Université de Montréal

Adaptation de la levure à la suite des perturbations du mécanisme de contrôle de qualité de l'ARN

Par

Louis Gendron

Département de biochimie et médecine moléculaire, Faculté de médecine

Mémoire présenté en vue de l'obtention du grade de maîtrise
en bio-informatique

5 septembre 2019

Université de Montréal

Département de biochimie et médecine moléculaire, Faculté de médecine

Ce mémoire intitulé

Adaptation de la levure à la suite des perturbations du mécanisme de contrôle de qualité de l'ARN

Présenté par

Louis Gendron

A été évalué) par un jury composé des personnes suivantes

Franz Bernd Lang
Président-rapporteur

Sebastian Pechmann
Directeur de recherche

Étienne Caron
Membre du jury

Résumé

Le cycle de vie des ARN est déterminé par différentes étapes permettant à la cellule d'exporter et de traduire un transcrit codant. La cellule a développé un mécanisme incroyablement complexe pour s'assurer de l'intégrité des étapes de maturation de l'ARN. Le mécanisme de contrôle de qualité balance la biosynthèse et la dégradation de différents transcrits, ce qui ajout un niveau de régulation au système de l'expression génique. L'exosome est une pièce centrale du mécanisme de contrôle de qualité de l'ARN alors qu'elle dégrade une grande partie des transcrits aberrants ou non-fonctionnels dans le noyau et le cytoplasme. Ce projet caractérise et souligne la réponse cellulaire à la suite de la mutation de composantes du mécanisme de contrôle de qualité de l'ARN chez *Saccharomyces cerevisiae*. Ces perturbations comportent des composantes fonctionnelles du complexe de l'exosome (*Csl4* et *Dis3)*, un cofacteur de l'exosome nucléaire (*Rrp6*), une protéine essentielle pour la maturation des pré-ARNr (*Enp1*) et une composante de la machinerie d'export de l'ARN (*Srm1*). Ici, je présente des approches bio-informatiques pour caractériser la réponse cellulaire au niveau de l'expression des transcrits et de la taille des segments polyadénylés. La réponse au stress cellulaire intégré dans le profil d'expression du génome est très similaire entre les mutants. Ce travail suggère une réponse générique à la suite de la perturbation de différentes composantes du mécanisme de contrôle de qualité de l'ARN.

Mots-clés : bio-informatique, transcriptomique, contrôle de qualité de l'ARN, exosome, polyadénylation.

Abstract

The life-cycle of RNA is determined by several processing steps, which allow the cell to export and translate a coding transcript. The cell has developed an astonishingly complex mechanism to ensure the integrity of RNA processing steps. The quality control mechanism of RNA balances the biosynthesis and degradation of various transcripts, adding another layer of gene regulation to the complex system of gene expression. The exosome is a central piece of the RNA quality control mechanism as it degrades many of the aberrant or non-functional RNAs in the nucleus and the cytoplasm. This project characterizes and highlight a response to mutation of components from the RNA quality control mechanism in *Saccharomyces cerevisiae*. These perturbations include functional components of the exosome (*Csl4* and *Dis3*), a cofactor of the nuclear exosome (*Rrp6*), an essential protein for pre-rRNA processing (*Enp1*) and a component of RNA export machinery (*Srm1*). Here, I present bioinformatics approaches to characterize the cellular response at a level of transcript expression and polyadenylation size. The stress response embedded in the gene expression profile is highly similar between the mutants. This work suggests a generic response to a failure in different components of the RNA quality control machinery.

Keywords: bioinformatic, transcriptomic, RNA quality control, exosome, polyadenylation.

Table des matières

Liste des figures

Liste des sigles et abréviations

ARN: Acide ribonucléique

ARNm: ARN messager

ARNr: ARN ribosomique

ARNt: ARN de transfert

ARNsn: petit ARN nucléaire

ARNsno: petit ARN du nucléole

ARNnc : ARN non codant

lncARN : long ARN non codant

CUTs : Transcrits cryptiques instables

CBC : complexe nucléaire liant la coiffe

CPSF : facteurs de clivage et de polyadénylation spécifique

SUTs : Transcrits stables non annotés

XUTS : Transcrits instable de *Xrn1*

NUTs : Transcrits instables de *Nrd1*

NMD : dégradation des ARN non-sens

NSD : dégradation des ARN sans arrêt

NGD : dégradation des ARN sans départ

NGS : Séquençage de nouvelle génération

PAP : Polymérase de polyadénylation

PABPS : protéines liant les régions polyadénylées

DNBC : Classificateur Bayésien dynamique naïf

HMM : modèle de Markov à états cachés

IR : rétention d'intron

ORF : cadre de lecture ouvert

À Camille, Gilles et Jacinthe …

Remerciements

J’aimerais premièrement remercier mon directeur de recherche Sebastian Pechmann pour son soutien moral et technique. J’aimerais également remercier tous mes collègues du laboratoire Pechmann famille pour m’avoir épaulé et conseillé durant la réalisation de ce projet de maîtrise : Pedro, Savandara, Musa, Nazli, Shamim, Amruta. Finalement, j’aimerais remercier ma famille et mes proches qui m’ont suivi et encouragé lors de mes études graduées.

1 Introduction

Le premier chapitre de mémoire est divisé en deux sections. La première section de l'introduction de mémoire est une revue de la littérature permettant de comprendre les processus biologiques en relation avec le projet de recherche. Cette section sera suivie d'une introduction de la méthodologie permettant d'expliquer les différentes technologies et approches bio-informatiques en lien avec les données analysées dans ce projet.

1.1 Introduction du contexte biologique : transcriptome

Afin de synthétiser et de réguler ses protéines, la cellule a développé un système étonnamment complexe pour transcrire la séquence d'ADN d'un gène en une chaîne d'acides ribonucléiques (ARN). Ce processus de transcription permet à la cellule d'exporter la 'recette' d'une protéine encodée sous forme de séquence d'ARN messager (ARNm) du noyau vers le cytoplasme. Le processus de transcription est suivi par la seconde étape du dogme central de la biologie moléculaire, la traduction. Cette seconde étape décrit la manière dont le ribosome décode la séquence d'ARNm afin de produire une protéine.

Toutefois, la diversité des espèces d'ARN est vaste alors qu'on estime que seulement 2 % de du génome achemine une information génétique sous forme d'ARNm vers le ribosome pour produire des protéines [Matticks, 2001]. Les transcrits associés à ces régions spécifiques du génome qui encodent pour des gènes sont des ARN messagers.

D'autres régions encodent pour des transcrits fonctionnels qui ne sont pas traduit en protéine. Par exemple, les régions spécifiques aux sous-unités du ribosome (ARNr) ou aux ARN de transfert (ARNt) produisent des molécules qui permettent de faire la connexion entre l'ARNm et la synthèse d'une chaine d'acides aminés. La famille des ARN non-codants (ARNnc) est principalement composée de petits ARN nucléaires (ARNsn) [Castle, 2010], de petits ARN nucléolaires (ARNsno) [Castle, 2010], de micro ARN (miARN) [Bissels, 2009] et de longs ARN non-codants (lncARN) [Menet, 2012]. Ces espèces d'ARN que l'on comprend moins bien, même s'ils

n’encodent pas pour des gènes, ont possiblement d’importantes fonctions de régulation sur l’expression génique.

Les transcrits cryptiques font partie d’une classe spéciale d’ARN non-codant. Ces transcrits normalement dégradés par différentes nucléases sont normalement indétectable chez une cellule saine. Tout comme les différentes espèces d’ARN non-codant, il est possible que les transcrits cryptiques aient un impact sur la régulation de l’expression du transcriptome.

1.1.1 La maturation de l’ARN

Plusieurs mécanismes de régulation de l’expression génique, tels que l'inhibition de la traduction, la rétention nucléaire des ARNm ou la dégradation des transcrits visent à contrôler la quantité d’ARNm fonctionnelle disponible pour la traduction. La concentration d’ARNm cellulaire est le résultat de la synthèse de transcrits et de leur dégradation. Ensemble, ces deux phénomènes contribuent à la portée dynamique de l’expression d’une protéine atteint par une cellule sous différentes conditions. De plus, les précurseurs d’ARN nouvellement synthétisés subissent plusieurs modifications avant d’être considéré comme étant « mature ». Cette étape inclut en autre les modifications des deux extrémités de l’ARN ainsi que le retrait de segments non-codants au sein du corps de l’ARN. Ces étapes de régulation donnent à la cellule encore plus de contrôle sur la régularisation de l’expression d’un gène, notamment en contrôlant le taux auquel un ARNm spécifique est dégradé.

Les étapes importantes de ce processus de maturation consistent à l’ajout d’une coiffe à l’extrémité 5’ du transcrit, au retrait des introns couplés à liaison des exons (épissage) et au clivage de l’extrémité 3’ couplé de l’ajout successif d’adénosines (polyadénylation) [Bentley, 2014]. Cette extension en 3’ est normalement dégradé pour générer un ARNm mature qui peut être transporté du noyau au cytoplasme pour la traduction. La dégradation 3’→5’ assurée par différentes nucléases joue un rôle important dans les voies métaboliques permettant la régulation du niveau d’ARNm et dans les mécanismes de contrôle de qualité comme la dégradation des ARNm non-sens (NMD) [Houseley, 2006].

1.1.1.1 Modification en 5': la coiffe

Chez les eucaryotes, la suite de réactions permettant l'ajout du groupement 7-methylguanosine à l'extrémité 5' du transcrit est primordiale. L'ajout de la coiffe prend place durant l'étape de transcription, une fois que les 25-30 premiers ribonucléotides sont ajoutés au transcrit naissant [Moteki, Price, 2002]. En plus de stabiliser le précurseur d'ARN, cette modification post-transcriptionnelle permet de coordonner plusieurs étapes du cycle de vie de l'ARNm. La coiffe permet de recruter le complexe nucléaire liant la coiffe (CBC) jouant un rôle crucial pour l'assemblage du spliceosome, l'export l'ARN, les futures modifications en 3', la protection du transcrit contre l'action d'une ribonucléase et la dégradation potentielle d'un ARN non-sens [Topisirovic, 2011]. Dans le contexte de dégradation chez *Saccharomyces cerevisiae*, on retrouve deux systèmes de contrôle de qualité de la coiffe. Le complexe *Rai1-Rat1* [Jiao X, 2010] et la protéine *Dxo1* [Chang JH, 2012] sont tous deux capables de retirer la coiffe et de catalyser la dégradation 5'→3'.

1.1.1.2 Modification en 3' : polyadénylation

La polyadénylation est une modification post-transcriptionnelle très importante pour la stabilisation, le transport nucléocytoplasmique et la régulation de l'expression de plusieurs types d'ARN. Ce processus décrit une réaction de dégradation 3'→5' couple à l'addition d'une suite d'adénosines. Le segment poly(A) qui est ajouté en 3' est une cible d'interaction pour plusieurs protéines liant les régions polyadénylés (PABPs) qui sont connues pour leur implication à la régulation de l'expression [Goss DJ, 2013]. Puisque cette étape de maturation est importante pour la stabilisation, l'initiation de la traduction et l'export de l'ARN, la vaste majorité des ARNm chez l'eucaryote subissent l'étape de polyadénylation. Une défaillance du système de polyadénylation, que ce soit une mutation au niveau du motif de reconnaissance de la polyadénylation ou une mutation des facteurs de polyadénylation, peut être relié à des maladies telles que le diabète de type I et II [Locke, 2001] et certains cas de tumeurs MCL [Wiestner, 2007].

Chez la levure comme chez les mammifères, le site de polyadénylation est marqué par un motif spécifique initialement reconnu par les facteurs de clivage et de polyadénylation spécifique (CPSF) [Dichtl, 2001]. Une partie de l'extrémité 3' est dégradée, puis la polymérase poly(A) (PAP) débute l'élongation du segment poly(A). Lorsque le segment poly(A) atteint entre 10 à 12 résidus, la

protéine *Pabpn*1 se lie à la chaîne d'adénosines [Nemeth A, 1995]. L'interaction directe de cette protéine avec la polymérase poly(A) permet de stabiliser le complexe jusqu'à ce que le segment poly(A) atteigne une longueur critique. À ce moment, la conformation de *Pabpn1* avec la chaîne poly(A) change de façon à déstabiliser l'interaction entre CPSF et PAP, ce qui met un terme à l'élongation du segment poly(A). Une étude [Keller R.W., 2000] par microscopie d'électron visant à caractériser le complexe *Pabpn1* avec la chaîne poly(A) a permis d'observer la formation de structures sphériques. Chez l'homme, ces particules permettent d'accueillir 200 à 300 nucléotides de la chaîne poly(A) où celle-ci pourrait être repliée lors de l'élongation du segment de sorte que l'interaction entre CPSF et PAP soit maintenue. Une fois que la longueur critique du segment poly(A) est atteinte et que la particule ne permet plus d'accueillir d'autres nucléotides, l'interaction entre CPSF et PAP pourrait être compromise menant à la terminaison de l'élongation du segment poly(A) [Kühn & Gündel, 2009].

Initialement mesurée entre 150 à 250 résidus [Jacobson A, 1996], la taille des régions poly(A) a été revue à la baisse par plusieurs études récentes [Choi YH. 2003, Meijer HA. 2007. Chang H, 2014] qui estiment que la population de segment poly(A) varie entre 50 à 100 résidus. Puisque certains mécanismes de dégradation ont été associés à la taille du segment poly(A), la régulation de la taille de ces segments est sans doute importante pour la cellule. Lorsqu'un ARNm s'approche de la fin de son cycle de vie, la queue poly(A) est raccourcie par une déadénylase menant à la dissociation des PABPs. Les ARN sans PABPs avec des courtes queues poly(A) (< 20) sont ciblés par les uridylyltransférases TUT4 et TUT7 pour l'uridylation [Lim J, 2014]. Cette modification similaire à la polyadénylation consiste à l'ajout de quelques uridines en région 3'. L'uridylation est connu pour faciliter la dégradation 5'→3' et 3'→5' des ARNm déadénylés [Lim J, 2014].

Toutefois, la relation entre la taille du segment polyadénylé et le niveau d'expression du transcrit n'est pas exactement claire. Alors que le dogme de la déadénylation induit une inhibition de la traduction et une dégradation de l'ARNm, des études récentes sont parvenues à observer que les transcrits fortement exprimés possèdent des segments poly(A) relativement courts (environ 30 nucléotides) chez plusieurs eucaryotes [Lima, 2017] [Subtelny AO, 2014]. Il n'est donc toujours pas clair comment la cellule détermine la taille critique du segment poly(A) et comment ce

phénomène régule potentiellement l'interaction du transcrit avec d'autre protéines pouvant avoir un impact sur le cycle de vie ou l'expression de l'ARN.

Il a été démontré que les transcrits hypoadénylés chez la levure sont retenus dans le noyau et ciblés pour la dégradation [Saguez, 2008] alors que les transcrits hyperadénylés observés chez les cellules infectées du virus de l'herpès sont également retenus dans le noyau pour y être dégradé [Lee YJ, 2009]. La manière dont la cellule arrive à synthétiser et à mesurer précisément la taille du segment polyadénylé reste toutefois incertaine. L'interaction de la protéine *Pabpn1* liant le segment poly(A) avec le transcrit pourrait possiblement participer à expliquer ce phénomène. Une diminution l'expression de *Pabpn1* chez les myoblastes de la souris induit un raccourcissement des segments poly(A) et une accumulation de ces transcrits dans le noyau [Apponi LH, 2010]. Un phénomène similaire est observable chez les cellules infectées du virus Influenza A où la protéine virale *NS1* séquestre la protéine *Pabpn1* [Chen Z, 1999]. D'autres facteurs in vivo ont également été observés, la protéine *NPM1* qui s'associe aux ARNm après l'étape de polyadénylation en fait partie. La mutation de cette protéine induit une hyperadénylation et une rétention nucléaire des ARNm. Alors que *Pabpn1* semble être un acteur important dans le système permettant d'optimiser la taille des segments polyadénylés, il est fort probable qu'il y existe d'autres cofacteurs permettant d'expliquer le fonctionnement de ce phénomène.

1.1.1.3 Épissage

L'épissage est une étape primordiale pour la diversité du protéome d'une espèce, surtout pour les organismes complexes. Le processus d'épissage consiste à retirer les parties non-codantes (introns) d'un ARNm et d'ensuite joindre les parties codantes (exons). L'étape d'épissage est kinétiquement plus lente que l'ajout de la coiffe et la polyadénylation, c'est pourquoi le substrat de cette étape est généralement un ARN linéaire muni d'une coiffe et d'une queue poly(A) [Padgett, 1986]. Chez la levure, tout comme chez les mammifères, cette réaction est catalysée par l'interaction du spliceosome et d'un vaste ensemble de cofacteurs avec le précurseur d'ARNm dans le noyau [Stark, 2006]. Alors que le génome de la levure ne contient que peu de gènes introniques, dont la majorité ne contient qu'un seul intron, les organismes plus complexes possèdent beaucoup de gènes multi-exoniques.

Le phénomène d'épissage alternatif décrit la manière dont les différents exons d'un gène sont utilisés ou retirés du corps du transcrit. Ce processus très fréquent chez les organismes plus complexes permet de créer plusieurs isoformes du transcrit à partir du même gène, ce qui contribue grandement à la diversité du protéome. On estime que près de 95% des transcrits chez l'humain sont issus du processus d'épissage alternatif [Pan, 2008] [Wang, 2008]. Le retrait des introns et la réorganisation des exons est un processus important qui est contrôlé de près par la cellule.

Une défaillance au niveau de l'étape d'épissage peut produire un ARN aberrant. Ce transcrit peut produire plusieurs isoformes incorrectes de la protéine ce qui pourrait avoir un impact négatif sur des processus cellulaires. On associe le phénomène de rétention d'intron aux transcrits dont les introns n'ont pas été correctement retirés du corps l'ARN. Ce phénomène est généralement un indicateur de la défaillance de processus d'épissage. Étant donné l'importance du processus d'épissage, la cellule a développé un système de contrôle de qualité de l'épissage afin de cibler et dégrader les transcrits où l'étape d'épissage ne s'est pas déroulée adéquatement. L'une des composantes de ce mécanisme de contrôle de qualité repose sur la dégradation des transcrits aberrants par l'action de nucléase.

La levure a mise au point des facteurs de rétention pour prévenir l'export d'ARNm aberrants vers le cytoplasme [Dziembowski, 2004]. Lorsqu'un transcrit aberrant est retenu dans le noyau, la coiffe du transcrit est retirée par la protéine *Xrn2* [Davidson, 2012] puis le transcrit est dégradé par une nucléase, généralement le complexe de l'exosome. En plus des mécanismes généraux de dégradation des ARN, certaines nucléases peuvent cibler des groupes de transcrits spécifiques, tel que l'endonucléase *Rnt1p* avec les précurseurs d'ARNr n'ayant pas subi l'étape d'épissage [Danin-Kreiselman, 2003].

Il arrive parfois que les transcrits dont le processus d'épissage s'est déroulé incorrectement soient exportés vers le cytoplasme. La rétention de l'intron au sein du corps de l'ARN peut introduire un changement de phase du cadre de lecture ou même l'ajout prématuré d'un codon de terminaison de la traduction. Il a même été établi que les régions introniques étaient sous haute pression sélective pour induire un codon de terminaison de la traduction au sein de la séquence [Jaillon,

2008]. La traduction de ces transcrits pourrait produire une grande quantité non-fonctionnelle de protéines tronquées pouvant avoir un effet néfaste sur certains processus cellulaires. Ce type de transcrit aberrant est pris en charge par le mécanisme de dégradation des ARNm non-sens (NMD) où la dégradation des transcrits est assurée par l'exosome cytoplasmique ou par l'exonucléase *Xrn1* [Chang YF, 2007]. Une défaillance de la voie NMD induit une accumulation générale des ARNm non-épissés nous indiquant que ce système contribue en grandement partie au mécanisme de contrôle de qualité de l'épissage [Egecioglu, 2011]. L'épissage est donc un processus très important dans la cellule dont l'intégrité du fonctionnement est contrôlée de près, autant dans le noyau que dans le cytoplasme.

1.1.2 Mécanisme de contrôle de qualité de l'ARN

Lors du cycle de vie d'un ARN, le transcrit nouvellement synthétiser lors de l'étape de transcription subit une variété de modifications avant d'être ultimement traduit en protéine. L'intégrité du fonctionnement des mécanismes apportant les modifications post-transcriptionnelles de l'ARN sont cruciales pour la cellule. Alors que le processus de maturation comporte plusieurs étapes telles que les modifications post-transcriptionnelles et l'export nucléocytoplasmique, cette section vise à présenter les principaux acteurs liés aux mécanismes de contrôle de qualité de l'ARN.

