Interaction entre protéines et acides nucléiques et régulation de l'expression des gènes

Gwenaël Badis

Habilitation à Diriger les Recherches

Faculté de BIOLOGIE

Sorbonne Université

UFR des Sciences de la Vie

Gwenael BADIS-BREARD

Interaction entre protéines et acides nucléiques et régulation de l'expression des gènes

Soutenance le 23 octobre 2019

Devant un jury composé de :

Rapporteurs :

Françoise Stutz, Professeure à l'Université de Genève (excusée)

Nicolas Leulliot, Professeur à l'Université Paris Descartes

Lionel Benard, Directeur de recherche CNRS

Examinateurs :

Odil Porrua, Chargée de recherche CNRS

Dominique Weil, Directrice de recherche CNRS

Benoit Palancade, Directeur de recherche CNRS

Frédéric Devaux, Professeur à Sorbonne Université

TABLE DES MATIERES

TABLE DES MATIERES 2

TABLE DES ILLUSTRATIONS 3

REMERCIEMENTS 4

SUMMARY 5

RESUME 8

CURRICULUM VITAE 11

LISTE DES PUBLICATIONS 17

RAPPORT DE RECHERCHE : 21

Interaction entre protéines et acides nucléiques et régulation de l'expression des gènes 21

I Etude du métabolisme des ARN chez *S. cerevisiae* : analyse fonctionnelle et caractérisation d'ARNs non codants. 21

Thèse : Décembre 1999-juin 2004 21

Développement technologique d'outils pour l'analyse fonctionnelle des gènes de levure *21*

Comprendre les interactions ARNs/protéine et leur rôle : le cas d'EDC3. ..... *22*

Découverte de nouveaux ARN non codants à l'ère post génome de la levure. *23*

II Etude des spécificités de reconnaissance des facteurs de transcription eucaryotes ... 27

Post-doctorat : Juillet 2004-Décembre 2008 27

Les interactions ADN-protéines et la régulation de la transcription. *28*

II Les interactions ARN-protéines et la régulation du transcriptome de levure 33

Depuis janvier 2009 33

Régulation génique par interférence transcriptionnelle médiée par la transcription d'antisens non codants. *33*

Les interactions ARN-protéine et la régulation du transcriptome. *34*

Améliorations techniques du CRAC *35*

Etude de facteurs impliqués dans le contrôle qualité des ARNs aberrants *36*

Facteurs liés au NSD/NGD *36*

III PERSPECTIVES : Etude des interaction ARN-protéines dans les mécanismes de contrôle qualité des ARNs. 38

Un nouveau facteur impliqué dans le recyclage des ribosomes. *39*

Un nouveau facteur associé à l'exosome. *40*

Facteurs liés au NMD et au delà du NMD. *41*

Conclusion *42*

REFERENCES 43

PUBLICATIONS MAJEURES 48

TABLE DES ILLUSTRATIONS

Figure 1 : modèle révisé de la dégradation du mRNA de RPS28B par Edc3 (extrait de He et al., 2014). 23
Figure 2 : Comparaison des transcriptomes d'une souche sauvage et Δrrp6. 24
Figure 3 : Conservation et structure de snR86. 25
Figure 4 : Enrichissement des 8 mers associés à Lhx2 et Lhx4 appartenant à la famille des facteurs de transcription à homéodomaine. 31
Figure 5 : Représentation schématique du protocole du CRAC actuel. 36
Figure 6: Dégradation ciblée des mRNP aberrantes par le NMD, NGD, et NSD (d'après Roy and Jacobson, 2013) 37
Figure 7 : Localisation des sites de fixation dans la structure tridimensionnelle de la sous-unité ribosomique. 38

REMERCIEMENTS

« La vie n'est facile pour aucun de nous. Mais quoi, il faut avoir de la persévérance, et surtout de la confiance en soi. Il faut croire que l'on est doué pour quelque chose, et que, cette chose, il faut l'atteindre coûte que coûte. »

Madame Curie, Eve Curie, éd. Gallimard, 1938, p. 131

Marie Curie résume quelque chose de très important dans notre métier et dans la vie en général, mais croire en soi et en ses capacités n'est pas si simple et peut prendre des années (ou faire gagner beaucoup de temps à certains). Néanmoins, ce que j'aime dans notre métier, c'est qu'effectivement avec un peu de curiosité, beaucoup de persévérance et en gardant son cap, on finit souvent par arriver à quelque chose.
Je ne pensais pas à 20 ans que je serais capable de réaliser autant de choses et rencontrer toutes ces personnes passionnantes. Ne pas avoir peur de penser différemment, de confronter ses idées à celles des autres, de défendre ses opinions (quitte à remettre en question des dogmes bien établis quand cela est nécessaire), et de toujours garder l'esprit critique. C'est pourtant bien le socle de notre métier.
Je suis toujours étonnée par cet émerveillement de la découverte, et après plus de 20 ans au laboratoire, de continuer à prendre autant de plaisir à apprendre, à comprendre et à transmettre.

Je remercie toutes les personnes qui m'ont formé, accompagné et appris durant mon parcours de chercheur, que ce soit au niveau professionnel ou personnel.

C'est un travail qui ne peut se faire qu'en équipe, et je suis aussi très reconnaissante envers les personnes avec qui j'ai collaboré.

Merci à tous!

SUMMARY

When I started my PhD, the yeast genome has just been published and more than half of yeast genes still had unknown functions. I participated to the functional analysis of factors involved in mRNA decay in Alain Jacquier's laboratory. In particular, I characterized the role of Edc3, a new enhancer of decapping, involved in the regulation of *RPS28B* mRNA. In parallel I characterized several unknown H/ACA snoRNAs and also the first Cryptic Unstable transcripts (CUTs). Aside, I developed technological tools to improve synthetic lethal screens.

During my post-doctoral fellowship in Timothy Hughes Laboratory, I worked on decoding the mouse "regulome", *i.e.* the characterization of Mouse transcription factors DNA-binding motifs. I set up and orchestrated the project in mouse, and we could clone ~1200 DNA binding domains fused to the GST tag from the ~2000 mouse transcription factor we selected. Out of these clones, we have been able to purify ~800 proteins and identify ~600 DNA motifs, using the Protein Binding Microarray technology, in collaboration with Martha Bulyk's laboratory at Harvard Medical School.

A similar approach in yeast enabled us to identify 112 DNA binding motifs from the ~200 cloned yeast transcription factors. These resources constituted major advances in the knowledge of the regulatory regions recognized by transcription factors in both yeast and mouse genomes.

Since 2008 when I came back in Alain Jacquier's laboratory, I have been interested in a mechanism of transcriptional interference mediated by pervasive transcription, antisense of genes. In this work, we show that gene repression resulting from antisense transcriptional interference overlapping gene promoter constitutes a new mechanism of transcriptional regulation and concerns a large set of genes in yeast. This mechanism involve a combination of chromatin modifiers and can be reciprocal in many cases – which means that a gene repressed by antisense transcription can become itself a repressor of the antisense transcription in condition where it becomes expressed, such as in quiescence state.

In parallel, I was interested in RNA-protein interactions characterization, and I implemented and improved the CRAC technic (UV Cross Linking and cDNA analysis)

in the laboratory, in order to characterize RNA-binding specificities of cytoplasmic surveillance factors bound to rRNA, mRNA or ncRNA.

This approach allowed me to characterize the targets of 5 rRNA binding factors, and I am now studying the RNA-binding specificities of several helicases involved in cytoplasmic surveillance pathways and capable to bind genome wide many RNA targets. I am also interested in the development of tools to distinguish sub-populations of RNA that we named "split CRAC".

Indeed, our laboratory expertise is protein complexes involved in cytoplasmic surveillance pathways and Micheline's and Cosmin's teams in the lab characterized key factors of the No Go Decay (NGD)/ Non Stop Decay (NSD) and the Nonsense Mediated mRNA Decay (NMD) pathways. In the wave of this, I am interested in RNA-protein interactions related to these pathways, and in particular to the mode of action of certain helicases that belong to these processes.

The first factor is a putative helicase that we named Tac4 for "translation associated factor 4" because we found it associated to the ribosomal small subunit. Preliminary data suggest that Tac4 could be involved in ribosome recycling when they are blocked in translation such as in NGD or NSD mechanisms. In collaboration with Micheline Fromont and Olivier Namy we aim to characterize the precise role of Tac4 in this recycling mechanism.

The second factor is Ski2, the main SKI complex helicase, whose Micheline Fromont's team in collaboration with Roland Beckman and Elena Conti, recently found that it could be associated to the ribosome to degrade coding segments of RNAs. In parallel, she observed that another form of the SKI complex could be involved in 3' untranslated regions of RNA and comprises the Ska1 factor (for "SKI associated component"). The knowledge of Ska1 binding specificities should allow a better understanding of its precise role.

The third factor is Upf1, the main helicase of the NMD mechanism. Cosmin Saveanu's team has recently characterized two sub-complexes associated to Upf1: a "detector" complex (composed of Upf1, 2 and 3) and an "effector" complex (composed of Upf1 Nmd4, Ebs1 and the decapping machinery). In the aim to understand how Upf1 is associated to these two complexes, I realized preliminary experiments of RIP-seq and CRAC with Upf1, and started "split-CRAC" experiment with Upf1 and proteins of each sub-complexes. In the context of Lena Auderbert's Master 2 internship we performed

analysis of RIP and CRAC with Upf1, and preliminary results indicate that Upf1 preferentially bind 3' UTR of mRNAs and 5'UTR in a lesser extend. Surprisingly, in addition to the expected NMD targets mRNAs, Upf1 is bound to an important number of "non-NMD" mRNA substrates. Lena started her PhD in September 2018 on this topic, and I am co-directing her thesis with Cosmin.

Finally, using transcriptomic and biochemical technics, I am focusing on cytoplasmic surveillance pathways and in particular those involving Upf1, and on the understanding of mechanisms regulating subtle variations of poly(A) isoforms stability.

Considering our converging interests and the complementarity of our expertise, joining Cosmin Saveanu's team in September 2018 appeared as an evidence for me.

RESUME

Au début de ma thèse, le génome de S. cerevisiae vient d'être publié et plus de la moitié des gènes de la levure ont des fonctions inconnues. J'ai participé à l'analyse fonctionnelle de facteurs impliqués dans le métabolisme des ARN dans le laboratoire d'Alain Jacquier. J'ai en particulier caractérisé le rôle d'Edc3, un activateur de decapping impliqué dans la régulation du messager du gène RPS28B. J'ai également caractérisé un certain nombre d'ARN non codants tels que des snoRNA à boite H/ACA ainsi que les premiers « Cryptic Unstable Transcripts ». En parallèle, j'ai participé au développement technologique d'outils pour pouvoir réaliser des cribles de létalité synthétique de manière plus performante chez la levure.

Durant mon post-doctorat chez Timothy Hughes à Toronto, j'ai travaillé sur le déchiffrage du « regulome », c'est à dire l'identification des motifs reconnus par les facteurs de transcriptions. J'ai mis en œuvre et orchestré ce projet chez la souris, et nous avons pu cloner ≈ 1200 domaines de liaisons à l'ADN (DBD) des ≈ 2000 facteurs de transcription de souris en fusion avec la GST. Nous avons réussi à purifier de manière satisfaisante plus de 800 protéines et environ 600 motifs ont pu être caractérisés en réalisant des expériences de « Protein Binding Microarray » en collaboration avec Martha Bulyk à Harvard. Une approche similaire chez la levure nous a permis de caractériser 112 motifs de liaison à l'ADN parmi les ≈200 facteurs de transcription clonés. Ces ressources ont permis de faire une avancée majeure dans la connaissance des régions régulatrices des gènes qui sont reconnues par les facteurs de transcription dans ces espèces.

Depuis 2008 et mon retour au laboratoire d'Alain Jacquier, je me suis intéressée à un mécanisme d'interférence transcriptionnelle induite par de la transcription pervasive en antisens des gènes. Dans cette étude nous montrons que la répression par interférence transcriptionnelle résultant de la transcription d'ARN antisens chevauchant le promoteur des gènes constitue un mécanisme de régulation qui concerne un grand nombre de gène chez la levure. Ce mécanisme d'interférence transcriptionnelle fait intervenir une combinaison de facteurs de modification de la chromatine et peut être réciproque dans beaucoup de cas, c'est à dire qu'un gène réprimé par la transcription d'un antisens pervasif peut devenir répresseur de la transcription pervasive en antisens quand il est exprimé, comme par exemple en quiescence.

En parallèle, je me suis intéressée à la caractérisation des interactions ARN-protéines et j'ai implémenté au laboratoire et amélioré la technique du CRAC afin de caractériser les spécificités de reconnaissance de facteurs liant les ARN ribosomiques, ARNm et/ou ARNnc. Cette approche a permis de caractériser les cibles de 5 facteurs liant l'ARN ribosomique et je suis actuellement intéressée par certaines hélicases qui lient un grand nombre d'ARNm ou d'ARNnc dans l'ensemble du génome et qui sont impliquées dans des mécanismes de surveillance cytoplasmique. J'ai également participé au développement des outils qui permettent de distinguer les sous populations d'ARNs associés à des sous-complexes spécifiques.

Notre laboratoire s'intéresse depuis plusieurs années aux complexes impliqués dans les mécanismes de surveillance cytoplasmique. En effet, les équipes du laboratoire ont caractérisé plusieurs des complexes associés aux mécanismes de contrôle qualité des ARNs. L'équipe de Micheline Fromont a participé à la découverte des facteurs clés du Non Stop Decay (NSD) et celle de Cosmin Saveanu a caractérisé récemment deux sous-complexes distincts associés à UPF1, l'hélicase majeure du Nonsense Mediated mRNA Decay (NMD). Plus globalement, je m'intéresse aux interactions ARN-protéines dans les mécanismes de contrôle qualité des ARNs, qui impliquent généralement l'action d'hélicases particulières au sein de complexes ribonucléoprotéiques dont le rôle précis dans les mécanismes de contrôle qualité des ARN n'est pas bien compris à ce jour. Je propose d'essayer de répondre à cette question par des approches génomiques pour caractériser les spécificités de reconnaissance de certaines de ces hélicases au sein de leurs complexes protéiques respectifs, et d'essayer de comprendre les mécanismes mis en jeu au sein des sous-complexes. Pour cela, j'utilise une version améliorée du protocole du CRAC pour étudier des facteurs qui lient des ARN cellulaires modérément abondant. En couplant cette approche à des expériences de purifications biochimiques et de transcriptomique, nous disposons d'outils puisant pour l'analyse fonctionnelle de ces gènes.
Le premier facteur en cours d'étude est l'hélicase putative que nous avons baptisée TAC4 (translation associated component 4) car il s'associe à la petite sous unité du ribosome. Les résultats préliminaires suggèrent que Tac4 pourrait intervenir dans le recyclage des ribosomes lorsque ceux-ci sont bloqués sur les ARNm lors des mécanismes de NGD (« No Go Decay ») ou de NSD. En collaboration avec Micheline

Fromont et Olivier Namy, nous nous proposons de caractériser le rôle de Tac4 dans ce mécanisme de recyclage.

Le deuxième facteur étudié est le gène SKI2 , l'hélicase majeure du complexe SKI, dont l'équipe de Micheline Fromont au laboratoire, en collaboration avec Roland Beckmann et Elena Conti, a récemment trouvé qu'il pouvait être associé au ribosome pour dégrader les parties traduites des ARNs. En parallèle, elle a aussi observé qu'une autre forme du complexe SKI mais incapable de s'associer au ribosome pouvait être impliquée dans la dégradation des régions non traduites des ARNm, et que ce sous complexes fait intervenir la protéine baptisée Ska1 (Ski associated component 1). Connaître les spécificités de reconnaissance des différents sous-complexes associés au complexe SKI devrait permettre de mieux comprendre le rôle de Ska1.

Enfin, le troisième facteur est Upf1, l'hélicase majeure du NMD, qui est trouvé dans deux sous-complexes : un complexe « détecteur » et un complexe « effecteur ». Afin d'essayer de comprendre de quelle manière Upf1 est associé à ces deux sous complexes, j'ai réalisé des expériences préliminaires de RIPseq et de CRAC avec Upf1 et démarré, à l'occasion de l'encadrement de Léna Audebert lors de son stage de Master 2, des expériences de « split-CRAC » (ou les deux étiquettes utilisées sont sur des protéines différentes ce qui permet de sélectionner des sous complexes) dans plusieurs contextes associés à Upf1. Les résultats préliminaires indiquent qu'Upf1 se lie préférentiellement dans les 3'UTR des transcrits et de façon plus modérée aux extrémités 5'. De manière très surprenante, en plus de trouver les cibles attendues du NMD, Upf1 se lie à un grand nombre de transcrits « non-NMD », et Léna Audebert a commencé en octobre 2018 une thèse que je co-dirige avec Cosmin Saveanu et dont le sujet consiste à comprendre le rôle joué par Upf1 et ses partenaires dans la régulation post- transcriptionnelle de ces substrats.

Précisément, je concentre donc actuellement mon intérêt principalement sur la compréhension des mécanismes qui régulent les variations subtiles de stabilité des différentes isoformes des queues poly(A) des transcrits en utilisant des approches transcriptomiques et biochimiques. Compte tenu de la convergence de nos sujets de recherche et de la complémentarité de nos expertises, j'ai naturellement rejoint en septembre 2018 le groupe de Cosmin Saveanu.

CURRICULUM VITAE

BADIS-BREARD Gwenael 43 ans. Nationalité Française. Mariée 3 enfants.

Adresse personnelle: 5 rue Jacques Brel, 95240 Cormeilles en Parisis

Adresse professionnelle:

Unité de Génétique des Interactions Macromoléculaires
CNRS URA 3525
Institut Pasteur
25, Rue du Dr Roux
75724 Paris Cedex 15
Tel: 01 40 61 33 31
e.mail: gbreard@pasteur.fr

DIPLOMES

2003 – Doctorat de Biochimie et Biologie Moléculaire, option Microbiologie
1999 – DEA de Microbiologie, option Microbiologie générale
1998 – Maîtrise de Biologie Cellulaire et Physiologie, option Génétique Moléculaire et cellulaire.
1992 - Baccalauréat série D

EXPERIENCE PROFESSIONNELLE

2009- présent	Chargé de recherche CR1 CNRS (CRCN depuis septembre 2017). Unité de génétique des interactions macromoléculaires (UMR3525 Alain Jacquier). Institut Pasteur Paris
2008-2009	Chargé de recherche contractuel. Unité de génétique des interactions macromoléculaires (Alain Jacquier –Bourse Roux). Institut Pasteur Paris
2004-2008	*Post doctorat*. Terrence Donnelly CCBR (Timothy Hughes) Thème : Déchiffrage du "Regulome" de la souris et de la levure. (Bourse CIHR-IRSC Juillet 2005-Déc 2008. Bourse Best Institute Juillet 2004-Juin 2005)
1999-2003	*Doctorat*. Unité de génétique des interactions macromoléculaires. Institut Pasteur (Alain Jacquier). Thème : régulation des ARNs et génétique chez *S. cerevisiae*.
1998-1999	Unité de génétique moléculaire des levures. Institut Pasteur (Bernard Dujon). Thème : Analyse fonctionnelle du génome de *S. cerevisiae*.
1993-1997	Agent contractuel. Informatisation du restaurant administratif du campus de Jussieu. Université Paris VI Paris

FORMATION

2012 Atelier 215 INSERM diversité des transcriptomes non codants révélés par RNA–seq (3 jours)
2014 Notions Fondamentales en statistiques (3 jours)
2014 Introduction au logiciel R pour les statistiques (3 jours)
2015 Aide à la rédaction de projets scientifiques en anglais (2 jours)
2015 Présentation orale de projets scientifiques en anglais (2 jours)

ACTIVITES D'ENCADREMENT DE LA RECHERCHE

1- Encadrement de stage (1ere ou 2eme année d'études universitaires)

2004 Ginny Costanzo, étudiante en 2ème année d'IUT (2 mois) – *1 publication*
2005-2006 Agatha Cheung, Melissa Chan et Sacha Bhinder (10 mois chacun)
2006-2007 Dimitri Terterov et Faïqua Khalid(10 mois chacun)
2007-2008 David Coburn et Rochelle Goldstein (6 mois chacun)

3- Encadrement de stages de Master

2005-2007 Shaheynoor Talukder (2 ans) – *7 publications*
2018 Léna Audebert (6 mois) – *Obtention d'une bourse de thèse*

4- Encadrement de thèse

2014-2017 Alicia Chery-Faleme (3 ans) – *1 publication – co encadrement avec Alain Jacquier*
2018-présent Léna Audebert – *co-encadrement avec Cosmin Saveanu*

5- Encadrement de technicien et bio-informaticien

2004-2005 Frédéric Bréard, bio-informaticien (1 an)
2005-2006 Sanie Mnaimneh, technicienne (2 ans)
Paul Qureshi, technicien (1 an)
2014-2018 Antonia Doyen, technicienne (4 ans)

ACTIVITES D'ENSEIGNEMENT

11 décembre 2008 Cours UE M2 Criblage Génomique / Génomique Fonctionnelle, Paris XI Orsay (3 heures). « Interactions acides nucléiques-protéines et régulation de l'expression des gènes chez les eucaryotes »

12 novembre 2009 Cours UE M2 Criblage Génomique / Génomique Fonctionnelle, Paris XI Orsay (3 heures). « Interactions acides nucléiques-protéines et régulation de l'expression des gènes chez les eucaryotes »

17 décembre 2009 Cours Pasteur Génétique cellulaire et moléculaire" (2 heures) « Deciphering the semantics of eukaryotic regulome »

01 mars 2011 Cours Pasteur Multiple Roles of RNAs (2 heures) : Protein-nucleic acid interactions and information it conveys :
Part I - DNA-protein interactions to understand transcription regulation
Part II- RNA-protein interactions to decipher RNA metabolism

2011 à 2012 Cours Pasteur Multiple Roles of RNAs (70 heures/an– 2 semaines): Encadrement du cours (public: étudiants en Master ou doctorants)

Depuis 2013 Cours Pasteur Multiple Roles of RNAs (70 heures/an – 2 semaines): Co-direction du cours (public: étudiants en Master ou doctorants)

J'ai pris la co-direction du cours Pasteur « Roles Multiple des ARNs » à la suite de Micheline Fromont-Racine qui avait monté ce cours en 2009 et que j'avais rejoint en 2011 pour encadrer ce cours. C'est un cours pratique de deux semaines qui permet aux étudiants d'apprendre quelques rudiments du travail avec les ARNs : RNAseq, RT-qPCR, et Northern blot, à travers l'études des ARNs ciblés par le Nonsense mediated mRNA decay. Depuis 2016, j'ai fait évoluer ce cours en implémentant des expériences de RNAseq (préparation des librairies et analyse bio-informatique des séquences) et en changeant la thématique (anciennement les CUTs). J'ai coordonné l'organisation des interventions notamment des 9 experts qui animent ce cours et en particulier la réalisation et l'analyse des expériences de RNAseq.

