identification des transcrits cryptiques initiés dans une région codante et dans le même sens que le gène, il suffit de détecter les ARNs en utilisant des probes complémentaires aux régions 3' des gènes. S'il y a bel et bien transcription cryptique, au moins deux bandes devraient apparaître sur le gel correspondant au transcrit pleine longueur et au transcrit cryptique qui est plus court. C'est donc ainsi que les premiers transcrits cryptiques ont été identifiés (Kaplan et al., 2003a). Cette méthode est grandement utilisée et elle permet de valider la présence de transcrits cryptiques identifiés par des méthodes de séquençage à haut débit du génome entier.

Les puces d'ADN

Afin d'observer le niveau d'expression de plusieurs gènes en même temps, les puces d'ADN sont couramment utilisées. Cette technique, basée sur l'hybridation d'ADN, d'ADNc ou d'ARN sur des plaques de verres contenant des sondes d'ADN complémentaire permet de quantifier l'expression des régions désirées du génome. Lorsqu'un transcrit s'hybride sur la plaque, les sondes émettent une fluorescence proportionnelle à la quantité de matériel hybridé. Les puces à ADN peuvent être utilisées pour comparer le niveau d'expression des gènes entre deux conditions. Ainsi, il est possible de détecter la transcription cryptique en comparant les ARNs totaux issus de cellules sauvages avec ceux issus de cellules mutantes pour des facteurs permissifs à la transcription cryptique tels que Spt6, Spt16 et Set2 (Lickwar et al., 2009; Winston et al., 1984). L'apparition de transcrits cryptiques corrélant avec un enrichissement du signal pour les sondes correspondant aux régions 3' du gène par rapport à ceux des régions 5'. Il est ainsi possible d'évaluer la présence de transcrits cryptiques à partir de puces à ADN, en comparant le niveau d'expression aux extrémités 5' et 3' des gènes.

Le séquençage d'ARN

Le séquençage d'ARN est une technique qui permet d'avoir la séquence des ARNs présents dans un échantillon donné. Pour ce faire l'ARN doit être extrait des cellules, transformé en ADN complémentaire grâce à une transcriptase-inverse et coupé en fragments qui sont par la suite séquencés (Figure 2). Le nombre de « reads » obtenus sera proportionnel à l'abondance des ARNs correspondants. Selon le type d'ARN que l'on vise à analyser, plusieurs méthodes sont utilisées pour la création des librairies de séquençage. L'une d'entre elle consiste à enrichir les ARNs polyadénylés permettant ainsi l'étude d'ARNs matures c'est-à-dire stables et non ciblés pour la dégradation. Une autre méthode consiste à faire une déplétion des ARNs ribosomaux grandement enrichis dans les cellules, permettant ainsi l'analyse d'ARNs polyadénylés ou non-polyadénylés présents dans le transcriptome (Figure 2). Le séquençage de fragments peut se faire de deux manières. La première, dite « single-end », consiste à séquencer uniquement une extrémité d'un fragment et la deuxième dite de « paired-end » permet de séquencer les deux extrémités du fragment. L'utilisation de la

méthode « paired-end » a l'avantage de permettre un meilleur assemblage des fragments sur le génome de référence.

Le séquençage peut se faire en tenant compte du brin dont le transcrit est issus. L'avantage d'utiliser des protocoles brin-spécifiques permet de prendre en compte les transcrits sens et antisens par rapport aux gènes. Il existe plusieurs méthodes pour produire des librairies brin-spécifiques. L'une d'entre elles, nommée « dUTP method », consiste à remplacer les dTTP par des dUTP lors de la synthèse du brin complémentaire. L'ADNc riche en dUTP est par la suite digéré pour conserver l'information sur le brin d'origine (Figure 2). Le séquençage d'ARN brin-spécifique a permis l'étude de transcrits cryptiques sur les brins sens et antisens chez la levure (DeGennaro et al., 2013).

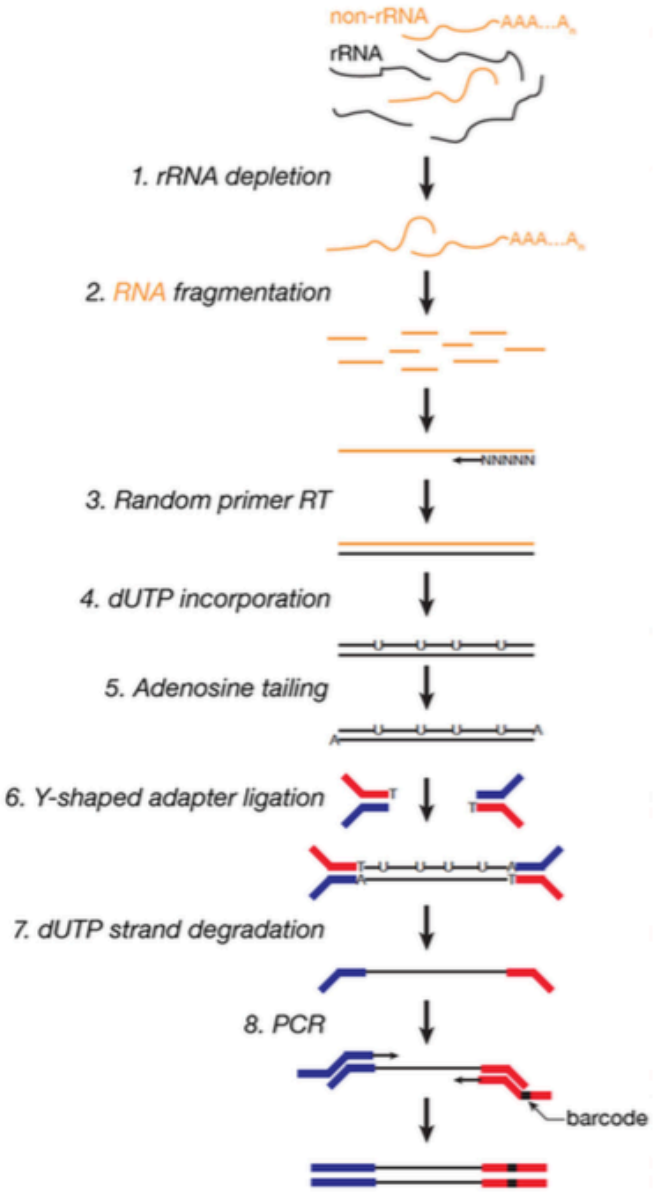

Figure 2. Protocole de création d'une librairie de séquençage.

Production d'une librairie de séquençage brin-spécifique à partir d'ARNs obtenus suite à la déplétion des ARNs ribosomaux (Zhang et al., 2012).

Les méthodes computationnelles d'identification des transcrits cryptiques à partir de données de séquençage à haut débit

Les transcrits cryptiques intragéniques sont, par définition, transcrits à partir d'un promoteur cryptique situé à l'intérieur d'un gène. Lorsque ces transcrits sont dans le même sens que le gène, on observe un enrichissement du signal dans la région 3' du gène (Figure 3B). Cette caractéristique permet de développer des algorithmes afin d'assigner un score cryptique aux gènes. Ce score désigne si le gène en question contient un promoteur cryptique actif. Le développement de ces méthodes est nécessaire puisque l'expression du transcrit cryptique peut être masqué dans l'expression du transcrit pleine longueur. Deux méthodes ont été développées à cette fin, soit la méthode du ratio 3'/5' (Cheung et al., 2008) et la méthode de l'enrichissement 3' (DeGennaro et al., 2013). Dans cette section, nous allons voir leurs implémentations.

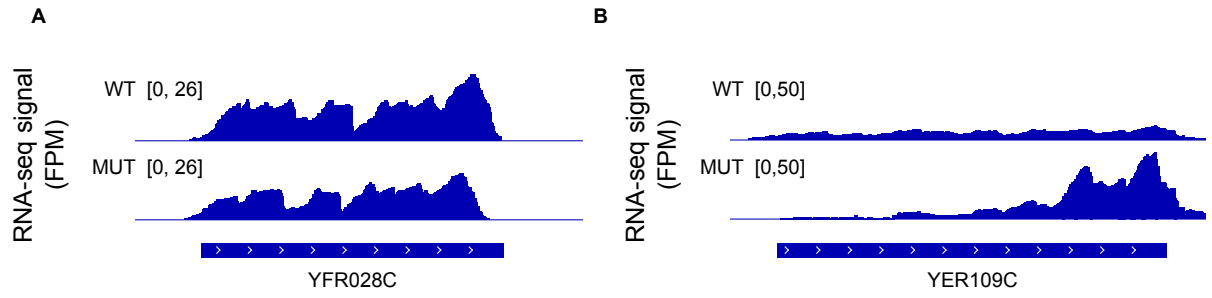

Figure 3. Exemple de signaux obtenus par RNA-Seq.
Niveau de signal RNA-Seq pour un gène (A) qui n'exprime pas de transcrit cryptique (*CDC14*) ou (B) qui exprime un transcrit cryptique (*FLO8*).
FPM : Fragments par million de pairs de bases.

La méthode du ratio 3'/5'

Dans cette méthode, le score cryptique pour un gène donné, correspond au ratio du nombre de « reads » obtenus dans une fenêtre de taille prédéterminée située dans la région 3'

sur une fenêtre de même taille située dans la régions 5’ du même gène. Pour chaque gène, ce ratio est calculé à la fois dans les cellules mutantes et dans les cellules sauvages. Le score cryptique est obtenu en divisant le ratio provenant des cellules mutantes par celui des cellules sauvages. Les gènes qui expriment un transcrit cryptique auront un ratio plus grand que 1, alors que ceux qui n’en expriment pas auront un ratio proche de un. Des contrôles sont produits en appliquant la même méthode mais en comparant les réplicas entre eux (mutant $_{\text{réplica 1}}$/mutant $_{\text{réplica 2}}$ et WT $_{\text{réplica 1}}$/WT $_{\text{réplica 2}}$).

La méthode de l’enrichissement 3’

Comparativement à la méthode du ratio 3’/5’, la méthode de l’enrichissement de l’extrémité 3’ du gène prend en compte toutes les données du gène puisqu’elle a été développée pour analyser des données de RNA-Seq. Pour calculer le score, une valeur d’expression est assignée de façon cumulative à chaque position du gène en commençant par son extrémité 3’. Ainsi, pour chaque position du gène, une valeur correspondant au nombre de « reads » retrouvés entre cette position et la fin du gène est calculée. Les valeurs ainsi calculées à partir des données dans les cellules mutantes sont corrélées à celles issues des cellules sauvages. S'il n'y a pas d'évènements d'initiation cryptique dans le gène, une corrélation positive et linéaire sera observée entre les données des cellules mutantes et les données des cellules sauvages. Par contre, s'il y a un évènement d'initiation cryptique, les valeurs pour le mutant seront plus élevées et donc situées au-dessus de la diagonale. Le score cryptique correspond au ratio entre l'aire en dessous de la courbe décrite et l'aire en dessous de la diagonale. Les contrôles sont obtenus en appliquant la même méthode mais en comparant les réplicas (mutant $_{\text{réplica 1}}$/mutant $_{\text{réplica 2}}$ et WT $_{\text{réplica 1}}$/WT $_{\text{réplica 2}}$).

Rationnel/Hypothèses/Objectifs

La transcription cryptique est un concept assez nouveau. Malgré le fait que plusieurs études aient permis d'identifier différents facteurs permissifs pour la transcription cryptique, le mécanisme par lequel cette transcription est régulée n'est pas encore clair. L'objectif principal de ce travail vise à étudier la transcription cryptique à l'échelle génomique dans le mutant *spt6-1004* afin de mieux comprendre ce phénomène. Spt6 est une chaperonne d'histone grandement étudiée pour son rôle dans le contrôle de la transcription. Plusieurs études ont déjà démontré son implication dans la régulation du niveau de nucléosomes le long des gènes afin de réprimer la transcription cryptique. Nous avons donc choisi d'étudier la transcription cryptique dans ce mutant pour deux raisons : premièrement parce que la transcription cryptique y est observée et deuxièmement parce que l'état actuel de la littérature ne permet pas d'apprécier le phénomène à l'échelle génomique.

Nous avons réalisé un séquençage à haut débit d'ARNs provenant de cellules *spt6-1004* de manière brin-spécifique afin d'étudier à la fois les transcrits sur les brins sens et antisens. Suite à cela, le premier objectif de ce travail était de développer un outil bio-informatique afin d'identifier les transcrits cryptiques à partir de données de RNA-seq. Ceci a été motivé par un manque apparent d'outils permettant l'analyse des transcrits cryptiques à l'échelle génomique. Le développement de méthodes précises pour l'identification des transcrits cryptiques est capital car, étant initiés à l'intérieur des gènes codants, ces transcrits sont difficilement observables puisque leur signal peut être brouillé par le signal des gènes dont ils sont issus. Le deuxième objectif de cette étude était de mettre au point une méthode permettant d'identifier les sites d'initiation des transcrits cryptiques afin de mieux caractériser les promoteurs cryptiques. Enfin, le dernier objectif de cette étude visait à caractériser les terminateurs des transcrits cryptiques. Pour ce faire, nous avons utilisé les données de RNA-seq afin d'identifier les sites de terminaison des transcrits cryptiques et par la suite, identifié les séquences propres à ces régions afin de mieux comprendre comment ces transcrits cryptiques sont terminés. Pour compléter ces expériences bio-informatiques, nous avons expérimentalement validé les résultats.