Le système de contrôle de qualité de l'ARN correspond à l'ensemble des molécules qui supervisent les différentes étapes de maturation et du cycle de vie des transcrits. Ce système permet de prévenir l'accumulation d'ARN qui ne sont pas fonctionnels et permet également de réguler l'expression génique. Ce niveau de régulation est obtenu par l'équilibre dynamique entre la synthèse et la dégradation des ARN, ce qui permet de contrôler le niveau d'ARN fonctionnellement disponible. L'origine d'un ARN aberrant peut être multiple puisque celui-ci participe à plusieurs réactions dans différentes régions de la cellule. L'introduction prématurée d'un codon de terminaison de la transcription ou toutes autres erreurs dans le processus de maturation sont généralement suffisantes pour que le transcrit soit ciblé pour la dégradation. La transcription d'une région intergénique produisant des transcrits cryptiques peut également produire des transcrits généralement dépourvus de fonction qui sont dégradés. Ce système de

contrôle de qualité permet aussi de dégrader des transcrits provenant d'un rétrotransposons ou d'un virus [Goodier, 2016].

La dégradation des transcrits peut se produire dans le noyau et dans le cytoplasme. À l'intérieur du noyau, la protéine *Xrn2* catalyse la dégradation 5'→3' ainsi que le retrait de la coiffe. Elle agit de pair avec le complexe de l'exosome et du cofacteur Rrp6 qui assurent la dégradation 3'→5' [Houseley, 2006]. La figure 1 offre une représentation simplifiée des mutations analysés dans ce projet dans le contexte de dégradation des ARN aberrants tels que les transcrits cryptiques.

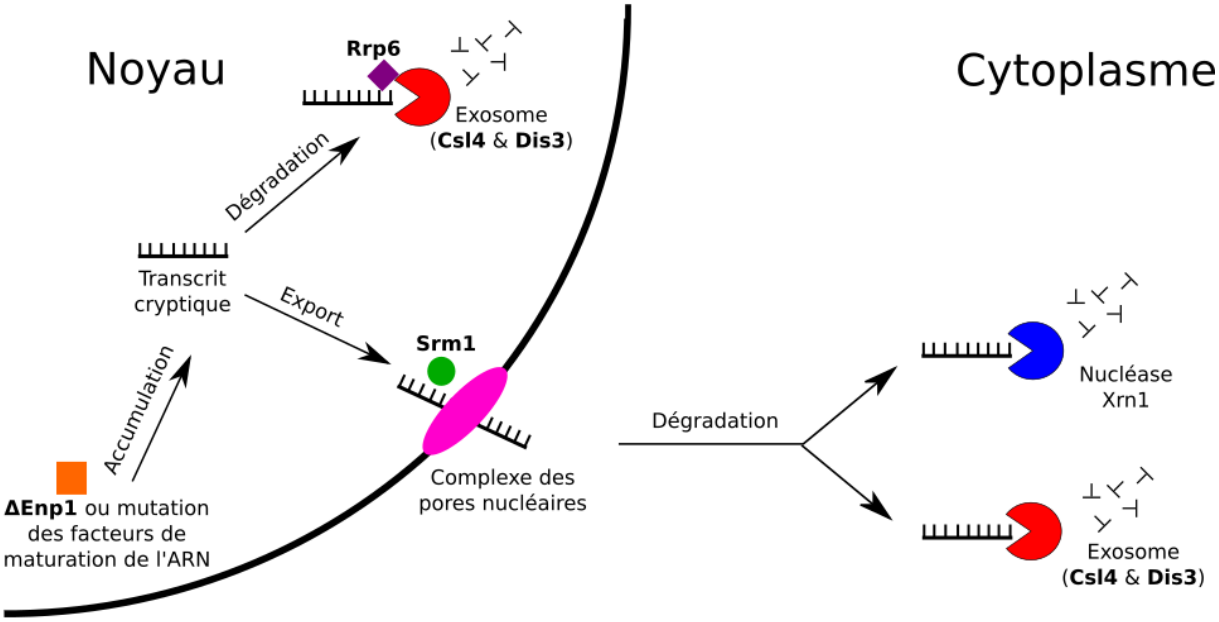

Figure 1. – PRINCIPAUX MUTANTS ANALYSÉS DANS LE CONTEXTE DE DÉGRADATION DES TRANSCRITS CRYPTIQUES. Le mécanisme de contrôle de qualité de l'ARN peut être déstabiliser en affectant différentes les composantes du complexe de l'exosome (*Csl4*, *Dis3* et le cofacteur *Rrp6*), du mécanisme de transport nucléocytoplasmique (*Srm1*) et du mécanisme de maturation des ARNr (*Enp1*).

Dans le cytoplasme, plusieurs mécanismes permettent de cibler les ARN aberrants pour la dégradation. La plupart d'entre eux utilisent des protéines qui interagissent avec la machinerie de traduction pour discriminer les ARN aberrants du reste des transcrits. Par exemple, le mécanisme de dégradation des ARN non-sens (NMD) utilise les protéines *Upf* pour cibler les codons de terminaison prématurés puis dégrade le transcrit aberrant [Isken O, 2007]. Le mécanisme de dégradation des ARN sans arrêt (NSD) permet de cibler les transcrits sans codon de terminaison. L'action de la protéine *Ski7* avec le site A du ribosome durant l'élongation permet de reconnaître lorsque ce site n'est pas occupé et permet de cibler le transcrit pour la dégradation. Dans le cas du mécanisme sans départ (NGD), si le ribosome s'arrête trop longtemps durant l'élongation, les protéines *Hbs1* et *Dom3* interagissent avec le ribosome pour initier le clivage du transcrit par une endonucléase [Harigaya, 2010]. Bien que le mécanisme sous-jacent reste inconnu, il a été démontré que les ARNr et la ARNt défectueux sont dégradés plus rapidement [LaRiviere, 2006] [Alexandrov, 2006]. Bien que les mécanismes soient multiples, le principe reste toujours assez similaire: les transcrits aberrants sont d'abord reconnus parmi le reste des transcrits par l'interaction d'une protéine ou d'un complexe puis ils sont ciblés pour la dégradation. La dégradation des transcrits dans le cytoplasme est assurée par la protéine *Xrn1* (figure 1) qui retire la coiffe et dégrade l'ARN en 5'→3' ainsi que l'exosome qui assure la dégradation dans le sens inverse.

1.1.3 Exosome

La majorité des réactions de dégradation 5'→3' est assurée par le complexe de l'exosome. Celui-ci est conservé dans l'évolution à travers le taxon de la plupart des procaryotes jusqu'aux eucaryotes. Premièrement découvert en 1997, le complexe de l'exosome était décrit par un ensemble de cinq exonucléases essentielles (*Rrp4, Rrp41, Rrp42, Rrp43* et *Rrp44*) à la maturation des régions 3' des ARNr 5.8 S [Mitchell, 1997]. Le rôle de *Rrp4* et *Rrp41* dans le mécanisme de dégradation des ARNm a permis de suggérer que ce complexe opérait dans le noyau et dans le cytoplasme [Anderson, 1998]. Six sous-unités additionnelles (*Rrp6, Rrp40, Rrp45, Rrp46, Csl4* et *Mtr3*) ont ensuite été reconnues comme membres du complexe alors qu'une augmentation des ARNr 5.8 S avec une extension en 3' était observable pour toutes levures mutées pour les nouvelles sous-unités tout comme les cinq exonucléases essentielles [Allmang, 1999].

L'échafaudage du complexe est composé de six protéines formant une structure d'anneau produisant un canal central qui reçoit l'ARN tandis que les sous-unités *Csl4*, *Rrp4* et *Rrp40* se lient au complexe en périphérie de l'entrée du canal central. Alors que ces neuf sous-unités essentielles forment la base du complexe de l'exosome, d'autres cofacteurs peuvent se lier au complexe. Par exemple, l'exosome nucléaire chez la levure possède comme cofacteur l'exonucléase *Rrp6* [Synowsky, 2009] tandis que le cofacteur *Ski7* s'associe à l'exosome cytoplasmique pour stimuler la dégradation d'un ARN viral [van Hoof, 2000].

Un acteur clef pour l'action de l'exosome dans le noyau est le complexe de polyadénylation TRAMP. Celui-ci est composé de la polymérase poly(A) (*Trf5*), de la protéine liant l'ARN (*Air1* ou *Air2*) et d'une hélicase (*Mtr4*). Ce complexe permet de stimuler la dégradation d'une courte partie de la région poly(A) de manière à créer une extension simple brin au transcrit. Cette modification permet de stimuler la réception du transcrit par l'exosome, favorisant ainsi la dégradation de l'ARN [LaCava, 2005]. Il n'est toutefois pas clair comment le complexe TRAMP cible spécifiquement un transcrit aberrant pour la polyadénylation et donc la dégradation.

Un autre complexe agissant étroitement avec l'exosome nucléaire est le complexe NSS. Celui-ci est composé de deux protéines liant l'ARN pour des séquences spécifiques (*Nrd1 et Nab3*) ainsi qu'une hélicase (*Sen1*). Ce complexe cible les transcrits courts (<1000 nucléotides) pour induire la terminaison de la transcription et possiblement stimuler la dégradation du transcrit par l'entremise de son interaction avec le complexe TRAMP [Tudek, 2014]. Parmi les cibles du complexe NNS, on retrouve principalement des ARNs non-codants stables comme les ARNsno ou les transcrits cryptiques instables (CUTs) qui sont observables seulement lorsqu'on observe une défaillance du mécanisme de contrôle de qualité de l'ARN.

Alors qu'il existe plusieurs façons de déstabiliser le fonctionnement de l'exosome, la mutation de certaines protéines clés du complexe permettent d'étudier le rôle de cette nucléase (figure 1). La protéine *Rrp44*/*Dis3* est la seule nucléase active du complexe de l'exosome chez la levure où elle se lie dans le bas de la structure en anneau vers la fin du canal central [Liu J, 2016]. Cette protéine à domaines multiples est une homologue de l'ARNase R du E. Coli de la famille des ARNase II qui se caractérise par l'activité d'exonucléase 3'→5'. Bien que la protéine *Dis3* ne possède que 19 %

d'homologie de séquence avec l'ARNase II, les sites actifs sont grandement conservés et les domaines sont arrangés de façon similaire le long de la séquence.

Les protéines *Csl4* et *Rrp4* sont également très importantes pour le fonctionnement de l'exosome. Situées à l'entrée du canal central de l'exosome, ces protéines confèrent un domaine de liaison de l'ARN permettant de cibler l'extension 3' simple brin d'un ARN et de le faire parvenir vers l'intérieur de la structure pour la dégradation [Lorentzen, 2007]. Ces deux protéines conservées chez les archéobactéries et chez les eucaryotes pourraient également favoriser la sélection de différents substrats. Une étude sur la spécificité des substrats menée sur l'archéobactérie *Sulfolobus solfataricus* a permis d'observer que l'exosome assemblé avec la protéine *Csl4* favorisait les transcrits peu adénylés alors que Rrp4 donnait plus de spécificité pour les transcrits polyadénylés et les longs ARNs [Evguenieva-Hackenberg, 2009].

La protéine *Rrp6*, présente seulement chez l'exosome nucléaire de la levure, est importante pour la maturation d'ARN (ARNr, ARNsn et ARNsno) ainsi que pour la dégradation de transcrits cryptiques (CUTs). Ce cofacteur de l'exosome se situe au-dessus du domaine S1 permettant l'ouverture sur le canal central et permet d'améliorer l'accès de l'ARN au canal central en favorisant l'élargissement de l'entrée du canal [Wasmuth, 2012]. Une mutation du gène *Rrp6* entraîne une surélongation de plusieurs transcrits dépendant du complexe NSS. Ces transcrits se manifestent comme une interférence de la transcription du brin antisens sont également observables chez les mutants du gène *Nrd1* du complexe NNS [Fox, 2015].

1.2 Introduction de la méthodologie

Cette seconde section de l'introduction est une revue des technologies de séquençage d'ARN et des approches bio-informatiques reliées à ce projet. Cette section permettra de comprendre la manière dont les données étudiées dans ce projet sont générées, normalisées et modélisées en plus d'expliquer la façon dont ces approches nous permettent d'étudier la défaillance du mécanisme de contrôle de qualité de l'ARN.

1.2.1 Séquençage de nouvelle génération

Toutes les données analysées dans ce projet sont issues de la technologie de séquençage de nouvelle génération (NGS). Cette technologie permet de déterminer l'ordre des nucléotides d'un fragment d'ADN. La technologie de séquençage de nouvelle génération a révolutionné le domaine de la génétique en diminuant grandement le coût et en accélérant le processus du séquençage. Permettant le séquençage du génome, cette technique permet aussi de séquencer l'ARN (RNA-Seq), ce qui permet d'obtenir les séquences de la faible fraction du génome qui encode pour des protéines.

Les transcrits sont séquencés par fragment puisque que la précision du séquenceur diminue fortement lorsque la taille du transcrit augmente. Par exemple, pour les séquenceurs de type Illumina, la taille maximale des fragments est limitée à 300 nucléotides. Les expériences de séquençage de type pairé consistent à séquencer seulement les extrémités du fragment, sur habituellement 50 nucléotides. Cette approche fournit assez d'informations pour cartographier le transcrit sur un génome de référence et permet de générer des fragments plus courts qui sont séquencés avec un score de qualité élevé.

Avant le séquençage, les ARN sont clivés en courts segments puis amplifiés par PCR pour obtenir plusieurs copies du même fragment. Des adaptateurs sont ensuite ajoutés aux deux extrémités des fragments afin que ceux-ci puissent s'associer aux oligos de la plaque de séquençage. Une fois que les fragments simple brin sont liés à la plaque de séquençage, l'étape de séquençage débute. Cette étape repose sur le principe d'émission du signal de fluorescence propre à chaque type de nucléotide. Le séquençage débute par l'ajout d'une séquence complémentaire à l'adaptateur qui vient s'hybrider à l'extrémité du fragment pour former le début d'une structure double brin. Le reste de l'étape de séquençage se fait sous forme cyclique où un nucléotide fluorescent complémentaire au nucléotide du fragment s'hybride à la structure simple brin puis, une lecture du signal de fluorescence est effectuée. Ce cycle produit une série de signaux fluorescents dont l'ordre chronologique permet de savoir l'ordre dans lequel les nucléotides ont été ajoutés, ce qui correspond à la séquence complémentaire du fragment. Cette technique permet alors de séquencer rapidement des millions des séquences génomiques alors que le cycle de séquençage se fait simultanément sur les vastes étendues d'oligo disposées sur la cellule de

séquençage. Alors que le séquenceur produit des lectures de fragments, plusieurs étapes bio-informatiques sont nécessaires pour analyser ces données.

1.2.1.1 Étapes bio-informatiques d'analyse des données de NGS

Plusieurs étapes bio-informatiques sont nécessaires à l'analyse des données de NGS. Les adaptateurs fixés aux extrémités des fragments peuvent sont d'abord être retirés des séquences à analyser. Les fragments dont la qualité du score de séquençage est plus faible sont ensuite retirés de l'ensemble des fragments. On peut ensuite cartographier les fragments séquencés sur le génome de référence de l'espèce à l'étude. Les fragments cartographiés sont rapportés dans un fichier d'alignement SAM, pour « Sequence Alignment Map », dont les coordonnées de cartographie peuvent être utilisées pour associer un gène pour chaque fragment à l'aide d'un fichier d'annotation du génome. On peut également aligner nos fragments à une liste de possibles contaminants afin de retirer ces séquences de nos données. Si l'on désire évaluer l'expression des gènes, les données devront toutefois être normalisées afin de contrôler différents biais tels que la taille du gène, la taille et la composition de la libraire de séquençage, le contenu GC, etc.

1.2.1.2 Normalisation et analyse d'expression différentielle

Une expérience RNA-Seq permet de séquencer un grand volume d'ARN afin d'analyser l'expression du transcriptome. Lors de ces expériences, on séquence généralement le même type de cellule dans les mêmes conditions afin de produire des réplicas nous permettant d'évaluer la variance des données du transcriptome. On séquence aussi des cellules sous d'autres conditions afin d'évaluer le changement de l'expression d'un gène à travers le temps ou sous différentes conditions. Lors de cette étape d'analyse différentielle, les données des différents réplicas sous différentes conditions sont normalisées pour corriger divers biais.

L'approche de normalisation de l'outil DESeq2 [Love, 2014] consiste à déterminer la moyenne géométrique de chaque gène parmi tous les échantillons. Pour chaque échantillon, le compte du gène est divisé par cette moyenne géométrique produisant un ratio pour chaque gène. La médiane de ces ratios chez un échantillon est utilisée comme facteur de normalisation. Cette métrique nous permet de normaliser nos données pour la taille de librairie de séquençage ainsi que le biais de composition des ARNs séquencés. Un tel biais est observable lorsque peu de gènes

sont très fortement exprimés dans un réplica d'une condition et pas dans les autres réplicas. Les comptes de gènes bruts sont alors divisés par le facteur de normalisation propre à l'échantillon afin de produire les comptes normalisés.

À partir de ces comptes normalisés, on procède à l'analyse de l'expression différentielle où l'on applique un test statistique afin de déterminer si un gène est exprimé différemment ou non. L'approche d'expression différentielle de DESeq2 est basée sur un modèle de régression binomiale négative dont la moyenne est approximée par méthode de ratio de médianes et la dispersion est approximée par rétrécissement bayésien. L'hypothèse que le gène soit différemment exprimé est ensuite testée par le test de Wald. Celui-ci nous permet de tester pour un contraste de coefficients entre deux conditions afin de produire une valeur-P. La magnitude de cette métrique nous indique la probabilité que l'hypothèse soit nulle, ce qui veut dire que le gène n'est pas différemment exprimé dans une certaine condition. Les gènes dont la valeur-p de Wald est sous 0.05 sont conservés puis corrigés pour des tests multiples par les méthodes de Benjamini and Hochberg [Benhamini Y, 1995].

1.2.1.3 Ontologie des gènes

Le projet d'ontologie des gènes [Ashburner, 2000] vise à annoter la fonction des gènes parmi le génome d'un grand nombre d'espèces. L'ontologie (GO) est organisée en trois catégories : processus biologiques, fonction moléculaire et composante cellulaire. Chacune de ces trois catégories est représentée par un graphe dirigé acyclique où chaque annotation présente dans le graphe possède une relation avec un ou plusieurs autres termes du même domaine. Par exemple, la fonction moléculaire « liaison à l'ADN endommagé » est liée à celle de « liaison de l'ADN » qui elle est liée à « liaison à un acide nucléique ». Le graphe est donc organisé de manière hiérarchique de façon que les liens entre les annotations soient dirigés vers des niveaux de complexité plus faibles juste qu'à atteindre l'un des trois termes globaux selon l'ontologie en question. Puisque ces annotations sont disponibles pour presque la totalité des gènes chez plusieurs espèces, il est très facile d'évaluer la surreprésentation d'un terme parmi une liste de gènes. Dans le cas d'une analyse d'expression différentielle, on pourrait par exemple évaluer la surreprésentation d'une fonction moléculaire ou d'un processus biologique parmi les gènes

différemment exprimés afin d'émettre des hypothèses sur les mécanismes biologiques affectés par la condition à l'étude.

1.2.2 Analyse de la polyadénylation par NGS

La section 1.2.2 présente une variante du NGS permettant l'étude des régions 3' du transcriptome, les difficultés de l'analyse de ces données ainsi que des modèles computationnels permettant de surmonter les problématiques liées à cette technologie.

1.2.2.1 TAIL-Seq

À partir de la technique classique du séquençage de nouvelle génération, un ensemble de variantes de cette expérience ont été développées afin d'analyser des aspects spécifiques du transcriptome. La technique de séquençage TAIL-Seq permet d'investiguer les propriétés des régions 3' du transcriptome malgré les difficultés techniques reliées au séquençage de segments répétitifs. L'approche de cette technique consiste à séquencer le transcriptome de manière pairée afin d'utiliser le fragment 5' pour cartographier l'ARN au génome de référence et le fragment 3' pour étudier, en autre, la taille des régions polyadénylées.

Initialement, les ARN ribosomiques sont retirés du reste des ARN, puis les ARN restant sont liés à des adapteurs 3' biotinylés avant d'être digérés partiellement par une ARNase T1. Les fragments sont ensuite phosphorylés puis purifiés par gel pour sélectionner les fragments d'ARN de 500 à 1000 nucléotides. Un adaptateur est ajouté en 5' avant d'effectuer la transcription inverse et l'amplification par PCR des ARN préalables au séquençage Illumina. Cette étape permet d'obtenir une grande quantité d'ADN double-brin à partir de fragments d'ARN. Lors du séquençage pairé, le segment polyadénylé en 3' est séquencé à partir du brin complémentaire produit par la transcription inverse, le segment de la queue poly(A) devient alors un segment poly(T).