RESPONSABILITE ADMINISTRATIVES ET EXPERTISES

- Depuis 2013: Correspondante hygiène et sécurité de l'unité
- Expert pour des appels à projet (Réseau International des Instituts Pasteur)
- Depuis 2015: Participation aux comités de thèse de 3 étudiants en tant qu'expert scientifique (Hilal Yieter, Drice Challal et Jian Bai).

AUTRES RECOMPENSES

2007 Lauréate du programme « Initiative Post-doc ».

SEMINAIRES

2002 SifrARN (SFBBM), Nancy, October 14-17. « A new mechanism of post-transcriptional auto-regulation that recruits the decapping machinery ».

2003 Levure modèle et outil VI, Geneva,Swiss, April 7-9. « New tools to perform synthetic lethal screen in yeast ».

2004 Club Levure Ile de France (CLIF), Paris, France, January 21. « Targeted mRNA degradation by deadenylation-independent decapping»

2005 Departemental seminar. CCBR, Toronto, Canada, October 5th. « Exploring the Yeast DNA-Protein Interactome »

Departemental seminar Structure et Dynamique des Genomes, Institut Pasteur, Paris, France, December 1st. « A biochemical toolkit for decoding the mammalian regulome »

2016 10e congrès du SifrARN (SFBBM) 9 mars. « Gene regulation by antisense-mediated transcriptional interference »

2017 Club noyau d'ile de France. 28 juin. « Antisense transcriptional interference mediates tight gene repression in budding yeast »

POSTERS

1997 Levure modèle et outil IV, Arcachon, France November. G Badis , J Boyer, and O Ozier-Kalogeropoulos; « Functional analysis of Saccharomyces cerevisiæ genome: analysis of dominant negative gene by overexpression.

2001 Levure modèle et outil V, Brussel, Belgium April 7-9. G Badis and A Jacquier; « Analyse fonctionnelle de deux gènes reliés dans des réseaux de double hybride au métabolisme des ARN ».

2002 7th annual meeting of the RNA society, Madison, USA, May 28th - June 2nd. G Badis, M Fromont-Racine and A Jacquier; « A new Saccharomyces cerevisiae H/ACA snoRNA that guides universally conserved pseudouridylation in 25S rRNA ».

2003 8th annual meeting of the RNA society, Vienna, Austria, July 1-6. G Badis, M Fromont Racine and A Jacquier. « A new mechanism of post-transcriptional auto-regulation by recruitment of the decapping machinery »

2005 System Biology: Global regulation of gene expression, Cold Spring Harbor, USA. March 17-20. G Badis, S Talukder, E Chan, S Bhinder, A Cheung, M Chan, S Mnaimneh, F Bréard, TR Hughes." A biochemical toolkit for decoding the mammalian regulome».

Chromatin and transcription, FASEB 2005, Snowmass Colorado, USA, July 9-14. G Badis, S Talukder, S Mnaimneh, S Bhinder, A Cheung, M Chan, S Agnihotri, E Chan, F Bréard and TR Hughes. "A Biochemical toolkit for decoding the Mammalian regulome"

Gene regulation: from chromatin remodelling to transcription, Paris, France. November 25-26. G Badis, S Talukder, E Chan, S Bhinder, A Cheung, M Chan, S Mnaimneh, F Bréard, TR Hughes." A biochemical toolkit for decoding the mammalian regulome».

2006 System Biology: Global regulation of gene expression, Cold Spring Harbor NY, USA March 23-26. T Hughes, G Badis, M Berger, S Talukder, A Philippakis, E Chan, S Mnaimneh, P Qureshi, A Gehrke, and M Bulyk. « Toward explaining global regulation of vertebrate gene expression »

System Biology: Global regulation of gene expression, Cold Spring Harbor NY, USA March 23-26. S Talukder, G Badis, M Berger, A Philippakis, L Pena-Castillo, E Chan, A Gehrke, M Bulyk and T Hughes « Seeking new cis-regulatory regions via the DNA-binding specificity of tissue specific transcription factors »

ORFeome meeting 2006, Boston, USA. November 15-18. G Badis, S Talukder, M Berger, A Gehrke, A Philippakis, X Chen, D Coburn, J Holroyd, F Bréard, S Mnaimneh, QD Morris, MD Bulyk and TR Hughes. «Toward decoding the mammalian regulome».

2007 CIFAR meeting, 25th anniversary, Seattle, USA. September 7-8. G Badis. «Deciphering the semantics of eukaryotic regulome».

2014 EMBO conference Gene transcription in yeast. San Feliu Spain. June14-19. Transcriptional interference in regulation of quiescence specific genes in yeast. Gwenael Badis, Antonia Doyen and Alain Jacquier

2017 TERM meeting, San Feliu Spain 26-28 Apr. Antisense transcriptional interference mediates tight gene repression in budding yeast. Nevers A , Doyen A, Malabat C, Jacquier A, Badis G

EMBO meeting Eukaryotic mRNA turnover, Oxford, UK. 10-13 July. Antisense transcriptional interference mediates tight gene repression in budding yeast. Nevers A , Doyen A, Malabat C, Jacquier A, Badis G

PRINCIPALES SOURCES DE FINANCEMENT

ANR CLEANMD (2014 - porteur de projet Alain Jacquier)
ANR SKA (2017- porteur de projet Micheline Fromont-Racine)
ANR RIBO RESCUE (2017 - porteur de projet Micheline Fromont-Racine)
ANR DEFineNMD (2018 - porteur de projet Cosmin Saveanu)

COLLABORATIONS NATIONALES ET INTERNATIONALES

-Frédéric DEVAUX, UPMC Sorbonne Université, Paris, France
Publications afférentes à cette collaboration : ***Publications n°4 et n°5***

- Bertrand SERAPHIN, IGBMC, Illkirch, France
Publications afférentes à cette collaboration : ***Publications n°4***

- Domenico LIBRI, Institut Jacques Monod, Paris, France
Publication afférente à cette collaboration : ***Publication n°4***

- Olivier GADAL, LBME UMR5099, Toulouse, France
Publication afférente à cette collaboration : ***Publication n°6***

- Andrew EMILI, University of Toronto, Toronto, Canada
Publication afférente à cette collaboration : ***Publication n°7***

- Corey NISLOW, University of Toronto, Toronto, Canada
Publication afférente à cette collaboration : ***Publication n°9***

- Lourdes PENA-CASTILLO, Memorial University of Newfoundland, St. John's, Canada
Publications afférentes à cette collaboration : ***Publication n°8,9 et 16***

- Wyeth WASSERMAN, Centre for Molecular Medicine and Therapeutics, Vancouver, Canada
Publication afférente à cette collaboration : ***Publication n°11***

- Rob SLADEK, McGill University, Montréal, Canada
Publication afférente à cette collaboration : ***Publication n°11***

- Marta BULYK, Harvard Medical School, Boston, USA
Publications afférentes à cette collaboration : ***Publication n°8, 10, 12, 13 et 14***

- Jussi Taïpale, Karolinska Institutet, Stockholm, Suède
Publication afférente à cette collaboration : ***Publication n° 14***

- Jesus De la Cruz, Instituto de Biomedicina de Sevilla (IBiS), Séville,
Publication afférente à cette collaboration : ***Publication n° 15***

Collaborations en cours

- Olivier NAMY, I2BC, Université Paris Sud, Orsay, France
  Financement afférent à cette collaboration : ANR RIBO RESCUE

- Roland BECKMANN, Gene Center and Department of Biochemistry Ludwig-Maximilians-Universität München, Munich, ALLEMAGNE
  Financement afférent à cette collaboration : ANR SKA

- Hervé Le Hir, IBENS, Ecole Normale Supérieure, Paris , France.
  Financements afférents à cette collaboration : ANR CLEANMD et DEFineNMD

LISTE DES PUBLICATIONS

17 publications (incluant 1 chapitre de livre en tant que dernier auteur, 6 en tant que premier ou co-premier auteur, 1 comme auteur correspondant), h index : 14 (nombre moyen de citation par article = 141)

2003

1) **Badis G**, Fromont-Racine M and Jacquier A. A snoRNA that guides the two most conserved pseudouridine modifications within rRNA confers a growth advantage in yeast, ***RNA*** Jul;9(7):771-779.

Ce premier article de ma thèse à consisté en la caractérisation du snoRNA à boite H/ACA caché dans l'intron de NOG2, gène que Micheline étudiait. J'ai réalisé toutes les expériences montrées dans ce manuscrit (excepté le premier northern blot), la plupart des analyses et j'ai participé à l 'écriture.

2004

2) **Badis G**, Saveanu C, Fromont-Racine M and Jacquier A. Targeted mRNA degradation by deadenylation-independent decapping, ***Mol Cell***. Jul 2;15(1):5-1

Dans cet article nous décrivons un mécanisme d'autorégulation de l'ARNm RPS28B par le recrutement de la machinerie de deccaping/dégradation en 5' via une épingle située dans le 3'UTR du mRNA et reconnue par Rps28b. J'ai préparé et réalisé toutes les expériences (excepté la construction d'une souche), réalisé les analyses et participé à l'écriture du manuscrit.

3) Boyer J, **Badis G**, Hantraye F, Talla E, Koszul R, Fairheaid C, Hennequin C, Ozier-Kalogeropoulos O, Lafontaine I, Fabre E, Fischer G, Ricchetti M, Richard GF, Thierry A & Dujon, B. Large-scale exploration of overexpression-dependent dominant negative phenotypes in *Saccharomyces cerevisiæ*, ***Genome Biology***, 5(9):R72

Cet article de ressource est un crible génétique pour rechercher les gènes et fragments de gènes dont l'expression est toxique en surexpression. J'ai initié ce projet durant un stage d'un an (à mi-temps) en maitrise (construction de la banque et criblage d'environ 10 000 clones et identification par séquençage de 126 clones sélectionnés). Jeanne Boyer a repris ce projet après mon départ du laboratoire de Bernard Dujon en 1999.

2005

4) Wyers F, Rougemaille M, **Badis G**, Rousselle JC, Dufour ME, Boulay J, Devaux F, Namane A, Libri D, Séraphin B & Jacquier A. Cryptic Pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase, ***Cell***, Jun 3;121(5):725-737

Ma contribution dans cet article est modeste néanmoins importante. C'était à la fin de ma thèse, quand j'ai caractérisé par hasard ce qui allait devenir les premiers « Cryptic Unstable Transcripts » (CUTs), en analysant les résultats de puces Affymetrix qui contenaient par chance des primers « contrôles » en dehors des phases codantes (dans des intergènes notamment). L'étude du transcriptome dans une souche rrp6Δ comparée à une souche sauvage m'a permis de caractériser un certain nombre de régions non codantes qui contenaient un transcrit stabilisé dans la souche Δrrp6 (quand l'exosome nucléaire était inactivé). Devant partir en post-doc, j'ai transmis mes résultats préliminaires et mes tentatives de northerns blots laborieux à Mathieu Rougemaille et Domenico Libri, et en

collaboration avec B. Séraphin nos trois équipes ont découvert l'existence des CUTs qui a donné ce bel article extrêmement cité.

5) Torchet C, **Badis G**, Devaux F, Costanzo G, Werner M & Jacquier A. The complete set of H/ACA snoRNAs that guide rRNA pseudouridylations in *Saccharomyces cerevisiae*, ***RNA***, Jun;11(6):928-938.

Après le travail sur snr191, nous avons décidé de caractériser le set complet de snoRNA à boite H/ACA. Ce travail a été initié par Claire Torchet qui a réalisé l'immunoprécipitation de Gar1 et Nhp2 et qui a construit la plupart des souches. Avec la participation de Giny Costanzo que j'encadrai, nous avons ensuite caractérisé la plupart des positions de l'ARNr (dont 14 totalement nouvelles) dont la pseudouridylation est guidée par chacun des 28 snoRNAs de la levure.

2007

6) Berger AB, Decourty L, **Badis G**, Nehrbass U, Jacquier A, Gadal O. (2007) Hmo1 Is Required for TOR-Dependent Regulation of Ribosomal Protein Gene Transcription, ***Mol Cell Biol***, Nov;27(22):8015-26.

Dans cet article, j'ai construit une partie des outils qui ont permis de réaliser les premiers cribles de létalité synthétique avec la banque systématique de délétion (les plasmides pGID1 et pGID2).

2008

7) Sandhu C* Hewel JA*, **Badis G,** Talukder S, Hughes TR and Emili A . Evaluation of data-dependent versus targeted shotgun proteomic approaches for monitoring transcription factor expression in breast cancer. ***J Proteome Res***. Apr;7(4):1529-41.

Dans cet article ma contribution y a été mineure. J'ai réalisé le clonage, l'expression et la purification dans E. coli du domaine de liaison à l'ADN du facteur NF-κB2 et Stat1 en utilisant la stratégie du projet « facteur de transcription de souris ».

8) Berger M*, **Badis G***, Gehrke A*, Talukder S*, Philippakis S, Pena-Castillo L, Alleyne TM, Mnaimneh S, Jaeger S, Chan E, Botvinnik O, Khalid F, Zhang W, Morris QD, Bulyk MD and Hughes TR. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences, ***Cell*** Jun27; 133(7):1266-76

Premier article majeur de mon post-doctorat dans lequel nous avons initié le projet en collaboration avec le laboratoire de Martha Bulyk et réalisé les premières identifications avec les facteurs de transcription appartenant à la famille des facteurs de transcription à Homéodomaines. . J'ai conçu et mis en œuvre la stratégie pour cloner les ≈ 168 facteurs de transcription de souris appartenant à cette famille (vérifier par séquençage et purifier les protéines recombinantes correspondantes). J'ai également encadré les 4 stagiaires, 2 techniciens et 1 bio-informaticien qui ont participé à ces étapes. Les expériences de PBM ont été réalisées par le laboratoire de M. Bulyk à Harvard et les analyses informatiques par les 4 bio-informaticiens de nos deux équipes. J'ai aussi participé à l'analyse des résultats et à l'écriture de l'article.

9) **Badis G**, Chan E, van Bakel H, Peña-Castillo L, Tillo D, Tsui K, Warren C, Gossett A, Xuan YZ, Carlson C, Gebbia M, Hasinoff M, Talukder S, Yang A, Mnaimneh S, Terterov D, Coburn D Clarke N, Lieb J, Ansari A, Nislow C and Hughes TR. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. ***Mol Cell.*** 2008 Dec 26;32(6):878-87.

Cet article est issu du projet sur les facteurs de transcriptions de souris. Nous avons décidé que puisque nous avions été capable d'identifier près d'un millier de motifs pour les souris, il serait aisé de le faire pour les ≈ 200 facteurs de transcription de levure. Contrairement au projet murin où nous collaborions avec le laboratoire de M. Bulyk, nous sommes compétiteurs sur ce projet. J'ai ainsi conçu et orchestré la réalisation de toute la partie sur l'identification des sites de liaisons à l'ADN de 112 facteurs de transcription de levure avec l'aide de deux stagiaires et une technicienne (clonage, purification des protéines recombinantes) et j'ai réalisé moi même toutes les expériences de protein-binding microarray. J'ai également participé à l'analyse des résultats (avec l'aide de 4 bioinformaticiens), la fabrication des figures et à l'écriture de l'article.

2009

10) Alleyne TM, Peña-Castillo L, **Badis G**, Talukder S, Berger MF, Gehrke AR, Philippakis A, Bulyk ML Morris Q and Hughes TR. Predicting the Binding Preference of Transcription Factors to Individual DNA k-mers. ***Bioinformatics***. 2009 Apr 15;25(8):1012-8. doi: 10.1093/bioinformatics/btn645.

Cet article de bioinformatique présente un outil pour prédire les motifs associés aux issus des expériences de PBM. J'ai fourni les résultats d'expériences de PBM et participé à la réflexion et à l'écriture de l'article.

11) Fulton DL, Sundararajan S, **Badis G**, Hughes TR, Wasserman WW, Roach JC and Sladek R. TFCat: The curated catalog OF Mouse and Human transcription factors. ***Genome Biology*** Mar 12;10(3):R29 doi: 10.1186/gb-2009-10-3-r29.

Dans cet article nous nous sommes organisés avec plusieurs laboratoires afin d'établir un catalogue complet et normalisé des facteurs de transcriptions humains et murins. Nous nous sommes répartis chacun une liste de facteurs pour lesquels nous avons vérifié dans la littérature chacun des paramètres permettant de le définir comme un facteur de transcription, de quelle famille, catégorie, etc.

12) **Badis G***, Berger M*, Philippakis A*, Talukder S*, Gehrke A*, Jaeger S*, Chan E*, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger D, Morris QD, HughesTR and Bulyk ML. Diversity and Complexity in DNA Recognition by Transcription Factors. ***Science*** 2009 Jun 26;324(5935):1720-3. doi: 10.1126/science.1162327.

Cet article est le troisième article majeur de mon post-doctorat. J'ai conçu et mis en œuvre la stratégie pour cloner les ≈ 1500 facteurs de transcription de souris appartenant à toutes les familles de domaines (vérifier par séquençage et purifier les protéines recombinantes correspondantes). J'ai également encadré les 4 stagiaires, 2 techniciens et 1 bio-informaticien qui ont participé à ces étapes. Les expériences de PBM ont été réalisées par le laboratoire de M. Bulyk à Harvard et les analyses informatiques par les 4 bio-informaticiens de nos deux équipes. J'ai également participé à l'analyse des résultats, la fabrication des figures et à l'écriture de l'article.

2010

13) Santos MA, Turinsky AL, Ong S, Tsai J, Berger MF, **Badis G**, Talukder S, Gehrke AR, Bulyk ML, Hughes TR, Wodak SJ. Objective sequence-based subfamily classifications of mouse homeodomains reflect their in vitro DNA-binding preferences. ***Nucleic Acids Res***. 2010 Dec;38(22):7927-42. doi: 10.1093/nar/gkq714.

Cet article est issu de celui sur les facteurs de transcription à homéodomaines. Je n'ai fait que fournir les données de PBM.

14) Wei GH, **Badis G**, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, Yan J, Talukder S, Turunen M, Taipale M, Stunnenberg HG, Ukkonen E, Hughes TR, Bulyk ML, Taipale J. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. ***EMBO*** J. 2010 Jul 7;29(13):2147-60. doi: 10.1038/emboj.2010.106.

Cet article est issu du projet sur les facteurs de transcriptions de souris. Ici nous avons collaboré avec Jussi Taïpale pour caractériser la spécificité des facteurs de transcription de la famille ETS. Ma contribution y a été mineure.