Méthodologie

Adaptation de la méthode du ratio 3'/5' aux données de RNA-Seq

La méthode du ratio 3'/5' peut être adaptée à des données de RNA-Seq qui couvrent toute la longueur du gène. La taille des régions 3' et 5' et leurs positions exactes sont des éléments importants à prendre en compte. Initialement l'algorithme a été appliqué sur des données de puces à ADN avec 6 sondes couvrant les régions codantes. Ainsi le score comparait l'intensité du signal de la 6ème sonde à celui de la 1ère sonde. Afin d'adapter la méthode aux données RNA-Seq, il suffit de définir des fenêtres aux extrémités des gènes. Une fenêtre prédéfinie correspond à un nombre de paires de bases fixe à utiliser pour les régions 3' et 5'. En utilisant une fenêtre de 100 pb à chaque extrémité du gène, cette méthode nous a permis de bien identifier les gènes exprimant un transcrit cryptique (voir la prochaine section).

Développement et validation d'une nouvelle méthode d'identification des transcrits cryptiques (méthode probabiliste).

Avant-propos

Cette section décrit, sous forme d'article, la méthode probabiliste qui est une alternative aux méthodes déjà existantes pour l'identification de transcrits cryptiques intragéniques sur le brin sens. Son rendement face aux méthodes déjà existantes y est évalué. Mon apport dans le manuscrit est d'avoir implémenté le package R et de l'avoir utilisé afin de comparer les trois méthodes. J'ai également contribué à la rédaction de l'article.

yCrypticRNAs: an R package for cryptic transcription analysis

Nicole Uwimana[1], Benjamin Haïbe-Kains[2] and François Robert[1,3]

[1]Institut de recherches cliniques de Montréal, 110 avenue des Pins Ouest, Montréal, H2W 1R7, Québec, Canada [2]Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada [3]Département de médicine, Faculté de médicine, Université de Montréal, 2900 Boulevard Édouard-Montpetit, Montréal, H3T 1J4, Québec, Canada

Abstract

Summary: The yCrypticRNAs package implements methods for cryptic transcription detection from RNA-Seq data. We have developed a probabilistic method based on a cumulative distribution function in order to identify cryptic transcripts genome-wide. This method takes into account the variations in gene expression and gene length due to a randomisation of the data. Applying our method to RNA-Seq data from *Saccharomyces cerevisiae* Spt6 mutant cells shows that our method outperforms previously published methods, providing a good compromise between false positives and false negatives. Importantly, while previous methods only identify genes within which cryptic transcription occurs, our method -using a bootstrap approach- allow for the prediction of the position of the cryptic transcription start sites.

Availability: The package is implemented in R and is available from CRAN. It works on LINUX/UNIX platforms. Its source code is available from Bitbucket.

Introduction

Transcription normally initiates at defined locations in the genome. In adverse conditions such as in mutants causing defects in chromatin structure, however, transcription sometimes initiates from aberrant locations within genes (Smolle and Workman, 2013). This so-called cryptic transcription is traditionally studied using single gene approaches such as

Northern blots or RT-qPCR (Cheung et al., 2008; Jeronimo et al., 2015; Kaplan et al., 2003a). More recently, methods have been developed to identify cryptic transcripts from tiling array and RNA-Seq data, enabling the study of cryptic transcription at a genome-wide level (Cheung et al., 2008; DeGennaro et al., 2013; Xu et al., 2009). While higher resolution techniques such as 3'-Long-SAGE (Neil et al., 2009) would be best suited to identify cryptic transcripts, reliably identifying cryptic transcripts from RNA-Seq data remains important as RNA-Seq is the most widely used method for gene expression data analysis.

Because cryptic transcription is, by definition, initiated from within genes, an increase in the RNA-Seq (or tiling array) signal towards the 3' end of a gene is indicative of a cryptic initiation site within that gene. This was exploited by Cheung et al. (Cheung et al., 2008) who successfully identified cryptic transcripts by calculating a 3'/5' ratio from tiling array data. Later, DeGennaro et al. (DeGennaro et al., 2013) used the shape of the cumulative RNA-Seq signal from the 3' end of genes to identify cryptic transcript from RNA-Seq data. Although both these methods efficiently identified genes within which cryptic transcription occurs, none of them predicts the position of the cryptic transcription start site (cTSS). In addition, our own analysis using these methods revealed limited overlap between both methods. This motivated us to develop a two-in-one method that can predict the occurrence of cryptic transcripts and their cTSS. The method assesses cryptic initiation using a probabilistic approach and the permutation of the RNA-Seq data allows eliminating bias related to gene length and fluctuation in gene expression.

Methods

Identification of cryptic intragenic transcripts

For each position of a gene, the cumulative RNA-Seq signal is calculated by summing the number of reads/fragments between the given position and the previous position, starting at the 5' end, in the wild-type and mutant samples (Figure 4A, second panel). The cumulative values from the mutant are then subtracted from those of the wild-type (Figure 4A, third panel). The resulting differential cumulative values are then used to calculate, for each

position of the gene, the perpendicular distance (f value) between the cumulative values and a diagonal linking the first and last data points (Figure 4A, fourth panel). The f score for a gene is then obtained by taking the maximum f value minus the minimum f value. In principle, a high f value should correlate with the presence of a cryptic transcript as it indicates the presence of excess RNA-Seq reads in the 3' end of the gene in the mutant compared to the WT. The f value, however, is also influenced by the expression level and the length of the gene. In order eliminate these biases, and in order to add a statistical score to the method, the RNA-Seq values are randomly permutated several times and the f score re-calculated after each permutation. Those simulated f scores are used to calculate a z score representing the distance between the observed f score and the distribution of the permutated f scores. That z score represents the probability that cryptic transcription is initiated somewhere within the tested genes.

Identification of cryptic transcription start sites (cTSS)

For each position of a gene, an f value is calculated as described above. The position where the maximum f (f max) value is reached represents the position where the cryptic transcript is initiated (cTSS). The exact position of the f max, however, is influenced by local noise in the RNA-Seq data. In order to identify the position of cTSS in a probabilistic manner, the data is re-sampled several times (each time randomly eliminating a fraction of the values) and the f max and its position is recalculated for each round. This allows for the identification of a cryptic zone, a region within the gene where a cryptic transcript is likely to have initiated. We have defined five ways of identifying cryptic zones. For all the methods, an observed f value is calculated using the real data and a simulated distribution of f max is calculated by re-sampling the data several times. For method A, the various positions of the genes where the corresponding f value is within the simulated distribution defines the cryptic zone. For method B, the mean f value at each position of the gene is calculated using the simulated data. Then the position of the gene for which the simulated mean f value is within the simulated distribution defines the cryptic zone. The method C is similar to method B but instead of taking the mean simulated f value, the method C considers the positions for which the simulated f values are within the simulated distribution at each round. Method D is different from other methods as it only considers the positions where the simulated maximum f value is

reached to define the cryptic zone. Lastly, the method C Gaussian is similar to method C, but the Gaussian mean and standard deviation are calculated in order to define the positions where there is a higher possibility to observe a cryptic initiation event. All the methods differ in the number or the length of the cryptic zones identified (Figure 5).

Case study

The 3'/5' ratio method (Cheung et al., 2008), the 3'enrichment method (DeGennaro et al., 2013) and the probabilistic method implemented in the package were used to identify genes with cryptic transcripts in *spt6-1004* cells. All methods successfully identified a set of 11 previously validated cryptic transcripts (Cheung et al., 2008; Kaplan et al., 2003a) (data not shown). Compared to the number of genes identified by the probabilistic method (1,760 genes), the 3' enrichment method identified fewer genes (446 genes) while the 3'/5' ratio method identified a larger number of genes (2,151 genes) (Figure 4B). To evaluate the accuracy and sensitivity of the methods, 200 genes were randomly selected from those identified by either method (also including negative controls). The gene list was randomized and submitted to three lab members who visually inspected the RNA-Seq data on a genome browser. For each of the 200 genes, each curator had to determine whether they considered the gene to contain a cryptic transcript or not (Figure 4C). Perhaps as expected, a combination of any two methods gives the best prediction with more than 70% of the identified genes being considered positives by the curators. When considering gene identified exclusively by one method, however, the probabilistic method outperformed the two others with fewer false positives and better sensitivity than the two other methods. Thus, the probabilistic method provides a good compromise between false positives false negatives.

For the set of 630 genes identified by the probabilistic method as having an intragenic cryptic transcript with a stringent z score of at least 20, we used the "method C Gaussian" to predict position of the transcription start site of the cryptic transcript accurately determined. First, a sharp increase of the RNA-Seq signal is apparent around the cTSS, similar to what is observed at the canonical TSS (Figure 4D, left panel). Second, motifs search around cTSS showed enrichment of TATAWAWR motif, a well-known promoter element (Figure 4D, right panel).

Conclusions

The yCrypticRNAs package provides methods for genome-wide identification of cryptic transcription from RNA-Seq data. The probabilistic method implemented in the package allows the identification of cryptic transcripts and their cryptic transcription start sites. The package also implements previously published methods allowing easy study of cryptic transcription.

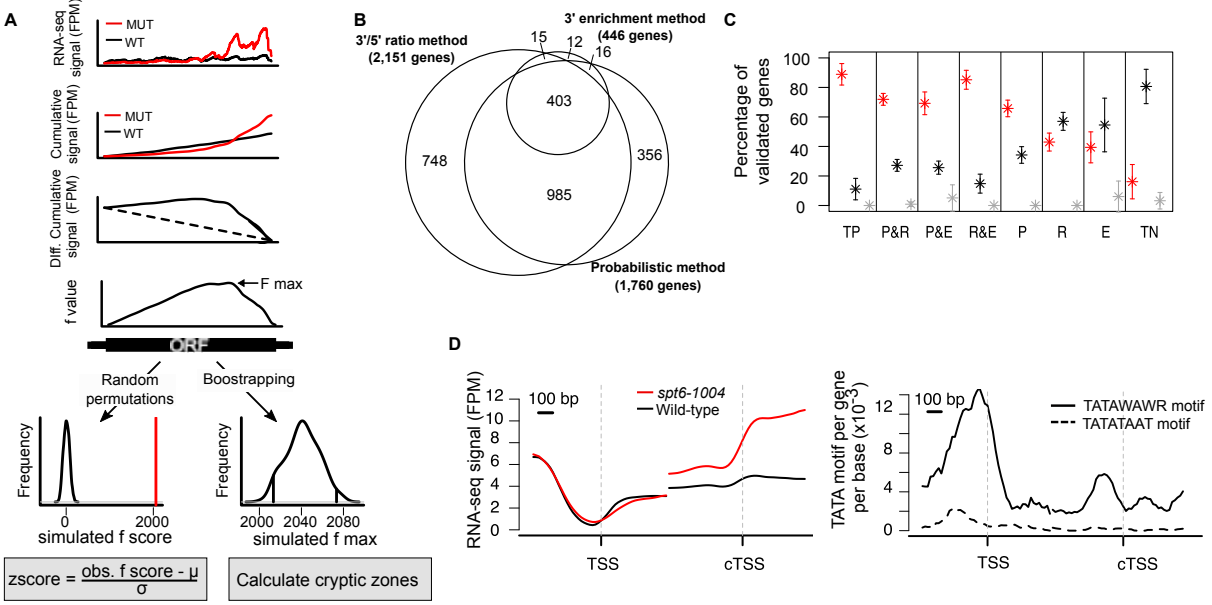

Figure 4. Probabilistic method. (A) Schematic representation of the probabilistic method workflow. (B) Venn diagram comparing the number of genes with an intragenic cryptic transcript identified by the 3'/5' ratio, the 3' enrichment and the probabilistic methods with allowing 1% of false discovery rate. (C) Expert validation of a set of 200 genes, by visual inspection of RNA-Seq data on a genome browser. TP: triple positive; P&R: probabilistic and 3'/5' ratio methods; P&E: probabilistic and 3' enrichment methods; R&E: 3'/5' ratio and 3' enrichment methods; P: probabilistic method-only; E: 3' enrichment method-only; R: 3'/5'

ratio method-only; TN: triple negative. (D) Meta-gene analysis around canonical and cryptic transcription start site. TATATAAT motif was used as a control sequence (dashed line).