Le défi technique du séquençage de segments polyadénylés repose sur la chimie utilisée par Illumina et la nature homopolymérique du segment poly(A). L'appareil de séquençage détermine le type du nucléotide à partir d'un signal de fluorescence propre à la nature du nucléotide. L'identité d'un nucléotide est déterminée à partir du signal de fluorescence qui prédomine le signal des autres types de nucléotides. Le problème avec les séquences répétitives est que le

signal de fluorescence provenant de la thymine tend à s'accumuler après plusieurs cycles dû au clivage incomplet du fluorophore avec la thymine. Les nucléotides séquencés directement après un long segment répétitif de thymine seront alors faussement identifiés comme étant des thymines. Le développement d'une approche computationnelle est alors crucial pour l'analyse des données TAIL-Seq.

1.2.2.2 Chaînes de Markov à états cachés

Les modèles de chaînes de Markov à états cachés (HMM) sont très fréquemment utilisés pour modéliser une séquence d'informations, c'est pourquoi on les retrouve à la base d'une multitude d'outils de reconnaissance de la parole ou de texte [Khorsheed, 2007] [Zhou, 2002]. L'approche d'un HMM permet d'analyser une donnée de façon séquentielle afin d'attribuer un état caché à chaque étape d'observation de la donnée en question. Par exemple, un HMM pourrait servir à déterminer la nature lexicale (nom, verbe, déterminant, adjectif, etc) des mots au sein d'un texte. Dans cet exemple, chaque mot représente une observation dont la nature lexicale que l'on tente de déterminer correspond à un état caché. Les HMM sont également utilisés dans plusieurs outils bio-informatiques dû à la nature séquentielle du génome. [Söding, 2005] [Käll, 2005] [Cawley, 2003]

Ce type de modèle permet de représenter une distribution de probabilité à travers une séquence d'observation selon deux propriétés. La première stipule qu'une observation Y a un temps t est le résultat d'un processus qui correspond à un état caché ou classe S_t parmi un ensemble discret de classes. La seconde propriété stipule que sachant la classe de l'état précédent S_{t-1}, la classe de l'état actuel S_t est indépendante de tous les autres états précédent t - 1. De ces deux propriétés, la distribution jointe d'une séquence d'états cachés S et d'observation Y peut être décrite comme étant :

$$P(S_{1:T}, Y_{1:T}) = P(S_1) \cdot P(Y_1|S_1) \cdot \prod_{t=2}^{T} P(S_t|S_{t-1}) \cdot P(Y_t|S_t) \quad (1)$$

Avec $P(S_t|S_{t-1})$ étant la probabilité de transition de l'état caché de l'observation précédente à l'état caché de l'observation actuelle selon une matrice de transition d'états. La probabilité $P(Y_t|S_t)$ correspond à la probabilité d'une observation selon l'état caché actuel obtenue à partir de la matrice d'émission. À partir d'un jeu de données où les séquences d'observations et d'états

cachés sont connues, les deux matrices de probabilités peuvent être apprises à l'aide de l'algorithme de maximum de vraisemblance.

1.2.2.3 Algorithme de Viterbi

L'algorithme de programmation dynamique de Viterbi permet de trouver la séquence d'états cachés la plus probable, soit le chemin de Viterbi, produit par une séquence d'observations. On définit la probabilité du chemin de Viterbi terminant par l'état k avec l'observation Y comme étant:

$$P_A(Y, x) = P(Y|A) \cdot max_k[P_k(j, x-1) \cdot P_{kA}] \quad (2)$$

Avec P(Y|A) étant la probabilité d'émission, c'est-à-dire la probabilité d'observer l'élément Y dans l'état caché A. La probabilité P_k(j, x-1) correspond au chemin de plus probable terminant en position x-1 dans l'état caché k avec l'observation de l'élément j. Le terme P_{kA} correspond à la probabilité de transition de l'état caché A à k. Le calcul de probabilité de la séquence d'observations devient alors un produit de plusieurs probabilités, c'est pourquoi nous utilisons le logarithme naturel de la probabilité afin d'obtenir une sommation des logarithmes des probabilités. Dans le cadre d'une approximation du segment poly(A), on utilise la taille du chemin associé à l'état caché qui correspond à l'état polyadénylé.

1.2.2.4 TailSeeker

L'outil TailSeeker a été mis au point par les créateurs du séquençage TAIL-Seq afin d'approximer la taille des segments poly(A). Cet outil est basé sur un modèle de chaînes de Markov à états cachés combiné à un mélange gaussien afin de déterminer la transition entre le segment polyadénylé et le reste de l'ARN. Ce modèle est composé de quatre états cachés: corps du segment poly(A), transition du segment poly(A), transition du segment 3' UTR et corps du segment 3' UTR. L'algorithme TailSeeker utilise directement les données de fluorescence du séquenceur qui sont normalisées afin que le modèle analyse le signal relatif de « T » par rapport signal de fluorescence des autres nucléotides. Le modèle est d'abord entrainé en utilisant l'algorithme de Baum-Welch sur un ensemble de séquences synthétiques. Cet entraînement permet d'approximer les probabilités des matrices d'émissions et de transitions du HMM qui représentent le mieux les séquences du jeu de données synthétiques dont la taille des fragments

est connue. Les séquences produites par l'expérience Tail-Seq sont ensuite analysées à l'aide du modèle TailSeeker utilisant l'algorithme de Viterbi pour déterminer la séquence d'états cachés la plus probable. La taille du segment polyadénylé est obtenue en combinant les régions appartenant aux états cachés du corps du segment poly(A) et de la région de transition du segment poly(A). Bien que l'outil performe mieux que les autres alternatives plus simplistes qui tentent de résoudre ce problème, ce modèle nécessite les données brutes de séquençages (signal de fluorescence) qui ne sont pas disponibles pour la plupart des nouveaux appareils de séquençage.

1.2.2.5 DNBC

Le modèle utilisé pour approximer la longueur des segments polyadénylés dans ce projet est une variante du HMM que l'on nomme classificateur bayésien dynamique naïf (DNBC). Ce modèle se distingue du HMM traditionnel par le fait que l'observation n'est pas issue d'une seule variable, mais plutôt d'un ensemble de variables. Dans le cadre de ce projet, le séquençage des ARN polyadénylés génère deux séquences d'observations, soit la séquence de nucléotides ainsi que le score de qualité de chacun d'entre eux. Cette modification dans la conceptualisation de l'observation nous permet de reformuler la probabilité d'observation $P(Y_t|S_t)$ de la formule 1 comme étant :

$$P(Y_t|S_t) = P(Nucléotide|S_t) * P(Qualité|S_t) \quad (3)$$

La séquence d'état la plus probable pour une séquence d'observation est prédite à l'aide de l'algorithme de Viterbi, permettant ainsi d'approximer la longueur du segment poly(A).

1.2.3 Score de rétention d'intron

L'intégrité du déroulement de l'étape d'épissage est maintenue par les mécanismes de contrôle de qualité de l'ARN de façon à cibler les transcrits aberrants pour la dégradation. Si les mécanismes de contrôle de qualité sont déstabilisés, il est probable que l'on observe une stabilisation des transcrits dont les régions introniques n'ont pas été correctement retirées du corps de l'ARNm. Le phénomène de rétention d'intron ne se produit pas seulement dans des maladies rares où une mutation ponctuelle interfère avec l'étape d'épissage, mais également

dans plusieurs cas de cancer. Pour le cas du cancer de la prostate et de l'œsophage, la rétention de l'intron 4 du proto-oncogène *CCND1* induit un isoforme tronqué de la cycline *D1b* associée à plusieurs types de cancer [Solomon, 2003]. Chez les trois principaux sous-types de cancer du sein (triple négatif, non-triple négatif et *HER2* positif), le phénomène de rétention d'intron a été observé chez 2038 gènes alors que leur impact biologique n'a pas encore été caractérisé [Eswaran, 2013].

Le score de rétention d'intron (IR) permet de quantifier l'intensité du signal émis par ces transcrits aberrants par rapport au niveau de transcrits correctement modifiés par le processus d'épissage. Alors qu'il y existe plusieurs outils pour quantifier le score IR [Middleton, 2017] [Pimentel, 2016] [Bai Y,2015] [Li HD, 2017], la plupart d'entre eux représentent le score IR comme étant le ratio de signal intronique par rapport la somme du signal intronique et exonique :

$$IR = \frac{Signal\ intronique}{Signal\ exonique + Signal\ intronique} \quad (4)$$

Le signal intronique correspond aux séquences qui débutent dans une région exonique et terminent dans une région intronique, ou vice-versa. Le signal exonique correspond aux séquences qui débutent dans un exon et qui terminent dans un autre exon. La méthode implémentée par IRfinder [Middleton, 2017] est adaptée pour prendre en considération les ARN non-codants fortement exprimés comme les micro-ARNs ou les ARNsno présents dans certaines régions introniques qui pourraient faussement amplifier le score IR.

1.2.4 Classification des transcrits cryptiques

Les transcrits cryptiques ou envahissants sont de courts ARN non-codants issus de la transcription antisens d'un gène ou de la transcription d'une région intergénique. Pratiquement indétectables dans la levure de type sauvage, ces transcrits ont initialement été découverts à la suite de la mutation du gène *Rrp6*, un cofacteur important de l'exosome nucléaire. Il existe actuellement quatre différents groupes de transcrits cryptiques chez la levure qui ont été nommés à partir du gène muté ayant permis d'observer l'expression de ces régions. L'origine de ces transcrits est liée à la bidirectionnalité des régions promotrices. Alors que le génome de la levure est assez dense,

la transcription antisens initiée dans une région dépourvue de nucléosome peut produire des transcrits non-codants des régions intergéniques ou d'un transcrit antisens d'un gène avoisinant.

Il n'est toujours pas clair si les transcrits cryptiques ont un impact fonctionnel sur l'expression génique. Certains transcrits cryptiques, comme le transcrit antisens du gène *Pho84*, ont un effet répressif sur l'expression du gène qu'ils chevauchent en stimulant la désacétylation des histones avoisinantes [Camblong J, 2007]. D'autres transcrits cryptiques, tel que le transcrit antisens du gène *TY1* avec les rétrotransposons, permettent la répression de l'expression de transcrits en s'associant par complémentarité à ARNm. L'impact des transcrits cryptiques sur l'intégrité du transcriptome a été identifié pour quelques transcrits spécifiques, il n'existe toutefois pas encore un consensus sur les mécanismes d'action des différents types de transcrits cryptiques sur le contrôle de l'expression génique.

1.2.4.1 CUTs et SUTs

Les deux premiers groupes de transcrits cryptiques ayant été découverts sont issus de la mutation du gène *Rrp6* [Xu, 2009]. Le groupe des CUTs pour « Cryptic Unstable Transcripts » décrit un groupe initial de 925 ARN non-codants transcrit par l'ARN polymérase II auquel est ajouté une coiffe et une queue polyadénylée. L'étape de polyadénylation est effectuée par le complexe NSS et TRAMP avant que le transcrit soit dégradé par l'exosome nucléaire. Une défaillance à diverses étapes du cycle de vie des CUTs peut entraîner une stabilisation de ces transcrits dans le noyau. La mutation du gène Rrp6 compromettant l'activité de nucléase de l'exosome est l'une des mutations pouvant induire la stabilisation des CUTs. Une étude plus récente [Vera, 2016] a permis d'étendre la liste des CUTs à un ensemble de 1412 régions non-codantes.

Le groupe des SUTs pour « Stable Unannotated Transcripts » comporte un ensemble de 847 régions qui sont moins sensibles à la mutation Rrp6, c'est-à-dire qu'elles sont observables chez le mutant Rrp6 ainsi que chez la levure de type sauvage. Tout comme les CUTs, le site d'initiation de la transcription des SUTs est enrichi pour l'absence de nucléosome. Toutefois, les deux groupes se distinguent par le mécanisme de dégradation en action. Alors que les CUTs sont pris en charge par les complexes NNS et TRAMP dans le noyau, la dégradation des SUTs est principalement

assurée par l'action du complexe *Dcp1-Dcp2* qui retire la coiffe et l'exoribonucléase Xrn1 qui dégrade le transcrit dans le sens 5'→3' dans le cytoplasme [Marquardt, 2011].

1.2.4.2 XUTs

Le troisième groupe de transcrits cryptiques a été découvert à la suite de la mutation de l'exoribonucléase *Xrn1* [Van Dijk, 2011]. Le groupe des XUTs pour « *Xrn1* unstable transcripts » décrit un vaste ensemble de 1658 régions sensibles à la mutation de *Xrn1*. Également transcrit par l'ARN polymérase II, il est estimé que 66 % des SUTs correspondent à des transcrits anti-sens au gène qu'ils chevauchent. Alors que la protéine *Xrn1* permet la dégradation des SUTs dans le cytoplasme, c'est également cette exoribonucléase qui est responsable de la dégradation de la plupart des XUTs. Il existe alors un certain chevauchement entre les régions décrites par le groupe des SUTs et des XUTs.

1.2.4.3 NUTs

Le quatrième de groupe de transcrits cryptiques correspond au NUTs [Schulz, 2013] pour « Nrd1 Unstable Transcripts ». Issue de la mutation du gène *Nrd1*, la transcription des 1526 régions de ce groupe débute également par l'action de l'ARN polymérase II aux endroits dépourvus de nucléosomes. Ces transcrits cryptiques peuvent également avoir un impact sur l'expression génique en s'associant à l'ARNm ou par interférence de la traduction de l'ARNm. La protéine associée au gène *Nrd1* est l'une des trois protéines du complexe NSS situé dans le noyau. Ce complexe est connu pour son implication dans la terminaison de la transcription, la polyadénylation et ultimement, la dégradation de courts ARN comme les ARNsno. Avec l'interaction de son partenaire *Nab3*, la protéine *Nrd1* s'associe à une séquence spécifique d'un ARN naissant pour initier la terminaison de la transcription [Carroll, 2007]. Il a même été démontré que ces séquences spécifiques à l'interaction de *Nrd1* avec son substrat étaient enrichies chez les ARN non-codants et faiblement représentées chez les ARNm [Schulz, 2013]. La défaillance du gène *Nrd1* induit une stabilisation d'ARN non-codants quatre fois plus longue que les autres groupes de transcrits cryptiques. Il existe un chevauchement entre les NUTs et les autres groupes de transcrits cryptiques puisque la défaillance du mécanisme du complexe NNS est en amont des mécanismes de dégradation par l'exosome ou par *Xrn1*.

2 Méthodologie

Ce chapitre détaille l'utilisation des outils liés à l'analyse d'expression différentielle du transcriptome de la levure ainsi que le modèle utilisé pour mesurer les segments polyadénylés. Les cultures de *Saccharomyces Cerevisiae* ont été produites à partir de la souche BY4741. Les cinq mutations analysées dans ce projet sont réalisées à partir d'allèle sensible à la température. Les données de levures mutantes ont été produites à la suite d'une phase de croissance de 90 minutes à 37 °C.

2.1 Données de séquençage RNA-Seq

Les ARNs extraits sont purifiés selon deux méthodes de purification. La méthode *Ribo* - permet de retirer l'ARN ribosomique du total d'ARN à séquencer alors que la méthode *Poly(A) +* consiste à sélectionner les ARN polyadénylés par affinité à un oligo. Le séquençage pairé (50/50) des ARN est effectué par le système Illumina Hiseq 2500 pour la levure de type sauvage ainsi que les mutants *Rrp6*, *Dis3*, *Csl4*, *Enp1* et *Srm1.*

2.1.1 Alignement et quantification

Les fragments d'ARN séquencés sont alignés au génome de référence de la levure (R64-1-1) à l'aide de l'outil *Hisat2* [Daehwan, 2019]. Les séquences pairées sont alignées en conservant seulement le meilleur alignement possible pour chaque fragment. Les fichiers d'alignements binaires (BAM) sont ensuite classés en ordre et indexés à l'aide de l'outil *SamTools* [Li H, 2009]. En moyenne, l'outil *Hisat2* trouve un alignement unique pour près de 95 % des transcrits.

L'outil *featureCounts* de la librairie *SubRead* [Liao Y, 2013] disponible sur R nous permet ensuite de compter le nombre de séquences alignées qui sont associées à un gène. L'outil sera forcé d'assigner les deux fragments pairés à une région annotée en activant les paramètres *isPairedEnd* et *requireBothEndsMapped*. Nous indiquons à l'outil qu'un fragment doit chevaucher une région annotée sur au moins 10 nucléotides à l'aide du paramètre *minOverlap.* Les fragments chimériques où les deux parties de la séquence pairée sont alignées sur des chromosomes différents seront omis en désactivant le paramètre *countChimericFragments*. Afin d'évaluer

l'expression différentielle des transcrits cryptiques, nous avons ajouté au fichier d'annotation du génome de la levure les coordonnées des régions des 4 familles de transcrits cryptiques (CUT, SUT, NUT et XUT). Globalement, 90 à 95 % des séquences alignées au génome de référence correspondent à des régions annotées chez la levure.

2.1.2 Normalisation et expression différentielle

L'expression différentielle est quantifiée à partir de la libraire *DESeq2* disponible sur *R*. Cet outil nous permet d'utiliser directement la matrice de comptes de gènes non-normalisés générés par *featureCount*. Les données sont ensuite normalisées par la méthode de ratio des médianes afin de contrôler pour la taille de la librairie de séquençage et pour quelconque biais dans la composition des ARNs séquencés. La quantification de l'expression différentielle est ensuite effectuée à l'aide d'un modèle binomiale négatif puis l'importance de la variation de l'expression génique est évaluée à l'aide du test de Wald. Les transcrits ayant une valeur-P en-dessous de 0.05 pour les cinq mutants ont été conservés pour l'analyse d'expression différentielle.

2.1.3 Analyse par composantes principales

L'analyse d'enrichissement des termes GO a été effectuée en utilisant la version 2.0 du fichier d'association de gènes (gaf) de la Saccharomyces Genomes Database. Les 236 gènes régulés à la hausse ont été sélectionnés à partir d'un seuil significatif de 1 (en log2) à travers les trois mutants de l'exosome (*Csl4, Rrp6 et Dis3*). De la même manière, les 325 gènes régulés à la baisse ont été sélectionnés à partir d'un seuil de -2 (en log2) à travers les trois mutants de l'exosome. Puisque la majorité du génome est régulé à la baisse, un seuil significatif de -1 aurait permis de sélectionner plus de la moitié du génome, c'est pourquoi nous utilisons une valeur plus stricte de -2. Les termes GO les plus significatifs ont été sélectionnés à partir de loi hypergéométrique en sélectionnant les termes avec une valeur-P en dessous de 0.05.

2.2 Classificateur Bayésien dynamique naïf

Cette section vise à présenter l'approche utilisée pour surmonter les difficultés techniques de l'analyse des données de séquençage TAIL-Seq. Lors du séquençage 3' d'ARNs, le signal de fluorescence de la thymine se perd dans la transition entre le segment polyadénylé et le reste du

corps de l'ARN. Si l'on tente d'estimer la longueur du segment à partir de la séquence produite par le séquenceur, la taille du segment sera surestimée. Le début du segment poly(A) sera séquencé avec un score de qualité de séquençage nettement supérieur à la région surestimée comme l'indique la figure 2.

Afin d'estimer la longueur réelle des segments poly(A), nous avons mis au point un modèle DNBC combinant la séquence de nucléotides et le score de qualité de séquençage.