2013

15) Babiano R*, **Badis G***, Saveanu C, Namane A, Doyen A, Díaz-Quintana A, Jacquier A, Fromont-Racine M, de la Cruz J. Yeast ribosomal protein L7 and its homologue Rlp7 are simultaneously present at distinct sites on pre-60S ribosomal particles. ***Nucleic Acids Res***. 2013 Nov;41(20):9461-70. doi: 10.1093/nar/gkt726.

Pour cet article j'ai réalisé les expériences de CRAC qui ont permis de définir de manière déterminante les sites de liaison des protéines Rlp7 et Rpl7. J'ai également formé Reyes Babiano à cette technique. Enfin, j'ai participé à l'analyse des résultats et à l'écriture de l'article.

2016

16) Peña-Castillo L, **Badis G**. Systematic Determination of Transcription Factor DNA-Binding Specificities in Yeast***. Methods Mol Biol***. 2016;1361:203-25. doi: 10.1007/978-1-4939-3079-1_12.

Pour ce chapitre de livre, nous avons souhaité expliquer la méthode de « Protein Binding Microarray » que j'ai utilisé en post-doc pour identifier les spécificités de liaison à l'ADN des facteurs de transcription. J'ai réalisé toute l'écriture de ce chapitre, mis à part la partie bio-informatique qui a été écrite par Lourdes Peña-Castillo.

2018

17) Nevers A, Doyen A, Malabat C, Néron B, Kergrohen T, Jacquier A* and **Badis G***. (2017) Antisense transcriptional interference mediates condition-specific gene repression in budding yeast. ***Nucleic Acids Res***. 2018, May 18 https://doi.org/10.1093/nar/gky342. (*BioRxiv 2017* 169730; doi: https://doi.org/10.1101/169730)

Cet article est le résultat du projet que j'ai initié en 2012, puis sur lequel Alicia Nevers et Antonia Doyen m'ont rejoint en 2014. Dans cet article nous observons à grande échelle et validons que la transcription en antisens des promoteurs des gènes a un effet répresseur sur la transcription de ces gènes et que cet effet peut être réciproque. J'ai conçu et participé à la réalisation des expériences, analysé les résultats et écrit l'article. C'est le premier article « peer review » que je signe en tant qu'auteur correspondant.

* co-premier auteur
** les étudiants que j'ai encadrés sont soulignés.

RAPPORT DE RECHERCHE :

Interaction entre protéines et acides nucléiques et régulation de l'expression des gènes

I Etude du métabolisme des ARN chez *S. cerevisiae* : analyse fonctionnelle et caractérisation d'ARNs non codants.

Thèse : Décembre 1999-juin 2004

J'ai débuté ma thèse à la fin de l'année 1999, quelques années après le séquençage complet du génome de la levure (Goffeau et al., 1996) . A cette époque, plus de la moitié des gènes de cet organisme avaient encore des fonctions inconnues et la génétique moléculaire et la biochimie étaient en plein essor, avec le développement d'outils comme le TAP-tag (Rigaut et al., 1999) ouvrant de nouvelles possibilités pour étudier les complexes protéiques. C'est aussi à cette époque que les notions de génomique émergent, et le domaine de la génomique fonctionnelle, avec le développement d'une multitude de collections de mutants systématiques chez la levure (délétion, TAP, GFP, …) et les premiers articles associés.

Un terrain vierge de 3000 gènes de fonctions inconnues vient de s 'ouvrir à nous, et encore, il ne s'agit que des gènes codants les protéines. Des efforts importants dans les laboratoires travaillant avec la levure sont menés pour caractériser les fonctions de ces gènes. Nous apprendrons plus tard qu'aux 6000 gènes de levure codant des protéines il faut ajouter un degré de complexité supplémentaire constitué par les ARNs non codants aux fonctions également inconnues.

C'est dans ce contexte que je débute ma thèse dans le laboratoire d'Alain Jacquier qui s'intéressait au métabolisme des ARNs de façon globale. En plus de vouloir caractériser la fonction de quelques gènes inconnus, nous nous rendons vite compte que la plupart des nombreux outils de génétique des levures qui ont été développés depuis des décennies pour réaliser l'analyse fonctionnelle des gènes et qui ont permis le défrichage d'un grand nombre de voie métaboliques ne sont pas adaptés à des études à grande échelle.

Développement technologique d'outils pour l'analyse fonctionnelle des gènes de levure

En effet, il n'existe pas encore par exemple de système assez rapide et performant pour réaliser des cribles de létalité synthétique chez *S. cerevisiæ* dans un temps raisonnable. Afin de palier ce manque, j'ai testé les systèmes classiques de cribles de

létalité synthétique et participé au développement du système "GID" (pour Genetic Interaction of Deletions), permettant de réaliser des cribles de létalité synthétique de façon efficace et relativement exhaustive pour les gènes non essentiels car il utilise les mutants de la banque systématique de délétion (Winzeler et al., 1999). Ce système présente l'avantage de pouvoir utiliser également la mutagenèse classique pour cibler les gènes essentiels. Ce système a notamment été utilisé dans un travail en collaboration avec le groupe d'Olivier Gadal (Berger et al., 2007) et a ensuite été amélioré par plusieurs membres du laboratoire après mon départ en post doctorat pour aboutir à une technique plus performante (Decourty et al., 2008).

Comprendre les interactions ARNs/protéine et leur rôle : le cas d'EDC3.

Par la caractérisation de réseaux d'interactions protéique à l'aide de cribles double – hybride itératifs, Micheline venait de relier YEL015w, un gène de fonction inconnue, aux protéines Lsm et à la machinerie de dégradation et de decapping (Fromont-Racine et al., 2000). Ce gène a été baptisé Edc3 pour « enhancer of decapping ». J'ai pu caractériser une de ses fonctions grâce à l'utilisation de puces Affymetrix qui m'ont permis de trouver une cible privilégiée d'Edc3, l'ARNm de *RPS28B*, codant une des copies de la protéine ribosomale Rps28. J'ai en effet mis en évidence un nouveau mécanisme d'autorégulation post-transcriptionnelle médié par Edc3 et par un élément ARN en *cis* dans le messager de *RPS28B*. Dans ce mécanisme de régulation, j'ai montré que c'est le décoiffage qui est l'étape limitante et régulée dans la dégradation, et non la déadénylation comme c'est le cas pour la plupart des ARNm (Beelman and Parker, 1995 pour revue). D'autres données expérimentales suggèrent qu'Edc3 pourrait intervenir sur d'autres cibles et être impliquée à un niveau plus général dans la dégradation des ARNm (Badis et al., 2004).

Dans ce travail, j'ai utilisé le triple hybride pour montrer la liaison (indirecte) entre Rps28 et l'épingle dans le 3'UTR de son messager. Plus récemment, notre modèle de 2004 a été révisé dans une étude qui montre que c'est Edc3 qui lie directement l'épingle dans le 3' UTR de *RPS28B* et non Rps28, sous la forme d'un dimère assemblé avec Rps28p en excès et donc libre dans le cytoplasme (Figure 1 et He et al., 2014).

De plus, il a été mis en évidence depuis que l'homologue humain d'Edc3 est une composante du complexe de decapping (Arribas-Layton et al., 2013 pour revue).

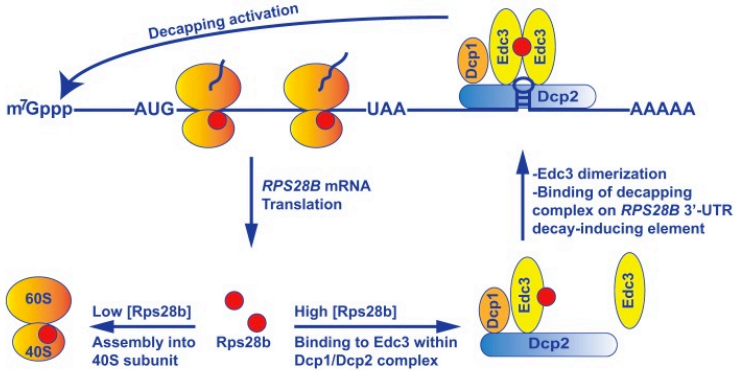


Figure 1 : modèle révisé de la dégradation du mRNA de RPS28B par Edc3 (extrait de He et al., 2014).

Ce premier volet de ma thèse m'a permis de mettre en évidence un rôle de la partie non codante d'un ARNm (ici une épingle dans le 3'UTR), dans un mécanisme d'autorégulation post-transcriptionnelle.

En parallèle, nous nous intéressions aux ARNs non codants en général, et si la connaissance du génome de la levure venait de dévoiler la plupart des gènes codant les protéines que l'on connaît actuellement, il restait encore beaucoup de choses à découvrir concernant les ARNs non codants et leurs fonctions.

Découverte de nouveaux ARN non codants à l'ère post génome de la levure.

L'autre sujet de recherche que j'ai développé pendant ma thèse a ainsi porté sur la recherche d'ARN non codants chez *S. cerevisiæ*.

En 2000, les découvertes foisonnent en matière d'ARNnc, et dans tous les organismes une grande diversité d'ARNnc est révélée : snmRNA (small non messenger RNA), sRNA (small RNA), fRNA (functional RNA), miRNA (microRNA), siRNA (small interfering RNA), stRNA (small temporal RNA), etc...

Curieusement, la levure *S cerevisiae* reste relativement « silencieuse » et seuls quelques nouveaux ARNnc sont identifiés : 9 « RNA of unknown function » ou RUFs (McCutcheon and Eddy, 2003) découverts par une approche phylogénétique et 15 ARN non codants mais « non répertoriés » sont cités dans (Olivas et al., 1997). Ces derniers ont été trouvés en explorant les intergènes « vides d'ORF » dans le génome.

S. cerevisiae ne possède pas la machinerie du RNAi et les outils de transcriptomique de l'époque ne permettent pas encore de regarder chaque position du génome de façon exhaustive.

Il est frappant de constater avec le recul d'aujourd'hui comment la transcriptomique a évolué, et bien que selon les méthodes de fabrication des banques il reste toujours des biais expérimentaux qui filtrent un certain nombre d'éléments, les techniques de séquençages actuelles ont un spectre beaucoup moins restreint et une profondeur de lecture incomparable à celle des années 2000.

Néanmoins, c'est en utilisant des puces Affymetrix que nous ferons les découvertes d'ARNs non codants les plus intéressantes : de nouveaux snoRNAs mais surtout les « Cryptic Unstables Transcripts » (CUTs).

Par un heureux hasard en effet, la technologie des puces Affymetrix qui consiste à sonder le génome avec une série d'oligonucléotides pour chaque gène comporte aussi des sondes dans des régions spécifiques de certains ARNnc (ARNr, ARNt, snoRNA et snRNA en particulier), ainsi qu'un bon nombre de sondes situées dans les grandes régions intergéniques (>1Kb) et dans les régions intergéniques suspectées de contenir des transcrits (SAGE, pour « Serial Analysis of Gene Expression » décrits dans (Velculescu et al., 1997).

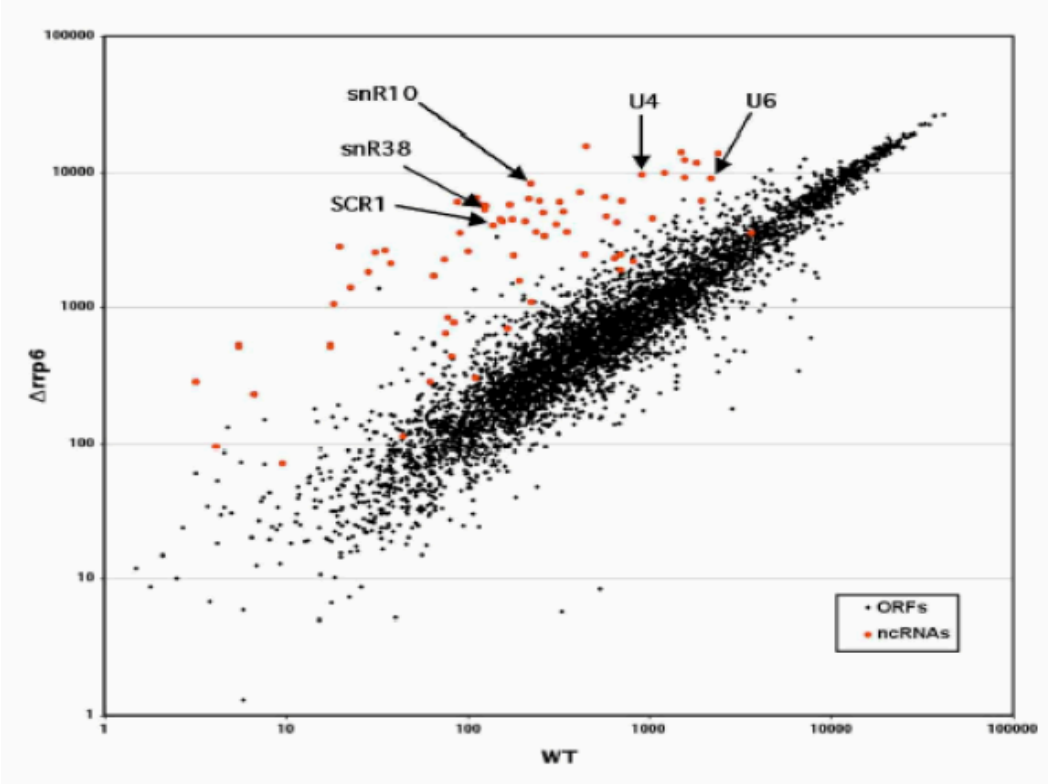


Figure 2 : Comparaison des transcriptomes d'une souche sauvage et Δrrp6.

Les ARN totaux de souches sauvage et *Δrrp6* ont été utilisés pour préparer les ADNc marqués avec des oligonucléotides poly-dT. Les points rouges correspondent aux signaux des ARNnc tandis que les points noirs correspondent aux transcrits des ORFs. Certains ARNnc connus sont montrés à titre d'exemple, illustrant comment les ARNnc répondent à ce test.

A partir des observations de (Allmang et al., 1999; van Hoof et al., 2000) montrant que la plupart des petits ARNnc ont des extrémités 3' étendues et polyadénylées dans un contexte génétique muté pour des gènes de l'exosome, comme par exemple dans une souche *Δrrp6*, nous émettons l'hypothèse selon laquelle nous pourrions identifier la plupart des ARNnc concernés par cette maturation en utilisant les puces Affymetrix pour l'analyse du transcriptome d'une souche *Δrrp6* en comparaison avec une souche sauvage (Figure 2).

Dans un contexte sauvage, les formes polyadénylées des snoRNAs et des snRNAs ne sont pas visibles, mais l'absence de *rrp6* entraine une accumulation des intermédiaires polyadénylés et étendus en 3' car Rrp6 avec l'exosome nucléaire participent à leur maturation.

Cette première expérience va me permettre de caractériser snr86, un snoRNA à boite H/ACA atypique car il a une structure et une taille inhabituelle pour un snoRNA (Figure 3). Ce travail sera intégré à une publication ultérieure avec Claire Torchet (*cf.* plus bas).

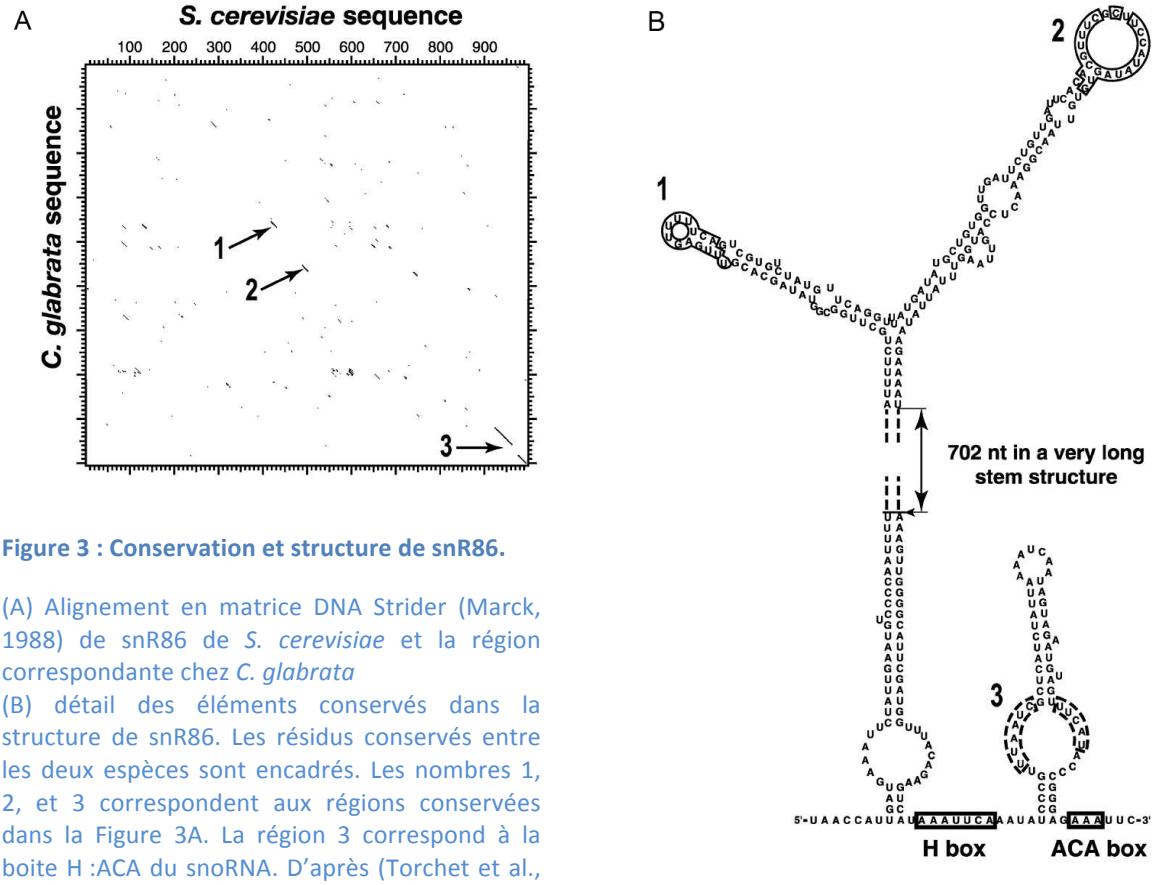


Figure 3 : Conservation et structure de snR86.

(A) Alignement en matrice DNA Strider (Marck, 1988) de snR86 de *S. cerevisiae* et la région correspondante chez *C. glabrata*

(B) détail des éléments conservés dans la structure de snR86. Les résidus conservés entre les deux espèces sont encadrés. Les nombres 1, 2, et 3 correspondent aux régions conservées dans la Figure 3A. La région 3 correspond à la boite H :ACA du snoRNA. D'après (Torchet et al., 2005).

Bien que les deux boucles soient bien conservées chez les levures, la fonction de la grande tige boucles de snR86 reste inconnue à ce jour.

Parallèlement à ce travail, je vais caractériser un deuxième snoRNA à boite H/ACA qui est hébergé par l'intron du facteur préribosomique NOG2, lui-même étudié par d'autres membres du laboratoire.

Ce snoRNA guide la modification dans l'ARNr des deux pseudouridines les plus conservées de l'évolution dans l'ARN ribosomique 25S (Badis et al., 2003) .

Pour finaliser la caractérisation des snoRNA et de leurs cibles dans l'ARN ribosomique, j'ai poursuivi ce travail ultérieurement en collaboration avec Claire Torchet. Une co-précipitation des ARNs associés aux protéines Gar1 et Nhp2, deux protéines associées spécifiquement à ce type de snoRNA, suivi de leur quantification à l'aide de microarray, nous a permis d'identifier l'ensemble des snoRNA à boîte H/ACA chez la levure. Par délétion systématique et cartographie des positions modifiées nous avons caractérisé tous les guides ARN spécifiant les modifications en pseudouridines de l'ARNr cytoplasmique chez la levure (Torchet et al., 2005). Ce travail a également révélé que des ARNm se lient à ces deux protéines, suggérant d'autres fonctions régulatrices possible.

Enfin, en plus des snoRNA, l'analyse des résultats des puces Affymetrix dans le mutant de Rrp6 m'a permis d'identifier dans des régions intergéniques un certain nombre de transcrits plus ou moins isolés et fortement stabilisés quand l'exosome nucléaire était muté (les points rouges Figure 2 qui ne correspondent pas à des snoRNAs).

J'ai passé les 6 derniers mois au laboratoire à essayer de caractériser plusieurs de ces éléments que nous nommerons ultérieurement les CUT pour « Cryptic Unstable Transcripts ». Devant partir en post-doctorat et travailler sur un autre sujet, je vais donner tous mes résultats préliminaires à l'équipe de Domenico Libri et ce travail aboutira à la publication de la découverte des CUTs (Wyers et al., 2005).