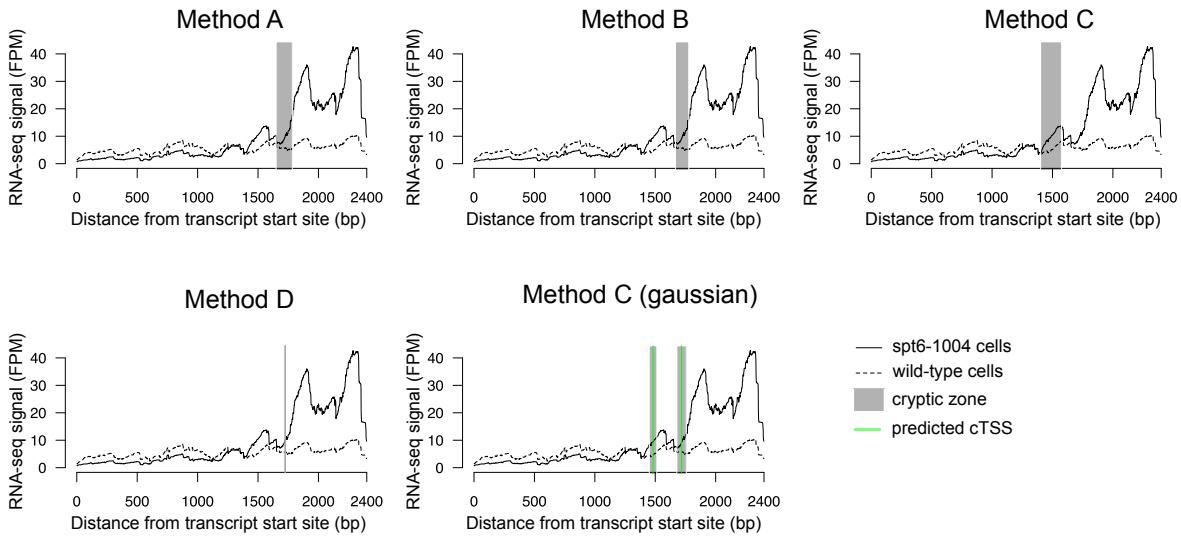

Figure 5. Cryptic zones. Cryptic zones (in grey) identified depending on the method used for the FLO8 gene.

Résultats

Avant-propos

Cette section présente un article en préparation. Mon apport dans le manuscrit est d'avoir participé à la conception de cette étude, réalisé les analyses bio-informatiques servant à l'identification des transcrits cryptiques ainsi que les analyses d'enrichissement des séquences (Figures 6, 7, 8, 9 et 10A,B). J'ai également contribué à la rédaction de l'article. Pierre Collin a conçu et réalisé les expériences permettant de valider expérimentalement la terminaison des transcrits sur le brin antisens (Figure 10C-E). Célia Jeronimo et François Robert ont préparé les ARNs utilisés pour produire les jeux de données de RNA-seq. François Robert a conçu l'étude et écrit le manuscrit principal.

Bidirectional terminators in *Saccharomyces cerevisiae* prevent cryptic transcription from invading neighbouring genes

Nicole Uwimana[1], Pierre Collin[1], Célia Jeronimo[1], Benjamin Haibe-Kains[2] and François Robert[1,3,*]

[1]Institut de recherches cliniques de Montréal, Montréal, Québec, H2W 1R7, Canada [2]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, M5G 2M9, Canada [3]Département de médecine, Faculté de médecine, Université de Montréal, Québec, H3T 1J4, Canada

*To whom correspondence should be addressed. Tel: 514-987-5737; Email: francois.robert@ircm.qc.ca

Abstract

Transcription can be quite disruptive for chromatin so cells have evolved mechanisms to preserve chromatin integrity during transcription, hence preventing the emergence of cryptic transcript from spurious promoter sequences. How these transcripts are regulated and processed by cells remains poorly characterized. Notably, very little is known about the termination of cryptic transcription. Here we used RNA-Seq to identify and characterize cryptic transcripts in Spt6 mutant cells (*spt6-1004*) in *Saccharomyces cerevisiae*. We found polyadenylated cryptic transcripts running both sense and anti-sense relative to genes in this mutant. Cryptic promoters were enriched for TATA boxes, suggesting that the underlying DNA sequence defines the location of cryptic promoters. While intragenic sense cryptic transcripts terminate at the terminator of the genes that host them, we found that anti-sense cryptic transcripts preferentially terminate at the 3'-end of upstream genes. This led us to demonstrate that most terminators in yeast are bidirectional, leading to termination and polyadenylation of transcripts coming from either direction. We propose that *S. cerevisiae* has evolved this mechanism in order to prevent spurious transcription from invading neighbouring genes, a feature particularly critical for organisms with small compact genomes.

Introduction

Transcription initiation occurs at promoters, which in eukaryotes are defined by small DNA motifs that direct the assembly of a preinitiation complex containing an RNA polymerase and its associated factors (Lenhard et al., 2012). In recent years, it became evident that, in addition to DNA sequence, chromatin structure plays critical roles in defining promoter regions. In *Saccharomyces cerevisiae*, where chromatin structure has been studied extensively, promoters are characterized by a nucleosome-free region flanked by well positioned nucleosomes carrying specific histone post-translational modifications and the histone variant H2A.Z (Albert et al., 2007). Outside promoters, different epigenetic signatures decorate nucleosomes, allowing for the prediction of regulatory regions based epigenetic signatures (Ernst and Kellis, 2010). These chromatin states are highly dynamics, notably during transcription elongation where histone chaperones, histone acetyltransferases, histone

deacetylases, histone methyltransferases and chromatin remodelers coordinate with RNA polymerase II and allow for maintaining proper chromatin structure and epigenetics over genes (Kwak and Lis, 2013; Smolle et al., 2013). Tampering with these chromatin modifying enzymes leads to the emergence of cryptic transcription, initiated within genes (Smolle and Workman, 2013). While this phenomenon is best described in yeast, a recent study suggests that cryptic transcription initiated within genes also occurs in cancer cells (Muratani et al., 2014). Maintaining proper chromatin structure and epigenetic state during transcription is therefore critical for transcription initiation fidelity and may play direct role in human diseases.

The first evidence that histone chaperones are important for preventing intragenic cryptic transcription came from the Winston group (Kaplan et al., 2003b). In the course of confirming microarray data by Northern blots, Kaplan and colleagues made the surprising finding that a mutant for the transcription-associated histone chaperone Spt6 expresses short transcripts, initiated from within gene bodies (Kaplan et al., 2003b). This was accompanied with hypersensitivity to nuclease, leading them to propose that Spt6 prevents cryptic transcription by maintaining proper chromatin structure during elongation (Kaplan et al., 2003b). Subsequently, similar phenotypes were shown in mutants for other histone chaperones, for genes involved in proper expression of histone genes, for histone methyltrasferase, for demethylases and for chromatin remodelers (Carrozza et al., 2005; Cheung et al., 2008; Chu et al., 2007b; Du et al., 2008; Du and Briggs, 2010; Hainer and Martens, 2011; Imbeault et al., 2008; Joshi and Struhl, 2005b; Kaplan et al., 2003b; Li et al., 2009; Radman-Livaja et al., 2012; Silva et al., 2012; Smolle et al., 2012; van Bakel et al., 2013). In essence any mutant that contributes to make gene body chromatin looking like promoter chromatin may contribute to the emergence of intragenic cryptic transcription. This include, nucleosome loss (Kaplan et al., 2003b), H2A.Z mislocalization (Jeronimo and Robert, 2016), histone acetylation (Carrozza et al., 2005) and increased histone turnover (Smolle et al., 2012).

While many of the factors involved in repressing cryptic transcription and their associated mechanisms are known, the repertoire of cryptic transcripts emerging in these mutants remains ill-defined. In addition, how these transcripts are terminated and processed

has to our knowledge never been formally investigated. Here, we addressed these questions by RNA-Seq profiling of *spt6-1004* cells. Interestingly, we found that anti-sense cryptic transcription often terminates at the terminator of the adjacent gene, thanks to the previously underestimated bi-directionality of most yeast terminators.

Material and methods

RNA-Seq

Cells from *spt16-1004* and its respective WT strain were grown to an OD_{600} of 0.5 at 30°C and shifted to 37°C for 80 min before RNA extraction. Total RNA was extracted using the hot phenol method. Prior to library preparation, total RNA was either depleted for ribosomal RNA using the Ribo-zero Gold yeast kit (MRZY1306) from Epicentre-Illumina or enriched for polyadenylated RNA using the NEBnext Poly(A) kit from New England Biolabs. Strand specific RNA-Seq libraries were prepared using the KAPA stranded RNA-Seq library preparation kit prior to paired-end sequencing on an Illumina Hi-Seq2000. Reads were mapped to the sacCer3 assembly of the *S. cerevisiae* genome using Tophat2 (Kim et al., 2013). Intron length range was set at 50 to 1000 bp and a reference annotation file was provided to guide the assembly.

Identification of intragenic sense cryptic transcripts

Sense cryptic transcripts were detected from RNA-Seq data using a probabilistic method we recently developed (Uwimana et al, under review; See also Chapter on Methodology page 19). For each gene, we assessed the probability of cryptic transcription initiation (z score) using the probabilistic method implemented in the "genome_wide_scores" function using default parameters. The z scores were calculated using values from *spt6-1004* and WT cells for which values from replicates were merged together. As a control, we calculated the z score for each gene comparing the replicates in mutant ($spt6\text{-}1004_{rep1}$/$spt6\text{-}1004_{rep2}$) and wild-type (WT_{rep1}/WT_{rep2}) cells. Using the z scores obtained when comparing replicates, we determined a cutoff by allowing 1% of false discovery.

For those genes with a high z score, we identified the cryptic transcription start sites (cTSS). The probabilistic method is based on a cumulative representation of the RNA-Seq, data providing an f score for each position of genes. This f score is obtained by calculating the perpendicular distance from the differential cumulative values (*spt6-1004* cumulative values mines WT cumulative values). The position where the maximum f score is achieved represents the position where the cryptic transcript is initiated as it indicates the position where the RNA-Seq signal sharply increases. By resampling the data 200 times, each time removing 10% of the data, we re-calculated simulated values of the f max and its position. This allows for the identification of a cryptic zone (a region where cryptic transcription is likely to initiate) as well as the most likely coordinate of the cryptic initiation site. This was done using the "initation_sites" function with the Gaussian method (method = "methodC_gaussian") and all other arguments set to default.

Identification of intragenic anti-sense cryptic transcripts

Cryptic transcripts running on the anti-sense strand were detected using Stringtie (Pertea et al., 2015). The minimum assembled transcript length was set to 100 (-m 100), the minimum reads per bp coverage to consider for transcript assembly was set to 2 (-c 2), the gap between reads mappings triggering a new bundle was set to 5 (-g 5) and no reference annotation file was provided for guiding the assembly process. We next removed all the transcripts that overlapped with known annotation on the same strand to keep the novo anti-sense cryptic transcripts. The fragments per kilobase per million mapped reads (FPKM) for each anti-sense cryptic transcript was calculated in WT and in *spt6-1004* cells and anti-sense transcripts having at least 4 FPKM and a $\log_2$ fold change of at least 1.25 were kept for further analysis.

Promoter sequence analyses

The yeast genome was scanned for the TATA box consensus sequence in *S. cerevisiae* (TATAWAWR) and a very close sequence with only one mismatch (TATATAAT) using HOMER (Heinz et al., 2010). A score of 1 was set to each position where a motif was found. The resulting scores were used to look at the sequence distribution around canonical and cryptic promoters using VAP (Brunelle et al., 2015; Coulombe et al., 2014).

Terminator motif analyses

The yeast genome was scanned for the "efficiency" (UAUAUA, UACAUA, UAUGUA) and "positioning/A-rich" (AAUAAA, AAAAAA) motifs as described in (Graber et al., 1999). Because these A/T-rich motifs are frequently found in the genome but require a stereotypical organization to be functional (Graber et al., 1999), we required each positioning motif to be within 50bp upstream of a positioning motif and vice-versa. A score of 1 was set to each position where motifs were found. The resulting scores were used to look at the motifs distribution relative to genes on both sense and anti-sense strands using VAP (Brunelle et al., 2015; Coulombe et al., 2014).

Identification of bidirectional or unidirectional terminators

Gene terminators that are challenged by a cryptic transcript were classified as unidirectional or bidirectional as follow. Unidirectional terminators are defined as terminators that allow termination of their corresponding sense transcripts but are inefficient at terminating anti-sense transcription while bidirectional terminators are defined as terminators allowing for both sense and anti-sense transcription termination. Terminators were classified as bidirectional if they allow termination of anti-sense transcription within a 500 bp window around their transcription termination sites (TTS). Which means that an anti-sense transcript terminating 250 bp before or after a gene's TTS was considered to be terminated at that terminator. This method identified 579 bidirectional terminators. However, we noticed many anti-sense transcripts for which the RNA-Seq signal drastically decreased, but did not completely disappear, around TTS. These are indication of terminators that are bidirectional but with weaker activity. To identify those weak bidirectional TTS, we used a probabilistic method similar to that used to identify cryptic transcripts. For each gene, we calculated the probability of a termination event in the -250 and +250 bp region around their TTS on the anti-sense strand, using both the f score and the z score. We found that genes with an anti-sense cryptic transcript terminating in their 3'-end tend to have very low f and z scores. We thus selected bidirectional promoters having f and z scores values smaller than -94.47 and -11.35 respectively. These cutoffs allow 15% and 5 % of false discovery respectively, based on

WT replicates. Using this approach, we identified 97 weak bidirectional terminators. Unidirectional terminators were identified by selecting TTS that overlap an anti-sense cryptic transcript, but are located at least 250 bp away of the cryptic transcription termination site. We found 150 unidirectional terminators using this criterion.