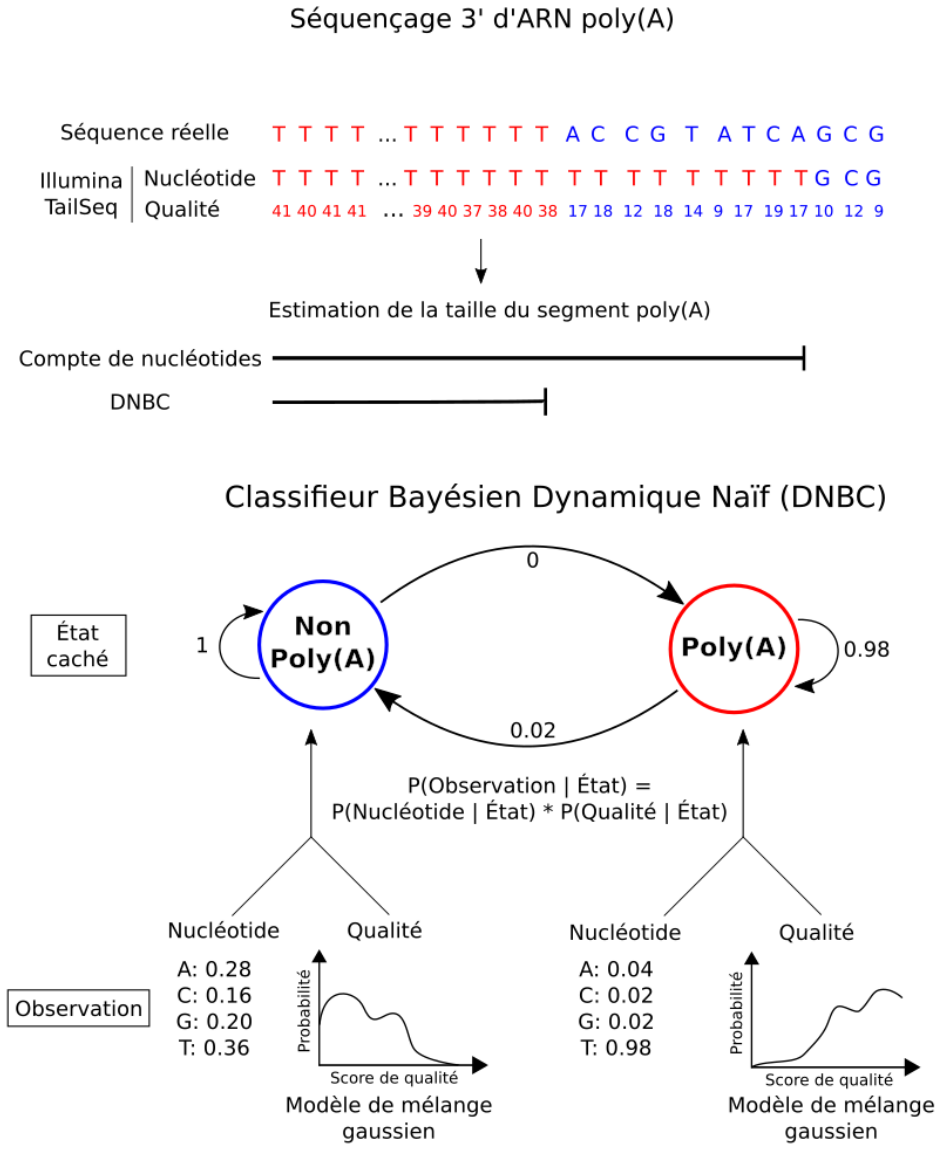

Figure 2. – ESTIMATION DE LA TAILLE DES SEGMENTS POLYADÉNYLÉS À L'AIDE DU DNBC. La longueur du segment polyadénylé est surestimée lorsque l'on observe unique la séquence de nucléotide. L'approche du modèle DNBC utilise le nucléotide séquencé et son score de qualité pour évaluer la probabilité qu'il appartienne à un état caché, soit une région polyadénylée ou non.

2.2.1 Estimation par méthode de compte

Alors qu'il n'existe pas une manière de savoir parfaitement la longueur des segments poly(A), nous sommes toutefois en mesure de faire une estimation à partir de la séquence de nucléotides seulement. Cette méthode bien simpliste, que l'on nommera la « méthode par compte », consiste à mesurer la plus longue séquence continue de « T » en tolérant 15 % de taux d'erreurs. Bien que cette méthode ne soit pas très précise, elle nous permet de générer un jeu de données sur lequel le DNBC peut être entraîné.

2.2.2 Entraînement du DNBC

La phase d'entraînement consiste à apprendre trois ensembles de probabilités. Le premièrement étant la probabilité d'émission d'un nucléotide, soit la probabilité d'observer un nucléotide sachant la classe à prédire. Le second ensemble correspond aux probabilités de transitions, soit la probabilité que l'observation actuelle provient de la classe X alors que la classe de l'observation précédente provient de la classe Y. Ces deux ensembles de probabilités sont obtenues en entraînant un HMM sur le jeu de données d'entraînement prédit par la méthode par compte. Nous utilisons dans ce cas un HMM discret, soit l'objet *dtHMM* de la librairie HMMs sur Python. Les paramètres du modèle sont estimés par l'algorithme de maximum de vraisemblance.

Le troisième ensemble de probabilités à apprendre est l'émission du score de qualité, soit la probabilité d'observer un score de qualité X pour un nucléotide sachant à quelle classe il appartient. Nous utiliserons le score de qualité des nucléotides prédit par méthode de compte pour entraîner un modèle de mélange gaussien pour chacune des deux classes. Pour l'implémentation du modèle de mélange gaussien, nous utilisons l'objet *GaussianMixture* de la librairie *SKlearn* déployé sur Python. Chaque modèle (polyA et non-polyA) utilise un mélange de deux courbes gaussiennes sphériques dont les paramètres sont estimés par maximum de vraisemblance.

Une fois que les trois ensembles de probabilités ont été appris, le DNBC peut prédire la séquence de classes la plus probable pour une séquence d'observation à l'aide de l'algorithme de Viterbi. L'algorithme a été modifié de sorte que la probabilité d'émission est remplacée par la probabilité

jointe d'émission d'un nucléotide et celle du score de qualité de séquençage tel qu'indiqué à la formule (2).

2.3 Données de séquençage TAIL-Seq

Le séquençage des ARNs polyadénylés par méthode TAIL-Seq est de type pairé, ce qui signifie que pour chaque ARN séquencé, nous obtenons une séquence en 3' de la queue poly(A) que nous analysons avec le modèle DNBC ainsi qu'une séquence en 5' que nous alignons avec le génome de référence de la levure (R64-1-1). L'alignement des séquences 5' est effectué à l'aide de l'outil *Hisat2* en conservant seulement le meilleur alignement possible pour chaque séquence. Les séquences dont le fragment 5' s'aligne à 1000 nucléotides et moins de la fin du gène ont été conservées pour la suite des analyses de polyadénylation.

2.3.1 Première expérience TAIL-Seq

Lors de l'expérience de séquençage TAIL-Seq, les ARNs sont séquencés de manière pairés, de sorte que le fragment séquencé en 5' permet d'associer un gène au fragment polyadénylé qui est séquencé par le fragment en 3'. Ce séquençage se fait normalement avec 50 nucléotides pour le fragment 5' et 250 nucléotides pour le fragment 3'. Toutefois, les données TAIL-Seq fournies initialement par nos collaborateurs avait été obtenues en séquençant 150 nucléotides pour les deux fragments. Le problème majeur de cette d'expérience est qu'il limite le pouvoir de résolution de notre modèle. De ces 150 nucléotides, les 20 premiers correspondent aux adaptateurs, ce qui implique que notre modèle ne pourra pas mesurer efficacement toutes les séquences poly(A) de plus de 130 nucléotides. Parmi ces longues séquences poly(A), on retrouve principalement des séquences poly(A) qui sont réellement d'au plus de 130 nucléotides, mais également des séquences plus courtes que notre modèle DNBC n'arrive pas à classifier correctement lorsqu'il surestime la longueur du segment. Ce phénomène se produit généralement lorsque le score de qualité de séquençage ne diminue pas de manière significative dans la région de transition poly(A) à non-poly(A).

2.3.2 Deuxième expérience TAIL-Seq

Puisque les longues séquences occupent une grande partie des séquences poly(A), les levures mutées ainsi qu'un ensemble de séquence synthétique ont été séquencés une deuxième fois avec une approche de séquençage 50/250. Toutefois, lors de l'analyse de ces séquences 3', nous avons remarqué que les segments poly(A) que l'on observe comme une suite de thymines dû à la transcription inverse étaient suivis uniquement d'adénosine chez plus de 90% des séquences poly(A). La présence d'adénosines sur ce fragment complémentaire correspond alors à un segment d'uracile provenant du fragment d'ARN original.

Il est possible de retrouver de courts segments d'uracile après un court segment poly(A) [Chang, 2014], mais rien n'a été publié jusqu'à maintenant à propos de longs segments d'uracile entre la queue poly(A) et la fin du gène encodé par l'ARN. Nous en sommes venus à la conclusion qu'il y a probablement eu un problème avec cette expérience de séquençage. Nous avons donc décidé d'effectuer l'analyse des séquences poly(A) à partir des données de la première expérience sans utiliser les séquences dont le segment poly(A) atteint les 130 nucléotides.

2.4 Quantification du score IR

La quantification du score de rétention d'intron est effectuée avec l'outil IRFinder à partir des fragments d'ARN séquencés par la libraire *Ribo* -. Le score IR est calculé en prenant en considération les différents réplicas à l'aide du test de Audic et Claverie [Audic, 1997] recommandé pour les études avec peu de réplicas. Le *heatmap* du score IR des différentes régions introniques est généré à l'aide de la librairie *ComplexHeatmap* [Gu Z, 2016] sur R. L'analyse de corrélation entre le score de rétention et l'expression des gènes associés à ces introns a été fait à l'aide de la librairie *GGplot2*. Les valeurs d'expressions de gènes sont produites avec la méthode *counts* de la libraire *DESeq2* afin de normaliser les valeurs d'expressions.

3 Résultats

La dégradation des ARNs est une composante majeure des mécanismes de contrôle de qualité du transcriptome qui est principalement opérée par le complexe de l'exosome et ses cofacteurs. Afin de mieux comprendre le fonctionnement et l'interdépendance de ce complexe avec les autres composantes du mécanisme de contrôle de qualité de l'ARN, ce projet vise à caractériser la réponse au stress cellulaire induite par la défaillance de l'exosome. Ainsi, différentes expériences de transcriptomiques ont été effectuées afin d'analyser l'impact de ce stress sur l'expression du transcriptome, l'expression de transcrits cryptiques, l'intégrité du processus d'épissage et le changement de la taille des régions polyadénylés. Les données de séquençage présentées dans ce mémoire sont issues d'une collaboration avec le laboratoire du Dr. Ben Montpetit à l'Université de Californie à Davis et le laboratoire du Dr. Marlene Oeffinger à l'Institut de recherches cliniques de Montréal (IRCM).

3.1 Analyse transcriptomique

Le séquençage des ARNs est une excellente méthode pour déterminer les processus biologiques affectés par les mutations de l'exosome. En plus d'obtenir les séquences des ARNs transcrits, nous pouvons inférer le niveau d'expression de ceux-ci à partir du nombre de transcrits séquencés par gène. Cette méthode nous permettra d'évaluer l'expression différentielle de l'ensemble du transcriptome pour ensuite tenter de cibler les composantes biologiques affectées par un tel stress. Ce chapitre vise donc à présenter les changements d'expression du transcriptome de la levure à la suite des mutations des composantes du complexe de l'exosome (*Csl4, Dis3 et Rrp6*), d'une protéine permettant le transport nucléocytoplasmique des macromolécules (*Srm1*) et d'une protéine nécessaire à la maturation de précurseurs d'ARNr.

3.1.1 Analyse par composante principales

La procédure d'analyse par composantes principales (PCA) nous permet de réduire la variance des profils d'expression génique à deux variables. Cette méthode nous permet d'évaluer la variance entre chacun des réplicas et entre les mutants. La variance entre les réplicas devrait être

assez faible puisque la même mutation devrait avoir le même impact entre les cultures de levures. En revanche, la variance entre les mutants devrait être plus élevée puisque les différentes mutations devraient affecter l'expression du transcriptome de manière différente.

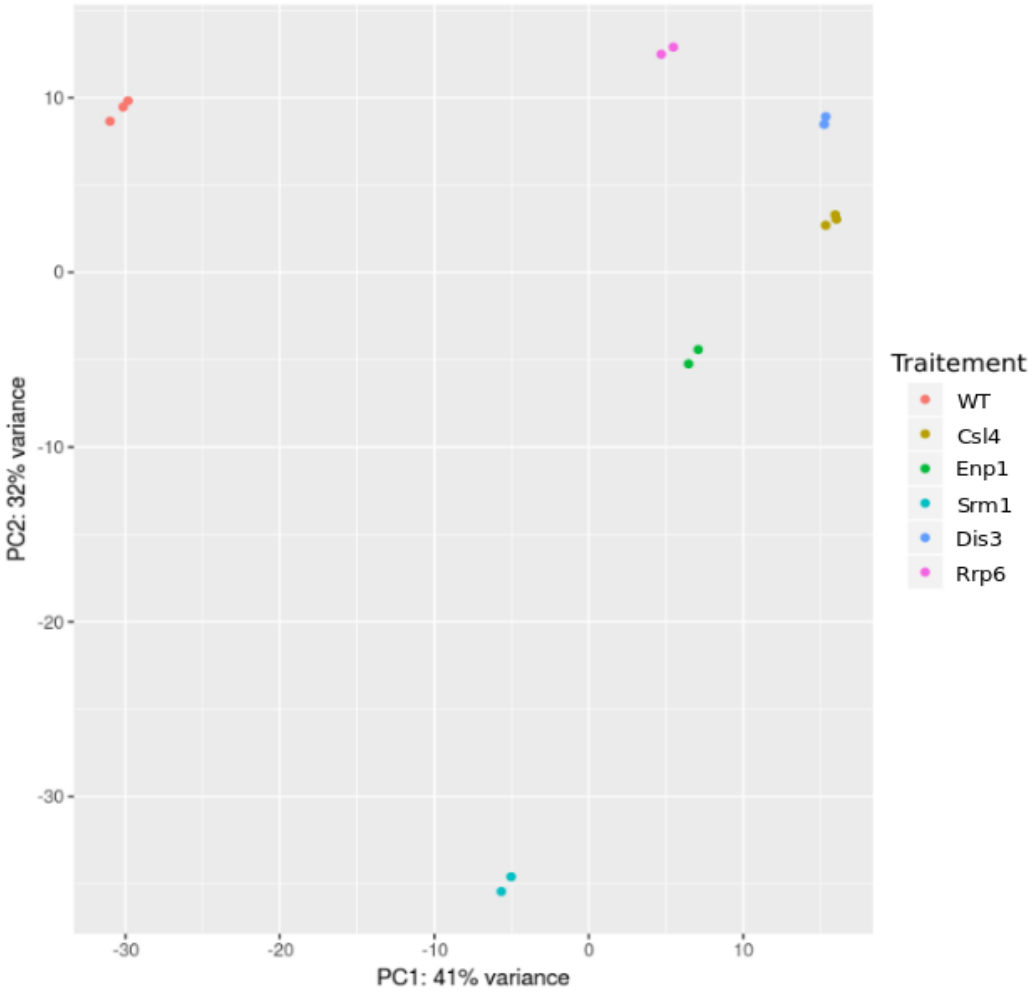

Figure 3. – ANALYSE PAR COMPOSANTES PRINCIPALES DU PROFIL D'EXPRESSION DES LEVURES MUTANTES. On y observe peu de variance entre les réplicas et largement plus de variance entre les différentes mutations.

Afin de visualiser la variance des échantillons, nous avons utilisé les deux premières composantes qui permettent de décrire 73 % de la variance des profils d'expressions de la librairie de séquençage *Ribo* – (figure3). On observe pour tous les types de mutants que les réplicas sont suffisamment proches les uns des autres alors que les différents mutants sont plus éloignés. Ceci nous confirme que la variance entre les réplicas est relativement faible alors que la variance entre

le profil des mutants est assez élevée. Les trois mutants de l'exosome (*Csl4, Dis3* et *Rrp6*) se regroupent tous dans la partie supérieure droite, ce qui nous laisse croire que la réponse cellulaire au stress chez les trois mutants est assez similaire. On observe également que les réplicas du mutant *Enp1* sont assez proches des trois mutants de l'exosome. La réponse cellulaire du stress induit par la mutation de *Enp1* pourrait donc être relativement semblable à celle causée par les mutations de l'exosome. La première composante permet de décrire la majorité de la variance entre ces quatre mutants et la levure *WT*. Le profil d'expression de *Srm1* est toutefois bien différent de toutes les autres levures. La réponse cellulaire à la suite du stress induit par une défaillance du mécanisme de transport des macromolécules entre le nucléole et le cytoplasme pourrait donc être légèrement différente à la réponse cellulaire à la suite d'une perturbation de l'exosome.

3.1.2 Expression des différentes classes d'ARNs

Pour déterminer l'impact des mutations de l'exosome, le changement de l'expression génique de la levure est obtenu par séquençage du transcriptome (RNA-Seq). Le transcriptome de six types de levure est séquencé selon deux méthodes de purification. La première méthode permet l'obtention d'échantillons purifiés sans présence d'ARN ribosomique (*Ribo* -). La seconde permet d'obtenir des échantillons riches en transcrit polyadénylés (*Poly(A) +*).

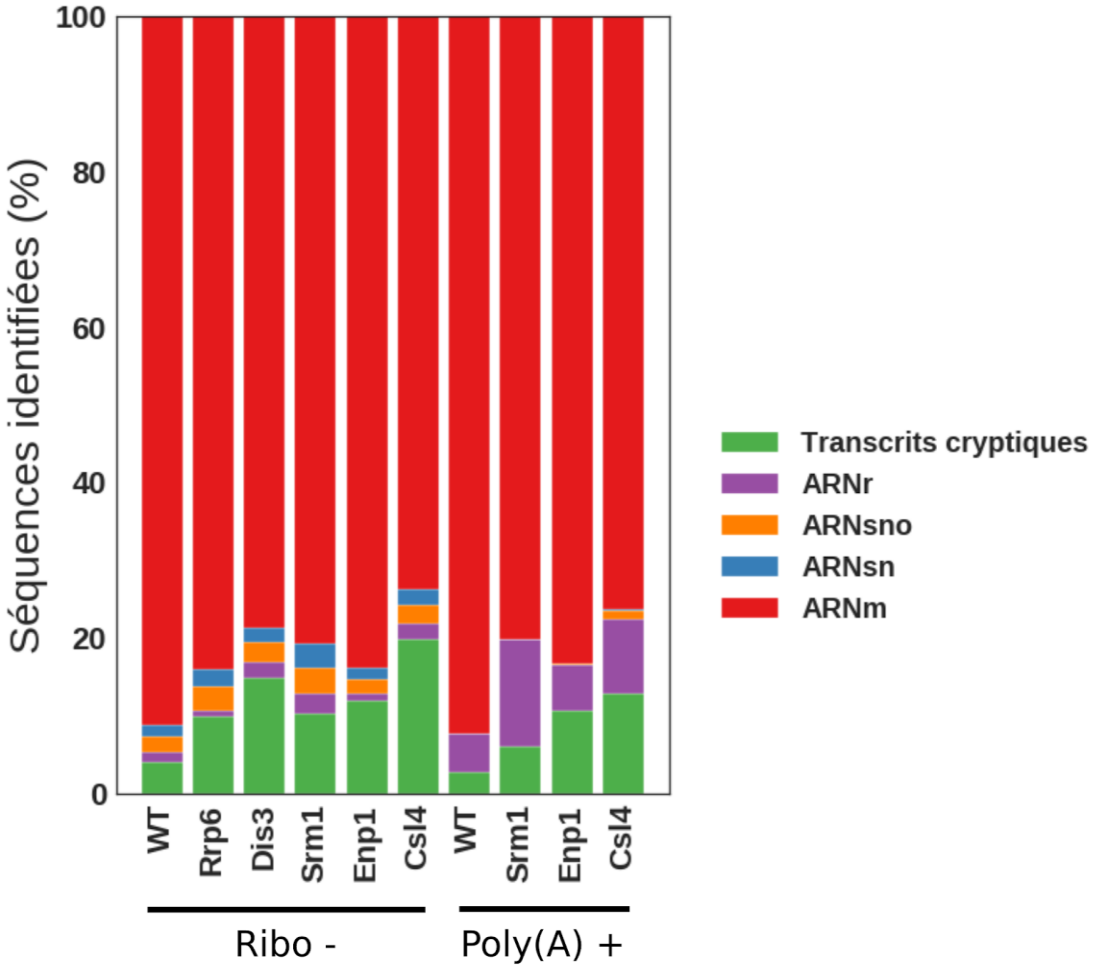

Figure 4. – RATIO DES DIFFÉRENTS TYPES D'ARNS SÉQUENCÉS. Les séquences associées au transcriptome sans ARNr sont identifiées par le terme *Ribo* - alors que les séquences associées au transcriptome polyadénylé sont identifiées par le terme *Poly(A)+*.

Les séquences d'ARNs alignées au génome de référence de la levure ont été regroupées selon les principaux sous-types d'ARNs. Pour les transcriptomes dépourvus d'ARNr (*Ribo* -), on observe que chez la levure de type sauvage (*WT)* près de 90 % des séquences appartiennent à la famille des ARNm alors que seulement 5% des séquences correspondent à des transcrits cryptiques. Chez les cultures de levures mutées pour des composantes directes l'exosome (*Rrp6*, *Dis3* et *Csl4*), on observe une accumulation des transcrits cryptiques qui se traduit en une augmentation de 5 à 15% du ratio de séquences associées à ces transcrits. Il existe un phénomène similaire pour les

mutants *Srm1* et *Enp1*, soit l'accumulation de transcrits cryptiques à la suite d'une défaillance générale des mécanismes de contrôle de qualité de l'ARN.

Pour le transcriptome polyadénylé de la levure *WT*, près de 90 % des transcrits correspondent aux ARNm alors que le 10 % restant est composé de transcrits cryptiques et d'ARNr. On y retrouve bien évidemment plus d'ARNr que pour les données séquençage *Ribo-* puisque celui-ci vise à appauvrir le transcriptome en ARNr. On observe également l'absence d'ARNsn et d'ARNsno dans le contexte d'ARN polyadénylé. Le transcriptome des trois mutants indique aussi une accumulation de 5 à 10 % de transcrits cryptiques. Il est possible qu'une grande partie de ces transcrits subissent des modifications post-transcriptionnelles comparables aux modifications que subissent les précurseurs d'ARNm. On observe également que le ratio d'ARNr double chez le mutant *Csl4* et triple chez le mutant *Srm1*. Le ratio d'ARNr reste toutefois stable chez le mutant *Enp1* qui est connue en autre pour son implication dans la maturation de précurseurs d'ARNr [Chen W, 2013].