Ces différents volets de mon travail de thèse m'ont permis de prendre conscience de la diversité des ARNs, et des différents niveaux de régulation.

La régulation de l'expression des gènes peut-être en effet transcriptionnelle ou post-transcriptionnelle, et il faut garder à l'esprit que ce qu'on observe en général est le

résultat de la transcription, du passage au travers les nombreuses étapes de contrôle qualité, et de la dégradation de chacun des transcrits.

II Etude des spécificités de reconnaissance des facteurs de transcription eucaryotes

Post-doctorat : Juillet 2004-Décembre 2008

Dans le contexte du formidable essor de la génomique des années 2000, j'aspire à essayer de comprendre les mécanismes de régulation génique avec une vision globale des mécanismes. Au moment de commencer mon stage post doctoral, je souhaitais étudier la régulation post transcriptionnelle de l'expression des gènes avec une approche plus systématique, afin d'essayer d'en déchiffrer la logique de régulation. Je souhaitais également apprendre le maniement d'expériences et l'analyse de données génomiques et je me suis naturellement orientée vers le groupe de Timothy Hughes qui a accueilli ma candidature avec enthousiasme.

Deux évènements vont m'amener à modifier ma proposition de sujet de recherche. En premier, une limitation technique. En effet, les techniques disponibles alors, et permettant d'identifier des motifs ARN constituant des sites de fixation pour les protéines (comme le SELEX), étaient délicates à mettre en œuvre, peu sensible et peu applicables à une étude génomique à grande échelle. Il était - et il est toujours - difficile d'associer un motif de reconnaissance aux diverses protéines liant l'ARN et impliquées dans la régulation post-transcriptionnelle des gènes, ce qui rendait l'approche génomique incertaine. En second, le projet de recherche que nous soumettrons sur le sujet, ne sera pas accepté par le CIHR-IRSC (Instituts de recherche en santé du Canada).

Nous décidons alors de modifier notre perspective et de nous porter sur la régulation transcriptionnelle et les interactions ADN-protéines. En effet, il est possible d'envisager l'étude de la régulation transcriptionnelle par une approche à grande échelle en identifiant les interactions ADN-protéine de façon systématique avec un outil très performant qui avait été mis au point dans le laboratoire de Martha Bulyk à Harvard en 2004. En effet, la technique de "Protein Binding Microarray" ou PBM (Mukherjee et al., 2004) permet d'identifier *in vitro* les motifs ADN reconnus par une protéine d'intérêt fusionnée à un tag GST rendant l'étude exhaustive des interactions ADN-protéine plus accessible même à l'échelle d'un génome de mammifère tel que celui de la souris.

Les interactions ADN-protéines et la régulation de la transcription.

J'ai ainsi réorienté mon étude vers le projet ambitieux d'essayer de décoder la « grammaire » expliquant la régulation transcriptionnelle chez la souris, en commençant par une identification systématique des motifs ADN reconnus par tous les facteurs de transcription murin. J'ai écrit et proposé ce projet conjointement avec Timothy Hughes et l'équipe de Martha Bulyk à Harvard. Nous avons obtenu un financement pour ce projet collaboratif du CIHR (Canadian Institute for Health Research) pour 4 ans en 2005. J'ai également obtenu en parallèle une bourse post doctorale du CIHR me finançant pour 3 ans.

Nous avons tout d'abord du dresser une liste la plus exhaustive possible des facteurs de transcription murin. Pour cela, nous nous sommes rapprochés de 3 autres groupes canadiens afin de se repartir l'inventaire des facteurs de transcription murins et humains. Cet inventaire à été établi à partir d'un criblage de la littérature afin de catégoriser les facteurs de transcription (FT): FT certain (avec des preuves expérimentales de leur liaison à l'ADN et de leur régulation de gènes), FT probable (avec des motifs très conservés) ou FT putatif (car contenant un domaine de liaison à l'ADN ou ressemblant à un FT connu). Ce catalogue sera publié en 2009 (Fulton et al., 2009).

Afin de mettre en œuvre le projet d'identification des sites de liaison à l'ADN des facteurs de transcription murins, j'ai adapté une approche qui permet de simplifier le clonage (« ligation independent cloning ») aux gènes de la souris, et j'ai combiné cette technique à un système utilisant le transfert bactérien *in vivo* par recombinaison homologue des clones dans une collection de plasmides exprimant différentes fusions protéiques (un système de type « Gateway» mais non commercial adapté de (Li and Elledge, 2005)). J'ai développé un set de 6 vecteurs compatibles avec ce système et permettant de reconstituer 6 types de protéines chimériques distinctes s'exprimant dans différents contextes. J'ai d'abord sélectionné les 40 types de domaines protéiques ayant des propriétés d'interaction avec l'ADN ("DNA-Binding Domain" ou DBD) présents chez les mammifères et comportant des gènes connus pour être impliqués dans la régulation de la transcription. J'ai préalablement validé cela expérimentalement avec quelques exemples bien caractérisés, avant de me lancer dans l'étude de l'ensemble des facteurs de transcription. J'ai ensuite dirigé le design des ~2000 paires de primers flanquant les DBD des FT, réalisé les expériences

préliminaires vérifiant que la région DBD définie est nécessaire et suffisante pour conférer les propriétés de liaison à l'ADN, tout en diminuant les contraintes de clonage ou d'insolubilité. À partir d'ARN totaux issus de 55 tissus différents de souris (Zhang et al., 2004), j'ai réalisé ~2000 transcriptions réverse et dirigé le clonage de plus de 1200 FT murins (tous validés par séquençage), puis transférés dans le vecteur (T7) GST Nterm receveurs permettant d'exprimer une fusion avec une protéine GST en N terminal. J'ai conçu le projet avec l'aide de Tim Hughes, puis développé les outils, et enfin organisé sa mise en œuvre avec l'aide de nombreux stagiaires (« undergraduate students » pour la plupart et une étudiante en Master), un bio informaticien et quelques techniciens (*cf*. la rubrique « activité d'encadrement de la recherche » dans mon *curriculum vitae* durant la période 2004-2008). Dans ce projet collaboratif, nous mettions en œuvre le clonage et la production des domaines de liaison à l'ADN des facteurs de transcription en fusion avec la GST, puis l'équipe de Martha Bulyk à Harvard réalisait les expériences de PBM, et enfin les analyse bio-informatiques en découlant étaient faites conjointement aux deux laboratoires.

Ce système *in vitro* et universel est applicable à n'importe quel organisme et pour des études expérimentales variées permettant par exemple de réaliser des immonuprécipitations grâce à différentes protéines (Fusions GST, MBP, 3Myc, TAP), ou dans différents systèmes d'expression (bactérien, levure ou cellules de mammifère). Un autre vecteur permet de tester chez la levure des interactions de type ADN-protéine, protéine-protéine ou ARN-protéine avec respectivement le simple hybride (1H), le double hybride (2H) et le triple hybride (3H) (fusion GAL4 AD). Ce système et/ou un système dérivé à d'ailleurs été utilisé depuis avec d'autres organismes eucaryotes tels que le nématode ou l'homme (Narasimhan et al., 2015; Weirauch et al., 2014 respectivement pour exemples).

De ces 1200 gènes, nous avons réussi à purifier de manière satisfaisante plus de 800 protéines qui ont été envoyées à nos collaborateurs. Ils ont réalisé des expériences de PBM avec des puces universelles (Berger et al., 2006). Ce projet a permis d'identifier pendant la période de mon post doctorat plus de 600 motifs issus de résultats satisfaisant d'expériences de PBM et à conduit à la publication d'un chapitre de livre (Peña-Castillo and Badis, 2016) ainsi que 7 articles dont 3 publications majeures en tant que premier ou co-premier auteur (Badis et al., 2008, 2009; Berger et al., 2008).

Les résultats obtenus indiquent que la plupart des membres de chacune des classes de DBD peuvent reconnaître leur séquence cible, même dans ce système *in vitro*, avec une grande fidélité par rapport aux données obtenues *in vivo*. Nous avons en effet retrouvé la signature de la plupart des FT connus. Nous avons validé cette approche avec les 40 différentes classes de FT et ce travail présentant 104 nouveaux motifs reconnus par 22 des 40 différentes classes de FT issus de ce système a été publié en 2009 (Badis et al., 2009). De manière plus ou moins attendue, nous avons observé que les familles de domaines tendent à reconnaître des séquences contenant des motifs similaires, mais les résultats ont également permis de révéler une grande variété de motifs secondaires nouveaux, qui reflètent les différences de spécificité de chaque facteur même au sein des familles de facteurs de transcription ayant des motifs consensus très conservés (comme cela est représenté dans la Figure 4). Dans cette étude, nous avons non seulement pu prouver que la technique utilisée permettait de révéler ces différences subtiles d'affinité des FT appartenant aux familles à consensus commun, mais nous proposons également une nouvelle façon de représenter ces motifs, en « *k mer* » afin d'en garder toute la complexité, c'est-à-dire l'information des motifs liés ou non par un FT donné, qui est très souvent peu ou mal représentée par les matrices de position pondérée (Position Weight Matrices ou PWM) utilisées de façon plus classique. La Figure 4 illustre bien ce point car si on voit que les PWM des facteurs Lhx2 et Lhx4 se ressemblent, du fait de leur motif principal « TAATTA » commun et dominant dans la représentation, on remarque que la représentation en *k mer* permet d'identifier deux paires de motifs secondaires différentes (« TAATCA » et « TAATCT » pour Lhx4 et « TAATGA » et « TAACGA » pour Lhx2).

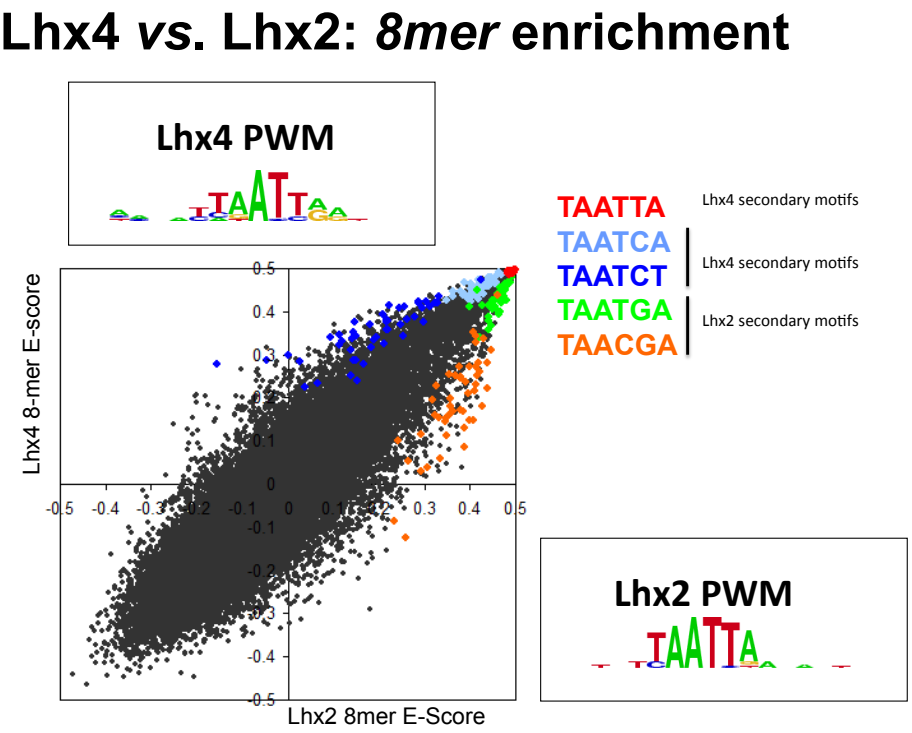


Figure 4 : Enrichissement des 8 mers associés à Lhx2 et Lhx4 appartenant à la famille des facteurs de transcription à homéodomaine.

Dans un autre article de ressource publié dans la revue Cell, nous présentons les motifs de 168 homéodomaines murins, qui constituent la deuxième plus grande famille de FT de cette espèce et l'une des mieux conservées de l'évolution (Berger et al., 2008). La caractérisation de la grande majorité des membres de cette famille a également permis de mettre au point un outil informatique permettant de prédire avec précision les spécificités des FT à partir de leurs séquences d'acides aminés et de trouver le code entre séquence du DBD et motif reconnu sur l'ADN. Parallèlement à ce travail de ressource et de déchiffrage des spécificités de reconnaissance des facteurs de transcription, nous avons collaboré avec plusieurs équipes sur des études plus spécifiques, et ces collaborations ont donné lieu à plusieurs publications (Sandhu et al., 2008; Wei et al., 2010).

En conclusion, nous avons pu déterminer au total environ 800 motifs pour les 1200 DBDs clonés. Ceci à permis de commencer à définir les bases régissant les modules régulateurs en cis (CRM ou "cis regulatory modules") qui peuvent être reconstitués en croisant ces résultats aux données d'expression des FT (ou s'expriment ils ?) ainsi qu'avec les données d'interactions protéiques (quels sont les partenaires de ces FT ?).

La caractérisation des modules cis régulateurs permet ainsi de commencer à "décortiquer" la "grammaire régulatrice" (« the regulatory lexicon »), comme par exemple cela avait été décrit avec les facteurs de transcription musculaires (Wasserman et al., 2000). La poursuite de cette analyse par l'équipe de Timothy Hughes a permis depuis de définir les premières grandes règles définissant la régulation transcriptionnelle dans plusieurs organismes (Najafabadi et al., 2015; Narasimhan et al., 2015; Weirauch et al., 2014 par exemple).

Compte tenu de la propriété « universelle » du système mis en place pour les facteurs de transcription murins et de la performance obtenue à l'échelle du génome de la souris, dans un deuxième temps de mon post doctorat j'ai appliqué une méthodologie similaire aux 200 facteurs de transcription de *S. cerevisiae.* Après avoir cloné et transférés les 200 DBD de levure et exprimé les 200 fusions GST correspondantes, j'ai réalisé moi-même les 200 expériences de PBM et obtenu 112 motifs de bonne qualité, portant à plus de 80% la fraction actuelle des facteurs de transcription de levure pour lesquels on connaît désormais le motif associé.

Parmi ces facteurs se trouve Rsc3, et nous avons pu également démontrer qu'en absence de Rsc3, les nucléosomes envahissent des centaines de promoteurs contenant le motif de liaison à Rsc3, alors qu'il n'y a pas d'effet significatif au niveau des promoteurs dépourvus de ce motif. Ces observations supportent l'hypothèse selon laquelle Rsc3 est responsable de l'exclusion des nucléosomes sur les promoteurs à un niveau général (Badis et al., 2008).

Durant les 4 ans passés à l'Université de Toronto, j'ai appris à utiliser les outils génomiques et à diriger un projet impliquant l'encadrement d'équipes de techniciens et de stagiaires qui se sont succédés pour m'aider. J'ai également compris que mon intérêt pour l'étude des ARN restait entier et que compte tenu de mon expertise, je pouvais apporter un savoir faire pour faire avancer la technologie et les connaissances dans ce domaine. Mener un projet à cette échelle comporte des aspects enthousiasmants car cela permet de mettre en lumière des résultats inattendus qui ne sont pas décelables autrement. Cependant, approfondir davantage par la caractérisation de mécanismes biologiques fonctionnels me semble indispensable et souvent plus stimulant. J'ai donc souhaité garder l'approche génomique comme outil mais d'avoir aussi la possibilité d'approfondir l'analyse fonctionnelle sur des

mécanismes spécifiques liés à l'ARN. La levure demeure un modèle de choix à la fois pour les études à grande échelle et pour l'analyse fonctionnelle de ces facteurs.

II Les interactions ARN-protéines et la régulation du transcriptome de levure.

Depuis janvier 2009

De retour de post-doctorat, je reviens dans le laboratoire d'Alain Jacquier avec un poste au CNRS et l'ambition de faire avec les protéines liant l'ARN dans la levure ce que j'avais réussi à faire pour les facteurs de transcription. Après une période de transition à essayer de mettre au point la technique « RNAcompete » pour mon projet (Ray et al., 2009), il s'avère que les expériences préliminaires avec des RNA binding aux motifs connus n'ont pas donné les résultats escomptés.

C'est à peu près à cette période que Sander Granneman développe le CRAC (UV CRoss linking and Analysis of cDNA) dans le laboratoire de David Tollervey (Granneman et al., 2009), et bien que son utilisation à grande échelle soit encore difficile à mettre en œuvre, je vais implémenter cette technique au laboratoire. Je consacrerais par la suite du temps au développement et à l'amélioration de cette technique pour explorer plusieurs nouveaux facteurs agissant sur l'ARN ou non connus (*cf*. plus bas).

En parallèle, dans la continuité de la découverte des CUT et de leur caractérisation à l'échelle génomique par le laboratoire (Neil et al., 2009), je m'intéresse à l'éventuel rôle régulateur de la transcription pervasive et des transcrits pervasifs, dans la régulation des gènes. Un niveau important de transcription pervasive peut être révélé lorsque les machineries de surveillance cytoplasmique et nucléaires sont inactivées (Malabat et al., 2015). L'observation de la persistance de certains de ces transcrits en phase stationnaire, alors que la transcription générale est globalement massivement réprimée, m'a motivé à essayer de comprendre le rôle de cette transcription dans la régulation des gènes et en particulier entre la phase exponentielle et la quiescence (G0).

Régulation génique par interférence transcriptionnelle médiée par la transcription d'antisens non codants.

Afin de caractériser l'incidence de ces observations à l'échelle génomique, nous avons analysé le transcriptome en phase exponentielle ou G0 dans un contexte génétique *upf1Δ*, dans lequel la machinerie de surveillance cytoplasmique est inactivée. Ces expériences nous ont permis d'identifier plusieurs centaines de transcrits antisens de gènes, dont la transcription s'étend au-delà du promoteur du gène correspondant. Les

gènes concernés sont principalement les gènes les moins exprimés en phase exponentielle, dont l'expression d'un grand nombre est induite après le shift diauxique ou en G0. Ces transcrits antisens sont détectables lorsque la transcription des ARNm sens est absente ou faible, et ils disparaissent quand les ARNm sont exprimés plus fortement (en G0 par exemple).

Quelques exemples d'interférence transcriptionnelle par des ARNs antisens avaient été décrits précédemment chez la levure, mais nous montrons que la répression médiée par la transcription d'ARN antisens chevauchant le promoteur des gènes constitue un mécanisme de régulation qui doit concerner un grand nombre de gène. Ce mécanisme concerne jusqu'à un tiers des gènes les moins exprimés, il agit en cis et l'interruption de la transcription antisens restaure l'expression de l'ARN sens. Ce mécanisme d'interférence transcriptionnelle fait intervenir une combinaison de facteurs de modification de la chromatine (tels que Set1, Set2 ou Hda1).

Nous montrons également que ce mécanisme de régulation existe aussi dans l'autre sens, c'est-à-dire que l'abolition de la transcription sens de certains ARNm en phase exponentielle (comme l'ARN du gène HIS1), permet de restaurer l'expression d'ARN antisens uniquement détectable en G0. Ce travail a fait l'objet d'une publication dans le journal Nucleic Acids Research et est l'objet de la thèse d'Alicia Nevers que j'ai encadrée (Nevers et al., 2018). Il aurait été intéressant de poursuivre ce travail, mais le besoin de me recentrer sur un seul sujet et de ne pas continuer à travailler de façon isolée dans le laboratoire m'ont poussé à abandonner la poursuite de ce projet.

Les interactions ARN-protéine et la régulation du transcriptome.

En parallèle de mon travail sur les ARNs antisens, j'ai implémenté au laboratoire depuis 2009 la technique du CRAC (UV CRoss linking and Analysis of cDNA, voir aussi Granneman et al., 2009).

Cette technique utilise une étiquette (6his-TEV-ProteinA) pour réaliser après cross link *in vivo* aux UV une double purification des protéines ainsi associées aux ARNs (IgG, élution TEV puis purification sur billes de Nickel en conditions dénaturante).

Le faible rendement de la technique originale ne m'a permis d'obtenir des résultats satisfaisants dans un premier temps qu'avec des protéines abondantes telles que les protéines ribosomiques ou pré-ribosomiques.

C'est dans le cadre d'une collaboration avec le laboratoire de Jesus de la Cruz et Micheline Fromont que nous publierons les premières expériences de ce type, avec

l'étude du positionnement du facteur pré-ribosomique Rlp7 et de la protéine ribosomique Rpl7. En effet, compte tenu de l'homologie entre les protéines Rpl7 et Rlp7, on pouvait penser alors que le facteur pré ribosomique Rlp7 prenait la place de Rpl7 dans la particule préribosomique, de la même manière que d'autre facteurs (comme par exemple Rlp24/Rpl24).