Termination assays

Putative terminator sequences were tested using a modified version of the assay developed by Carroll et al to test snoRNA terminators (Carroll et al., 2004). In this system, the *HIS3* gene is expressed from a plasmid under the control of the *ADH1* promoter and putative terminators are cloned between the promoter and the *HIS3* gene allowing for termination to be monitoring by growth on plates lacking histidine. To adapt this assay to protein-coding gene terminators, we inserted the 5' UTR (627 bp) from the *YNR051C* gene between the promoter and the *HIS3* open reading frame. This creates space between the promoter and the cloning site for the putative terminators, therefore mimicking the promoter-terminator organization found in typical yeast protein-coding genes. Putative terminators (500bp fragments) were then cloned downstream of this 5' UTR and termination was monitored by Northern blotting using a single stranded RNA probe corresponding to the *YNR051C* 5' UTR. Northern blotting was used instead of growth on histidine minus plates since the terminators tested often contain ATG codons. Northern blots were performed using fluorescent probes as described previously (Jeronimo et al., 2015).

Results

A probabilistic method for the identification of intragenic sense cryptic transcripts from RNA-Seq

The genome-wide identification of intragenic sense cryptic transcripts is not trivial since these transcripts are embedded within genes. Previous studies have identified genes hosting a cryptic transcript by seeking for genes with excessive signal in the 3' portion of genes using either tiling arrays or RNA-Seq (Cheung et al., 2008; Li et al., 2007b). This

approach suffers from the fact that RNA-Seq signal can be quite "wavy", leading to the introduction of randomness in the 3'/5' ratio measurements. More recently, DeGennaro et al looked at the cumulative RNA-Seq signal starting from the 3'-end to identify genes with excessive signal in the 3'-end in *Schizosaccharomyces pombe spt6-1* cells relative to WT (DeGennaro et al., 2013). This method generates fewer false positives than the 3'/5' ratio method but suffers from a lack of sensitivity (Uwimana et al, under review). In addition, neither of these methods allows for the mapping of the 5'-end of cryptic transcripts (i.e. cryptic promoters) as they only predict which genes are hosting a cryptic transcript.

In order to alleviate these limitations, we developed a probabilistic method for the identification of intragenic cryptic transcripts from RNA-Seq experiments (Figure 6A; See also Chapter on Methodology page 19). Briefly, for each position of a gene, the cumulative RNA-Seq signal is calculated by summing the number of reads per fragments between the given position and the previous position (from 5' to 3'). The cumulative values from the WT are subtracted from those of the mutant and the perpendicular distance (f value) between the differential cumulative values and a diagonal linking the first and last data points is calculated. The f score for a gene is then obtained by taking the maximum f value minus the minimum f value. High f values correlate with the presence of a cryptic transcript as it indicates the presence of excess RNA-Seq reads in the 3'-end of the gene in the mutant compared to the WT. In order to add a statistical score to the method, and also to remove biases from gene length and gene expression levels, which both can impact on the f score, the RNA-Seq values are randomly permutated several times and the f score re-calculated after each permutation. Those simulated f scores are used to calculate a z score representing the distance between the observed f score and the distribution of the permutated f scores. The z score represents the probability that cryptic transcription is initiated somewhere within the tested genes. In addition, the method allows for the identification of the position where the cryptic transcript is initiated by considering the position where the maximum f (f max) value is reached. This f max value is computed several times by re-sampled the data (each time randomly eliminating a fraction of the values) allowing for the identification of a "cryptic zone", a region within the gene where a cryptic transcript is likely to have initiated, as well as the most likely cTSS coordinate. A more detailed description of the method, as well as its benchmarking with

previously published methods, is described elsewhere (Uwimana et al, under review; See Chapter on Methodology page 19). The method is embedded with the R package yCrypticRNAs available at (https://cran.r-project.org/web/packages/yCrypticRNAs/index.html).

Pervasive sense and anti-sense transcription in *spt6-1004*

Applying the method described above to RNA-Seq experiments generated from poly(A)-enriched RNA preparations identified 1,703 intragenic sense cryptic transcripts in *spt6-1004* cells (with a false discovery rate of 1%). Importantly, this list includes all the previously confirmed cryptic transcripts for this mutant, namely *FLO8*, *RAD18*, *SPB4*, *STE11*, *VPS72*, *APM2*, *DDC1*, *SYF1*, *OMS1*, *PUS4* and *CHS6* (Figure 7A).

Contrarily to intragenic sense cryptic transcripts, anti-sense cryptic transcripts are more easily identified since very little signal is detected on the anti-sense strand of genes in WT cells. We therefore used a standard assembler (Pertea et al., 2015) to identify de novo transcripts overlapping annotated genes on the anti-sense strand (Figure 6B). Using this approach, we identified 1,616 anti-sense cryptic transcripts, overlapping 1,491 genes, in *spt6-1004* cells. Tampering with Spt6 therefore leads to widespread cryptic transcription running both sense and anti-sense relative to genes, consistent with previous studies in budding (Cheung et al., 2008; van Bakel et al., 2013) and fission yeasts (DeGennaro et al., 2013). Interestingly, anti-sense cryptic transcripts are globally expressed at lower levels than those on the sense strand (Figure 7B). It is not clear whether this is due to differences in transcription or RNA stability, although cryptic transcripts from both strands appear to be polyadenylated (see below). Sense and anti-sense transcripts emerge from the same gene slightly more often than expected by chance but we often detect genes with a cryptic transcript running only in one direction (Figure 7C). Mapping the 5' ends of anti-sense cryptic transcripts relative to the position of the 5' ends of the sense cryptic transcripts revealed that both sense and anti-sense cryptic transcripts tend to initiate from the same region within a given gene (Figure 7D).

Cryptic promoters are enriched for TATA box sequences

We next asked whether cryptic promoters share sequence attributes with canonical promoters. The best characterized core promoter motif is the TATA box. We therefore

mapped the occurrence of the TATAWAWR sequence around predicted sense and anti-sense cryptic promoters. The motif was shown enriched 50-100 bp upstream of cTSS, a pattern similar to what is observed at the TSS of annotated genes (Figure 7E). A control motif with a single mismatch did not show any enrichment, demonstrating the specificity of the signal. TATA box enrichment, however, was lower at cTSS than at the TSS of annotated genes, suggesting that cryptic promoter may preferentially use different types of promoter elements. Alternatively, this may reflect imperfect mapping of some cTSS, especially in the sense direction. The presence of TATA elements within cryptic promoters suggest that these promoters may be regulated by the same mechanism as their canonical counterparts. This result is consistent with the fact that genes with an intragenic TATA box were shown to be three time more likely to express a sense cryptic transcript in *spt6-1004* cells (Cheung et al., 2008) and with mutational analyses which demonstrated that a cryptic transcript within the *FLO8* gene requires an intragenic TATA box (Kaplan et al., 2003b). Altogether, our data suggests that while nucleosome depletion is the driving force for the emergence of cryptic initiation, the underlying DNA sequence, such as the TATA box motif, is defining the location of the cryptic promoters.

Anti-sense cryptic transcripts terminate at the 3'-end of the adjacent gene via the poly(A)-dependent termination pathway

Intragenic sense cryptic transcripts naturally terminate at the terminator site of the gene hosting them, but where does anti-sense cryptic transcription terminate? Visual inspection of the data revealed that many anti-sense cryptic transcripts terminate near the 3' end of the upstream gene (See Figure 8A for an example). In order to systematically test whether anti-sense cryptic transcripts tend to terminate in this region, we mapped the 3'-end of anti-sense cryptic transcripts relative to annotated genes. Interestingly, anti-sense cryptic transcripts tend to terminate in the 3' region of the adjacent gene (Figure 8B). These regions contain terminators but these are on the opposite strand. This prompted us to investigate whether know DNA motifs involved in polyadenylation-dependent termination were present on both strands in terminator regions. In silico and experimental analyses have defined yeast terminators as an array of motifs often referred to as the "efficiency", "positioning/A-rich", "near upstream/U-rich", "polyadenylation site" and "near downstream/U-rich" motifs ((Graber et al., 1999) and

references therein). In order to look for evidence of terminators on the anti-sense strand we mapped the density of the "efficiency" and "positioning" motifs (the other motifs having poor information content on their own), relative to genes, on both stands. We only considered motifs occurring in the correct arrangement ("efficiency" being always upstream of "positioning"). As expected, when looking on the sense strand, we found enrichment for both motifs at the 3'-end of annotated genes (Figure 8C, D, blue). Surprising, however, both motifs were also enriched on the anti-sense strand (Figure 8C, D, gold). The position of these termination motifs is consistent with anti-sense cryptic transcript coming from the downstream gene, being terminated at these sites via the polyadenylation-dependent termination pathway.

Our RNA-Seq experiments were performed on poly(A)-enriched RNA, suggesting that anti-sense cryptic transcripts are indeed polyadenylated. We noticed, however, that the level of anti-sense cryptic transcripts is on average 1.6 (mean) fold less abundant than those on the sense strand (See Figure 7B), suggesting that they may be terminating via another pathway leading to less stable transcripts. In order to address this issue, we repeated the RNA-Seq experiments using ribosomal RNA depletion and poly(A)-enrichment in parallel starting from the same total RNA preparations. Quite strikingly, the expression level of anti-sense cryptic transcripts in both datasets is very similar in all aspects (Figure 8E). Notably, the expression level of anti-sense cryptic transcripts is similarly lower than that of sense cryptic transcripts using both methods (Figure 9), further supporting the idea that both the sense and anti-sense cryptic transcripts terminate via the polyadenylation-dependent termination pathway.

Most yeast terminators are bi-directional

The data shown above implies that many, perhaps most, yeast terminators are functionally bi-directional. Terminator bi-directionality has been described anecdotally for a few yeast genes (Egli and Braus, 1994; Egli et al., 1997; Irniger et al., 1991), but its prevalence was never thoroughly investigated. Among the eight terminators that were previously shown to be bi-directional (*ARO4*, *TRP1*, *TRP4*, *ADH1*, *CYC1*, *GAL1*, *GAL7*, *GAL10*), only one, *ARO4*, is facing a cryptic transcript coming from the anti-sense strand in *spt6-1004* cells. Satisfyingly, this cryptic transcript indeed terminates around the *ARO4* terminator (not shown) demonstrating that our data can capture terminator bidirectionality.

Because hundreds of anti-sense cryptic transcripts emerge in *spt6-1004* cells, we reasoned that this could be used as an opportunity to classify terminators as uni- or bidirectional. From the 1,616 anti-sense cryptic transcripts identified in *spt6-1004* cells, we could predict the directionality of 826 terminators. Of those, 676 were classifies as bidirectional and 150 as unidirectional (see Methods) (Figure 10A). Figure 10B shows examples of bidirectional (left) and unidirectional (right) terminators identified in our analysis. Visual inspection of these putatively unidirectional terminators suggests that many of them are likely to be bidirectional but were not captured by our directionality prediction algorithm. These analyses predict that more than 80% of yeast promoters are functionally bidirectional.

In order to challenge that prediction, we tested bi-directionality on a set of terminators from each group. For that, we modified a previously developed genetic system where candidate terminators are cloned downstream of a strong promoter driving the expression of a *HIS3* gene on a plasmid allowing terminator efficiency to be monitored by Northern blotting (Carroll et al., 2004) (See Figure 10C). As expected, all terminators tested were active in this assay when tested in their natural (sense) orientation (Figure 10D, E). When cloned in the inverted orientation (anti-sense) however, all terminators predicted to be bidirectional were efficient at terminating transcription when used in the reverse orientation, confirming their bidirectionality (Figure 10D). Among the terminator predicted to be unidirectional, some behaved as expected but others were active in the reverse orientation as well (Figure 10E), confirming that our predictions are overestimating the number of unidirectional terminators. These experiments validate our predictions and establish bidirectionality as a prevalent characteristic of most yeast terminators. Importantly, these analyses clearly show that terminator bidirectionality allows for the termination of anti-sense cryptic transcription from invading neighbouring genes.

Discussion

We provide a detailed analysis of cryptic transcription in *spt6-1004* cells using RNA-Seq. Consistently with previous work, we found cryptic transcripts running both sense and anti-sense relative to annotated genes in this mutant. Cryptic promoters are enriched for TATA motifs suggesting that DNA sequence is the main determinant of cryptic promoters, despite

chromatin structure disruption being the driving force for their usage. Surprisingly, we observed that anti-sense cryptic transcription tends to terminate near the 3'-end of the upstream gene. This prompted us to systematically predict the ability of yeast terminators to terminate transcription coming from the other direction. Quite strikingly, we found that most terminators (more than 80%) are efficient at inducing termination and polyadenylation of anti-sense transcripts. Consistently, we found that DNA motifs characteristic of yeast terminators are enriched on both strands in the 3'end of genes. We therefore conclude that most yeast terminators are functionally bidirectional.