Le stress induit par ces mutations modifie grandement la composition du transcriptome, autant chez les mutants de l'exosome que chez les mutants associés à des processus de l'ARN. Il pourrait donc y avoir une réponse générique à la suite des défaillances généralisées des mécanismes de contrôle de qualité de l'ARN.

Nous avons ensuite quantifié l'expression différentielle des mutants par rapport à la levure de type sauvage afin de caractériser davantage la réponse cellulaire causée par ce stress. Les données d'expressions différentielles des figures de cette section du chapitre ont été normalisées et quantifiées par l'outil DESeq2.

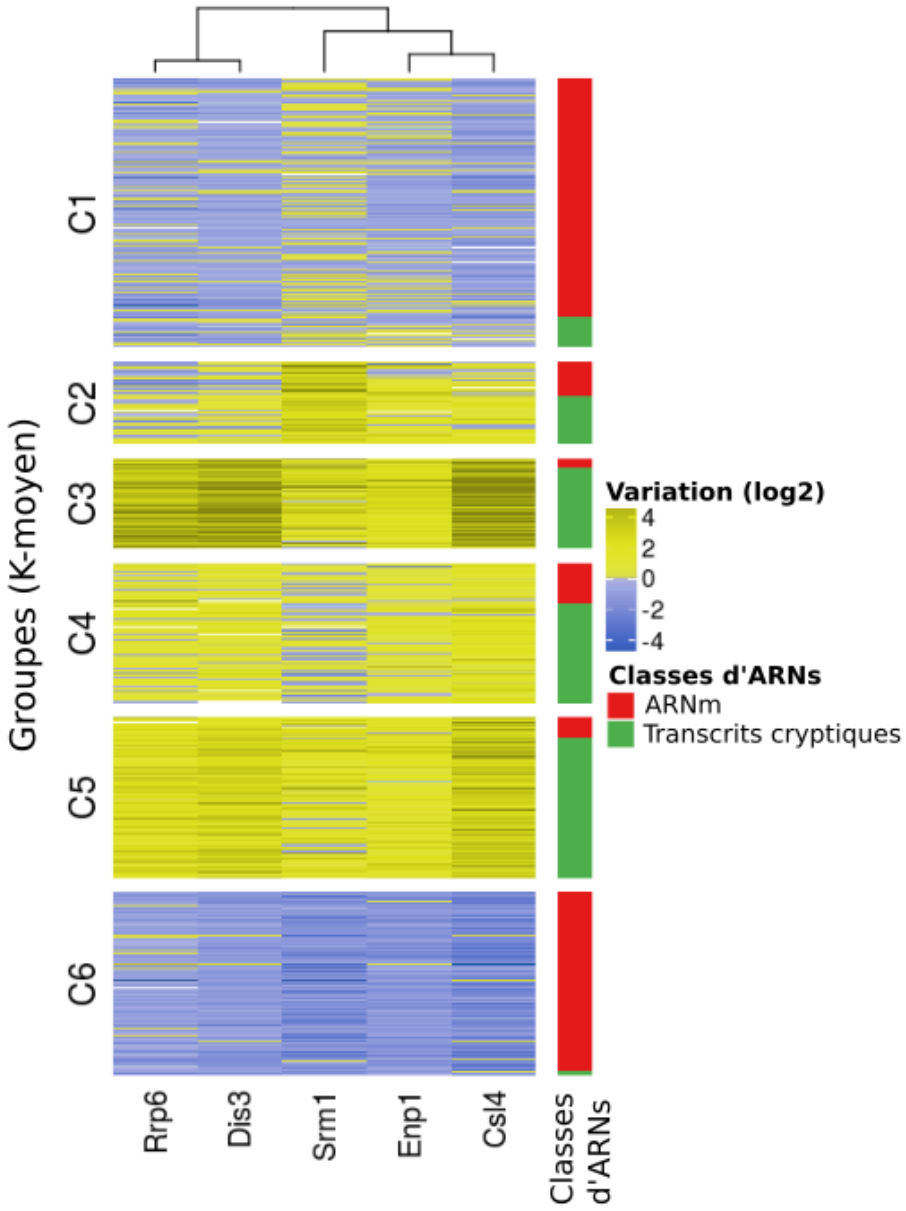

Figure 5. – LE HEATMAP DE L'EXPRESSION DIFFÉRENTIELLE DES ARNM ET DES TRANSCRITS CRYPTIQUES. Les transcrits ont divisés en 6 groupes générés par l'algorithme des K-moyens. Le changement de l'expression génique des mutants par rapport à levure *WT* est quantifié sur une échelle log2 à partir de la librairie DESeq2

L'ensemble du transcriptome a été regroupé en 6 groupes selon la similitude du profil d'expression entre mutants (figure 5). Ces groupes ont été déterminés par l'algorithme de K-moyen. Parmi les groupes de transcrits, on observe que les groupes C1 et C6 sont principalement composés d'ARNm régulés à la baisse alors que les autres groupes sont dominés par des transcrits cryptiques ainsi que quelques ARNm régulés à la hausse. L'accumulation des transcrits cryptiques semble être légèrement plus prononcée chez les mutants de l'exosome (*Rrp6, Dis3 et Csl4*) tel que l'indique le signal des groupes C3 et C5.

De manière globale, la perturbation des mécanismes de contrôle de qualité de l'ARN induit un ralentissement généralisé du régime de transcription du génome de la levure alors que les transcrits cryptiques s'accumulent à la suite de la défaillance des mécanismes de dégradation de l'ARN.

Afin d'évaluer l'expression différentielle des différentes catégories du transcriptome (figure 6), nous avons utilisé les 10 groupes de transcrits classifiés dans l'étude de *Tuck & Tollervey* [Tuck AC, 2013]. Lors de cette étude, le transcriptome a été divisé en 10 groupes selon le profil d'interaction aux protéines liant l'ARN. Le groupe I est principalement composé de CUTs, des transcrits cryptiques relativement courts, alors que les groupes II et III englobent 65.6 % des SUTs qui sont généralement plus long. Les groupes IV à X sont composés des ARNm ainsi qu'une partie de la famille des SUTs. De ces 10 groupes, nous avons observé une tendance générale où les trois premiers groupes sont régulés à la hausse tandis que les sept groupes suivants sont régulés à la baisse. Les dix groupes ont alors été simplifiés en deux classes, soit la première classe comportant les groupes I-III et la seconde classe comportant les groupes IV-X.

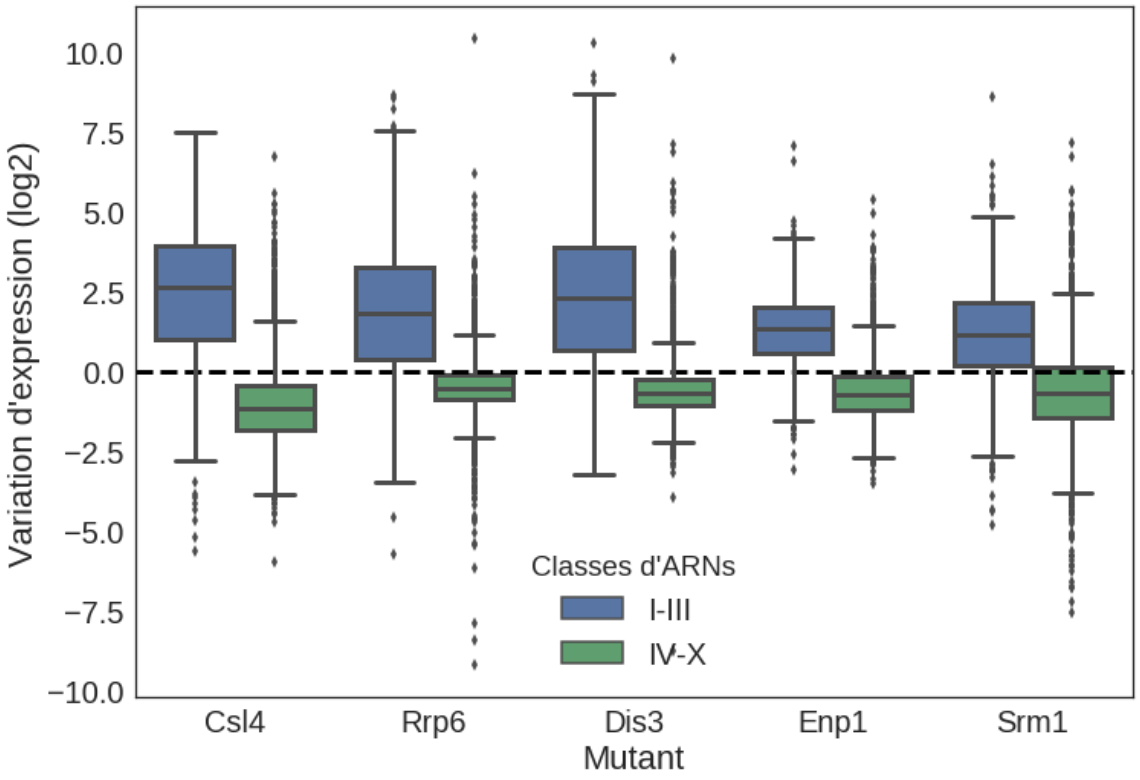

Figure 6. – EXPRESSION DIFFÉRENTIELLES DES GROUPES DE TRANSCRITS DE *TUCK & TOLLERVEY*. Ce diagramme de boîte à moustaches permet de visualiser la distribution du niveau d'expression différentielle de deux groupes de transcrits (I-III et IV-X) chez les différents mutants analysés.

Les dix groupes de transcrits par l'étude de *Tuck* et *Tollervey* ont été simplifiés en deux classes, la première(I-III) comportant principalement les transcrits cryptiques et le seconde (IV-X) comportant la plupart des gènes à ARNm. Globalement, on observe que le groupe des gènes à ARNm est régulé à la baisse alors que le groupe des transcrits cryptiques est régulé à la hausse. La surexpression de ces transcrits est légèrement plus marquée pour les composantes directes de l'exosome (*Csl4*, *Rrp6* et *Dis3*) que pour les mutants *Enp1* et *Srm1*.

Les mutants Enp1 et Srm1 démontrent des caractéristiques similaires aux mutants de l'exosome, c'est-à-dire la stabilisation de plusieurs types d'ARNs polyadénylés (ARNm, ARNnc et ARNr) ainsi qu'une accumulation de transcrits cryptiques [Paul B, 2016].

Pour mieux comprendre la similitude des profils d'expression des mutants *Enp1* et *Srm1* avec les mutants de l'exosome, nous avons analysé la corrélation de l'expression différentielle entre les différentes paires de mutants (figure 7).

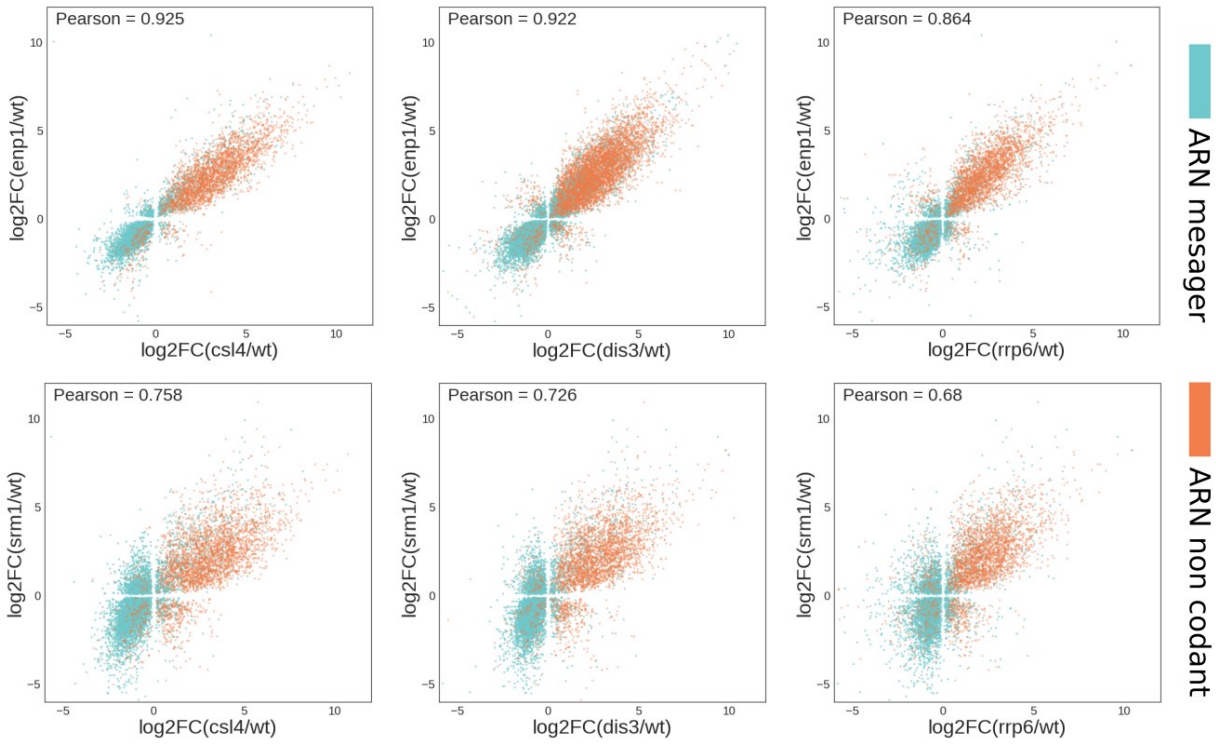

Figure 7. – DISPERSION DE L'EXPRESSION DIFFÉRENTIELLE DES ARNS CODANTS ET NON-CODANTS ENTRE LES MUTANT ENP1 ET SRM1 AVEC LES MUTANTS DE L'EXOSOME. Plus de 95 % des ARNs non-codants correspondent à des transcrits cryptiques.

Le profil d'expression du mutant *Enp1* est assez similaire à celui des mutants de l'exosome avec un coefficient de corrélation de Pearson allant de 0.86 à 0.92. On y observe encore une fois que la plupart des gènes à ARNm sont sous-exprimés tandis que les ARNs non-codants, principalement constitués de transcrits cryptiques, sont généralement sur-exprimés. Un phénomène similaire s'observe chez le mutant *Srm1* où les ARN non-codants sont principalement surexprimés. Les ARNm ne sont toutefois pas tous sous-exprimés puisqu'on observe un signal considérable de sous-expression chez un mutant de l'exosome et de surexpression chez le mutant *Srm1*. Comparativement avec le mutant *Enp1*, les profils d'expression de *Srm1* sont légèrement moins

corrélés avec les mutants de l'exosome alors que le coefficient de Pearson varie entre 0.68 et 0.76. De façon générale, le profil d'expression de transcriptome de la levure mutée pour le gène *Enp1* ou *Srm1* est considérablement similaire à celui d'une levure mutée pour une composante majeure du complexe de l'exosome.

Globalement, les mutations de l'exosome entraînent une accumulation généralisée des différents transcrits cryptiques. Étonnamment, la levure mutée pour le gène *Enp1*, connue pour son implication dans la maturation de précurseurs d'ARN ribosomique, produit un phénotype similaire à celui des mutants de l'exosome. La levure mutée pour le gène *Srm1* participant au transport des macromolécules du nucléole vers le cytoplasme permet également d'observer une accumulation des transcrits cryptiques ainsi qu'un profil d'expression génique assez similaire aux quatre autres mutants. La similitude de la réponse cellulaire au stress causé par la mutation des gènes *Enp1* et *Srm1* avec les mutants de l'exosome nous permet de supposer qu'il existe un mécanisme cellulaire systématique à la perturbation généralisée des mécanismes de contrôle de qualité de l'ARN.

3.1.3 Analyse d'enrichissement des termes GO

Afin d'émettre des hypothèses sur la réponse cellulaire induite par la défaillance de l'exosome, nous avons tenté d'identifier un enrichissement pour des processus biologiques au sein des gènes sur-exprimés et sous-exprimés chez mutants *Csl4, Dis3* et *Rrp6*.

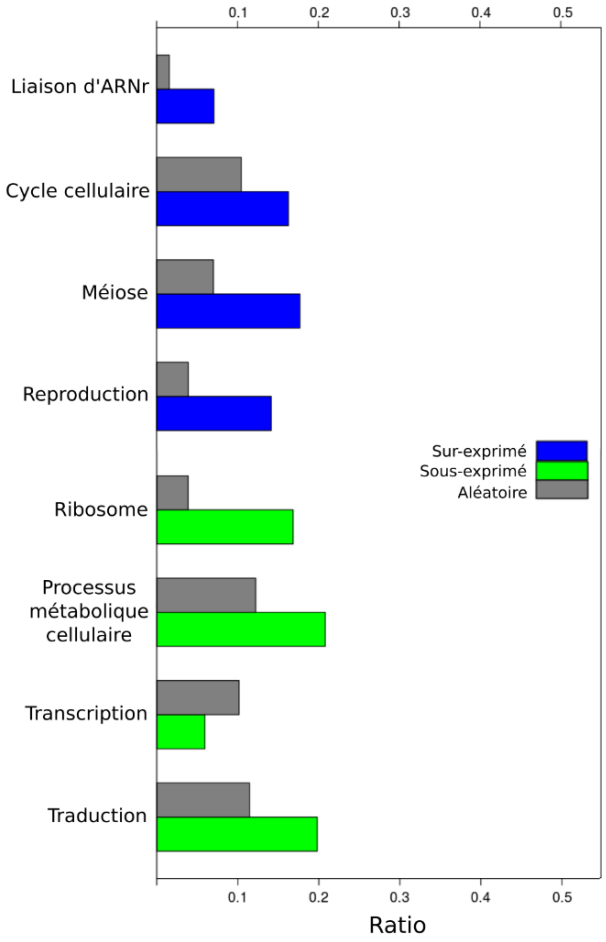

Figure 8. – ENRICHISSEMENT DES PROCESSUS BIOLOGIQUES CHEZ LES GÈNES SUR-EXPRIMÉS ET SOUS-EXPRIMÉS À LA SUITE DES MUTATIONS DE L'EXOSOME DE LA LEVURE. Les processus biologiques correspondent aux termes GO avec une valeur-P sous 0.05 obtenue à partir du test hypergéométrique.

La figure 8 présente les principaux processus biologiques qui sont sur-représentés parmi les 236 gènes sur-exprimés et les 325 gènes sous-exprimés. On observe chez les gènes sur-exprimés un enrichissement pour des protéines liant l'ARN, possiblement en réponse à l'accumulation des ARNm polyadénylés et des transcrits cryptiques dont la dégradation implique l'interaction de plusieurs protéines liant l'ARN. Les gènes sur-exprimés sont également enrichis pour les processus de cycle cellulaire, plus particulièrement celui de la méiose qui est favorisé par la levure lors d'un stress cellulaire [Honigberg, 2003]. Les transcrits sous-exprimés sont enrichis pour des gènes ribosomaux et des gènes reliés à la traduction. Cette observation concorde avec le

ralentissement général du régime de transcription (fig. 4). Dans ce cas, il se pourrait que la cellule utilise les ressources et l'énergie nécessaire à produire de nouvelles protéines pour gérer l'accumulation possiblement toxique d'ARN dans le noyau.

3.2 Expression des transcrits cryptiques

Chez la levure, les régions promotrices des gènes sont connues pour être bidirectionnelles. Lorsque l'ARN polymérase s'installe dans la région promotrice d'un gène, il est alors possible que celle-ci s'installe dans le sens opposé du gène. L'ARN polymérase produit alors un ARN aberrant qui est ciblé pour la dégradation par une endonucléase ou une exonucléase (figure 9). Ces transcrits cryptiques ne sont pas détectables chez la levure de type sauvage, mais tendent à s'accumuler à la suite des perturbations des mécanismes de contrôle de qualité de l'ARN. Ce sous-chapitre présente l'analyse de l'expression des quatre groupes de transcrits cryptiques chez les levures mutantes de la librairie de séquençage *Ribo* -.