Nous montrerons que, contrairement à ce que suggérait l'hypothèse de départ, Rlp7 se lie dans les ITS1 et 2 du pré-rRNA 60S et que Rlp7 et Rpl7 peuvent coexister sur la même particule préribosomique (Babiano et al., 2013).

Améliorations techniques du CRAC

En m'intéressant à des hélicases moins abondantes que les facteurs ribosomiques qui ciblent un grand nombre des transcrits de la cellule, je réalise que la sensibilité de la méthode de CRAC que nous utilisons est loin d'être satisfaisante et qu'elle induit certains biais techniques, et je vais modifier progressivement entre 2009 et 2018 un certain nombre d'étapes par rapport au protocole initial afin d'améliorer la sensibilité de la technique du CRAC et de la rendre réellement « génomique ». Le principe du CRAC ainsi que les modifications apportées au protocole sont récapitulés dans la Figure 5. Le traitement à la RNase 1 (étape n°3) permet de s'affranchir d'un biais en faveur des régions riches en poly(A) qui étaient sur-représentées dans les premières expériences avec les RNAse A et T1 anciennement utilisées. L'utilisation du système Gelfree nous affranchit du marquage radioactif (étape n°5) et surtout du transfert sur membrane de nitrocellulose (étape n°6), ce qui nous a permis de gagner au moins 6 à 10 cycles de PCR (soit une récupération de 64 à plus de 1000 fois plus de matériel par rapport aux premières expériences), reflétant la plus grande avancée dans l'amélioration du protocole. Enfin, l'ajout d'un traitement à l'Exonuclease I après la RT (étape n°7) permet d'éliminer les primers libre et en excès, et de réduire ainsi de façon importante la proportion de dimères d'adaptateurs vides dans la banque finale.

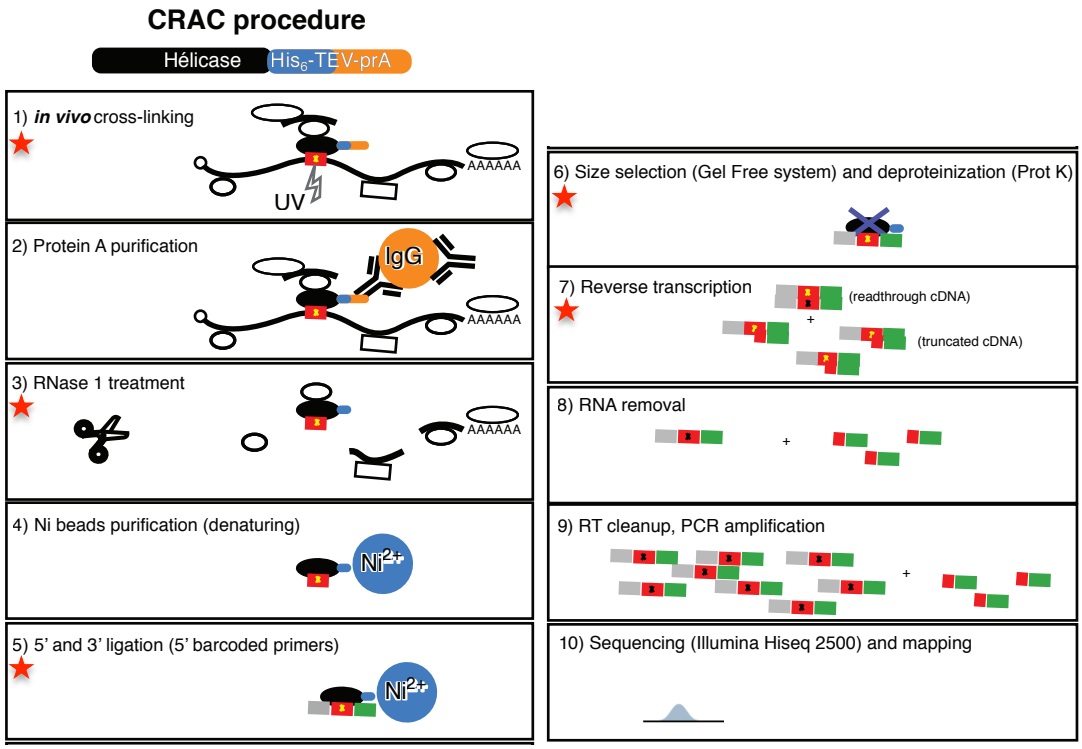


Figure 5 : Représentation schématique du protocole du CRAC actuel.

Les étoiles rouges indiquent les étapes qui ont été modifiées par rapport au protocole original.

Etude de facteurs impliqués dans le contrôle qualité des ARNs aberrants

Au laboratoire, nous nous intéressons au métabolisme des ARNs en général et en particulier aux divers mécanismes de surveillance cytoplasmique. En effet, la cellule possède un système sophistiqué de surveillance de l'intégrité du signal exprimé. Trois types de contrôle qualité ciblant les ARNm cytoplasmiques ont été décrits : le « nonsense-mediated decay » (NMD), le « nonstop decay » (NSD), et le « no-go decay » (NGD) (voir Figure 6).

Dans ces trois mécanismes, un évènement de traduction reconnu comme « aberrant » active la dégradation accélérée des ARNm concernés :

Facteurs liés au NSD/NGD.

Si un ARNm ne possède pas de codon stop, il va être pris en charge par le « Non Stop Decay » (NSD) et s'il est bloqué lors de la traduction, ce sera le « No Go Decay » (NGD) (Figure 6).

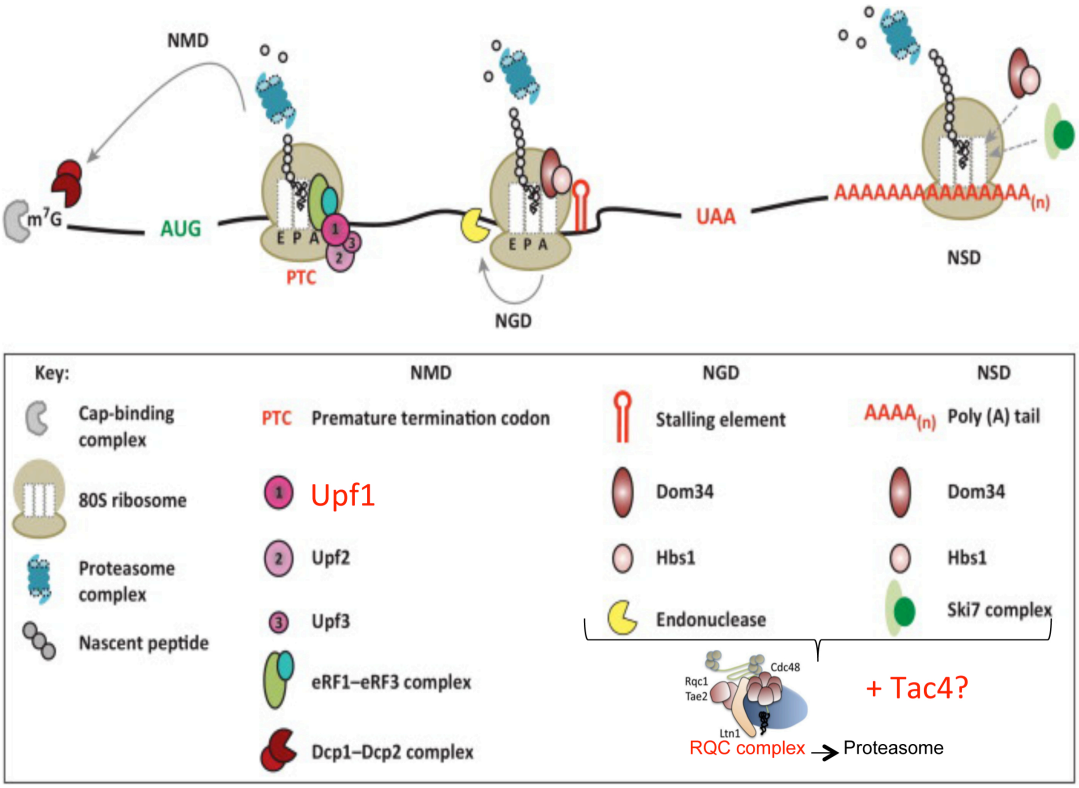


Figure 6: Dégradation ciblée des mRNP aberrantes par le NMD, NGD, et NSD (d'après Roy and Jacobson, 2013)

Micheline Fromont a animé récemment un groupe thématique axé sur le contrôle qualité cytoplasmique, les mécanismes de NSD et NGD. Elle a mis en évidence un nouveau complexe, le complexe RQC, impliqué dans l'élimination des peptides aberrants associés aux ribosomes bloqués par la traduction de messagers non conformes (Defenouillère et al., 2013). Ce complexe RQC constitué de trois facteurs, Ltn1, Rqc1 et Tae2, se lie à la sous-unité 60S contenant des peptides naissants aberrants après dissociation du ribosome (Figure 6). Cependant, la localisation de ce complexe sur la 60S n'était pas connue au début de ce travail. J'ai utilisé la technique du CRAC afin de caractériser de façon précise les éventuels sites de liaison des composants du complexe RQC sur le ribosome *via* l'ARN ribosomique. Les résultats pour Tae2 et Rqc1 sont présentés Figure 7 A, B et C.

Concomitamment à ce travail, la localisation de Ltn1 et de Rqc2 sur la sous-unité ribosomique 60S a été déterminée par cryo-électromicroscopie et publié par l'équipe de J. Weissman (Shen et al., 2015). Mes résultats sont en accord avec les résultats publiés.

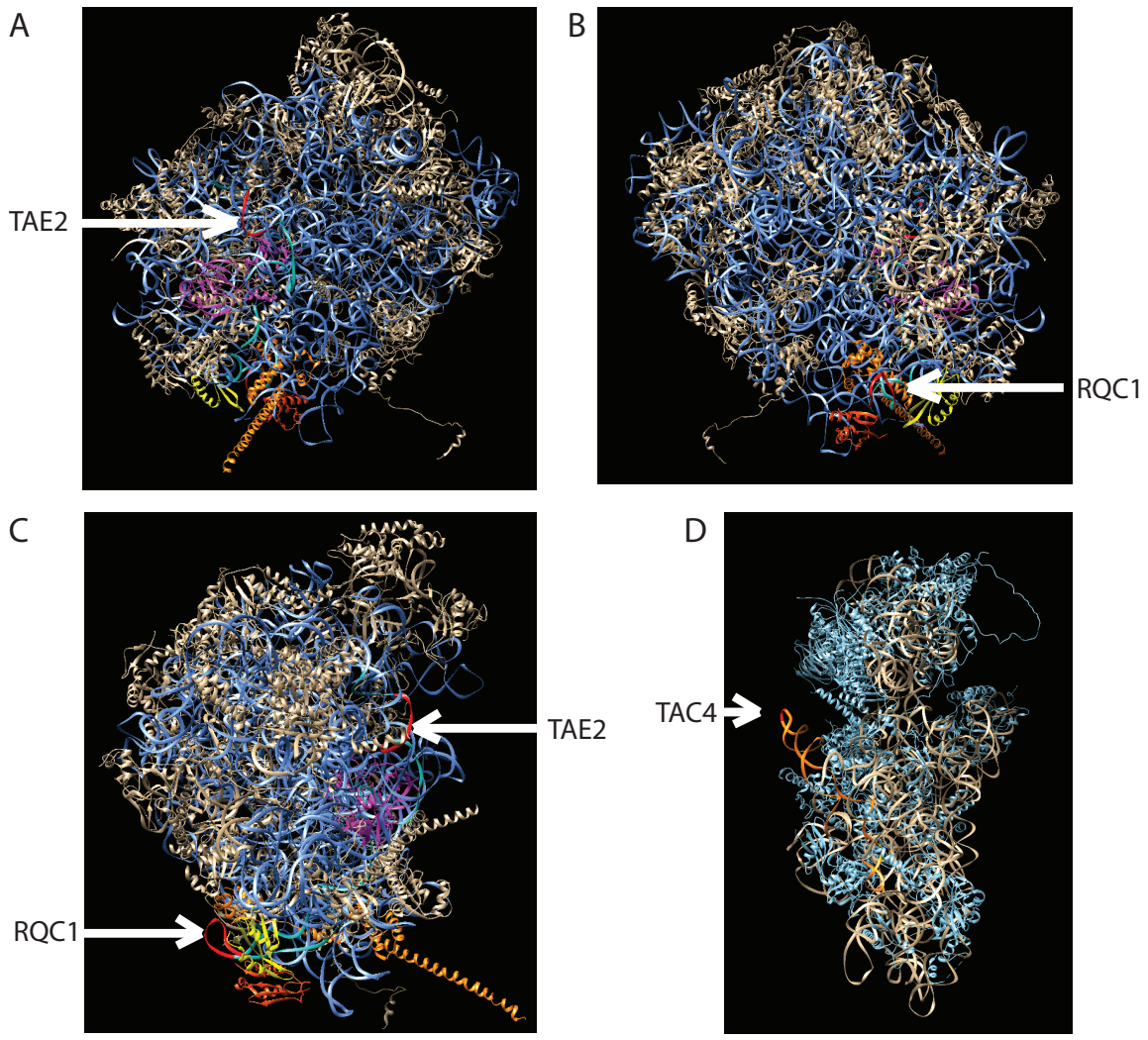


Figure 7 : Localisation des sites de fixation dans la structure tridimensionnelle de la sous-unité ribosomique.

Les sites de fixations des protéines identifiés par CRAC sont indiqués par des flèches et des résidus colorés en rouge. L'image a été obtenue avec le programme Chimera de l'UCSF (Pettersen et al., 2004) en utilisant les structures ribosomiques décrites dans Ben-Shem et al. (3U5I and 3U5H, (Ben-Shem et al., 2011)). Pour la 60S, (A, B et C) les protéines à proximité des sites identifiés sont colorées : L19 (orange), L22 (orange foncé) et L38 (jaune) pour Rqc1, et L2 (rose) pour Rqc2. A, B et C) Localisation des sites de fixation de Rqc1 et Rqc2 dans la structure tridimensionnelle de la sous-unité ribosomique 60S respectivement de face (A), dos (B) et coté(C). D) Localisation des sites de fixation de Tac4 dans la structure tridimensionnelle de la sous-unité ribosomique 40S.

III PERSPECTIVES : Etude des interaction ARN-protéines dans les mécanismes de contrôle qualité des ARNs.

Notre laboratoire s'intéresse depuis quelques années aux complexes impliqués dans les mécanismes de surveillance cytoplasmique. En effet, les équipes du laboratoire ont caractérisé plusieurs des complexes associés aux mécanismes de contrôle qualité des ARNs et des protéines synthétisées à partir d'ARNm aberrants. L'équipe de

Micheline Fromont a participé à la découverte des facteurs clés du NSD/NGD (Defenouillère et al., 2013, 2016, 2017) et celle de Cosmin Saveanu a caractérisé récemment deux sous-complexes distincts associés à UPF1, l'hélicase majeure du NMD (Dehecq et al., 2018). Plus globalement, je m'intéresse aux interactions ARN-protéines dans les mécanismes de contrôle qualité des ARNs, qui impliquent l'action d'hélicases particulières dont la spécificité de reconnaissance est plus ou moins bien caractérisée. En effet, en plus de permettre d'ouvrir des structures double brins dans l'ARN, les ARN hélicases sont connues pour pouvoir dissocier des protéines liées à l'ARN ou bien simplement se lier très fortement à l'ARN et de façon non spécifique d'une séquence particulière, sans nécessairement avancer sur le transcrit (Jarmoskaite and Russell, 2014 pour revue). Le rôle précis de chacune de ces ARN hélicases dans les mécanismes de contrôle qualité des ARN n'est pas bien compris à ce jour.

Je propose d'essayer de répondre à cette question par une approche génomique des spécificités de reconnaissance en me focalisant sur des hélicases impliquées dans la traduction et le contrôle qualité des ARNm.

Un nouveau facteur impliqué dans le recyclage des ribosomes.

Micheline Fromont a identifié par double hybride avec Rqc1 une hélicase putative baptisé *TAC4* pour « Translation associated complex » et je participe à son analyse fonctionnelle depuis 2014. Tac4 contient un motif ARN hélicase et est associé au ribosome.

En collaboration avec Varun Khanna du Hub bioinformatique de l'institut Pasteur, les analyses préliminaires d'expériences de CRAC que j'ai réalisé ont permis de localiser des cibles de Tac4 sur le génome entier, bien qu'il soit nécessaire d'obtenir de nouveaux résultats avec une version optimisée du protocole de CRAC pour avoir une meilleure profondeur. Ces résultats suggèrent que Tac4, en plus de se lier à l'hélice h16 de la sous-unité 40S du ribosome (Figure 7D), se lierait également aux 3'UTR des ARN messagers au niveau de leurs sites de polyadénylation (pA sites). Ces résultats sont tout à fait inattendus car l'homologue humain de ce gène, DDX29, qui lie exactement la même hélice dans la petite sous unité du ribosome, est décrit comme une hélicase potentiellement impliquée dans l'initiation de la traduction et le scan des ARNm dans leurs parties 5' UTR (des Georges et al., 2015; Hashem et al., 2013). Par ailleurs il a été

récemment montré par le laboratoire de Rachel Green que des ribosomes pouvaient passer le codon Stop et se retrouver dans le 3' UTR des messagers (Guydosh and Green, 2014; Young et al., 2015). Soit ils ré-initient la traduction et synthétise des petits peptides, avec une terminaison de la traduction normale, soit ils scannent le 3' UTR et restent bloqués, et ont dans ce cas une terminaison aberrante. Compte tenu de ces observations, cela suggère que Tac4 pourrait avoir un rôle dans les mécanismes de recyclage des ribosomes bloqués dans les mécanismes de NSD/NGD.

En collaboration avec l'équipe d'Olivier Namy, Micheline Fromont vient d'obtenir un financement de l'Agence Nationale pour la Recherche (ANR) auquel je suis associée pour poursuivre la caractérisation de Tac4 ainsi que d'autres facteurs impliqués dans l'éventuel recyclage des ribosomes lorsqu'ils ne sont pas dissociés aux stops et qu'ils restent associés aux 3'UTR des ARNm. La connaissance des transcrits spécifiquement associés à ce facteur et de la région particulière liée à ces transcrits sera cruciale pour comprendre la fonction moléculaire de Tac4 dans ce processus.

Un nouveau facteur associé au complexe Ski.

L'équipe de Micheline s'intéresse également à l'exosome cytoplasmique qui permet une dégradation 3'-5' des ARNs. Elle a identifié récemment un nouveau facteur Ska1, qui est biochimiquement associé au complexe SKI indépendamment du ribosome. De manière intéressante, la structure de l'exosome que Micheline Fromont a publié en collaboration avec Roland Beckman et Elena Conti (Schmidt et al., 2016) montre le complexe ski associé au ribosome, alors que le facteur Ska1 est associé à l'ARNm indépendamment du ribosome. Dans un article publié récemment (Zhang et al., 2019), elle montre que le complexe SKI peut être soit associé aux ribosomes dans les phase codantes, soit être associé au facteur baptisé « SKA1 » pour « SKI associated component 1 » lorsque l'exosome parcourt des régions dépourvues de ribosomes telles que les 3'UTR. L'étude des cibles de Ski2, l'hélicase communément trouvée dans ces deux sous-complexes, par une approche de type « split-CRAC » et « split-RIP » pourrait nous permettre de caractériser les deux sous-population de transcrits spécifiquement associées à Ski2. En effet, le fait que Ski2 soit partagé dans plusieurs sous complexes nécessite d'adapter la stratégie du CRAC classique en séparant les étiquettes permettant la purification de sous-complexes protéiques distincts : une première purification est effectuée avec un tag sur un des facteurs spécifique de l'un ou l'autre des sous-complexes, tandis que la seconde

purification se fait sur Ski2 en condition dénaturante. Une alternative est d'effectuer le CRAC directement avec Ska1, car les données préliminaires indiquent que Ska1 peut lier directement l'ARN. Selon le modèle proposé dans (Zhang et al., 2018), le complexe SKI spécifiquement associé à Ska1 devrait trouver des cibles préférentiellement localisées dans les 3' UTR des transcrits.

Ce travail s'inscrit dans un projet plus global de caractérisation de ce sous-complexe, pour lequel Micheline Fromont, en collaboration avec Roland Beckmann vient d'obtenir un financement par l'ANR auquel je suis associée.

Facteurs liés au NMD et au delà du NMD.