Promoters in *S. cerevisiae* are intrinsically bidirectional but divergent transcription is rapidly terminated via the Nrd1 pathway. This pathway prevents divergent transcription, initiated at canonical promoters from invading the upstream gene. From our data, it appears that contrarily to divergent transcription, cryptic anti-sense transcription is not efficiently terminated via this pathway. Indeed, these transcripts are not terminated in the promoter region of the gene that host them (as do divergent transcripts) but rather read through the intergenic region until they reach the terminator region of the upstream gene, where they terminate via the polyadenylation pathway. Why is the Nrd1 termination pathway inefficient at terminating anti-sense cryptic transcription is not known but the status of the C-terminal domain of RNAPII may be part of the answer. Indeed, the Nrd1 pathway relies mainly on P-Ser5, while the polyadenylation pathway proceeds via P-Ser2. When an RNAPII molecule transcribing an anti-sense cryptic transcript reaches the Nrd1 sites in the promoter region of the gene hosting them, it has already transcribed longer than an RNAPII molecules that would have initiated divergently from that promoter. The CTD phosphorylation status of these RNAPII is therefore likely not optimal for Nrd1-dependent termination (too low P-Ser5 and too high P-Ser2), perhaps explaining why termination via this pathway is inefficient.

Interestingly, our analysis of published RNA-Seq data from *spt6-1 Schizosaccharomyces pombe* cells shows that cryptic anti-sense transcripts in this organism preferentially terminate around promoter regions, rather than terminator region, suggesting that fission yeast has evolved different mechanisms for coping with the termination of anti-sense cryptic transcription (data not shown). Why both yeasts use different mechanisms is not

clear but differences with regards to transcription termination between these two species have been reported before.

Our analysis was performed on RNA-Seq data from cells that have functional RNA degradation pathways, which implies that the cryptic transcripts we detected are stable. Future experiments using combinations of Spt6 and exosome mutations, -or using technique such as NET-Seq or GRO-Seq that measure ongoing transcription- may reveal additional cryptic transcripts, perhaps terminated by alternative pathways in histone chaperone mutants. The use of higher resolution approaches to map the 5' and 3' ends of these transcripts shall also help decipher how cryptic promoters emerge and how cryptic transcription is terminated.

Another important aspect about cryptic transcription is the impact it may have on the expression of bona fide genes. This question is especially important because knowing that cryptic transcription occurs in cancer cells. It is increasingly recognized that non-coding transcripts and non-coding transcription can regulate gene expression through multiple mechanisms. It therefore appears likely that the emergence of massive spurious transcription as observed in *spt6-1004* cells would impact the expression of bona fide genes, notably those that host a cryptic transcript. Unexpectedly, however, we failed to show any significant effect of the presence of a cryptic transcript (should it be sense or anti-sense) on the expression of the gene hosting it. Indeed, while the expression of protein-coding genes is widely affected in *spt6-1004*, we found no correlation between these defects and the presence of cryptic transcription. This is not to say that cryptic transcription has no impact on gene expression, but simply that our data does not allow measuring it. We surmise that using mutants with less dramatic effect on chromatin structure (e.g. mutants in the Set2 pathway) may be better suited for addressing this important question.

Acknowledgement

We thank Craig D. Kaplan and Nicole J. Francis for helpful discussions and for critical reading of the manuscript. We are also thankful to Alexis Blanchette for bioinformatics support.

Funding

This work was supported by Canadian Institutes of Health Research [MOP-133648 to F.R.]. Funding for open access charge: Canadian Institutes of Health Research.

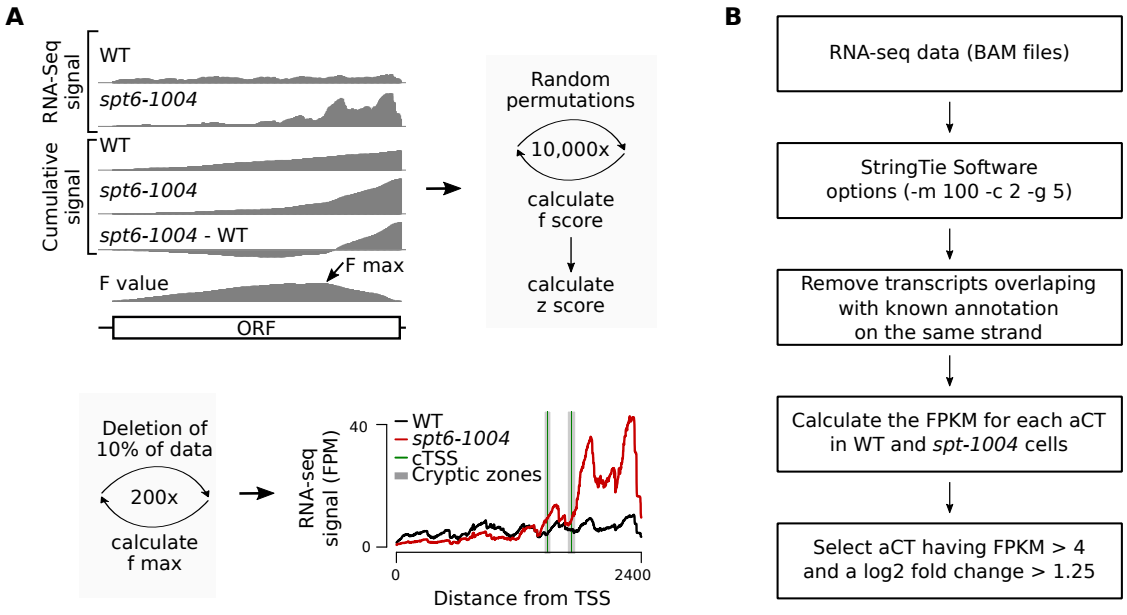

Figure 6. Identification of sense and antisense cryptic transcripts. (A) Schematic representation of the probabilistic method for the annotation of sense cryptic transcripts. (B) Workflow of how antisense cryptic transcripts were identified.

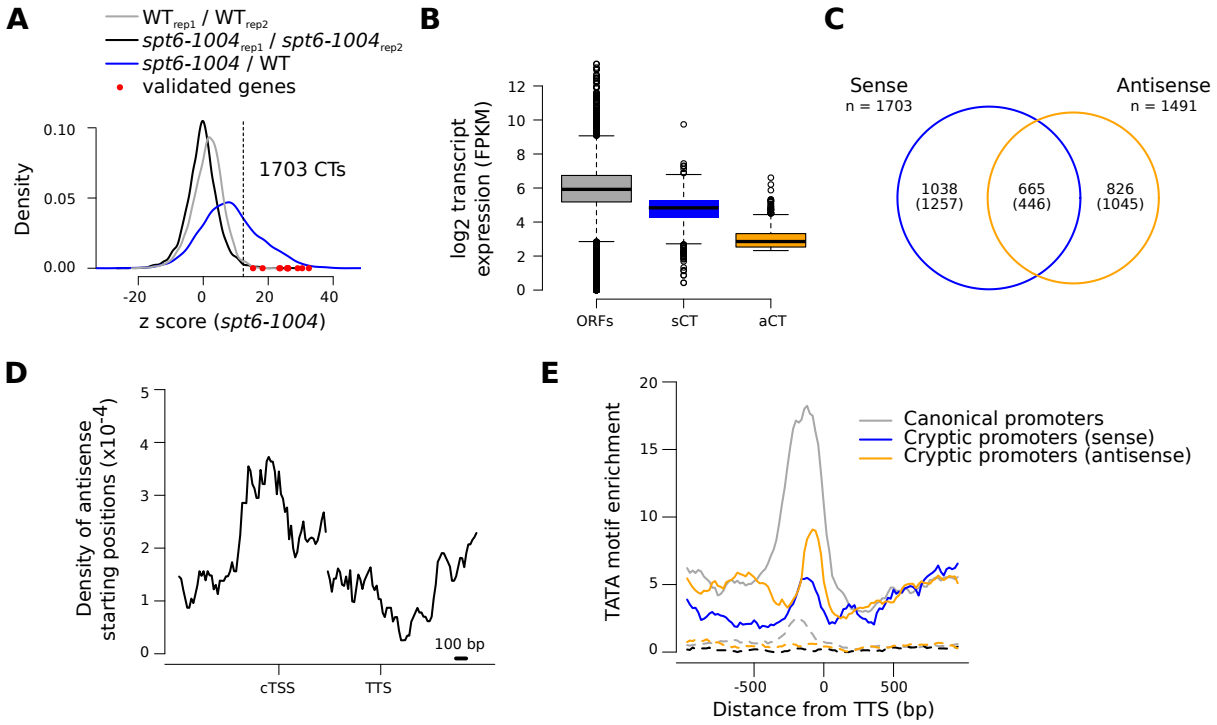

Figure 7. Sense and anti-sense cryptic transcription in *spt6-1004* cells. (A) Density of z scores obtained when comparing biological replicates as a control (black) or when comparing *spt6-1004* cells to WT cells (blue). Values for genes previously confirmed as having cryptic transcripts are represented by the red dots. n = 5,695 genes which exclude dubious and overlapping genes on the same strand. (B) Expression of transcripts in *spt6-1004* cells for three groups of genes; All 5,695 annotated ORFs (ORFs); 642 sense cryptic transcript; 1,616 anti-sense cryptic transcripts. (C) Venn diagram comparing the number of genes having intragenic sense cryptic transcripts and overlapping with anti-sense cryptic transcripts. Expected-by-chance overlap is in parenthesis. (D) Aggregate profiles of anti-sense cryptic transcripts starting positions over sense cryptic transcripts. (E) Aggregate profiles of TATAWAWR motif enrichment around canonical (grey), sense cryptic (black) and anti-sense cryptic promoters (gold). The dotted lines represent the enrichment of TATATAAT motif as a control.

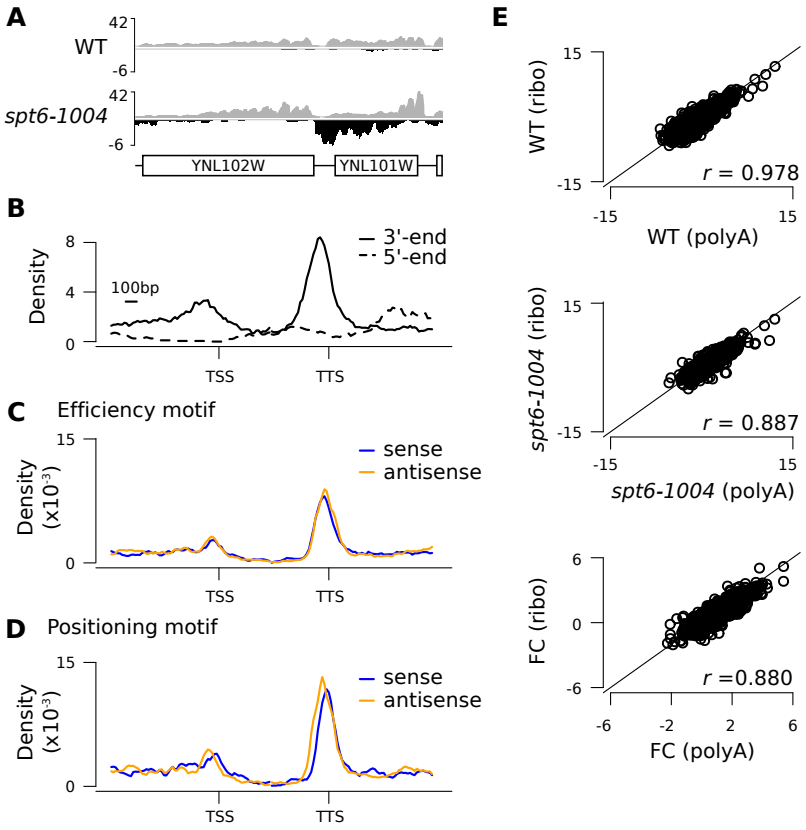

Figure 8. Anti-sense cryptic transcripts are polyadenylated and tend to terminate at the 3'-end of adjacent genes. (A) A genome-browser snapshot illustrating a terminator (*YNL102W*) efficiently terminating an anti-sense cryptic transcript. RNA-Seq signal on the Watson (grey) and Crick (black) strand are shown. (B) Aggregate profile of anti-sense starting (dashed line) and ending (plain line) positions over genes. (C) Aggregate profile of the efficiency motif (UAUAUA, UACAUA, UAUGUA) enrichment on the sense (black) or anti-sense (gold) strands over genes. (D) Aggregate profile of the positioning motif (AAUAAA, AAAAAA) enrichment on the sense (black) and anti-sense (gold) strands over genes. (B)(C)(D) The analysis are done over 1,408 yeast genes that are oriented in a divergent, tandem manner (← → →). (E) Anti-sense expression profiles in the poly(A)-enriched (polyA) and ribo-depleted (ribo) datasets. Log_2 FPKM values of each gene on the anti-sense strand in the ployA dataset are plotted against those in the ribo dataset in wild type cells (top panel) and in *spt6-1004* cells (middle panel). The bottom panel shows the log_2 fold change in anti-sense transcript expression upon Spt6 inactivation (*spt6-1004* vs WT). The r value indicates the Pearson correlation.