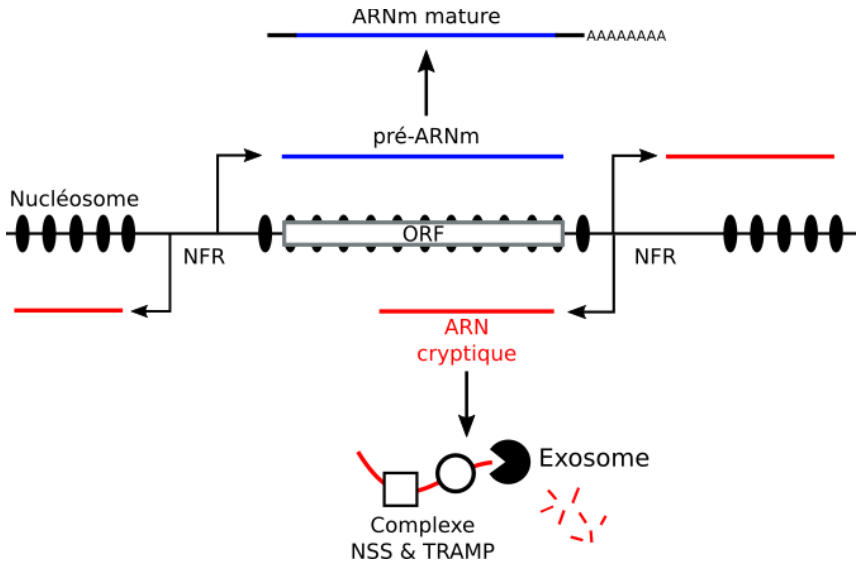

Figure 9. – CYCLE DE VIE DES TRANSCRITS CRYPTIQUES. La bidirectionnalité des régions promotrices sans nucléosome (NFR) qui flanquent les gènes donne lieu à la transcription d'ARN aberrants. Ces transcrits que l'on nomme « cryptiques » sont dégradés par des complexes d'endo/exonucléases, principalement par le complexe de l'exosome et de ses cofacteurs (NSS et TRAMP).

Bien que la plupart des transcrits cryptiques sont issus d'une mauvaise orientation de l'ARN polymérase dans la région promotrice d'un gène avoisinant, certains transcrits cryptiques sont possiblement produits par des ARN polymérases se fixant la région sans nucléosome (NFR) situé à l'extrémité 3' d'un cadre de lecture (ORF). Un exemple de ce genre de phénomène est observable par couverture de séquençage du transcrit cryptique SUT-509 et des gènes avoisinants à la figure 10.

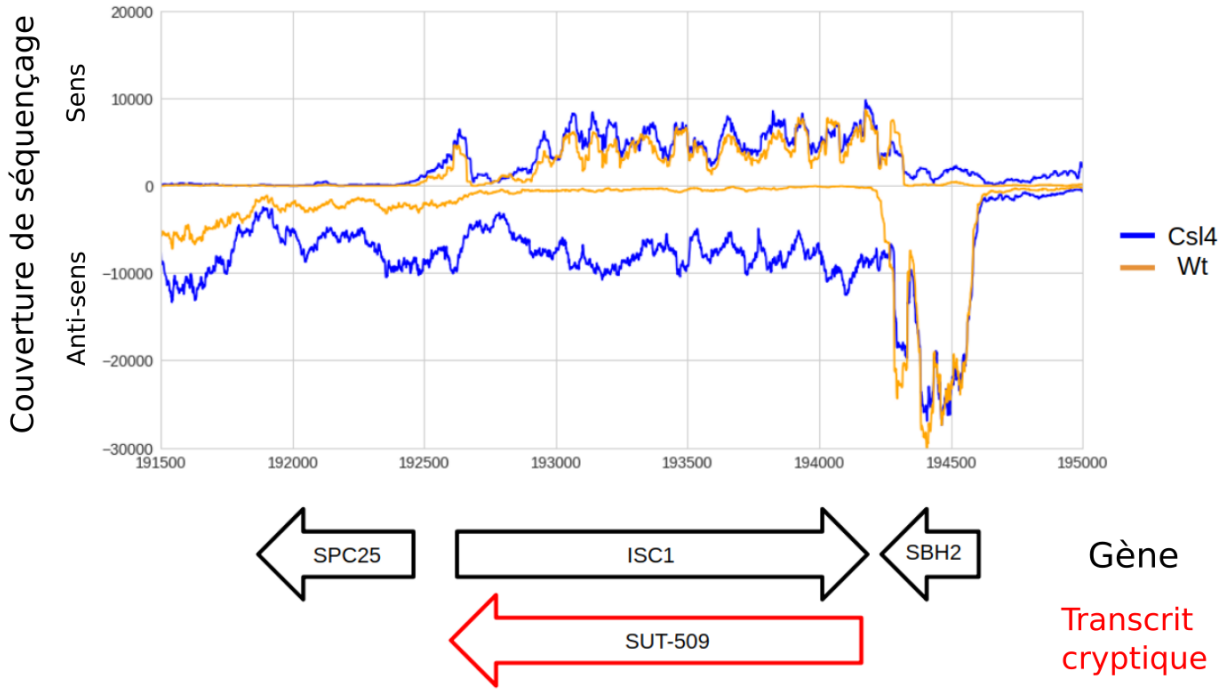

Figure 10. – COUVERTURE DE SÉQUENÇAGE DU TRANSCRIT CRYPTIQUE SUT-509 CHEZ LA LEVURE MUTANTE CSL4 ET LA LEVURE DE TYPE SAUVAGE (WT). L'expression des transcrits produits dans l'orientation anti-sens de cette région varient grandement entre les deux levures.

Les valeurs de couverture par nucléotide sont générées par l'outils *genomecov* de la librairie *bedtools*. Le transcrit SUT-509 est de type 3' antisens puisqu'il est issu de la transcription inverse du gène *ISC1*. La plupart du temps, les transcrits 3' antisens proviennent d'une région promotrice du gène suivant en 3'. Étonnamment, le transcrit SUT n'est pas issu de ce phénomène alors que la région promotrice du gène suivant (*SBH2*) ne se trouve pas entre *ISC1* et *SBH2* puisque les deux gènes sont antisens. La région 3' suivant le gène *ISC1* est toutefois dépourvue de nucléosome, ce

qui nous laisse supposer que la transcription de SUT-509 est produite par un ARN polymérase initialement fixé à une région sans nucléosome et non-promotrice.

La couverture de séquençage avoisinant le transcrit SUT-509 chez le mutant *Csl4* nous permet d'observer l'expression d'un transcrit normalement indétectable chez la levure de type sauvage. Le niveau d'expression de ce transcrit est relativement similaire à celui du gène qu'il superpose (*ISC1*) chez les deux levures.

Les transcrits cryptiques sont séparés en quatre familles (NUT, XUT, SUT et CUT) selon le gène muté ayant permis d'observer l'expression de ces régions normalement peu ou non exprimés. Les régions associées aux SUTs et aux CUTs sont mutuellement exclusives alors qu'elles ont été déterminées à partir du même mutant (*Rrp6*). Il existe alors un chevauchement considérable entre les régions associées aux NUTs, aux XUTs et celles associées aux CUTs et SUTs. Nous avons d'abord évalué la distribution des tailles de ces régions parmi les différentes familles afin de tenter de caractériser les quatre familles (figure 11).

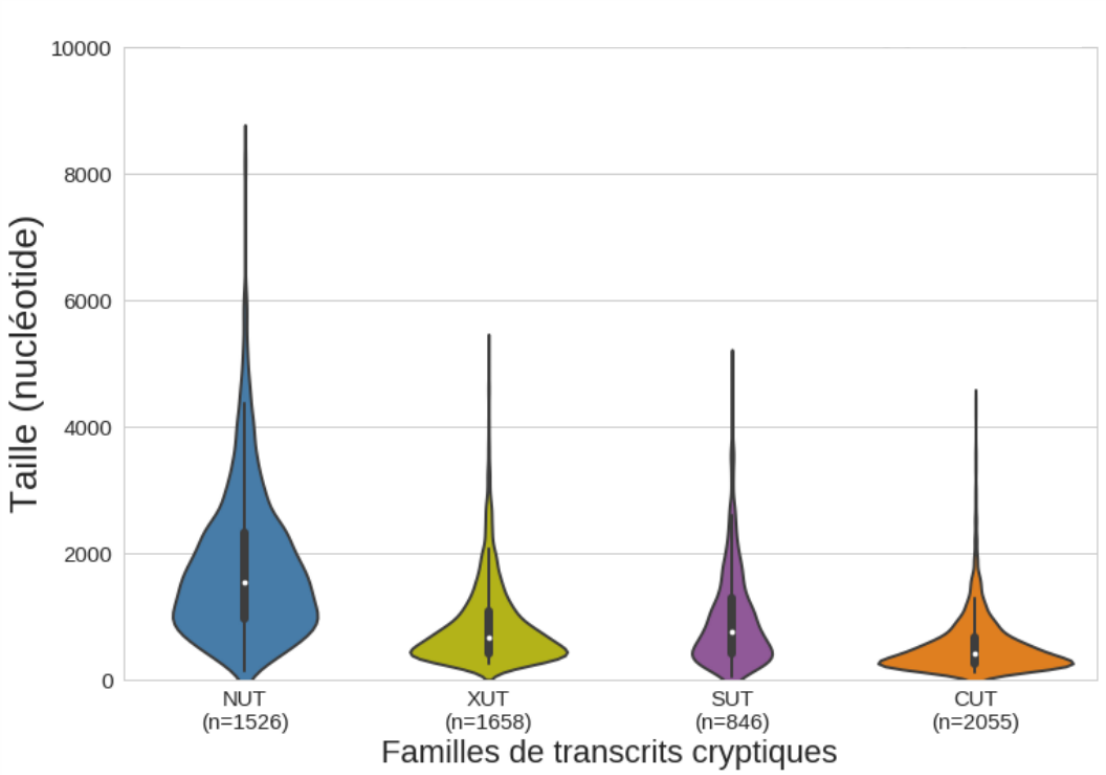

Figure 11. – DISTRIBUTION DE LA TAILLE DES TRANSCRITS CRYPTIQUES PARMI LES 4 FAMILLES. Ce graphique en violon représente l'estimation de la densité de ces distributions. Les NUT sont principalement des lncARNs alors que les autres groupes sont formés de courts transcrits.

Les groupes mutuellement exclusifs des CUT et SUT ainsi que les XUT correspondent à de courts ARNs qui sont généralement des segments antisens d'un gène à ARNm alors que les NUT sont principalement composés de longs ARNs non-codants. Il existe un chevauchement considérable des coordonnées des transcrits cryptiques des NUT et XUT avec les coordonnées des groupes mutuellement exclusifs des CUT et des SUT.

Nous avons ensuite investigué l'expression différentielle des groupes de transcrits cryptiques chez les différents mutants de la levure. Les distributions de changement d'expression sont rapportées sur une échelle log 2 en ne conservant que les transcrits ayant une valeur-p sous 0.05.

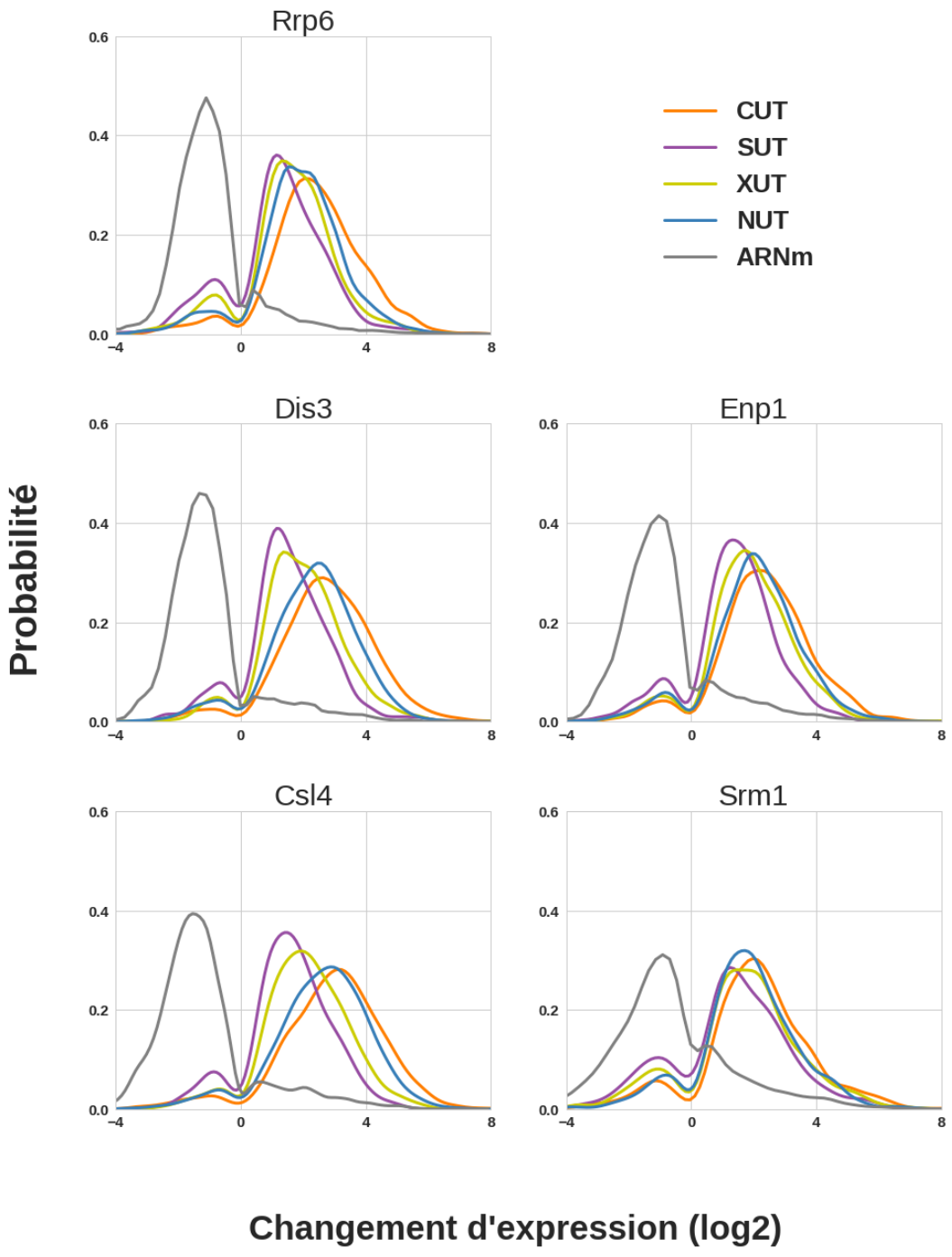

Figure 12. – DISTRIBUTION DU CHANGEMENT D'EXPRESSION DES GÈNES À ARNM ET DES QUATRE FAMILLES DE TRANSCRITS CRYPTIQUES CHEZ LES DIFFÉRENTS MUTANTS DE LA LEVURE. La distribution de l'expression différentielle des différentes catégories de transcrits nous permet d'observer l'accumulation des quatre familles de transcrits cryptiques.

La distribution du changement d'expression varie entre les principaux groupes d'ARNs à la suite des mutations des composantes du système de contrôle de qualité de l'ARN (figure 12). On observe chez tous les mutants une diminution générale de l'expression des ARNm ainsi qu'une augmentation de l'expression des différents groupes d'ARN cryptique. L'expression de certains groupes de transcrits cryptiques diffère légèrement entre certains groupes de mutants. Chez les cultures de levure mutées pour les gènes *Csl4* et *Dis3*, on observe que les CUTs et les NUTs sont légèrement plus sur-exprimés que les XUTs et les SUTs. D'une autre part, tous les groupes de transcrits cryptiques sont sur-exprimés de façon similaire pour les mutants *Enp1* et *Srm1*. Globalement, la défaillance des mécanismes de contrôle de qualité mène à un ralentissement général de la transcription des gènes à ARNm alors que les transcrits cryptiques s'accumulent. Étonnamment, les mutations des gènes *Srm1* et *Enp1* permettent également d'observer l'accumulation des transcrits cryptiques.

3.3 Rétention d'intron

Les mécanismes de régulation du transcriptome font intervenir plusieurs processus cellulaires. L'un de ces processus est l'épissage des introns où une région non-codante est retirée du corps de l'ARN. Le phénomène de rétention d'intron se produit lors qu'un intron n'est pas correctement retiré du précurseur d'ARN. Ce type d'ARN considéré comme étant aberrant est normalement dégradé par des endonucléases ou exonucléases tel que le complexe de l'exosome. Une défaillance de l'exosome pourrait alors mener à l'accumulation d'ARNs ayant échoués le processus d'épissage. Nous avons ici évalué le score de rétention d'intron (IR) pour les 317 régions introniques de la levure (figure 13).

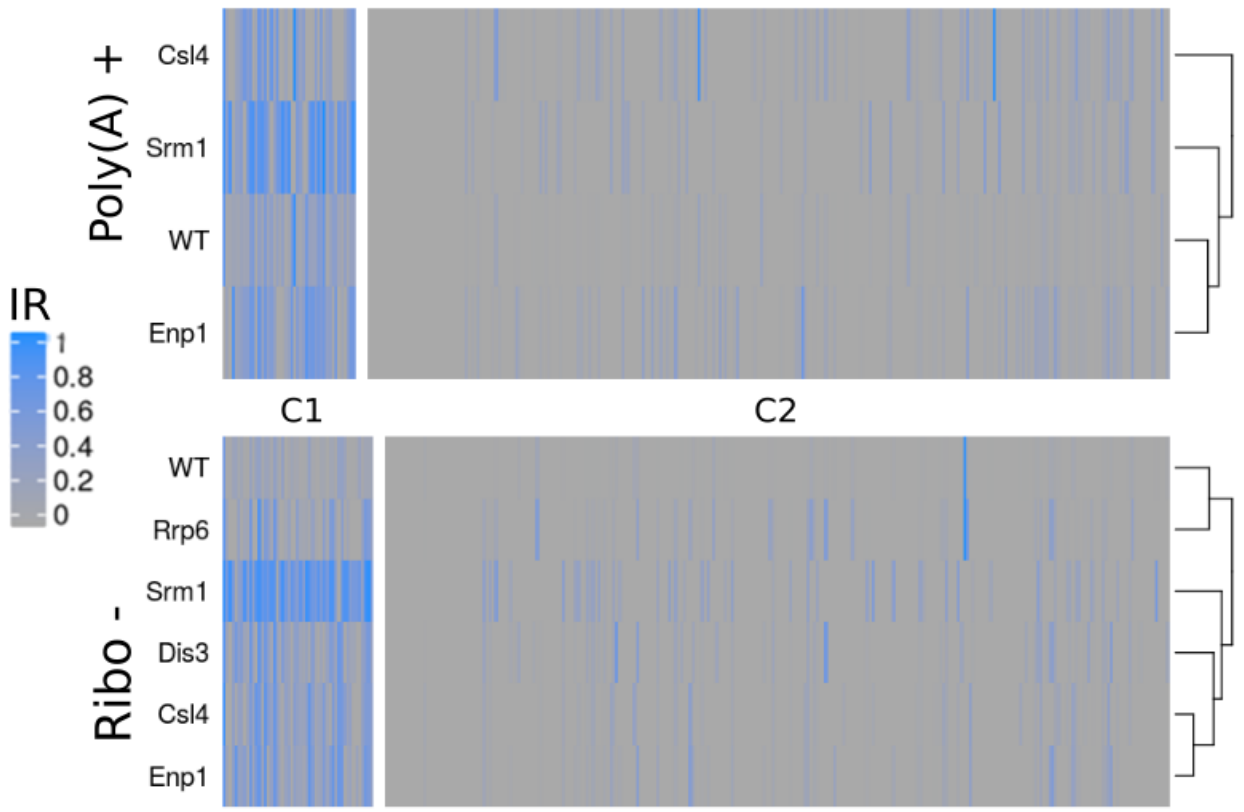

Figure 13. – HEATMAP DU SCORE DE RÉTENTION. Le score IR des 317 introns du génome de la levure sont par *IRfinder.* Les régions introniques ont été classées en deux groupes par l'algorithme des K-moyens.

Les 317 régions introniques ont été divisées en deux groupes selon la similitude du score IR à l'aide de l'algorithme de K-moyen. Pour les deux méthodes de purification de séquençage (*Ribo -* et *Poly(A) +*), on observe un signal de rétention d'intron chez les régions comprises dans le groupe C1. Les groupes C1 sont composés de 48 gènes introniques pour la méthode *Poly(A) +* et 51 gènes introniques pour la méthode *Ribo -*. La composition des deux groupes est assez similaire puisque 70 % des gènes du groupe C1 *Poly(A) +* font également partie du groupe C1 *Ribo -*. Le score IR est pratique nul pour les régions introniques de la levure *WT* où les ARNm aberrants devraient être ciblés pour la dégradation. La levure mutante *Srm1* démontre un signal plus prononcé que les autres mutants parmi les régions du groupe C1. Le phénomène de rétention d'intron est observable chez seulement ~15 % des régions introniques.

Nous avons évalué la corrélation entre le score de rétention d'intron et l'expression des gènes associés à ces introns. Nous voulons ici s'assurer que le signal IR ne soit pas causé par une faible expression du gène où la couverture de séquençage des régions exon/intron pourraient amplifier le score IR dû à la couverture de séquençage plutôt faible des régions exonique.