Lorsqu'un ARN possède un codon stop prématuré dans une ORF (par exemple parce qu'un intron n'a pas été épissé correctement), ou qu'il possède une petite ORF suivie d'un long 3'UTR (certaines uORF dans les 5'UTR des gènes par exemple), il devient un bon substrat pour être éliminé par le NMD orchestrée entre autre par l'ARN hélicase Upf1 (ou Nam7 chez la levure). Si Upf1 a été trouvé associé spécifiquement aux polysomes et donc aux ARNm en cours de traduction, la manière dont il est recruté sur les ARNm n'est pas connue.

Le groupe de Cosmin Saveanu, au laboratoire, vient de caractériser deux sous-complexes associés à Upf1 : un complexe « détecteur » comprenant Upf1, Upf2 et Upf3, et un complexe « effecteur » comprenant Upf1, Nmd4, Ebs1 et la machinerie de decapping et de dégradation : Dcp1, Dcp2, Edc3, Lsm1-7 et Hrr25 (Dehecq et al., 2018). Afin d'essayer de comprendre la manière dont Upf1 est associé à ces deux sous complexes, j'ai réalisé des expériences de RIPseq, de CRAC avec Upf1 et initié et encadré Léna Audebert en stage de Master 2 pour réaliser des expériences de « split-CRAC » dans plusieurs contextes associés à Upf1. Les résultats préliminaires semblent indiquer qu'Upf1 se lie préférentiellement dans les 3'UTR des transcrits et de façon plus modérée aux extrémités 5'. De manière très surprenante, en plus de trouver les cibles attendues du NMD, Upf1 se lie a un grand nombre de transcrit « non-NMD », et Léna Audebert commence actuellement une thèse que je co-dirige avec Cosmin et dont le sujet est d'essayer de comprendre le rôle d'Upf1 dans la régulation post-transcriptionnelle de ces substrats. En effet, la synthèse de ces résultats suggère qu'Upf1 pourrait être associé à une sous-population d'ARN « non-NMD », qui pourrait également impliquer les protéines Lsm. La voie de régulation post-transcriptionnelle mise en jeu reste à caractériser mais les

données préliminaires que nous avons obtenu indiquent que ces substrats « non NMD » d'UPF1 pourraient avoir leur demie vie et leurs taille de queues poly(A) affectées dans ce sous complexe.
Par des approches génomiques et moléculaires, nous essayons de comprendre comment cette sous population est prise en charge dans un mécanisme qui implique UPF1, et avec quels autres partenaires.

En résumé, de manière générale, je cherche à décortiquer plus précisément les mécanismes de contrôle qualité des ARNs avec des approches transcriptomiques (RNAseq ; mapping 5' ou 3', RIPseq, CRAC, …) replacées dans des contextes biochimiques complexes (les différentes mRNPs et sous-complexes).

Conclusion

Après avoir passé des années à essayer de simplifier les mécanismes dans le but de les expliquer, en représentant souvent à tord les ARNs ou l'ADN nus par exemple, ou décoré d'un facteur ou quelques facteurs, on oublie que la cellule présente une formidable complexité, avec une multitude de facteurs qui interagissent en permanence avec la multitude de molécules d'ADN et d'ARN. C'est cette complexité que je souhaite appréhender à l'avenir, en étudiant les variations subtiles des états de l'ARN (pour un même mRNA, 5', du 3', queue polyA ou produit de dégradation, …) selon son association avec tel ou tel facteur ou associations de facteurs.
La ligne conductrice de mon parcours a été d'essayer de comprendre un peu mieux comment les molécules interagissaient entre elles, au sein des complexes biochimiques formant les mRNPs. Que ce soit durant ma thèse avec l'étude d'EDC3 et RPS28B, durant mon post-doctorat avec la caractérisation des séquences de l'ADN reconnues par les facteurs de transcription murins et de levure, ou plus tard avec l'identification des ARNs ciblés par des facteurs « RNA binding » de surveillance cytoplasmique, ma curiosité à vouloir comprendre ces interactions et les mécanismes mis en jeux reste intacte. A chaque nouveau projet, je suis enthousiaste de pouvoir participer à lever un coin du voile.

REFERENCES

Allmang, C., Kufel, J., Chanfreau, G., Mitchell, P., Petfalski, E., and Tollervey, D. (1999). Functions of the exosome in rRNA, snoRNA and snRNA synthesis. EMBO J. *18*, 5399–5410.

Arribas-Layton, M., Wu, D., Lykke-Andersen, J., and Song, H. (2013). Structural and functional control of the eukaryotic mRNA decapping machinery. Biochim. Biophys. Acta *1829*, 580–589.

Babiano, R., Badis, G., Saveanu, C., Namane, A., Doyen, A., Díaz-Quintana, A., Jacquier, A., Fromont-Racine, M., and de la Cruz, J. (2013). Yeast ribosomal protein L7 and its homologue Rlp7 are simultaneously present at distinct sites on pre-60S ribosomal particles. Nucleic Acids Res. *41*, 9461–9470.

Badis, G., Fromont-Racine, M., and Jacquier, A. (2003). A snoRNA that guides the two most conserved pseudouridine modifications within rRNA confers a growth advantage in yeast. RNA N. Y. N *9*, 771–779.

Badis, G., Saveanu, C., Fromont-Racine, M., and Jacquier, A. (2004). Targeted mRNA degradation by deadenylation-independent decapping. Mol. Cell *15*, 5–15.

Badis, G., Chan, E.T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C.D., Gossett, A.J., Hasinoff, M.J., Warren, C.L., et al. (2008). A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. Mol. Cell *32*, 878–887.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science *324*, 1720–1723.

Beelman, C.A., and Parker, R. (1995). Degradation of mRNA in eukaryotes. Cell *81*, 179–183.

Ben-Shem, A., Garreau de Loubresse, N., Melnikov, S., Jenner, L., Yusupova, G., and Yusupov, M. (2011). The structure of the eukaryotic ribosome at 3.0 Å resolution. Science *334*, 1524–1529.

Berger, A.B., Decourty, L., Badis, G., Nehrbass, U., Jacquier, A., and Gadal, O. (2007). Hmo1 is required for TOR-dependent regulation of ribosomal protein gene transcription. Mol. Cell. Biol. *27*, 8015–8026.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep III, P.W., and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotech *24*, 1429–1435.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Peña-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. Cell *133*, 1266–1276.

Decourty, L., Saveanu, C., Zemam, K., Hantraye, F., Frachon, E., Rousselle, J.-C., Fromont-Racine, M., and Jacquier, A. (2008). Linking functionally related genes by sensitive and quantitative characterization of genetic interaction profiles. Proc. Natl. Acad. Sci. U. S. A. *105*, 5821–5826.

Defenouillère, Q., Yao, Y., Mouaikel, J., Namane, A., Galopier, A., Decourty, L., Doyen, A.,

Malabat, C., Saveanu, C., Jacquier, A., et al. (2013). Cdc48-associated complex bound to 60S particles is required for the clearance of aberrant translation products. Proc. Natl. Acad. Sci. U. S. A. *110*, 5046–5051.

Defenouillère, Q., Zhang, E., Namane, A., Mouaikel, J., Jacquier, A., and Fromont-Racine, M. (2016). Rqc1 and Ltn1 Prevent C-terminal Alanine-Threonine Tail (CAT-tail)-induced Protein Aggregation by Efficient Recruitment of Cdc48 on Stalled 60S Subunits. J. Biol. Chem. *291*, 12245–12253.

Defenouillère, Q., Namane, A., Mouaikel, J., Jacquier, A., and Fromont-Racine, M. (2017). The ribosome-bound quality control complex remains associated to aberrant peptides during their proteasomal targeting and interacts with Tom1 to limit protein aggregation. Mol. Biol. Cell *28*, 1165–1176.

Dehecq, M., Decourty, L., Namane, A., Proux, C., Kanaan, J., Le Hir, H., Jacquier, A., and Saveanu, C. (2018). Nonsense-mediated mRNA decay involves two distinct Upf1-bound complexes. EMBO J.

Fromont-Racine, M., Mayes, A.E., Brunet-Simon, A., Rain, J.C., Colley, A., Dix, I., Decourty, L., Joly, N., Ricard, F., Beggs, J.D., et al. (2000). Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. Yeast Chichester Engl. *17*, 95–110.

Fulton, D.L., Sundararajan, S., Badis, G., Hughes, T.R., Wasserman, W.W., Roach, J.C., and Sladek, R. (2009). TFCat: the curated catalog of mouse and human transcription factors. Genome Biol. *10*, R29.

des Georges, A., Dhote, V., Kuhn, L., Hellen, C.U.T., Pestova, T.V., Frank, J., and Hashem, Y. (2015). Structure of mammalian eIF3 in the context of the 43S preinitiation complex. Nature *525*, 491–495.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546, 563–567.

Granneman, S., Kudla, G., Petfalski, E., and Tollervey, D. (2009). Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. Proc. Natl. Acad. Sci. U. S. A. *106*, 9613–9618.

Guydosh, N.R., and Green, R. (2014). Dom34 rescues ribosomes in 3' untranslated regions. Cell *156*, 950–962.

Hashem, Y., des Georges, A., Dhote, V., Langlois, R., Liao, H.Y., Grassucci, R.A., Hellen, C.U.T., Pestova, T.V., and Frank, J. (2013). Structure of the mammalian ribosomal 43S preinitiation complex bound to the scanning factor DHX29. Cell *153*, 1108–1119.

He, F., Li, C., Roy, B., and Jacobson, A. (2014). Yeast Edc3 targets RPS28B mRNA for decapping by binding to a 3' untranslated region decay-inducing regulatory element. Mol. Cell. Biol. *34*, 1438–1451.

Jarmoskaite, I., and Russell, R. (2014). RNA helicase proteins as chaperones and remodelers. Annu. Rev. Biochem. *83*, 697–725.

Li, M.Z., and Elledge, S.J. (2005). MAGIC, an in vivo genetic method for the rapid construction of recombinant DNA molecules. Nat. Genet. *37*, 311–319.

Malabat, C., Feuerbach, F., Ma, L., Saveanu, C., and Jacquier, A. (2015). Quality control

of transcription start site selection by Nonsense-Mediated-mRNA Decay. ELife *4*.

Marck, C. (1988). "DNA Strider": a "C" program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. Nucleic Acids Res. *16*, 1829–1836.

McCutcheon, J.P., and Eddy, S.R. (2003). Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nucleic Acids Res. *31*, 4119–4128.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat. Genet. *36*, 1331–1339.

Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M., et al. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. *33*, 555–562.

Narasimhan, K., Lambert, S.A., Yang, A.W.H., Riddell, J., Mnaimneh, S., Zheng, H., Albu, M., Najafabadi, H.S., Reece-Hoyes, J.S., Fuxman Bass, J.I., et al. (2015). Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities. ELife *4*.

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

Nevers, A., Doyen, A., Malabat, C., Néron, B., Kergrohen, T., Jacquier, A., and Badis, G. (2018). Antisense transcriptional interference mediates condition-specific gene repression in budding yeast. Nucleic Acids Res. *46*, 6009–6025.

Olivas, W.M., Muhlrad, D., and Parker, R. (1997). Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. Nucleic Acids Res. *25*, 4619–4625.

Peña-Castillo, L., and Badis, G. (2016). Systematic Determination of Transcription Factor DNA-Binding Specificities in Yeast. Methods Mol. Biol. Clifton NJ *1361*, 203–225.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Ray, D., Kazan, H., Chan, E.T., Peña Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat. Biotechnol. *27*, 667–670.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nat. Biotechnol. *17*, 1030–1032.

Roy, B., and Jacobson, A. (2013). The intimate relationships of mRNA decay and translation. Trends Genet. TIG *29*, 691–699.

Sandhu, C., Hewel, J.A., Badis, G., Talukder, S., Liu, J., Hughes, T.R., and Emili, A. (2008). Evaluation of data-dependent versus targeted shotgun proteomic approaches for monitoring transcription factor expression in breast cancer. J. Proteome Res. *7*, 1529–1541.

Schmidt, C., Kowalinski, E., Shanmuganathan, V., Defenouillère, Q., Braunger, K., Heuer, A., Pech, M., Namane, A., Berninghausen, O., Fromont-Racine, M., et al. (2016). The cryo-EM structure of a ribosome-Ski2-Ski3-Ski8 helicase complex. Science *354*, 1431–1433.

Shen, P.S., Park, J., Qin, Y., Li, X., Parsawar, K., Larson, M.H., Cox, J., Cheng, Y., Lambowitz, A.M., Weissman, J.S., et al. (2015). Protein synthesis. Rqc2p and 60S ribosomal subunits mediate mRNA-independent elongation of nascent chains. Science *347*, 75–78.

Torchet, C., Badis, G., Devaux, F., Costanzo, G., Werner, M., and Jacquier, A. (2005). The complete set of H/ACA snoRNAs that guide rRNA pseudouridylations in Saccharomyces cerevisiae. RNA N. Y. N *11*, 928–938.

van Hoof, A., Lennertz, P., and Parker, R. (2000). Yeast exosome mutants accumulate 3'-extended polyadenylated forms of U4 small nuclear RNA and small nucleolar RNAs. Mol. Cell. Biol. *20*, 441–452.

Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Hieter, P., Vogelstein, B., and Kinzler, K.W. (1997). Characterization of the yeast transcriptome. Cell *88*, 243–251.

Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. (2000). Human-mouse genome comparisons to locate regulatory sites. Nat. Genet. *26*, 225–228.

Wei, G.-H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., et al. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. *29*, 2147–2160.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science *285*, 901–906.

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

Young, D.J., Guydosh, N.R., Zhang, F., Hinnebusch, A.G., and Green, R. (2015). Rli1/ABCE1 Recycles Terminating Ribosomes and Controls Translation Reinitiation in 3'UTRs In Vivo. Cell *162*, 872–884.

Zhang, E., Khanna, V., Namane, A., Doyen, A., Dacheux, E., Turcotte, B., Jacquier, A., and Fromont-Racine, M. (2018). A SKI subcomplex specifically required for the degradation of ribosome-free RNA regions. BioRxiv 409490.

Zhang, E., Khanna, V., Dacheux, E., Namane, A., Doyen, A., Gomard, M., Turcotte, B., Jacquier, A., and Fromont-Racine, M. (2019). A specialised SKI complex assists the cytoplasmic RNA exosome in the absence of direct association with ribosomes. EMBO J. *38*, e100640.

Zhang, W., Morris, Q.D., Chang, R., Shai, O., Bakowski, M.A., Mitsakakis, N., Mohammad, N., Robinson, M.D., Zirngibl, R., Somogyi, E., et al. (2004). The functional landscape of mouse gene expression. J. Biol. *3*, 21.

PUBLICATIONS MAJEURES

Molecular Cell, Vol. 15, 5–15, July 2, 2004, Copyright ©2004 by Cell Press

Targeted mRNA Degradation by Deadenylation-Independent Decapping

Gwenael Badis, Cosmin Saveanu, Micheline Fromont-Racine, and Alain Jacquier*
Génétique des Interactions Macromoléculaires
Institut Pasteur
25, Rue du Docteur Roux
75724 Paris cedex 15
France

Summary

Modulating the rate of mRNA degradation is a fast and efficient way to control gene expression. In a yeast strain deleted of *EDC3*, a component of the decapping machinery conserved in eukaryotes, the transcript coding the ribosomal protein Rps28b is specifically stabilized, as demonstrated by microarray and time course experiments. This stabilization results from the loss of *RPS28B* autoregulation, which occurs at the level of mRNA decay. Using mutants of the major deadenylase, we show that this regulation occurs at the level of decapping and bypasses deadenylation. Rps28b interacts with a conserved hairpin structure within the 3′UTR of its own mRNA and with components of the decapping machinery, including Edc3. We conclude that Rps28b, in the presence of Edc3, directly recruits the decapping machinery on its own mRNA. These findings show that specific modulation of the decapping efficiency on natural transcripts can control mRNA turnover.

Introduction

Although the regulation of transcription plays an essential role in controlling gene expression, the modulation of posttranscriptional events is also a key determinant of the control of gene expression. This is especially true when a very rapid response is needed or when the relative expression of different genes must be coordinated in order to ensure equimolar amounts of factors, for example because they are part of a common macromolecular complex. Virtually any posttranscriptional events can be regulated, but the most prominent ones are probably the regulation of mRNA turnover as well as the control of translation and protein degradation.

In higher eukaryotes, the decay rates of individual transcripts may vary by more than two orders of magnitude (see Tucker and Parker, 2000; Wilusz et al., 2001 for reviews). In the yeast *Saccharomyces cerevisiae*, a recent study showed that, under a single experimental condition, the half-life of individual transcripts varied from 3 to more than 90 min (Wang et al., 2002). Two general pathways of cytoplasmic mRNA degradation, both requiring deadenylation as an initiation step, have been characterized in yeast where this process has been best studied. In the major pathway, deadenylation is followed by decapping and rapid 5′ to 3′ exonucleolytic trimming of the mRNA. In a minor pathway, deadenylation is followed by 3′ to 5′ exonucleolytic degradation by the cytoplasmic form of the exosome complex (reviewed in Tucker and Parker, 2000; Wilusz et al., 2001). Many of the *cis*-acting elements that modulate the half-life of individual mRNAs influence the rate of the initial deadenylation step. Probably the best-studied examples of such *cis*-acting elements are the so-called AU-rich elements (AREs) present in the 3′ untranslated regions (3′UTRs) of a variety of eukaryotic mRNAs (Ross, 1996; Vasudevan and Peltz, 2001; Wilusz et al., 2001) and the sequences recognized by the factors of the PUF family (Wickens et al., 2002). Both types of elements exist in yeast (Duttagupta et al., 2003; Olivas and Parker, 2000; Tadauchi et al., 2001; Vasudevan and Peltz, 2001).

Striking examples of posttranscriptional regulation are found in the regulation of ribosomal protein translation in eubacteria and vertebrates (see for example Nomura and Meyuhas [1999] and Meyuhas [2000] for reviews). In rapidly dividing cells, ribosome biogenesis and translation use a major part of the cell energy and resources. This process involves the synthesis of large amounts of ribosomal proteins that must be produced in equimolar amounts with the ribosomal RNAs. This major synthesis effort thus requires tight regulation to modulate gene expression in order to coordinate ribosomal protein synthesis and to adapt it to the cell physiology. However, in yeast, in contrast to eubacteria or vertebrates, most of this regulation occurs at the transcriptional level (see Warner [1999] for review). Yet, a few examples have been described where ribosomal protein synthesis involves an additional layer of fine regulation. Until now, only four yeast ribosomal proteins have been found to autoregulate their synthesis by influencing different posttranscriptional steps. The first and best-studied case is Rpl30 (formely Rpl32) that binds to a pre-mRNA hairpin to inhibit splicing and binds to a stem-loop formed within the mature mRNA to inhibit translation (Eng and Warner, 1991; Li et al., 1996; Vilardell and Warner, 1994). Likewise, Rps14 binds to *RPS14B* pre-mRNA and inhibits splicing (Fewell and Woolford, 1999). Rpl4 (formerly Rpl2) regulates the *RPL4A* mRNA level by another feedback mechanism that involves an endonucleolytic cleavage followed by exonucleolytic degradation of the transcript (Presutti et al., 1995). Finally, Rps3 synthesis is regulated by a feedback mechanism that is likely to be posttranscriptional, but the molecular basis of this autoregulation remains unknown (Hendrick et al., 2001). In all cases, the autoregulation feedback loop uses the intrinsic RNA binding property of these ribosomal proteins to specifically recognize *cis*-regulatory elements within their own mRNAs.