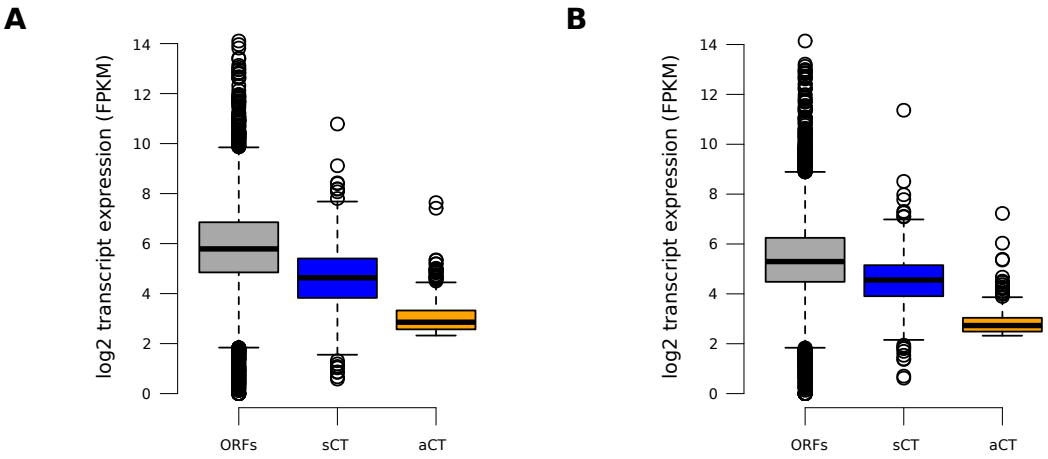

Figure 9. Expression of transcripts in *spt6-1004* cells for All annotated ORFs and sense and anti-sense cryptic transcripts. (A) Polyadenylated RNA enrichment dataset. (B) Ribosomal RNA depletion dataset.

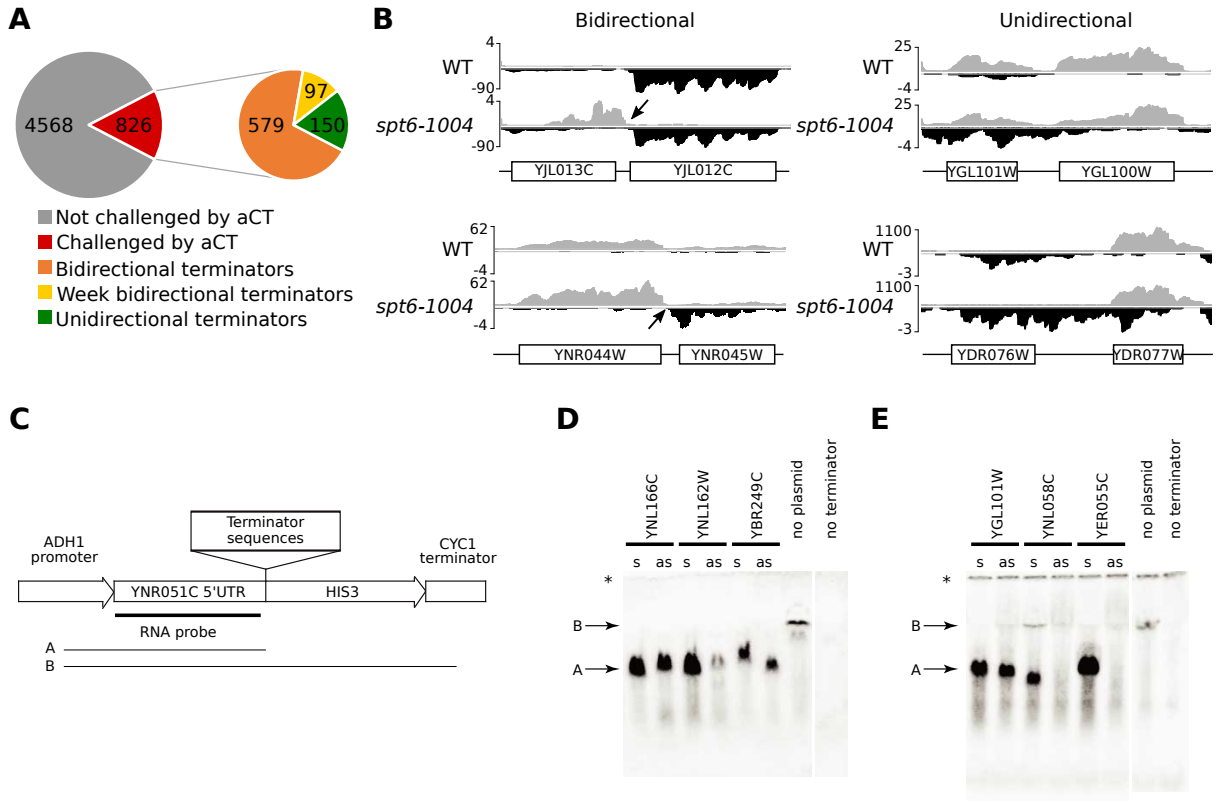

Figure 10. Yeast terminators are mostly bi-directional. (A) Pie chart displaying the terminators that are challenged by anti-sense cryptic transcription and a pie displaying the number of bidirectional, week bidirectional and unidirectional terminators. (B) Genome-browser snapshots illustrating examples of bi-directional (left panel) and unidirectional (right panel) terminators. RNA-Seq signal on the Watson (grey) and Crick (black) strand are shown. (C) Schematic representation of the terminator assay used in panels D, E. (D) RNA blot for the terminator assay testing for terminators predicted to be bidirectional (S, terminator clones in sense orientation; as, terminator cloned in anti-sense orientation). (E) Same as D but for terminators predicted to be unidirectional. * Endogenous transcript of the 5'-UTR of the reporter gene.

Discussion

La méthode probabiliste

La méthode probabiliste développée dans ce projet est un nouvel outil qui permet la détection de transcrits cryptiques à partir de données de séquençage d'ARN à haut débit. Nous avons démontré que cette méthode performe mieux que les méthodes antérieures, soit la méthode du ratio 5'/3' et la méthode d'enrichissement en 3'. Premièrement, nous avons trouvé que la méthode d'enrichissement en 3' est capable de détecter seulement 446 gènes, comparativement à la méthode du ratio 5'/3' et à la méthode probabiliste qui détectent 2151 et 1760 gènes respectivement. Ce nombre étant très petit comparé aux autres méthodes suggère que la méthode de l'enrichissement en 3' génère beaucoup de faux négatifs. Aussi, on pourrait être porté à croire que cette méthode en détecte peu mais qu'elle est très spécifique, c'est-à-dire que le peu de gènes identifiés sont en effet de vrai cas de gènes ayant des transcrits cryptiques dans leurs régions codantes. Toutefois, lors de la validation des gènes détectés par les différentes méthodes, les curateurs ont trouvé que 55 % des gènes détectés uniquement par la méthode de l'enrichissement en 3' étaient des faux positifs. Quant à la méthode du ratio 5'/3' nous avons trouvé qu'elle détecte un grand nombre de gènes comparativement à la méthode probabiliste, suggérant un manque de spécificité de cette méthode, c'est-à-dire qu'elle génèrerait beaucoup de faux positifs. En effet, lors de la validation, les curateurs ont trouvé 57 % de faux positifs comparativement à 34 % pour les gènes uniquement identifiés par la méthode probabiliste. Toutefois, il faut noter que la combinaison des trois méthodes permet de détecter les gènes ayant des transcrits cryptiques avec seulement 11% de faux positifs et la combinaison de deux méthodes permet également une bonne détection avec en moyenne 23% de faux positifs.

Bien que la combinaison des méthodes permette une meilleure spécificité pour la détection des gènes ayant des transcrits cryptiques, l'utilisation de la méthode probabiliste est tout de même plus avantageuse. La méthode probabiliste est en mesure d'identifier des gènes non identifiés par les deux autres méthodes et ce avec un faible taux de faux positif (34%). Cette spécificité est probablement due au fait que cette méthode tient compte des fluctuations

du signal d'ARN le long des gènes et des variations dans les longueurs des gènes grâce à la randomisation des données. De plus, la méthode probabiliste possède un deuxième volet qui permet d'identifier les sites d'initiation des transcrits cryptiques. Ainsi, en utilisant cette méthode, il est possible d'annoter les transcrits cryptiques intragéniques. L'annotation des régions où les transcrits cryptiques sont initiés pourrait grandement faciliter l'étude des mécanismes qui régulent la transcription cryptique, notamment dans la recherche des motifs d'ADN qui caractérisent les promoteurs cryptiques.

En utilisant la méthode probabiliste, nous avons détecté une région cryptique située entre +1671 et +1757 pb par rapport à l'ATG pour le gène *FLO8* (*YER109C*) avec un site d'initiation cryptique préférentiellement situé à +1714 pb. Les sites d'initiation cryptique dans le gène *FLO8* ont été déterminés par extension d'amorce en 2003 (Kaplan et al., 2003a). Ces régions sont +1679, +1684 et + 1685 pb et ils sont tous compris dans la région cryptique que nous avons identifiée. De plus, nous avons observé une réelle augmentation du signal d'ARN aux alentours des sites d'initiation cryptiques pour l'ensemble des gènes. Toutefois, il semblerait que notre méthode ait un biais dans la détection des zones cryptiques, celles-ci étant un peu plus en aval qu'observée par curation visuelle des résultats. Ceci est probablement dû à l'essence même de la méthode puisque celle-ci se base sur une cumulation des données le long des gènes. Une augmentation significative du signal n'est observée que quelques bases en aval de la position à laquelle il y a un gain important de signal d'ARN. Ainsi, bien que la méthode soit efficace, elle reste à améliorer. Une approche possible serait d'inclure une variable d'ajustement prenant en compte le niveau d'augmentation du signal et le nombre de paires de bases permettant cette augmentation.

La transcription cryptique sur les brins sens et antisens dans les cellules *spt6-1004*

Grâce à la méthode probabiliste, nous avons été en mesure d'identifier un grand nombre de gènes ayant un ou des promoteurs cryptiques dans leur région codante sur le brin sens dans des cellules mutantes pour la chaperonne d'histone Spt6. La totalité des gènes pour lesquels des expériences de northern blot ont confirmé la présence de transcrits cryptiques ont été identifiés

avec succès par la méthode probabiliste. Nous avons également pu identifier les promoteurs cryptiques sur le brin antisens par rapport aux gènes. Jusqu'à présent, la transcription cryptique sur le brin antisens dans les cellules *spt6-1004* n'est pas encore bien caractérisée. Ceci est sûrement dû au fait que ces transcrits sont difficilement détectables dans des expériences de puces à ADN. En séquençant de manière brin-spécifiques les ARNs totaux, nous avons été en mesure de détecter les ARNs cryptiques sur le brin antisens. La première observation que nous avons faite est que ces transcrits antisens sont moins exprimés que ceux sur le brin sens même si, en termes de nombre, ils sont aussi fréquents que ceux sur le brin sens. Nous avons ensuite voulu savoir si ces transcrits antisens étaient en réalité autant exprimés que les transcrits sur le brin sens mais que pour d'autres raisons nous n'étions pas en mesure de bien les détecter avec notre méthode. Les expériences étant faites dans un contexte où la machinerie de dégradation de l'ARN est fonctionnelle suggère que les transcrits cryptiques ne sont pas rapidement dégradés par l'exosome. La différence d'expression observée entre les transcrits cryptiques sur le brin sens et antisens pourrait-elle être dû à une différence de transcription plutôt qu'à une différence dans leur mode de dégradation ? Nous avons vu que les deux types de transcrits cryptiques sont riches en boîte TATA dans leur région promotrice suggérant qu'ils sont tous les deux régulés par les mêmes mécanismes qui régulent les promoteurs canoniques. Nous avons aussi cherché à savoir si ces transcrits antisens étaient polyadénylés ou non, ce qui pourrait expliquer le fait que l'on les détecte moins dans nos expériences. En re-séquençant les ARNs totaux mais cette fois en digérant les ARN ribosomaux plutôt que d'enrichir les ARN polyadénylés, nous n'avons remarqué aucune différence significative entre les deux ensembles de données en termes d'expression des transcrits antisens. Afin de bien comprendre les mécanismes qui régulent ces transcrits et pour pouvoir expliquer la différence d'expression sur les deux brins, d'autres travaux seront nécessaires.

Nous avons aussi découvert qu'il y a une relation entre la transcription cryptique sur les brins sens et antisens, excluant l'hypothèse que la transcription cryptique serait un phénomène complètement aléatoire. En effet, nous avons trouvé que les transcrits sur les brins sens et antisens ont tendance à initier à l'intérieur d'un même gène plus souvent qu'attendu par la chance. De plus, lorsqu'ils sont initiés dans le même gène, les transcrits cryptiques ont

le plus souvent tendance à initier dans la même région. Ceci suggère que la transcription cryptique n'est pas totalement aléatoire et qu'il existe un mécanisme faisant en sorte que le plus souvent la transcription cryptique se fait à partir d'une zone cryptique et ce sur les deux brins. Une hypothèse possible serait que plusieurs promoteurs cryptiques permettent l'initiation de la transcription de manière bidirectionnel comme c'est le cas pour certains promoteurs canoniques.