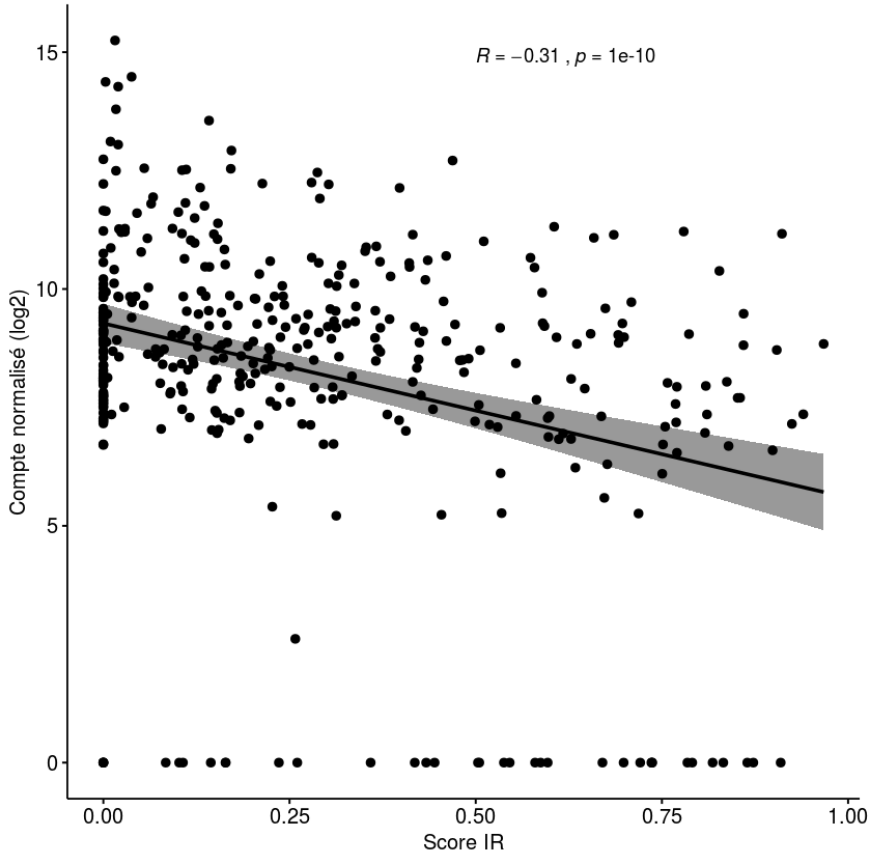

Figure 14. – CORRÉLATION ENTRE LE SCORE IR ET LE COMPTE NORMALISÉ DES GÈNES INTRONIQUES. Les comptes normalisés sont produits par DESeq2 puis transformés par log2.

On observe une très faible corrélation (Pearson = -0.31) entre les deux variables ce qui confirme que le signal de rétention d'intron n'est pas attribuable à une faible couverture des régions exoniques (figure 14).

De façon générale, les perturbations induites par les mutations analysées dans ce projet permettent d'observer un phénomène de rétention d'intron chez seulement 15 % des régions introniques de la levure. La perturbation du mécanisme de transport nucléocytoplasmique induite par la mutation *Srm1* semble être légèrement plus importante sur la magnitude du signal de rétention d'intron chez le peu de gènes affectés.

3.4 Analyse des données de séquençage TAIL-Seq

L'une des trois étapes principales à la maturation des ARNm consiste à l'ajout d'un segment d'adénosines à l'extrémité 3' de l'ARN. Ce processus que l'on nomme polyadénylation est nécessaire à la stabilisation et le transport des ARNs vers le cytoplasme. L'objectif de cette section est d'appliquer le modèle DNBC sur les données TAIL-Seq afin d'évaluer l'impact de la déstabilisation de l'exosome sur la longueur des segments polyadénylés. Le séquençage des régions 3' des transcrits polyadénylés a été effectué par la méthode de séquençage TAIL-Seq pour les trois mutants de l'exosome (*Csl4, Dis3* et *Rrp6*) ainsi que pour le mutant *Enp1*.

3.4.1 Comparaison du modèle DNBC avec l'outil TailSeeker

Nous avons testé notre modèle avec celui développé initialement avec la méthode TAIL-Seq. Comparativement à notre modèle, l'algorithme TailSeeker utilise directement les données brutes de fluorescence du séquenceur. Nous avons entraîné notre modèle avec les séquences poly(A) des deux transcriptomes mesurés et annotés par TailSeeker selon l'approche décrite au chapitre 2.2.2 pour ensuite prédire l'ensemble des séquences. Ces données sont disponibles sur la base de données GEO à l'aide des numéros d'accession GSE51299 et GSE54114.

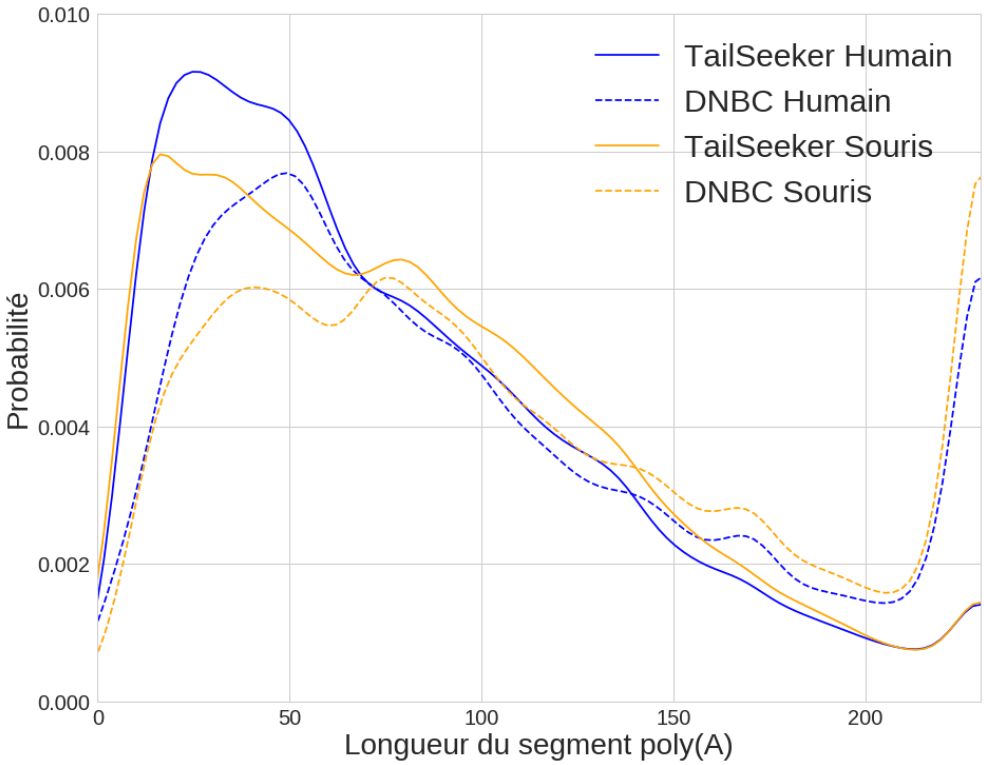

Figure 15. – COMPARAISON DES DISTRIBUTIONS DE LONGUEURS DE SEGMENTS POLY(A) PRÉDITS PAR LE MODÈLE DNBC ET PAR L'ALGORITHME TAILSEEKER. L'estimation de la taille des segments poly(A) est comparable entre les deux modèles bien que le DNBC surestime la taille de longs transcrits.

Nous comparons ici la distribution de la longueur des queues poly(A) mesurées par notre modèle sur deux jeux de données ayant été mesurés par l'algorithme TailSeeker (figure 15). Si l'on observe les queues poly(A) de 50 adénosines et moins, on observe que le modèle DNBC n'arrive pas à prédire efficacement la taille de ces segments. On observe également que notre modèle surestime plusieurs séquences poly(A) alors qu'on observe une très haute densité de probabilité pour les segments de 230 nucléotides.

Les courtes séquences qui ne sont pas correctement prédites sont possiblement de courtes séquences pour lesquelles le signal de fluorescence de la thymine ne s'est pas accumulé de façon significative. Dans ce cas, les nucléotides suivant la queue poly(A) sont séquencés avec un score de qualité convenable. La prédiction du DNBC sera alors largement supérieure à la taille réelle dû

au modèle DNBC qui restera dans l'état poly(A) tout au long de la séquence. Dans ce cas, le modèle DNBC reste dans l'état poly(A) puisque la probabilité qu'un nucléotide appartienne à une région poly(A) augmente lorsque le score de qualité est élevé. La relation est inverse pour les régions non-polyadénylées, alors que la probabilité augmente lorsque le score de qualité est faible. Il serait donc possible qu'une partie des segments estimée à 230 nucléotides soient en fait de courtes séquences pour lesquelles le score de qualité de séquençage reste élevé et stable pour toute la séquence.

3.4.2 Distribution des longueurs de segments poly(A)

Les données de séquençage de ce sous-chapitre sont issues des expériences de séquençage TAIL-Seq utilisant 150 nucléotides pour chacun des deux fragments pairés. Tel qu'il a été discuté au chapitre 2.3, les segments poly(A) de 130 nucléotides ont été omis lors de l'analyse des distributions des longueurs de fragments.

Nous avons d'abord comparé les distributions des longueurs de fragments polyadénylés des mutants de l'exosome obtenues par le modèle DNBC avec l'estimation de l'approche simpliste par compte de nucléotide tel que discuté au chapitre 2.2.1.

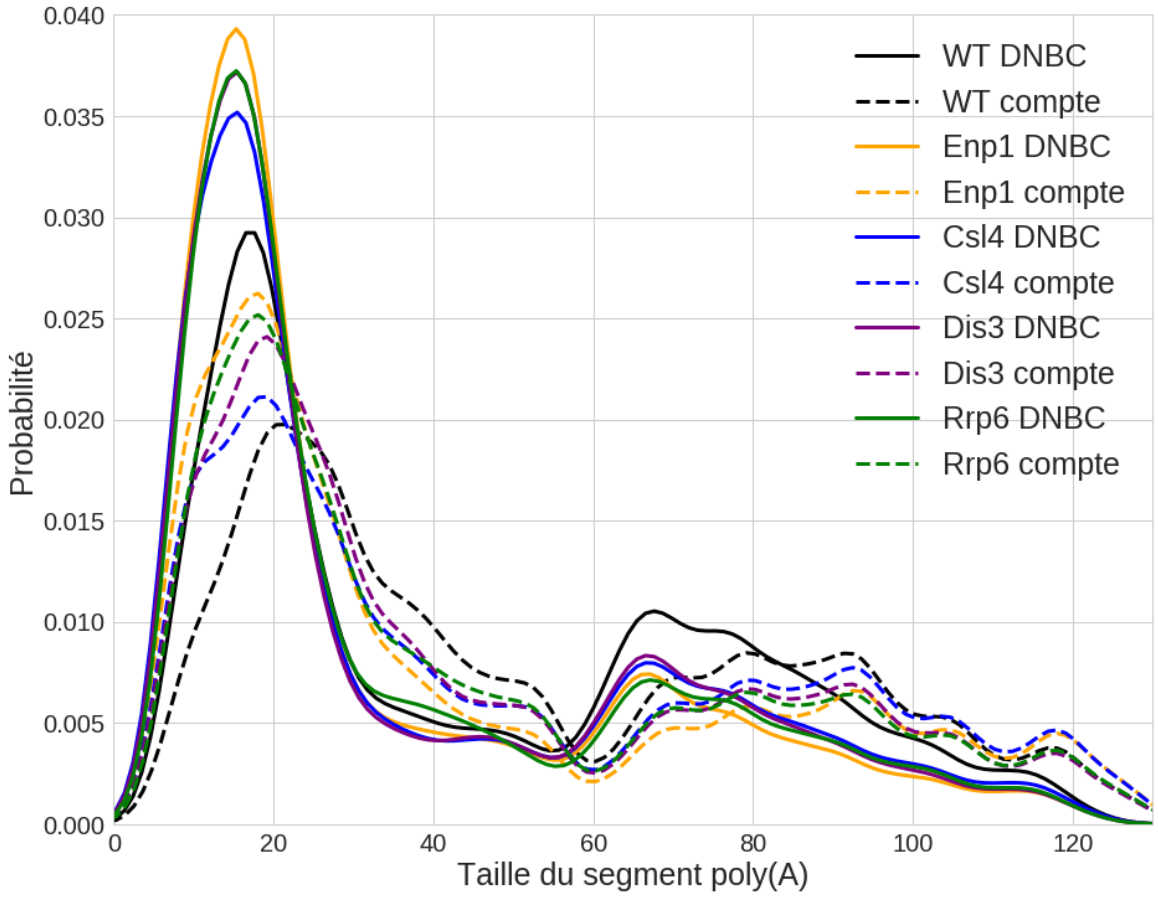

Figure 16. – COMPARAISON DES DISTRIBUTIONS DES TAILLES DE FRAGMENTS DÉTERMINÉS PAR LE MODÈLE DNBC ET PAR LA MÉTHODE DE COMPTE. La distribution des tailles de segments poly(A) produite par la méthode de compte pourrait indiquer que cette méthode produit des fragments plus long que le DNBC.

De manière générale, on observe qu'une grande partie des séquences mesurées sont de courts segments (~20 nucléotides). On observe également que les courbes de densité diminuent progressivement juste qu'à 60 nucléotides puis augmentent subitement avant de diminuer progressivement à nouveau (figure 16). Si l'on tente de comparer la méthode par compte avec celle du DNBC, on peut voir que la méthode par compte surestime légèrement la taille des segments. En effet, on observe moins de courtes séquences (< 20) avec méthode par compte puis une densité plus élevée pour les séquences de 20 à 60 nucléotides. L'allure du graphique pour les séquences de plus de 60 nucléotides est assez semblable à ce qui se produit au début du

graphique : la méthode DNBC produit un signal plus fort initialement suivi d'une décroissance progressive alors que la méthode par compte est légèrement déphasée avec un signal plus faible entre 60 à 80 nucléotides suivi d'une décroissance progressive avec une plus forte densité que celle du DNBC. Si l'on tente de comparer le profil de densité entre les mutants, on observe la même tendance entre la méthode par compte et celle du DNBC : on retrouve plus de courtes séquences (< 20) chez les mutants que chez la levure de type sauvage (*WT*), puis légèrement plus de séquences de longueur moyenne à longue pour la levure *WT* que pour les mutants.

Nous avons ensuite évalué la distribution des longueurs médianes par gène afin d'observer si la défaillance du l'exosome aurait un impact généralisé sur la taille des segments poly(A (figure 17).

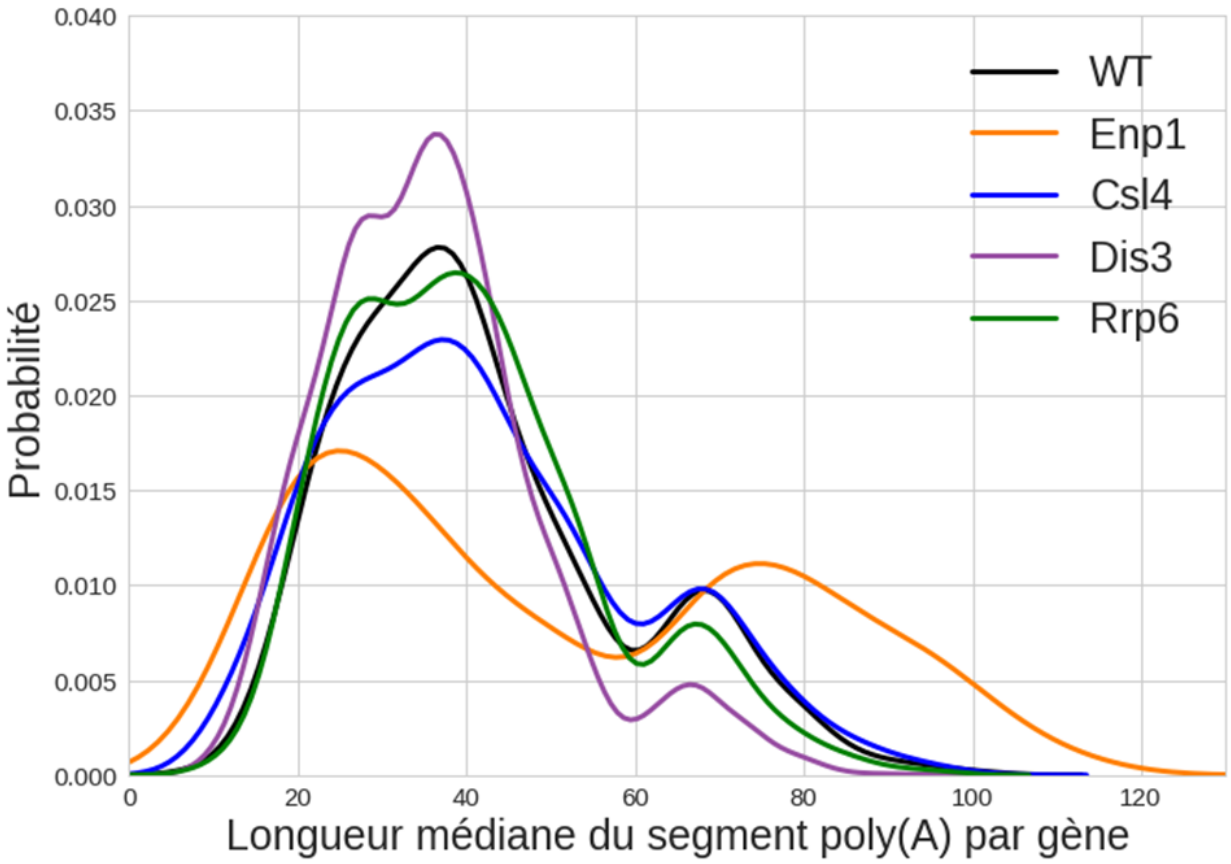

Figure 17. – DISTRIBUTION DES LONGUEURS MÉDIANE DES SEGMENTS POLY(A) PAR GÈNE. Seulement les gènes avec au moins 20 queues poly(A) séquencés dans chaque réplica ont été utilisés pour calculer les distributions.

On observe que la médiane des queues polyadénylées par gène est d'approximativement ~40 adénosines. La distribution de ces médianes est très similaire chez les levures mutées pour les composantes de l'exosome et pour la levure de type sauvage. Le profil de la levure mutée pour le gène *Enp1* semble différer légèrement des autres. On y observe un signal plus faible pour les médianes plus courtes (20 à 40 nucléotides) et un signal plus fort pour les médianes plus longues (60 nucléotides et plus). De manière générale, on observe le même phénomène qu'à la figure 3.3.1.1 où les distributions semblent être produites par deux signaux distincts, soit un avant et un après la marque des 60 nucléotides.

De manière générale, notre modèle réussi à estimer convenablement la longueur des queues poly(A) produites par séquençage TAIL-Seq à part pour les segments très courts. Nous avons réussi à mesurer ces segments pour l'ensemble du transcriptome de la levure. Nous sommes toutefois dans l'incapacité d'identifier un ensemble considérable de gènes dont la distribution des longueurs de queue poly(A) est significativement influencée par les mutations. En effet, la variance du signal entre les réplicas est trop élevée pour établir avec confiance un ensemble de gènes candidats.

4 Discussion

Le travail présenté dans ce mémoire a permis d'approfondir la compréhension de la réponse cellulaire de la levure à la suite des perturbations du mécanisme de contrôle de qualité de l'ARN. L'étude de ce mécanisme est une tâche complexe puisque les différentes composantes de ce mécanisme telles que les étapes de la maturation, le transport ou la dégradation des transcrits, sont en relation les unes entre les autres. Il peut donc être difficile d'associer un lien direct entre la perturbation de l'exosome et l'observation d'un certain phénotype. Il est possible que le phénotype observé soit l'effet secondaire ou tertiaire de la chaîne de réactions des différentes composantes du mécanisme de contrôle de qualité. Par exemple, la perturbation de l'exosome pourrait influencer une composante du réseau de régulation de l'expression qui permettrait d'induire un changement dans l'expression d'un vaste ensemble de gènes. Afin d'étudier ce mécanisme complexe, nous devons alors prendre en compte et caractériser les différents aspects du transcriptome tels que son niveau d'expression, la rétention de régions introniques chez les ARNm et le changement de la taille des régions polyadénylées. Nous avons toutefois été en mesure d'observer une similarité dans l'expression du transcriptome des levures à la suite des perturbations survenues à différents niveaux du mécanisme. En effet, les levures où le mécanisme d'export nucléocytoplasmique a été déstabilisé par la mutation du gène *Srm1* ainsi que les levures dont la maturation des précurseurs d'ARN ribosomiques est déstabilisée par la mutation du gène *Enp1* ont des phénotypes similaires aux levures mutées pour des composantes fonctionnelles de l'exosome.