In this report, we describe the fine control of expression that occurs at the level of mRNA decay for the yeast ribosomal protein gene *RPS28B*. This control occurs by a mechanism that is unique in two ways. First, the autoregulated protein Rps28 interacts both with a hairpin in the 3′UTR of its own mRNA and with components of the decapping machinery. Second, *RPS28B* mRNA

*Correspondence: jacquier@pasteur.fr

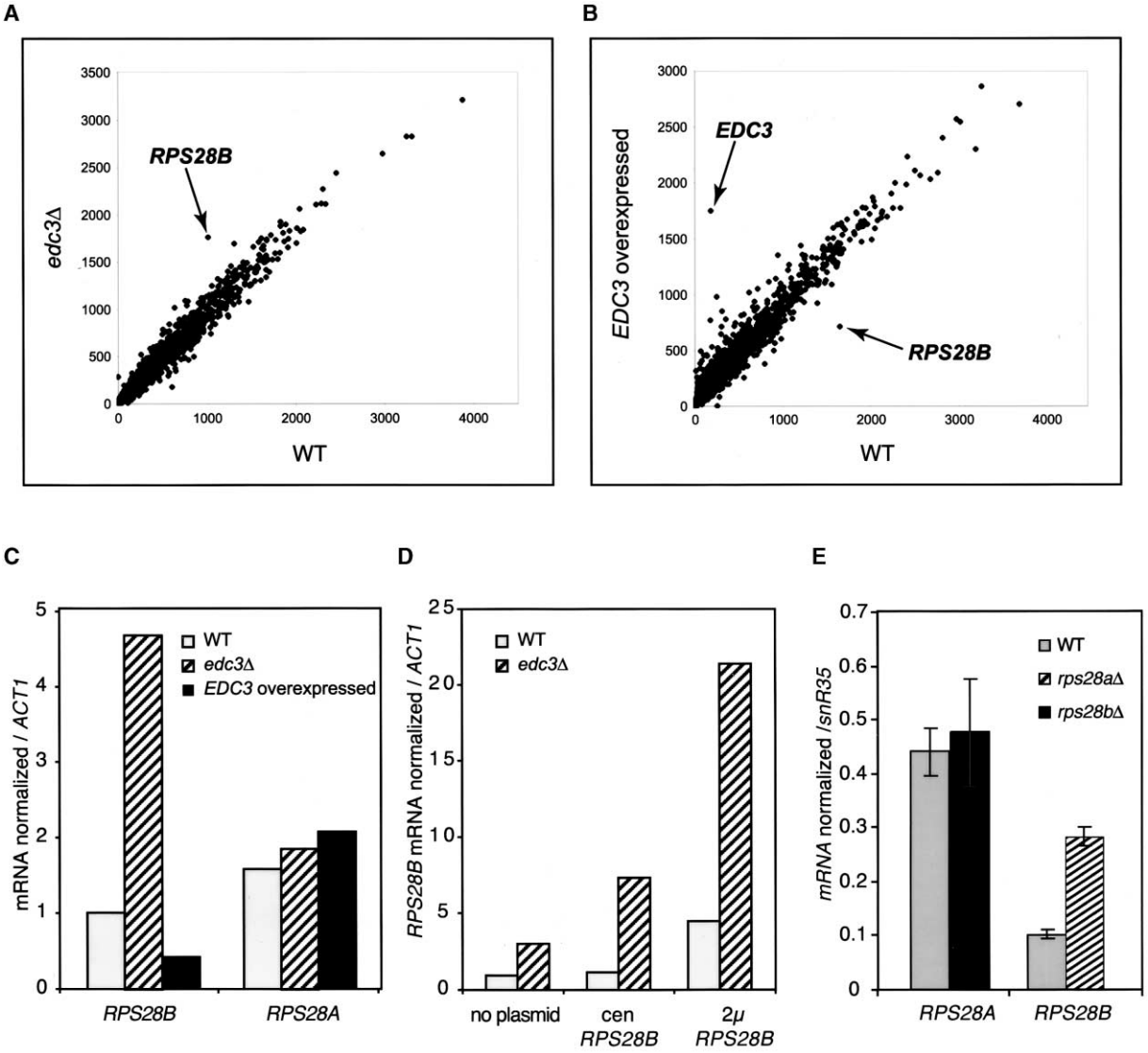


Figure 1. *EDC3* Mediates *RPS28B* Autoregulation

(A) Transcriptome comparison of mRNAs from wild-type and *edc3*Δ strains performed using Affymetrix DNA microarrays (YG-S98). The arrow points to the *RPS28B* transcript signal.

(B) Transcriptome comparison of mRNA from wild-type and *EDC3* overexpressing strains performed using Affymetrix DNA microarrays (YG-S98). The *RPS28B* and *EDC3* transcript signals are indicated by arrows.

(C) Quantitation of the *RPS28A* and *RPS28B* transcripts in wild-type, *edc3*Δ, and *EDC3* overexpressing strains. *RPS28B* and *RPS28A* mRNAs were analyzed by Northern blot using a radiolabeled oligonucleotide (GB096) that hybridizes with both *RPS28B* and *RPS28A* transcripts and were normalized with the *ACT1* transcripts (see Experimental Procedures).

(D) Northern blot analysis of *RPS28B* transcript in wild-type and *edc3*Δ strains, transformed with no, centromeric (+cen *RPS28B*), or multicopy (+2 μ *RPS28B*) plasmids carrying *RPS28B* gene (plasmid pRS315/*RPS28B* + 3′ and pRS425/*RPS28B* + 3′, respectively, see Experimental Procedures). The *RPS28B* transcript levels normalized to *ACT1* were determined using a Phosphorimager.

(E) Quantitation of the *RPS28A* and *RPS28B* transcripts in wild-type, *rps28a*Δ (LMA203), and *rps28b*Δ (LMA204) strains. *RPS28A* (left) and *RPS28B* (right) mRNAs were analyzed by Northern blot using a 5′ end-labeled oligonucleotide (GB096) that hybridizes with both *RPS28B* and *RPS28A* transcripts. The mRNAs were normalized relative to the *snR35* transcripts hybridized with oligonucleotide MFR523 (see Experimental Procedures).

degradation follows a deadenylation-independent pathway by direct stimulation of decapping that requires the newly identified enhancer of decapping Edc3. Unlike other deadenylation-independent mechanisms, like nonsense-mediated decay, this mechanism targets a natural transcript to add a supplemental layer of posttranscriptional regulation to an abundant mRNA.

Results

Rps28b Regulates the Decay of Its Own mRNA in an EDC3-Dependent Manner

Edc3 is a nonessential protein that we and others have shown to be physically linked to the decapping and degradation machinery. It was found to interact in two-

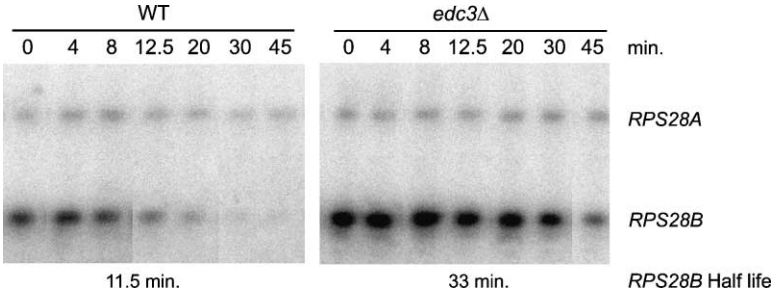


Figure 2. *EDC3* Influences the Stability of the *RPS28B* mRNA

The levels of the *RPS28B* mRNA were measured in a wild-type strain and in an *edc3*Δ strain by reverse transcription at different times after doxycycline-induced transcriptional repression. The plasmid pCM190/*RPS28B* + 3′UTR was introduced in strains deleted for the endogenous *RPS28B* gene in a wild-type or *edc3*Δ background (LMA204 and LMA225 strains). Doxycycline was added to the medium to turn off the expression of *RPS28B*. The resulting *RPS28B* transcripts and the endogenous *RPS28A* transcripts (used as a control) were analyzed by reverse transcription using oligonucleotide GB214 (see Experimental Procedures). These time courses were performed in triplicate and quantified with a Phosphorimager.

hybrid with Dcp1, Dcp2, Xrn1, Dhh1, and Lsm proteins (Fromont-Racine et al., 2000; Ito et al., 2001; Uetz et al., 2000). Edc3 also copurified with the Dcp1/Dcp2 complex isolated by tandem affinity purification (Gavin et al., 2002). These physical links suggested that Edc3 could be involved in the 5′ to 3′ degradation pathway of mRNAs. A strain deleted of *EDC3* shows no growth defect, whereas Edc3 overexpression is slightly toxic at 30°C (data not shown). The absence or the overexpression of Edc3 in otherwise wild-type cells did not result in any detectable effect on the stability of reporter mRNAs, such as the unstable *MFA2* transcript or the stable *PGK1* transcript (Roy Parker, personal communication; see also Kshirsagar and Parker, 2004). One possible explanation for the above results was that Edc3 is involved in mRNA decay but that its role is restricted to a subset of mRNAs. To test this hypothesis and to identify potential Edc3 mRNA targets, we used Affymetrix yeast DNA microarrays to search for transcripts that would be affected in their abundance by the presence of Edc3. The transcriptome profile of a wild-type strain and that of strains lacking or overproducing Edc3, respectively, were very similar (Figures 1A and 1B). One exception was obvious, the *RPS28B* transcript coding for the Rps28b ribosomal protein, the level of which was found significantly increased in the *edc3*Δ strain and significantly decreased in the strain overexpressing *EDC3* when compared to the wild-type strain. These microarray results were confirmed by Northen blotting (Figure 1C). Strikingly, the product of this mRNA, Rps28, was found to interact, in a two-hybrid assay, with Edc3 and Dcp1 (Ito et al., 2001). We independently confirmed the Edc3-Rps28b two-hybrid interaction (data not shown) and showed that an EDC3-Gal4 DNA binding domain fusion (EDC3-Gal4BD, expressed from pAS2ΔΔ/YEL015w; Fromont-Racine et al., 2000) is enriched in an Rps28b TAP purification from yeast, demonstrating that a biochemical complex contains both Rps28 and Edc3 in vivo (see Supplemental Figure S1 at http://www.molecule.org/cgi/content/full/15/1/5/DC1).

As for many ribosomal proteins in yeast, two genes, *RPS28A* and *RPS28B*, encode the Rps28 proteins Rps28a and Rps28b differing by only one amino acid. In contrast to *RPS28B*, the level of the *RPS28A* transcript remained unaffected by the presence or absence of *EDC3* (Figure 1C). Since the level of ribosomal proteins is tightly controlled, we tested if the observed *EDC3*-dependent variations of the *RPS28B* transcript level could reflect an autoregulation of this ribosomal protein. We found that doubling the copy number of *RPS28B* by introducing an extra copy of this gene on a single copy centromeric plasmid resulted in a 2-fold overexpression of the *RPS28B* transcript, relative to the actin mRNA taken as control, only when *EDC3* was absent. In contrast, no significant increase of the *RPS28B* mRNA was observed in wild-type cells (Figure 1D). This indicates that Rps28b is able to autoregulate the level of its own transcript in an Edc3-dependent manner. This autoregulation was also observed when *RPS28B* was expressed from a multicopy plasmid, the level of the *RPS28B* transcript being increased about 22-fold in *edc3*Δ cells, compared to only 5-fold in *EDC3* wild-type cells (Figure 1D, right). However, the fact that the level of RPS28B mRNA was not reduced to the level of the endogenous transcript when Edc3 was present suggests that the autoregulation mechanism is prone to saturation. Conversely, decreasing the overall amount of Rps28 proteins by deleting the *RPS28A* gene resulted in an increase of the *RPS28B* mRNA, while the deletion of *RPS28B* did not affect the amount of *RPS28A* mRNA (Figure 1E). Thus, the amount of *RPS28B* mRNA can be regulated by both Rps28a and Rps28b. In further experiments, the *RPS28A* transcript was taken as an internal control for the quantification of the *RPS28B* transcript. Altogether, these observations suggest that Rps28 regulates the level of *RPS28B* mRNA by forming a complex with Edc3 and the decapping machinery.

The physical links of Rps28b with the decapping/degradation machinery suggested that the autoregulation of the *RPS28B* mRNA occurs at the level of mRNA decay. We directly tested this hypothesis by measuring the half-life of the *RPS28B* transcript using a high-copy number plasmid in which *RPS28B* is under the control of a tetracycline-repressible transcription activator (plasmid pCM190/*RPS28B* + 3′UTR; see Experimental Procedures). The levels of overexpressed *RPS28B* and endogenous *RPS28A* mRNAs in strains deleted of the endogenous *RPS28B* gene and in the presence or absence of Edc3 were quantified by reverse transcription (Figure 2). The strong promoter of pCM190 and the multicopy nature of this plasmid resulted in a very strong overproduction of the *RPS28B* mRNA. Even in these conditions, the absence of Edc3 still resulted in a 2-fold increase of the *RPS28B* mRNA level, demonstrating that the *EDC3*-mediated regulation of *RPS28B* occurs independently of the nature of the promoter, consistent with a posttran-

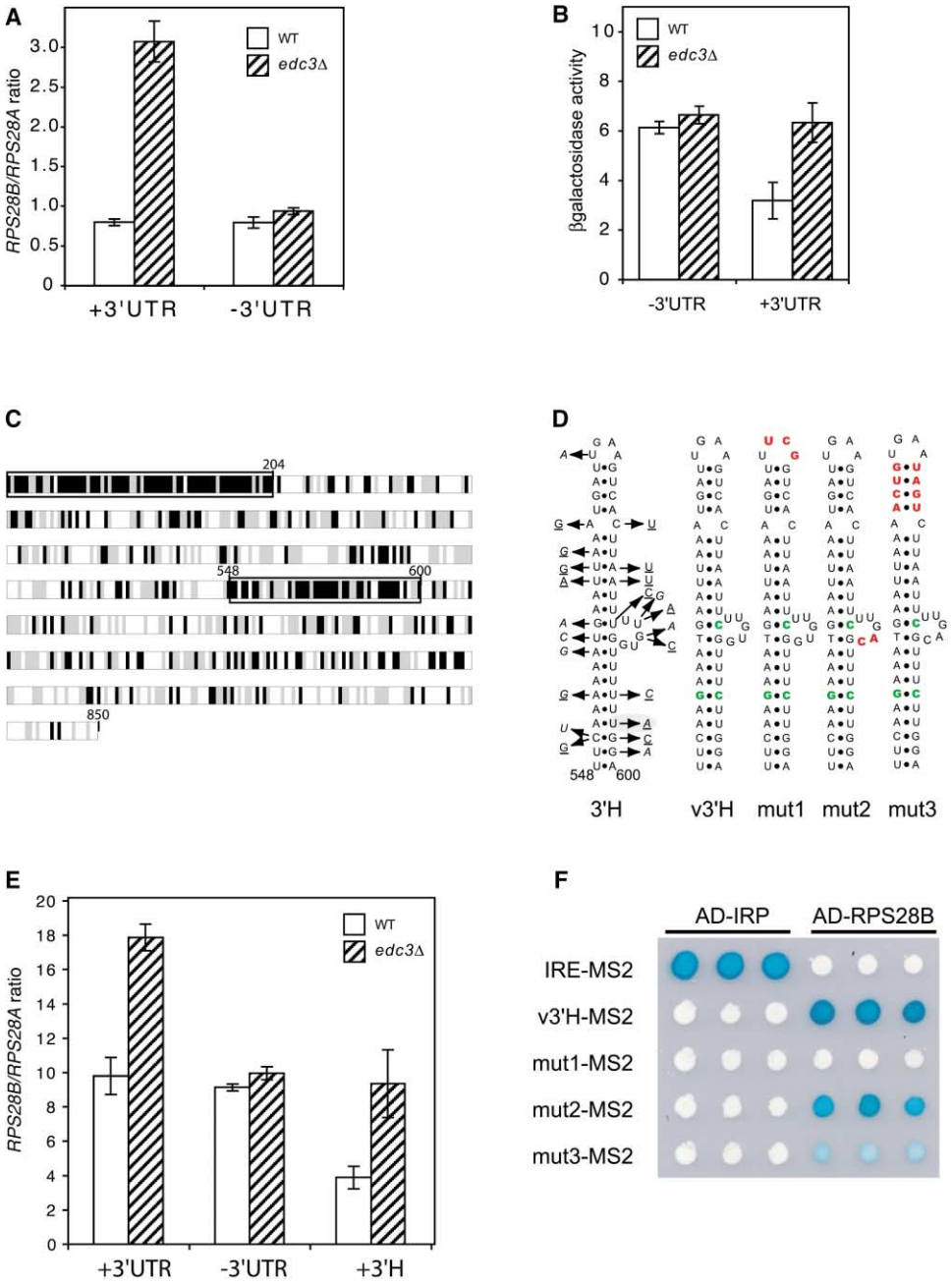


Figure 3. The RPS28B mRNA 3′UTR Contains a *cis* Regulatory Element

(A) *RPS28B* 3′UTR is required to promote the *EDC3*-mediated regulation of *RPS28B* mRNA. The ratios of the *RPS28B* over *RPS28A* transcripts were measured by primer extension (plasmids pRS315/*RPS28B* + 3′UTR and pRS315/*RPS28B* − 3′UTR in LMA204 and LMA225 strains; see Experimental Procedures). These experiments were performed in triplicate and quantified with a Phosphorimager.

(B) The *RPS28B* 3′UTR can modulate the expression of a *LacZ* reporter gene. Plasmids pCM190/LacZ + 3′UTR and pCM190/LacZ − 3′UTR (as control) were transformed in strains deleted for the endogenous *RPS28B* gene (but with endogenous *RPS28A* gene) in a wild-type or *edc3*Δ background. β-galactosidase activity was measured by a colorimetric assay using ONPG as described (Miller, 1972).

(C) Alignment of the *Saccharomyces cerevisiae RPS28B* 3′UTR sequence with the corresponding sequences in *Saccharomyces kluyveri*, *ZygoSaccharomyces rouxii* (from Souciet et al., 2000), *Saccharomyces bayanus*, and *Saccharomyces paradoxus* (from Kellis et al., 2003). The alignments were performed using the program *ClustalW* (Thompson et al., 1994). Nucleotides conserved in all five sequences are schematized by black boxes; nucleotides conserved in three or four species are schematized in gray.

(D) Secondary structure of the conserved region of *RPS28B* 3′UTR. The conserved region of *RPS28B* (from 548 to 600 downstream of the ATG) corresponds to a putative secondary structure predicted using the *Mfold* program (Walter et al., 1994) represented in the 3′H scheme (left). Arrows represent changes in other yeast species. Five out of eleven loop or bulge positions carry a substitution. Among the 21 base pairs that define the structure, five show compensatory base changes and two show conservative changes. Only one base pair exhibits noncompensatory changes (shaded in gray). Nucleotide substitution found in *Saccharomyces kluyveri* are shown as underlined characters and those found in *ZygoSaccharomyces rouxii* are in italics. v3′H schematizes the variant version of the *RPS28B* regulatory hairpin used during the three-hybrid experiment: naturally occurring substitutions, shown in green, were introduced in order to allow expression from PolIII

scriptional regulation mechanism (data not shown and Figure 2, time points 0). After turning off *RPS28B* transcription by addition of doxycycline, the half-life of the *RPS28B* transcript was significantly longer in the *edc3*Δ strain when compared to the wild-type strain (33 min versus 11.5 min). These results demonstrated that Edc3-dependent autoregulation of *RPS28B* occurs at the level of mRNA decay.

cis-Acting Elements Required for the Regulation of the RPS28B mRNA Decay

Since this autoregulation mechanism only affects the *RPS28B* mRNA and not the level of the *RPS28A* mRNA, we looked for differences in sequences between these two mRNAs. Northern blot analysis and sequencing of the cRT-PCR product of the *RPS28B* transcript (cRT-PCR is a method in which divergent RT-PCR is performed on circularized transcripts; see Couttet et al., 1997, and Experimental Procedures) showed that it starts 25 nucleotides upstream of the ATG initiation codon and that its cleavage/poly(A) termination site lies 646 nucleotides downstream of the stop codon (data not shown). The *RPS28A* transcript, which is not affected by Edc3, has a shorter 3′UTR (less than 150 nucleotides long). The unusual length of the *RPS28B* 3′UTR suggested that it could carry the *cis*-acting elements required for its *EDC3*-dependent autoregulation. To test this hypothesis, we cloned the *RPS28B* open reading frame (ORF) in a low-copy centromeric plasmid, under the control of its natural promoter but with or without its 3′UTR region (with an *ADH1* termination site). These plasmids were transformed in strains deleted of the endogenous *RPS28B* gene in a wild-type or *edc3*Δ background. The measure of the *RPS28B* over *RPS28A* transcripts ratio showed that the *EDC3*-dependent autoregulation of *RPS28B* requires its 3′UTR (Figure 3A). In addition, these data show that, in the absence of *EDC3*, the 3′UTR appears to stabilize the *RPS28B* transcript, pointing to a dual role of this region in transcript stabilization and as providing the target of the *EDC3*-dependent destabilization.

To further test the importance of the *RPS28B* 3′UT in the *EDC3*-dependent control of mRNA decay, we placed the 3′UTR downstream a *LacZ* coding frame under the control of a tetracycline-repressible transcription activator. We performed Northern blot analysis (data not shown) and measured the β-galactosidase activity, at steady state, to quantify *LacZ* expression in wild-type or *edc3*Δ cells transformed with these constructs. Figure 3B reports the β-galactosidase activity measured in conditions of induced expression (−Doxy) for the two *LacZ* constructs. An *EDC3*-dependent decrease of the β-galactosidase activity was observed when the *RPS28B*-3′UTR sequence was present downstream of the LacZ open reading frame but not when it was absent. This demonstrates that the *cis* regulatory element of the *RPS28B* mRNA is located within its 3′UTR.