Promoteurs cryptiques

Nous avons démontré que les promoteurs cryptiques étaient significativement enrichis en boîte TATA autour de 50 à 100 pb en amont des sites d'initiation prédits. Ces motifs étaient par contre moins enrichis dans les promoteurs cryptiques comparativement aux promoteurs canoniques. Ceci peut être dû au fait que les promoteurs cryptiques utilisent préférablement d'autres séquences que nous n'avons pas été en mesure de détecter. Une autre possibilité serait que ces promoteurs utilisent d'autres variantes de la séquence TATA (avec 1 ou plusieurs non-appariements) mais puisque ces motifs sont fortement enrichis dans le génome, il est difficile d'apprécier leur enrichissement dans les promoteurs cryptiques. Une autre possibilité serait que notre méthode ne permet pas de déterminer efficacement les sites d'initiations avec suffisamment de précision, biaisant ainsi l'enrichissement observé. Comme discuté plus tôt, notre méthode possède un léger biais en estimant les sites d'initiation quelques paires de base en aval des sites réels. Ainsi, avec une meilleure prédiction, nous serions en mesure d'apprécier d'avantage l'enrichissement des motifs TATA dans les promoteurs cryptiques et éventuellement d'identifier d'autres motifs spécifiques à ces promoteurs.

Ce que l'on peut dire pour l'instant, c'est que la présence des motifs TATA suggère que les promoteurs cryptiques sont régulés de la même manière que les promoteurs canoniques. Il ne fait aucun doute que la déplétion des nucléosomes dans les cellules *spt6-1004* est à la base de la transcription cryptique, puisqu'en absence de nucléosomes la machinerie de transcription est capable de lier des régions semblables à des régions promotrices pour initier les transcrits cryptiques. Toutefois, selon les résultats que nous avons, la déplétion seule des nucléosomes n'est pas suffisante puisque nous identifions un même ensemble de promoteurs cryptiques dans les cellules mutantes pour Spt6 et Spt16, alors que

les deux chaperonnes d'histones n'affectent pas le positionnement des nucléosomes de la même manière. Nous croyons que c'est plutôt la séquence en nucléotide d'une région codante (en occurrence la présence de motifs TATA) combinée à une ouverture de la chromatine (causée par l'absence de facteurs qui régulent le positionnement des nucléosomes) qui détermine la position des promoteurs cryptiques.

La bidirectionnalité des terminateurs chez *S. cerevisiae*

Étant initiés à l'intérieur d'un gène, les transcrits cryptiques sur le brin sens sont transcrits jusqu'au terminateur du gène auquel ils sont issus. Ces terminateurs étant actifs et efficaces pour terminer l'expression des gènes, il n'y a à priori aucune raison pour ne pas terminer les transcrits cryptiques initiés en amont de ceux-ci. Par contre, nous avons observé que les transcrits cryptiques sur le brin antisens empruntent également les terminateurs des gènes, en terminant préférentiellement au terminateur du gène en amont. Nos résultats démontrent également que ces transcrits antisens sont polyadénylés et stables, puisque détectés dans un contexte où la machinerie de dégradation est fonctionnelle. Ceci suggère que ces transcrits sont terminés par la voie poly(A)-dépendante comme c'est le cas pour les gènes. Les transcrits cryptiques médiés par Spt6 se comporteraient donc comme des SUTs plutôt que comme des CUTs. Premièrement, les transcrits observés dans cette étude sont stables et polyadénylés comme les SUTs alors que les CUTs sont rapidement dégradés. En effet, lorsque la machinerie de dégradation des ARNs n'est pas altérée, les CUTs sont rapidement dégradés par l'exosome nucléaire via la voie Nrd1/Nab3/Sen1, alors que les SUTs sont stables et terminent par la voie de clivage/polyadénylation Pcf11-dépendante. Deuxièmement, les CUTs mesurent entre 200 et 500 pb alors que les SUTs sont plus long avec une longueur médiane de 761 pb. Les transcrits antisens que nous observons sont beaucoup plus long que les CUTs avec une longueur médiane de 1092 pb. La voie Nrd1-dépedante étant bien connue pour agir sur des courts transcrits tels que les CUTs et les petits ARNs nucléolaires (snoRNA), il est peu probable qu'elle agisse sur les transcrits décris dans cette étude. Toutefois, l'utilisation de méthodes de séquençage de transcrits naissants, dans un contexte où les différentes voies de terminaison seraient inactivées, pourrait nous permettre de mieux comprendre le mécanisme de terminaison des transcrits cryptiques initiés sur le brin antisens aux gènes.

Conclusions

Dans le but de mieux comprendre la transcription cryptique dans les cellules *spt6-1004*, nous avons séquencé les ARNs totaux de cette souche de manière brin-spécifique. Nous avons développé une méthode probabiliste nous permettant l'identification des transcrits cryptiques ainsi que leurs sites d'initiation à l'échelle génomique. Nous avons démontré que cette méthode surpasse celles précédemment décrites dans la littérature en termes de sensibilité.

Nous avons pu identifier plusieurs transcrits cryptiques initiés sur les brins sens et antisens par rapport aux gènes. Ces transcrits sont enrichis en motifs TATA dans leurs régions promotrices ce qui suggère d'une part qu'ils sont régulés par les même mécanismes qui régulent les gènes et d'autre part, que c'est la séquence en nucléotides qui détermine la position des promoteurs cryptiques dans le génome.

Nous avons démontré que les transcrits cryptiques sur le brin antisens terminent préférentiellement aux extrémités 3' des gènes en amont. Ces transcrits antisens étant polyadénylés et stables, ceci nous amène à proposer que les terminateurs chez *S. cerevisiae* ont évolué pour terminer la transcription de manière bidirectionnelle afin d'empêcher la transcription cryptique d'envahir les gènes voisins.

Bibliographie

Albert, I., Mavrich, T.N., Tomsho, L.P., Qi, J., Zanton, S.J., Schuster, S.C., and Pugh, B.F. (2007). Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature *446*, 572-576.

Arigo, J.T., Eyler, D.E., Carroll, K.L., and Corden, J.L. (2006). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Mol Cell *23*, 841-851.

Batta, K., Zhang, Z., Yen, K., Goffman, D.B., and Pugh, B.F. (2011). Genome-wide function of H2B ubiquitylation in promoter and genic regions. Genes Dev *25*, 2254-2265.

Belotserkovskaya, R., Oh, S., Bondarenko, V.A., Orphanides, G., Studitsky, V.M., and Reinberg, D. (2003). FACT facilitates transcription-dependent nucleosome alteration. Science *301*, 1090-1093.

Bortvin, A., and Winston, F. (1996). Evidence that Spt6p controls chromatin structure by a direct interaction with histones. Science *272*, 1473-1476.

Brewster, N.K., Johnston, G.C., and Singer, R.A. (1998). Characterization of the CP complex, an abundant dimer of Cdc68 and Pob3 proteins that regulates yeast transcriptional activation and chromatin repression. J Biol Chem *273*, 21972-21979.

Brunelle, M., Coulombe, C., Poitras, C., Robert, M.A., Markovits, A.N., Robert, F., and Jacques, P.E. (2015). Aggregate and Heatmap Representations of Genome-Wide Localization Data Using VAP, a Versatile Aggregate Profiler. Methods Mol Biol *1334*, 273-298.

Carroll, K.L., Pradhan, D.A., Granek, J.A., Clarke, N.D., and Corden, J.L. (2004). Identification of cis elements directing termination of yeast nonpolyadenylated snoRNA transcripts. Mol Cell Biol *24*, 6241-6252.

Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.J., Anderson, S., Yates, J., Washburn, M.P*., et al.* (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell *123*, 581-592.

Chang, C.H., and Luse, D.S. (1997). The H3/H4 tetramer blocks transcript elongation by RNA polymerase II in vitro. J Biol Chem *272*, 23427-23434.

Cheung, V., Chua, G., Batada, N.N., Landry, C.R., Michnick, S.W., Hughes, T.R., and Winston, F. (2008). Chromatin- and transcription-related factors repress transcription from within coding regions throughout the Saccharomyces cerevisiae genome. PLoS Biol *6*, e277.

Chu, Y., Simic, R., Warner, M.H., Arndt, K.M., and Prelich, G. (2007a). Regulation of histone modification and cryptic transcription by the Bur1 and Paf1 complexes. EMBO J *26*, 4646-4656.

Chu, Y., Simic, R., Warner, M.H., Arndt, K.M., and Prelich, G. (2007b). Regulation of histone modification and cryptic transcription by the Bur1 and Paf1 complexes. The EMBO journal *26*, 4646-4656.

Clark-Adams, C.D., and Winston, F. (1987). The SPT6 gene is essential for growth and is required for delta-mediated transcription in Saccharomyces cerevisiae. Mol Cell Biol *7*, 679-686.

Coulombe, C., Poitras, C., Nordell-Markovits, A., Brunelle, M., Lavoie, M.A., Robert, F., and Jacques, P.E. (2014). VAP: a versatile aggregate profiler for efficient genome-wide data representation and discovery. Nucleic Acids Res *42*, W485-493.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006). A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci U S A *103*, 5320-5325.

Davis, C.A., and Ares, M., Jr. (2006). Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *103*, 3262-3267.

DeGennaro, C.M., Alver, B.H., Marguerat, S., Stepanova, E., Davis, C.P., Bahler, J., Park, P.J., and Winston, F. (2013). Spt6 regulates intragenic and antisense transcription, nucleosome positioning, and histone modifications genome-wide in fission yeast. Mol Cell Biol *33*, 4779-4792.

Dion, M.F., Kaplan, T., Kim, M., Buratowski, S., Friedman, N., and Rando, O.J. (2007). Dynamics of replication-independent histone turnover in budding yeast. Science *315*, 1405-1408.

Drouin, S., Laramee, L., Jacques, P.E., Forest, A., Bergeron, M., and Robert, F. (2010). DSIF and RNA polymerase II CTD phosphorylation coordinate the recruitment of Rpd3S to actively transcribed genes. PLoS Genet *6*, e1001173.

Du, H.-N., Fingerman, I.M., and Briggs, S.D. (2008). Histone H3 K36 methylation is mediated by a trans-histone methylation pathway involving an interaction between Set2 and histone H4. Genes & development *22*, 2786-2798.

Du, H.N., and Briggs, S.D. (2010). A nucleosome surface formed by histone H4, H2A, and H3 residues is needed for proper histone H3 Lys36 methylation, histone acetylation, and repression of cryptic transcription. J Biol Chem *285*, 11704-11713.

Egli, C.M., and Braus, G.H. (1994). Uncoupling of mRNA 3' cleavage and polyadenylation by expression of a hammerhead ribozyme in yeast. J Biol Chem *269*, 27378-27383.

Egli, C.M., Duvel, K., Trabesinger-Ruf, N., Irniger, S., and Braus, G.H. (1997). Sequence requirements of the bidirectional yeast TRP4 mRNA 3'-end formation signal. Nucleic Acids Res *25*, 417-422.

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol *28*, 817-825.

Felsenfeld, G., and Groudine, M. (2003). Controlling the double helix. Nature *421*, 448-453.

Fleming, A.B., Kao, C.F., Hillyer, C., Pikaart, M., and Osley, M.A. (2008). H2B ubiquitylation plays a role in nucleosome dynamics during transcription elongation. Mol Cell *31*, 57-66.

Graber, J.H., Cantor, C.R., Mohr, S.C., and Smith, T.F. (1999). In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. Proc Natl Acad Sci U S A *96*, 14055-14060.

Guillemette, B., Bataille, A.R., Gevry, N., Adam, M., Blanchette, M., Robert, F., and Gaudreau, L. (2005). Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. PLoS Biol *3*, e384.

Hainer, S.J., and Martens, J.A. (2011). Identification of histone mutants that are defective for transcription-coupled nucleosome occupancy. Mol Cell Biol *31*, 3557-3568.

Hardy, S., Jacques, P.E., Gevry, N., Forest, A., Fortin, M.E., Laflamme, L., Gaudreau, L., and Robert, F. (2009). The euchromatic and heterochromatic landscapes are shaped by antagonizing effects of transcription on H2A.Z deposition. PLoS Genet *5*, e1000687.

Hartzog, G.A., Wada, T., Handa, H., and Winston, F. (1998). Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in Saccharomyces cerevisiae. Genes Dev *12*, 357-369.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.

Hennig, B.P., Bendrin, K., Zhou, Y., and Fischer, T. (2012). Chd1 chromatin remodelers maintain nucleosome organization and repress cryptic transcription. EMBO Rep *13*, 997-1003.

Hennig, B.P., and Fischer, T. (2013). The great repression: chromatin and cryptic transcription. Transcription *4*, 97-101.

Huang, S., Zhou, H., Katzmann, D., Hochstrasser, M., Atanasova, E., and Zhang, Z. (2005). Rtt106p is a histone chaperone involved in heterochromatin-mediated silencing. Proc Natl Acad Sci U S A *102*, 13410-13415.

Imbeault, D., Gamar, L., Rufiange, A., Paquet, E., and Nourani, A. (2008). The Rtt106 histone chaperone is functionally linked to transcription elongation and is involved in the regulation of spurious transcription from cryptic promoters in yeast. J Biol Chem *283*, 27350-27354.

Irniger, S., Egli, C.M., and Braus, G.H. (1991). Different classes of polyadenylation sites in the yeast Saccharomyces cerevisiae. Mol Cell Biol *11*, 3060-3069.

Izban, M.G., and Luse, D.S. (1991). Transcription on nucleosomal templates by RNA polymerase II in vitro: inhibition of elongation with enhancement of sequence-specific pausing. Genes Dev *5*, 683-696.