Le phénotype des levures mutantes est caractérisé par la rétention d'ARN polyadénylés dans le nucléole, dont la localisation a été confirmée par une étude FISH sur oligo-dT [Biplab, 2016]. Cette même étude a permis d'observer une ségrégation des protéines liant l'ARN dans le nucléole. Par exemple, chez le mutant *Enp1*, une grande quantité d'ARN polyadénylé est retenue dans le nucléole forçant la relocalisation des facteurs de maturation dans cette région de la cellule [Biplab, 2016]. Ce phénomène est également observable dans notre étude du transcriptome où la liste de gènes surexprimés est enrichie pour des protéines liant l'ARN. L'accumulation des

transcrits cryptiques est également un aspect phénomène partagé par le phénotype des levures mutantes. Alors que l'expression de ces transcrits repose normalement sur l'inactivation des processus de dégradation, l'observation de ce phénomène chez les mutants *Srm1* et *Enp1* témoigne de l'interconnexion des différentes composantes du processus de maturation de l'ARN avec l'exosome. Ces résultats pourraient souligner l'importance des nucléases cytoplasmiques dans la voie de dégradation des ARN aberrants. Nous n'avons pas été capables de déterminer si une famille de transcrits cryptiques était particulièrement plus exprimée que les autres puisqu'il existe un chevauchement considérable des coordonnées des régions entre les familles. Mécaniquement, ces perturbations doivent probablement transmettre une information au système de régulation de l'expression alors que le phénomène de stabilisation des transcrits cryptiques semble être généralisé.

Un autre processus biologique qui pourrait être affecté de façon directe ou indirecte par les perturbations à l'étude est l'épissage. Des introns correctement retirés des transcrits qui s'accumulent dans le noyau pourraient indiquer un problème avec le système de dégradation de l'ARN alors que la rétention d'introns au sein des transcrits suggère un problème avec la machinerie de l'épissage. On ne retrouve que 317 régions introniques chez la levure qui possède approximativement 5800 gènes fonctionnels. Il y a donc très peu de gènes candidats pour tenter de tester cette hypothèse. L'observation d'un signal considérable de rétention d'intron chez seulement 50 des 317 régions introniques ne nous permet pas d'affirmer que la défaillance induite par les mutations ciblées dans ce projet affecte le processus d'épissage de manière généralisée. La redondance de la fonction de dégradation des différentes nucléases en action pourrait expliquer pourquoi nous obtenons un signal de rétention chez très peu de gènes. En effet, la dégradation des ARN aberrants dans le cytoplasme est assurée par l'exosome et par la protéine *Xrn1*. Puisque celle-ci n'est pas affectée directement par les mutations induites à la levure, elle pourrait maintenir la fonction de dégradation des ARN aberrants dans le cytoplasme. Les quelques gènes identifiés dans cette analyse pourraient être des gènes spécifiquement retenus dans le noyau où l'action de l'exosome serait déstabilisée. D'autre part, les mutations des gènes *Enp1* et *Srm1* pourraient également affecter l'homéostasie nucléaire ayant un effet négatif sur la maturation et l'export des transcrits. Par exemple, si le mécanisme de transport

nucléocytoplasmique ne fonctionne pas correctement, un transcrit dont la dégradation est dépendante de la nucléase cytoplasmique *Xrn1* ne pourrait donc pas être géré correctement par la cellule. Alors que ces transcrits aberrants liés à l'étape d'épissage sont ciblés pour la dégradation dans le noyau et dans le cytoplasme, les perturbations induites par les mutations étudiées dans le cadre de ce projet n'ont pas permis d'observer un signal clair de rétention d'introns chez la levure.

La polyadénylation est une étape cruciale pour la stabilisation, le transport et la régulation de l'ARN. Nous avons été en mesure d'établir un modèle probabiliste permettant d'approximer la taille des séquences poly(A) générée par séquençage TAIL-Seq. Toutefois, il est très difficile d'évaluer si les perturbations étudiées dans ce projet ont un impact sur la distribution des segments poly(A). Les difficultés techniques liées au séquençage des segments répétitifs et le manque de modèle permettant d'expliquer la régulation de la taille de ces segments ont grandement complexifié l'interprétation de nos résultats. Étant donné que le mécanisme lié à la régulation de la taille des segments poly(A) reste relativement incompris, il est difficile de déterminer si la distribution des segments poly(A) associée à un gène varie à la suite d'une mutation. Puisque le modèle permettant de décrire cette distribution est inconnu, nous avons tenté de caractériser ces distributions en utilisant la médiane des longueurs poly(A) par gène. La variance du signal entre les réplicas nous empêche d'établir avec certitude une liste de gènes dont la taille des segments poly(A) est significative influencée par les mutations testées. Alors que les transcrits hypoadénylés ou hyperadénylés sont généralement ciblés pour la dégradation, peu d'information est connue sur la magnitude de la taille du segment poly(A) en relation avec le cycle de vie du transcrit. Nous avons évalué à quel point ce changement dans la distribution des tailles de segments aurait un impact sur le transcrit, mais aucune corrélation n'a pu être établie entre la taille des régions poly(A) et le niveau d'expression du gène ou son temps de demi-vie.

Ensemble, les différentes analyses de transcriptomique de ce projet ont permis de caractériser la réponse générique de la défaillance du mécanisme de contrôle de qualité de l'ARN. Ce système complexe aux composantes interconnectées semble fonctionner de manière similaire à la suite de la déstabilisation de différentes composantes du mécanisme.

Si ce projet avait à se poursuivre, il serait intéressant d'étudier le phénomène de rétention d'intron en combinant les mutations de l'exosome étudiées dans ce projet avec une déstabilisation des autres complexes cytoplasmiques qui favorisent la dégradation des ARNs aberrants. Par exemple, une mutation affectant la fonctionnalité du complexe NMD nous permettrait d'évaluer la redondance de ces mécanismes cruciaux pour le maintien de l'intégrité du transcriptome. Davantage d'efforts pourraient également être alloués à combiner les coordonnées génomiques des différentes familles de transcrits cryptiques. Alors qu'il existe une certaine redondance entre ceux-ci, il serait intéressant d'établir une méthode d'unification de ces annotations afin de faciliter l'analyse de ces transcrits. Finalement, le projet pourrait se poursuivre en développant un modèle probabiliste pour trouver de nouveaux transcrits cryptiques. L'analyse des données de séquençage utilisées dans ce projet par un tel outil nous permettrait de valider les coordonnées des transcrits pour des mutants connus, tel que les SUT et les CUT avec le mutant *Rrp6*, mais aussi de trouver de nouvelles régions sur les mutants qui n'avait pas été étudiés dans le passé, comme le mutant *Srm1*.

5 Références bibliographiques

Alexandrov A, Chernyakov I, Gu W, Hiley SL, Hughes TR, Grayhack EJ, Phizicky EM. Rapid tRNA decay can result from lack of nonessential modifications. Mol Cell. 2006 Jan 6;21(1):87-96.

Allmang C, Petfalski E, Podtelejnikov A, Mann M, Tollervey D, Mitchell P. The yeast exosome and human PM-Scl are related complexes of 3' --> 5' exonucleases. Genes Dev. 1999 Aug 15;13(16):2148-58.

Anderson JS, Parker RP. The 3' to 5' degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3' to 5' exonucleases of the exosome complex. EMBO J. 1998 Mar 2;17(5):1497-506.

Apponi LH1, Leung SW, Williams KR, Valentini SR, Corbett AH, Pavlath GK. Loss of nuclear poly(A)-binding protein 1 causes defects in myogenesis and mRNA biogenesis. Hum Mol Genet. 2010 Mar 15;19(6):1058-65.

Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.

Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res. 1997 Oct;7(10):986-95.

Bai, Y., Ji, S. & Wang, Y. IRcall and IRclassifier: Two methods for flexible detection of intron retention events from RNA-Seq data. BMC Genomics 16, S9 (2015).

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289-300.

Bentley DL. Coupling mRNA processing with transcription in time and space. Nat Rev Genet. 2014;15(3):163–175.

Bissels U, Wild S, Tomiuk S, et al. Absolute quantification of microRNAs by using a universal reference. RNA. 2009;15(12):2375–2384.

Camblong J1, Iglesias N, Fickentscher C, Dieppois G, Stutz F. Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in S. cerevisiae. Cell. 2007 Nov 16;131(4):706-17.

Carroll KL, Ghirlando R, Ames JM, Corden JL. Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements. RNA. 2007 Mar;13(3):361-73

Castle JC, Armour CD, Löwer M, et al. Digital genome-wide ncRNA expression, including SnoRNAs, across 11 human tissues using polyA-neutral amplification. PLoS One. 2010;5(7):e11779. Published 2010 Jul 26.

Cawley, S. L. & Pachter, L. HMM sampling and applications to gene finding and alternative splicing. Bioinformatics 19, (2003).

Chang H, Lim J, Ha M, Kim VN. TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications. Mol Cell. 2014 Mar 20;53(6):1044-52.

Chang JH, Jiao X, Chiba K, Oh C, Martin CE, Kiledjian M, Tong L. Dxo1 is a new type of eukaryotic enzyme with both decapping and 5'-3' exoribonuclease activity. Nat Struct Mol Biol. 2012 Oct;19(10):1011-7.

Chang YF1, Imam JS, Wilkinson MF. The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem. 2007;76:51-74.

Chen W, Bucaria J, Band DA, Sutton A, Sternglanz R. Enp1, a yeast protein associated with U3 and U14 snoRNAs, is required for pre-rRNA processing and 40S subunit synthesis. Nucleic Acids Res. 2003;31:690–699

Chen Z, Li Y, Krug RM. Influenza A virus NS1 protein targets poly(A)-binding protein II of the cellular 3'-end processing machinery. EMBO J. 1999;18(8):2273–2283.

Cherry JM, Adler C, Ball C, et al. SGD: Saccharomyces Genome Database. Nucleic Acids Res. 1998;26(1):73–79. doi:10.1093/nar/26.1.73

Choi YH, Hagedorn CH. Purifying mRNAs with a high-affinity eIF4E mutant identifies the short 3' poly(A) end phenotype. Proc Natl Acad Sci U S A. 2003 Jun 10;100(12):7033-8. Epub 2003 May 30.

Daehwan Kim, Joseph M. Paggi, Chanhee Park, Christopher Bennett & Steven L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology volume 37, pages 907–915 (2019)

Danin-Kreiselman M1, Lee CY, Chanfreau G. RNAse III-mediated degradation of unspliced pre-mRNAs and lariat introns. Mol Cell. 2003 May;11(5):1279-89.

Davidson L, Kerr A, West S. Co-transcriptional degradation of aberrant pre-mRNA by Xrn2. EMBO J. 2012;31(11):2566–2578. doi:10.1038/emboj.2012.101

Dichtl B, Keller W. Recognition of polyadenylation sites in yeast pre-mRNAs by cleavage and polyadenylation factor. EMBO J. 2001;20(12):3197–3209. doi:10.1093/emboj/20.12.3197

Dziembowski A, Ventura AP, Rutz B, Caspary F, Faux C, Halgand F, Laprévote O, Séraphin B. Proteomic analysis identifies a new complex required for nuclear pre-mRNA retention and splicing. EMBO J. 2004 Dec 8;23(24):4847-56. Epub 2004 Nov 25.

Egecioglu DE, Chanfreau G. Proofreading and spellchecking: a two-tier strategy for pre-mRNA splicing quality control. RNA. 2011 Mar;17(3):383-9. doi: 10.1261/rna.2454711. Epub 2011 Jan 4.

Eswaran, J. et al. RNA sequencing of cancer reveals novel splicing alterations. Sci. Rep. 3, (2013).

Evguenieva-Hackenberg E, Klug G. RNA degradation in Archaea and Gram-negative bacteria different from Escherichia coli. Prog Mol Biol Transl Sci. 2009;85:275-317. doi: 10.1016/S0079-6603(08)00807-6.

Fox MJ, Gao H, Smith-Kinnaman WR, Liu Y, Mosley AL. The exosome component Rrp6 is required for RNA polymerase II termination at specific targets of the Nrd1-Nab3 pathway. PLoS Genet. 2015 Feb 13;11(2):e1004999. doi: 10.1371/journal.pgen.1004999.

Goodier JL. Restricting retrotransposons: a review. Mob DNA. 2016 Aug 11;7:16. doi: 10.1186/s13100-016-0070-z.

Goss DJ, Kleiman FE. Poly(A) binding proteins: are they all created equal? Wiley Interdiscip Rev RNA. 2013 Mar-Apr;4(2):167-79. doi: 10.1002/wrna.1151. Epub 2012 Dec 13.

Gu Z, Eils R, Schlesner M (2016). “Complex heatmaps reveal patterns and correlations in multidimensional genomic data.” Bioinformatics.

Harigaya, Y. and Parker, R. (2010), No-go decay: a quality control mechanism for RNA in translation. WIREs RNA, 1: 132-141.

Honigberg S, Purnapatre K. Signal pathway integration in the switch from the mitotic cell cycle to meiosis in yeast. Journal of Cell Science 2003 116: 2137-2147.

Houseley J, LaCava J, Tollervey D. RNA-quality control by the exosome. Nat Rev Mol Cell Biol. 2006 Jul; 7(7): 529–539.

Isken O, Maquat LE. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA funtion. Genes Dev. 2007 Aug 1;21(15):1833-56.

Jiao X, Xiang S, Oh C, Martin CE, Tong L, Kiledjian M. Identification of a quality-control mechanism for mRNA 5'-end capping. Nature. 2010;467(7315):608–611.

Jacobson A, Peltz SW. Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. Annu Rev Biochem. 1996;65:693-739.

Jaillon O. Translational control of intron splicing in eukaryotes. Nature. 2008 Jan 17;451(7176):359-62.

Käll, L., Krogh, A. & Sonnhammer, E. L. L. An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 21, 251–257 (2005).

Keller RW, Kühn U, Aragón M, Bornikova L, Wahle E, Bear DG. The nuclear poly(A) binding protein, PABP2, forms an oligomeric particle covering the length of the poly(A) tail. J Mol Biol. 2000 Mar 31;297(3):569-83.

Khorsheed, M. S. Offline recognition of omnifont Arabic text using the HMM ToolKit (HTK). Pattern Recognit. Lett. 28, 1563–1571 (2007).

Kühn U, Gündel M, Knoth A, Kerwitz Y, Rüdel S, Wahle E. Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. J Biol Chem. 2009 Aug 21;284(34):22803-14.

LaCava J, Houseley J, Saveanu C, Petfalski E, Thompson E, Jacquier A, Tollervey D. RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. Cell. 2005 Jun 3;121(5):713-24.

LaRiviere, F.J., Cole, S.E., Ferullo, D.J., Moore, M.J.(2006) A late-acting quality control process for mature eukaryotic rRNAs. Mol. Cell 24:619–626

Lee YJ, Glaunsinger BA. Aberrant herpesvirus-induced polyadenylation correlates with cellular messenger RNA destruction. PLoS Biol. 2009 May 5;7(5):e1000107

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

Li HD, Funk CC, Price ND. iREAD: A Tool For Intron. bioRxiv 135624.

Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013 May 1;41(10):e108.

Lim J, Ha M, Chang H, Kwon SC, Simanshu DK, Patel DJ, Kim VN. Uridylation by TUT4 and TUT7 marks mRNA for degradation. Cell. 2014 Dec 4;159(6):1365-76.

Lima SA, Chipman LB, Nicholson AL, et al. Short poly(A) tails are a conserved feature of highly expressed genes. Nat Struct Mol Biol. 2017;24(12):1057–1063.

Liu JJ, Niu CY, Wu Y, Tan D, Wang Y, Ye MD, Liu Y, Zhao W, Zhou K, Liu QS, Dai J, Yang X, Dong MQ, Huang N, Wang HW. CryoEM structure of yeast cytoplasmic exosome complex. Cell Res. 2016 Jul;26(7):822-37.

Locke JM1, Da Silva Xavier G, Rutter GA, Harries LW. An alternative polyadenylation signal in TCF7L2 generates isoforms that inhibit T cell factor/lymphoid-enhancer factor (TCF/LEF)-dependent target genes. Diabetologia. 2011 Dec;54(12):3078-82.

Lorentzen E1, Dziembowski A, Lindner D, Seraphin B, Conti E. RNA channelling by the archaeal exosome. EMBO Rep. 2007 May;8(5):470-6.

Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, 550.

Marquardt S, Hazelbaker DZ, Buratowski S. Distinct RNA degradation pathways and 3' extensions of yeast non-coding RNA species. Transcription. 2011 May;2(3):145-154.

Mattick JS. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep*. 2001;2(11):986–991.

Meijer HA, Radford HE, Wilson LS, Lissenden S, de Moor CH. Translational control of maskin mRNA by its 3' untranslated region. Biol Cell. 2007 May;99(5):239-50.

Menet JS, Rodriguez J, Abruzzi KC, Rosbash M. Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. Elife. 2012;1:e00011. Published 2012 Nov 13.

Middleton, R. et al. IRFinder: Assessing the impact of intron retention on mammalian gene expression. Genome Biol. 18, 1–11 (2017).

Mitchell P1, Petfalski E, Shevchenko A, Mann M, Tollervey D. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases. Cell. 1997 Nov 14;91(4):457-66.

Moteki S, Price D. Functional coupling of capping and transcription of mRNA. Mol Cell. 2002 Sep; 10(3): 599–609.

Nemeth A, Krause S, Blank D, Jenny A, Jenö P, Lustig A, Wahle E. Isolation of genomic and cDNA clones encoding bovine poly(A) binding protein II. Nucleic Acids Res. 1995 Oct 25;23(20):4034-41.

Padgett RA, Grabowski PJ, Konarska MM, Seiler S, Sharp PA. Splicing of messenger RNA precursors. Annu Rev Biochem. 1986;55:1119-50

Paul B, Montpetit B. Altered RNA processing and export lead to retention of mRNAs near transcription sites and nuclear pore complexes or within the nucleolus. Mol Biol Cell. 2016;27(17):2742–2756.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008 Dec;40(12):1413-5.

Pimentel, H. et al. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. Nucleic Acids Res. 44, 838–851 (2016).

Saguez C, Schmid M, Olesen JR, Ghazy MA, Qu X, Poulsen MB, Nasser T, Moore C, Jensen TH. Nuclear mRNA surveillance in THO/sub2 mutants is triggered by inefficient polyadenylation. Mol Cell. 2008;31:91–103.

Schulz D, Schwalb B, Kiesel A, Baejen C, Torkler P, Gagneur J, Soeding J, Cramer P. Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell. 2013 Nov 21;155(5):1075-87.

Söding, J. Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951–960 (2005).

Solomon, D. A. et al. Cyclin D1 splice variants: Differential effects on localization, RB phosphorylation, and cellular transformation. J. Biol. Chem. 278, 30339–30347 (2003).

Stark H, Lührmann R. Cryo-electron microscopy of spliceosomal components. Annu Rev Biophys Biomol Struct. 2006;35:435-57.

Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature. 2014 Apr 3;508(7494):66-71.

Synowsky SA, van Wijk M, Raijmakers R, Heck AJ. Comparative multiplexed mass spectrometric analyses of endogenously expressed yeast nuclear and cytoplasmic exosomes. J Mol Biol. 2009 Jan 30;385(4):1300-13.

Topisirovic I, Svitkin YV, Sonenberg N, Shatkin AJ. Cap and cap-binding proteins in the control of gene expression. Wiley Interdiscip Rev RNA. 2011 Mar-Apr;2(2):277-98.

Tuck AC, Tollervey D. A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs. Cell. 2013 Aug 29;154(5):996-1009. doi: 10.1016/j.cell.2013.07.047.

Tudek A, Porrua O, Kabzinski T, Lidschreiber M, Kubicek K, Fortova A, Lacroute F, Vanacova S, Cramer P, Stefl R, Libri D. Molecular basis for coordinating transcription termination with noncoding RNA degradation. Mol Cell. 2014 Aug 7;55(3):467-81.

Van Dijk EL, Chen CL, d'Aubenton-Carafa Y, Gourvennec S, Kwapisz M, Roche V, Bertrand C, Silvain M, Legoix-Né P, Loeillet S, Nicolas A, Thermes C, Morillon A. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature. 2011 Jun 22;475(7354):114-7.

Van Hoof A, Lennertz P, Parker R. Yeast exosome mutants accumulate 3'-extended polyadenylated forms of U4 small nuclear RNA and small nucleolar RNAs. Mol Cell Biol. 2000 Jan;20(2):441-52.

Vera JM, Dowell RD. Survey of cryptic unstable transcripts in yeast. BMC Genomics. 2016 Apr 26;17:305.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008 Nov 27;456(7221):470-6.

Wasmuth EV, Lima CD. Exo- and endoribonucleolytic activities of yeast cytoplasmic and nuclear RNA exosomes are dependent on the noncatalytic core and central channel. Mol Cell. 2012 Oct 12;48(1):133-44.

Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Wiestner A. Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. Blood. 2007 Jun 1;109(11):4599-606.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. Bidirectional promoters generate pervasive transcription in yeast. Nature. 2009 Feb 19;457(7232):1033-7.

Zhou, Guodong & Jian, Su. (2002). Named Entity Recognition using an HMM-based Chunk Tagger. proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 10.3115/1073083.107