Izban, M.G., and Luse, D.S. (1992). Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. J Biol Chem *267*, 13647-13655.

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet *10*, 833-844.

Jeronimo, C., and Robert, F. (2016). Histone chaperones FACT and Spt6 prevent histone variants from turning into histone deviants. Bioessays *38*, 420-426.

Jeronimo, C., Watanabe, S., Kaplan, C.D., Peterson, C.L., and Robert, F. (2015). The Histone Chaperones FACT and Spt6 Restrict H2A.Z from Intragenic Locations. Mol Cell *58*, 1113-1123.

Joshi, A.A., and Struhl, K. (2005a). Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. Mol Cell *20*, 971-978.

Joshi, A.a., and Struhl, K. (2005b). Eaf3 chromodomain interaction with methylated H3-K36links histone deacetylation to Pol II elongation. Molecular cell *20*, 971-978.

Kaplan, C.D., Holland, M.J., and Winston, F. (2005). Interaction between transcription elongation factors and mRNA 3'-end formation at the Saccharomyces cerevisiae GAL10-GAL7 locus. J Biol Chem *280*, 913-922.

Kaplan, C.D., Laprade, L., and Winston, F. (2003a). Transcription elongation factors repress transcription initiation from cryptic sites. Science *301*, 1096-1099.

Kaplan, C.D., Laprade, L., and Winston, F. (2003b). Transcription elongation factors repress transcription initiation from cryptic sites. Science (New York, N.Y.) *301*, 1096-1099

Kaplan, C.D., Morris, J.R., Wu, C., and Winston, F. (2000). Spt5 and spt6 are associated with active transcription and have characteristics of general elongation factors in D. melanogaster. Genes Dev *14*, 2623-2634.

Katan-Khaykovich, Y., and Struhl, K. (2011). Splitting of H3-H4 tetramers at transcriptionally active genes undergoing dynamic histone exchange. Proc Natl Acad Sci U S A *108*, 1296-1301.

Keogh, M.C., Kurdistani, S.K., Morris, S.A., Ahn, S.H., Podolny, V., Collins, S.R., Schuldiner, M., Chin, K., Punna, T., Thompson, N.J.*, et al.* (2005). Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. Cell *123*, 593-605.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14*, R36.

Knezetic, J.A., and Luse, D.S. (1986). The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. Cell *45*, 95-104.

Koster, M.J., Yildirim, A.D., Weil, P.A., Holstege, F.C., and Timmers, H.T. (2014). Suppression of intragenic transcription requires the MOT1 and NC2 regulators of TATA-binding protein. Nucleic Acids Res *42*, 4220-4229.

Krogan, N.J., Kim, M., Ahn, S.H., Zhong, G., Kobor, M.S., Cagney, G., Emili, A., Shilatifard, A., Buratowski, S., and Greenblatt, J.F. (2002). RNA polymerase II elongation factors of Saccharomyces cerevisiae: a targeted proteomics approach. Mol Cell Biol *22*, 6979-6992.

Krogan, N.J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., Richards, D.P., Beattie, B.K., Emili, A., Boone, C.*, et al.* (2003). Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. Mol Cell Biol *23*, 4207-4218.

Kuo, M.H., and Allis, C.D. (1998). Roles of histone acetyltransferases and deacetylases in gene regulation. Bioessays *20*, 615-626.

Kwak, H., and Lis, J.T. (2013). Control of transcriptional elongation. Annual review of genetics *47*, 483-508.

Lambert, J.P., Fillingham, J., Siahbazi, M., Greenblatt, J., Baetz, K., and Figeys, D. (2010). Defining the budding yeast chromatin-associated interactome. Mol Syst Biol *6*, 448.

Laribee, R.N., Krogan, N.J., Xiao, T., Shibata, Y., Hughes, T.R., Greenblatt, J.F., and Strahl, B.D. (2005). BUR kinase selectively regulates H3 K4 trimethylation and H2B ubiquitylation through recruitment of the PAF elongation complex. Curr Biol *15*, 1487-1493.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet *13*, 233-245.

Li, B., Gogol, M., Carey, M., Lee, D., Seidel, C., and Workman, J.L. (2007a). Combined action of PHD and chromo domains directs the Rpd3S HDAC to transcribed chromatin. Science *316*, 1050-1054.

Li, B., Gogol, M., Carey, M., Pattenden, S.G., Seidel, C., and Workman, J.L. (2007b). Infrequently transcribed long genes depend on the Set2/Rpd3S pathway for accurate transcription. Genes & development *21*, 1422-1430.

Li, B., Howe, L., Anderson, S., Yates, J.R., 3rd, and Workman, J.L. (2003). The Set2 histone methyltransferase functions through the phosphorylated carboxyl-terminal domain of RNA polymerase II. J Biol Chem *278*, 8897-8903.

Li, B., Jackson, J., Simon, M.D., Fleharty, B., Gogol, M., Seidel, C., Workman, J.L., and Shilatifard, A. (2009). Histone H3 lysine 36 dimethylation (H3K36me2) is sufficient to recruit the Rpd3s histone deacetylase complex and to repress spurious transcription. The Journal of biological chemistry *284*, 7970-7976.

Li, J., Moazed, D., and Gygi, S.P. (2002). Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation. J Biol Chem *277*, 49383-49388.

Lickwar, C.R., Rao, B., Shabalin, A.A., Nobel, A.B., Strahl, B.D., and Lieb, J.D. (2009). The Set2/Rpd3S pathway suppresses cryptic transcription without regard to gene length or transcription frequency. PLoS One *4*, e4886.

Lindstrom, D.L., Squazzo, S.L., Muster, N., Burckin, T.A., Wachter, K.C., Emigh, C.A., McCleery, J.A., Yates, J.R., 3rd, and Hartzog, G.A. (2003). Dual roles for Spt5 in pre-mRNA processing and transcription elongation revealed by identification of Spt5-associated proteins. Mol Cell Biol *23*, 1368-1378.

Lorch, Y., LaPointe, J.W., and Kornberg, R.D. (1987). Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. Cell *49*, 203-210.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251-260.

Luger, K., and Richmond, T.J. (1998). The histone tails of the nucleosome. Curr Opin Genet Dev *8*, 140-146.

Mason, P.B., and Struhl, K. (2003). The FACT complex travels with elongating RNA polymerase II and is important for the fidelity of transcriptional initiation in vivo. Mol Cell Biol *23*, 8323-8333.

Miura, F., Kawaguchi, N., Sese, J., Toyoda, A., Hattori, M., Morishita, S., and Ito, T. (2006). A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. Proc Natl Acad Sci U S A *103*, 17846-17851.

Muratani, M., Deng, N., Ooi, W.F., Lin, S.J., Xing, M., Xu, C., Qamra, A., Tay, S.T., Malik, S., Wu, J.*, et al.* (2014). Nanoscale chromatin profiling of gastric adenocarcinoma reveals cancer-associated cryptic promoters and somatically acquired regulatory elements. Nat Commun *5*, 4361.

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

Pattenden, S.G., Gogol, M.M., and Workman, J.L. (2010). Features of cryptic promoters and their varied reliance on bromodomain-containing factors. PLoS One *5*, e12927.

Pavri, R., Zhu, B., Li, G., Trojer, P., Mandal, S., Shilatifard, A., and Reinberg, D. (2006). Histone H2B monoubiquitination functions cooperatively with FACT to regulate elongation by RNA polymerase II. Cell *125*, 703-717.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol *33*, 290-295.

Quan, T.K., and Hartzog, G.A. (2010). Histone H3K4 and K36 methylation, Chd1 and Rpd3S oppose the functions of Saccharomyces cerevisiae Spt4-Spt5 in transcription. Genetics *184*, 321-334.

Radman-Livaja, M., Quan, T.K., Valenzuela, L., Armstrong, J.A., van Welsem, T., Kim, T., Lee, L.J., Buratowski, S., van Leeuwen, F., Rando, O.J.*, et al.* (2012). A key role for Chd1 in histone H3 dynamics at the 3' ends of long genes in yeast. PLoS Genet *8*, e1002811.

Raisner, R.M., Hartley, P.D., Meneghini, M.D., Bao, M.Z., Liu, C.L., Schreiber, S.L., Rando, O.J., and Madhani, H.D. (2005). Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. Cell *123*, 233-248.

Rufiange, A., Jacques, P.E., Bhat, W., Robert, F., and Nourani, A. (2007). Genome-wide replication-independent histone H3 exchange occurs predominantly at promoters and implicates H3 K56 acetylation and Asf1. Mol Cell *27*, 393-405.

Schwabish, M.A., and Struhl, K. (2006). Asf1 mediates histone eviction and deposition during elongation by RNA polymerase II. Mol Cell *22*, 415-422.

Silva, A.C., Xu, X., Kim, H.S., Fillingham, J., Kislinger, T., Mennella, T.A., and Keogh, M.C. (2012). The replication-independent histone H3-H4 chaperones HIR, ASF1, and RTT106 co-operate to maintain promoter fidelity. J Biol Chem *287*, 1709-1718.

Smolle, M., Venkatesh, S., Gogol, M.M., Li, H., Zhang, Y., Florens, L., Washburn, M.P., and Workman, J.L. (2012). Chromatin remodelers Isw1 and Chd1 maintain chromatin structure during transcription by preventing histone exchange. Nature structural & molecular biology *19*, 884-892.

Smolle, M., and Workman, J.L. (2013). Transcription-associated histone modifications and cryptic transcription. Biochim Biophys Acta *1829*, 84-97.

Smolle, M., Workman, J.L., and Venkatesh, S. (2013). reSETting chromatin during transcription elongation. Epigenetics *8*, 10-15.

Squazzo, S.L., Costa, P.J., Lindstrom, D.L., Kumer, K.E., Simic, R., Jennings, J.L., Link, A.J., Arndt, K.M., and Hartzog, G.A. (2002). The Paf1 complex physically and functionally associates with transcription elongation factors in vivo. EMBO J *21*, 1764-1774.

Steinmetz, E.J., Warren, C.L., Kuehner, J.N., Panbehi, B., Ansari, A.Z., and Brow, D.A. (2006). Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. Mol Cell *24*, 735-746.

Strahl, B.D., Grant, P.A., Briggs, S.D., Sun, Z.W., Bone, J.R., Caldwell, J.A., Mollah, S., Cook, R.G., Shabanowitz, J., Hunt, D.F.*, et al.* (2002). Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression. Mol Cell Biol *22*, 1298-1306.

Straka, C., and Horz, W. (1991). A functional role for nucleosomes in the repression of a yeast promoter. EMBO J *10*, 361-368.

Swanson, M.S., and Winston, F. (1992). SPT4, SPT5 and SPT6 interactions: effects on transcription and viability in Saccharomyces cerevisiae. Genetics *132*, 325-336.

Tanny, J.C., Erdjument-Bromage, H., Tempst, P., and Allis, C.D. (2007). Ubiquitylation of histone H2B controls RNA polymerase II transcription elongation independently of histone H3 methylation. Genes Dev *21*, 835-847.

Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J., and Libri, D. (2006). Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. Mol Cell *23*, 853-864.

Tirosh, I., Sigal, N., and Barkai, N. (2010). Widespread remodeling of mid-coding sequence nucleosomes by Isw1. Genome Biol *11*, R49.

van Bakel, H., Tsui, K., Gebbia, M., Mnaimneh, S., Hughes, T.R., and Nislow, C. (2013). A compendium of nucleosome and transcript profiles reveals determinants of chromatin architecture and transcription. PLoS Genet *9*, e1003479.

van Dijk, E.L., Chen, C.L., d'Aubenton-Carafa, Y., Gourvennec, S., Kwapisz, M., Roche, V., Bertrand, C., Silvain, M., Legoix-Ne, P., Loeillet, S.*, et al.* (2011). XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature *475*, 114-117.

Venkatesh, S., Smolle, M., Li, H., Gogol, M.M., Saint, M., Kumar, S., Natarajan, K., and Workman, J.L. (2012). Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes. Nature *489*, 452-455.

Weake, V.M., and Workman, J.L. (2008). Histone ubiquitination: triggering gene activity. Mol Cell *29*, 653-663.

Winston, F., Chaleff, D.T., Valent, B., and Fink, G.R. (1984). Mutations affecting Ty-mediated expression of the HIS4 gene of Saccharomyces cerevisiae. Genetics *107*, 179-197.

Wood, A., Schneider, J., Dover, J., Johnston, M., and Shilatifard, A. (2005). The Bur1/Bur2 complex is required for histone H2B monoubiquitination by Rad6/Bre1 and histone methylation by COMPASS. Mol Cell *20*, 589-599.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B.*, et al.* (2005). Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell *121*, 725-737.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033-1037.

Youdell, M.L., Kizer, K.O., Kisseleva-Romanova, E., Fuchs, S.M., Duro, E., Strahl, B.D., and Mellor, J. (2008). Roles for Ctk1 and Spt6 in regulating the different methylation states of histone H3 lysine 36. Mol Cell Biol *28*, 4915-4926.

Zhang, Z., and Dietrich, F.S. (2005). Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. Nucleic Acids Res *33*, 2838-2851.

Zhang, Z., Theurkauf, W.E., Weng, Z., and Zamore, P.D. (2012). Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly(A) selection. Silence *3*, 9.