Once the role of the *RPS28B* 3′UTR in regulating the transcript stability demonstrated, we searched for conserved sequences downstream the *RPS28B* ORF in other yeast species. A small region, comprised between nucleotides 548 to 600 downstream the ATG, appeared more conserved than the average sequences of the entire region (Figure 3C). Interestingly, RNA structure prediction using the *Mfold* program (Walter et al., 1994) showed that this sequence could adopt an RNA stem-loop structure. This potential RNA structure exhibits a large number of compensatory and conservative base pair changes, suggesting that it is under evolutionary constraint (Figure 3D, column 3′H). This sequence element was an obvious candidate for the *cis*-acting element mediating the *EDC3*-dependent regulation. To test its functional significance, we generated derivatives of the *RPS28B* transcript under control of a tetracycline-repressible promoter in which the 3′UTR region was deleted or replaced by the hairpin sequence alone. We next quantified the relative levels of these various constructs in comparison with the endogenous *RPS28A* transcript (Figure 3E). Consistent with the results presented in Figure 3A, we observed that the 3′UTR was required for the *edc3*Δ-dependent increase of the *RPS28B* transcripts generated from these constructs and that the 3′UTR appears to stabilize the mRNA in the absence of *EDC3*. When the hairpin alone replaced the 3′UTR, there was no stabilization of the *RPS28B* transcript in absence of *EDC3*, but the capacity of Edc3 to regulate the level of this transcript was restored. This demonstrates that the hairpin is a key element of the *EDC3*-mediated regulation of the *RPS28B* mRNA.

Rps28b Interacts with the Conserved RNA Hairpin in a Yeast Three-Hybrid System

In other examples of autoregulated ribosomal proteins, the protein was shown to directly interact with its own mRNA (Fewell and Woolford, 1999; Vilardell and Warner, 1994). If Rps28b autoregulates its mRNA at a posttranscriptional level, we expected that it would interact with its own mRNA. The above data also suggested that the conserved hairpin might be the element recognized by Rps28b. To test this hypothesis, we used a yeast three-hybrid system (Bernstein et al., 2002). Analogous to the two-hybrid system, this three-hybrid system depends upon the interaction of RNA and protein hybrids to reconstitute an active *trans*-activator for a reporter gene

(see Results). mut1, mut2, and mut3 carry mutations (in red) within the terminal loop (mut1), the internal bulge (mut2), or the four apical base pair of the stem (mut3) of the *RPS28B* regulatory hairpin.

(E) The hairpin regulatory element 3′H is sufficient to promote *RPS28B* autoregulation. The ratios of *RPS28B* over *RPS28A* were measured in LMA204 and LMA225 strains. The level of different *RPS28B* transcripts (from pCM190/*RPS28B* + 3′UTR, pCM190/*RPS28B* − 3′UTR, and pCM190/*RPS28B* + 3′H) relative to the level of the *RPS28A* transcripts were analyzed as in Figure 3A.

(F) The hairpin regulatory element interacts with Rps28b. Yeast three-hybrid assays for binding of IRP (as control) (AD-IRP) and Rps28b (AD-RPS28B) to four RNA hybrids (in triplicates): IRE-MS2, v3′H-MS2 carrying the natural variant of the *RPS28B* regulatory hairpin, and mut1-MS2, mut2-MS2, or mut3-MS2 that carry the mutations described above. An interaction between the protein and the RNA determines activation of the *LacZ* reporter gene and thus a blue color for the colonies.

in yeast. A number of specific RNA/protein interactions have already been demonstrated using this system, including IRE/IRP (SenGupta et al., 1996) that we used as control. *RPS28B* was fused in frame to the *GAL4* transcription activator domain in pACTIIst (AD-RPS28B). In the three-hybrid system that we used, the RNA hybrid is under the control of a polIII promoter, which does not allow the transcription through runs of five Ts (SenGupta et al., 1996). Therefore, we constructed a variant of the *RPS28B* regulatory stem-loop sequence, v3′H, by choosing mutations corresponding to naturally occurring substitutions within this element in other yeast species (nucleotides in green in Figure 3D). In addition, we generated three mutant versions of this construct, mut1, mut2, and mut3, which incorporated, respectively, substitutions of the conserved nucleotides in the terminal loop (we substituted the natural terminal tetraloop by the UUCG stable loop in order to conserve the hairpin structure), in the internal bulge, or in the terminal stem (nucleotides in red in Figure 3D). These sequences were cloned upstream the MS2 coat protein binding site in pIIIA/MS2-2 (Bernstein et al., 2002). The *LacZ* reporter gene is activated when either the IRE-MS2 and AD-IRP hybrids, used as positive controls, or the v3′H-MS2 and AD-RPS28B hybrids were present in the same cells (Figure 3F). In contrast, the combination of the IRE-MS2 RNA hybrid with the AD-RPS28B protein hybrid or the v3′H-MS2 and AD-IRP hybrids did not generate detectable β-galactosidase activity, even though the IRE element acquires, as the v3′H element, a stem-loop structure. In addition, Figure 3F shows that the substitution of three nucleotides of the conserved tetraloop (mut1-MS2) is sufficient to abolish the three-hybrid interaction, while the substitution of the two conserved nucleotides of the internal bulge (mut2-MS2) or in the four apical base pair of the stem (mut3-MS2) had only a partial detrimental effect for the interaction. We verified by Northern hybridization that these mutations did not affect the level of expression of the hybrid RNAs. These results indicate that Rps28b is able to interact specifically with the conserved hairpin present in its own mRNA and that the terminal loop of this hairpin is essential for this interaction.

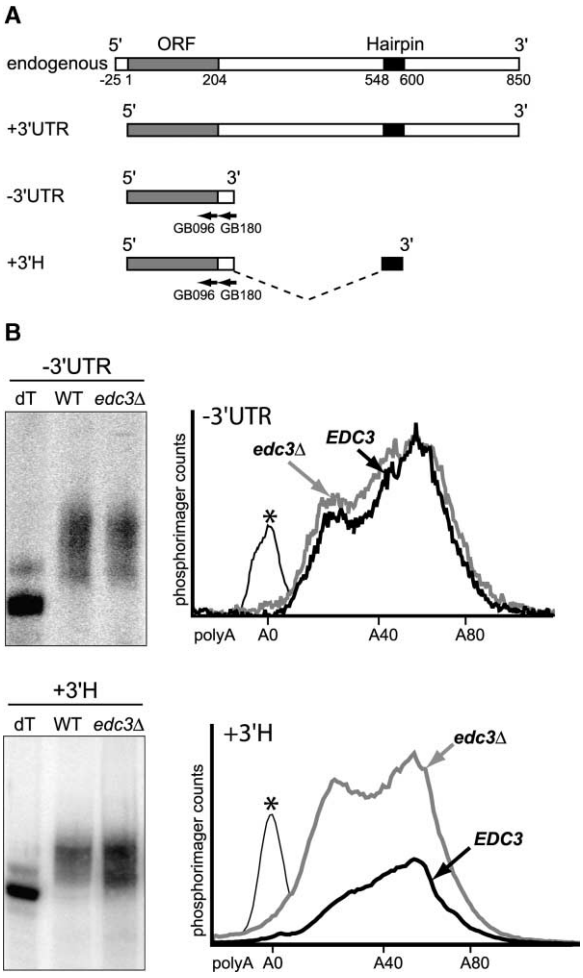


Figure 4. The *EDC3*-Mediated Decay of *RPS28B* mRNA Requires the Presence of the *cis* Regulatory Hairpin

(A) Schematic organization of the *RPS28B* transcripts. The oligonucleotides used to target the RNase H cleavage or used as [^{32}P]-labeled probes are shown by arrows.

(B) Analysis of poly(A) tail length of mutants of *RPS28B* 3′UTR mRNAs. The poly(A) tail average length of *RPS28B* with or without the conserved hairpin (+3′H or −3′UTR, see Experimental Procedures) was determined by RNase H treatment of mRNAs from wild-type (wt) or *edc3*Δ strains with GB096 primer and polyacrylamide Northern blotting using radiolabeled GB180 primer. Lanes marked dT correspond to wt samples treated with RNase H in presence of oligo dT18 to remove the poly(A) tail (as control). Phosphorimager profiles of the poly(A) tails from the mRNA expressed from constructs −3′UTR or +3′H, as indicated, in wt (*EDC3*, black line) or *edc3*Δ (gray lines) strains are shown in the right panels. The peak marked with an asterisk represents the profile of the mRNA after RNase H cleavage performed in the presence of oligo dT. The measures were quantified from the data shown in the left panel and normalized relative to the snRNA U4 probed with oligonucleotide MFR521 (data not shown).

The Edc3-Mediated RPS28B mRNA Decay Is Independent of Deadenylation

The *cis*-acting elements that modulate the half-life of individual mRNAs generally influence the rate of the initial deadenylation step. The physical links between Rps28 and Edc3 with the decapping/degradation machinery suggested that the Edc3-mediated *RPS28B* mRNA decay occurs by a direct enhancement of the decapping and not by affecting the rate of deadenylation. If Edc3 stimulates the decay of the *RPS28B* mRNA by enhancing its deadenylation, then the steady-state average length of the *RPS28B* mRNA poly(A) tail should be longer in an *edc3*Δ background compared to an *EDC3* background. In contrast, if Edc3 enhances the decapping rate, the average steady-state length of the *RPS28B* mRNA poly(A) tail should decrease following the deletion of *EDC3* because deadenylation would be less rate limiting relative to the subsequent decapping step or even bypassed. To directly address this question, we first compared the poly(A) tail length distribution of *RPS28B* mRNA transcripts deleted of the complete 3′UTR (−3′UTR) or retaining the conserved 3′ hairpin (+3′H; see Figure 4A), expressed from a high-copy number plasmid in either an *EDC3* or an *edc3*Δ strain. When the entire *RPS28B* 3′UTR was deleted, the presence or the absence of EDC3 did not change the profile of the poly(A)

tail length distribution. In contrast, when the conserved hairpin was present, the average length of the poly(A) tails was significantly longer in the wild-type versus the *edc3*Δ strain (Figure 4B). This observation is not compatible with the Edc3-mediated destabilization of the *RPS28B* mRNA resulting from an enhancement of the deadenylation rate. Moreover, it can also be seen on Figure 4B that, when the conserved hairpin is present, the deletion of *EDC3* not only results in the stabilization of the *RPS28B* transcripts with short poly(A) tails but also of the transcripts with long poly(A) tails. This suggests that, in the presence of the conserved hairpin, Edc3 can mediate the *RPS28B* transcripts degradation without deadenylation.

In order to directly test this hypothesis, we determined the poly(A) tail length profiles at different times after turning off transcription of *RPS28B* placed under the control of a tetracycline-repressible promoter and retaining the conserved hairpin as a 3′UTR (+3′H transcript in Figure 4A). In a wild-type strain, the decay of the transcript goes along with shortening of the poly(A) tails (Figures 5A and 5B). Thus, if the Edc3-mediated decay can bypass deadenylation, it does not preclude it. In absence of Edc3, short poly(A)-tailed transcripts accumulate before degradation, and the half-life of the mRNA is doubled, consistent with decapping being strongly limiting in this context. We thus tested the effects of changes in the deadenylation rates on the level and half-life of the *RPS28B* transcript. To this end, we placed *CCR4*, a core component of the major cytoplasmic deadenylation complex, under the control of a galactose-inducible promoter (*GAL::CCR4*). Strikingly, when *CCR4* is repressed in glucose, the *RPS28B* mRNA half-life is not significantly affected but its decay now occurs without poly(A) tail shortening (Figures 5A and 5B). This demonstrates that Edc3 mediates *RPS28B* mRNA decay by a mechanism that bypasses deadenylation. When both Edc3 and Ccr4 are absent, the half-life of the mRNA is strongly increased and slow shortening of the poly(A) tail is visible anew, once again consistent with the absence of Edc3 inhibiting decapping of this transcript.

Interestingly, when *GAL::CCR4* is expressed in the presence of galactose as the unique carbon source, the profile of the *RPS28B* mRNA poly(A) tails appeared significantly shifted toward the short poly(A) forms, consistent with deadenylation proceeding faster as a result of Ccr4 being overexpressed under these conditions. Remarkably, in this context, the absence of Edc3 resulted in the strong accumulation of *RPS28B* mRNAs with very short poly(A) tails (Figure 5C). In the absence of the conserved hairpin, the poly(A) tail profiles were identical whether Edc3 was present or not (data not shown). Once again, this observation can be explained by Edc3 being required for efficient decapping and 5′ to 3′ degradation of the *RPS28B* mRNA.

If Edc3 is required for decapping of the *RPS28B* mRNA, the absence of Dcp2, the catalytic subunit of the decapping enzyme, should have similar effects on this transcript as the absence of Edc3. Indeed, the *edc3*Δ and the *dcp2*Δ mutations have identical effects on the stabilization of the *RPS28B* mRNA and on the distribution of its poly(A) tail lengths (Figure 5D). These results suggest that, for this transcript, Edc3 is essential for efficient decapping.

In conclusion, our observations show that, in the presence of the conserved hairpin, Edc3 destabilizes the *RPS28B* mRNA by enhancing decapping independently of deadenylation.

Discussion

Degradation of specific mRNAs is an efficient mechanism to modulate the expression of given genes. Most of the genes known to be regulated at the level of mRNA turnover in *S. cerevisiae* are involved in major cellular pathways that require prompt adaptation to environmental changes (McCarthy, 1998). In this report, we show that the level of the *RPS28B* mRNA encoding the ribosomal protein Rps28b is subject to a fine level of autoregulation by a feedback mechanism acting on the rate of mRNA decay. This regulation involves a conserved *cis* regulatory hairpin, within the *RPS28B* 3′UTR, to which the Rps28 proteins are able to specifically bind. In addition, Rps28b not only binds this *cis*-acting regulatory RNA hairpin but also interacts with Lsm proteins Dcp1 and Edc3 (Ito et al., 2001; Uetz et al., 2000). We found that Edc3/Yel015w is required for the fine level of Rps28b regulation to take place. Like Dcp1, Edc3 is a factor physically associated with the decapping complex (Fromont-Racine et al., 2000; Ito et al., 2001; Uetz et al., 2000; Gavin et al., 2002). Very recently, Edc3 was shown to be an enhancer of decapping conserved in eukaryotes (Kshirsagar and Parker, 2004). Kshirsagar and Parker found that an *edc3*Δ mutation, like the *edc1*Δ and *edc2*Δ mutations, decreases the decapping of mRNA reporter transcripts when associated with a *dcp1* or a *dcp2* mutation that compromises the decapping activity and makes it rate limiting. They further bring compelling evidence that Edc3 is indeed a factor that enhances decapping.

Altogether, these observations suggested an attractive model in which Rps28b is able to directly recruit the decapping complex on its own mRNA, in an Edc3-dependent manner, in order to downregulate its expression by enhancing the decay of its transcript (Figure 6). Despite numerous efforts, we were unable to show a direct in vitro interaction between Rps28b and either the conserved RNA hairpin or Edc3. While these negative results might simply reflect our inability to produce a functional Rps28b protein in vitro, they prevent us to conclude that either of these interactions is direct. Nevertheless, the overall assumptions of this model remain valid, in particular the fact that the *RPS28B* mRNA turnover regulation results from a novel mechanism that modulates decapping. In support of this model, we show that, upon depletion of the major deadenylase Ccr4, the Edc3-mediated decay of *RPS28B* mRNA is retained, yet mRNA decay occurs without poly(A) shortening. This demonstrates that the Edc3-dependent degradation of the *RPS28B* mRNA bypasses deadenylation. In contrast, when Ccr4 is overexpressed, the absence of Edc3 results in the strong accumulation of deadenylated forms of the *RPS28B* mRNA. In addition, the distinctive stabilization and poly(A) tail length shift observed for the *RPS28B* mRNA in the absence of Edc3 are strikingly

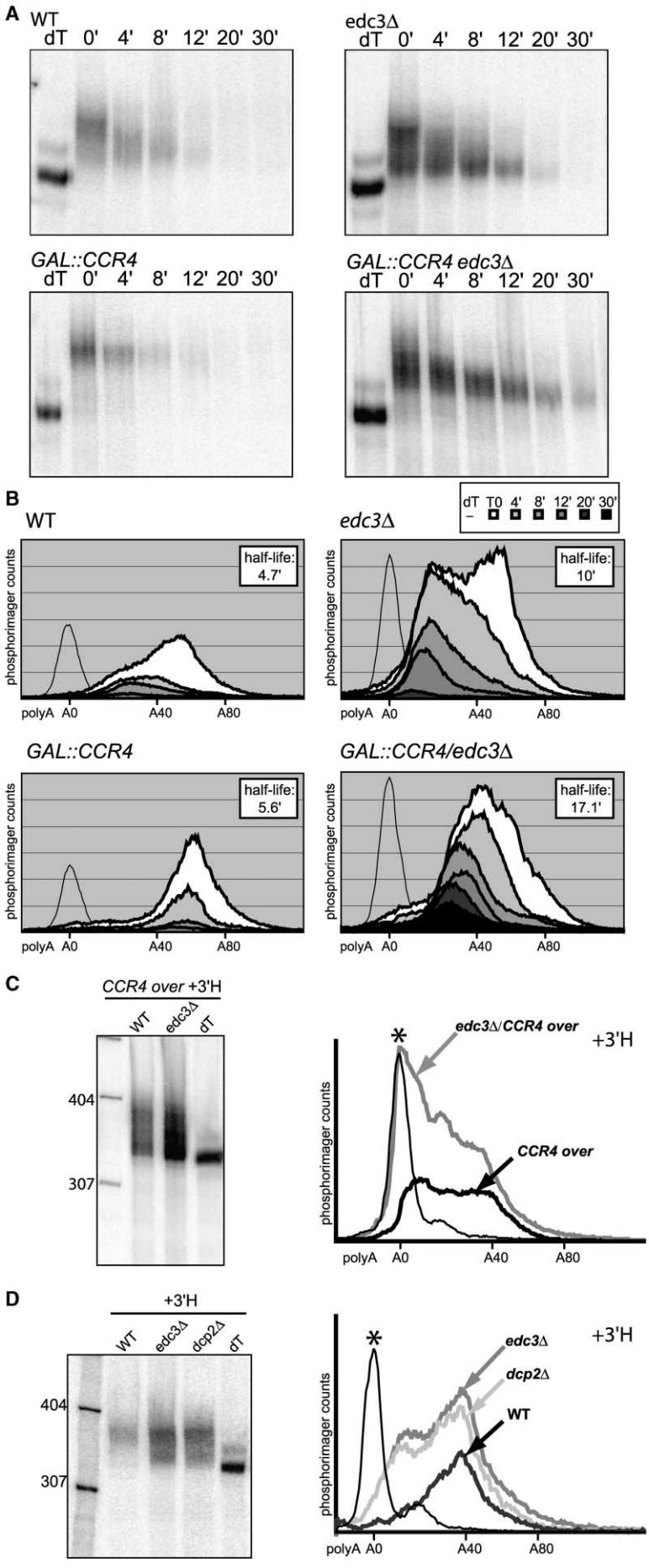


Figure 5. The *RPS28B* mRNA Regulatory Mechanism Is Independent of Deadenylation and Acts by Activation of Decapping

(A) Analysis of *RPS28B* mRNA poly(A) tail lengths after blocking transcription. The poly(A) tail average length of *RPS28B* transcript containing the hairpin (+3′H) was determined as described in Figure 4B. Total RNAs were extracted from cells taken at different time points after doxycycline-induced transcriptional repression. The plasmid pCM190/*RPS28B* + 3′H was introduced in *EDC3* wild-type or *edc3*Δ cells with or without a galactose-inducible promoter upstream the endogenous *CCR4* gene (wt, *edc3*Δ, *GAL::CCR4*, and *GAL::CCR4/edc3*Δ, corresponding to BMA64, LMA220, LMA328, and LMA329 strains, respectively; see Experimental Procedures). Strains *GAL::CCR4* and *GAL::CCR4/edc3*Δ were grown in galactose medium and transferred to glucose medium for 20 hr. Doxycycline was added to the medium to turn off the expression of *RPS28B*.

(B) Phosphorimager profiles of the poly(A) tails from the mRNA expressed from constructs containing the 3′UTR hairpin (3′H) in different strains with or without EDC3 (wt, *edc3*Δ, *GAL::CCR4*, and *GAL::CCR4/edc3*Δ). The signals were quantified from the data shown in (A) and normalized relative to the snRNA U4 as in Figure 4B. mRNA half-lives were measured from quantifications on reverse transcription products as described in Figure 2.

(C) Poly(A) tails distribution of *RPS28B* transcripts when *CCR4* is overexpressed. The poly(A) tail average length of *RPS28B* transcripts containing the hairpin (+3′H) was determined as described in Figure 4B. Total RNA was extracted from the *GAL::CCR4* and *GAL::CCR4/edc3*Δ strains transformed by pCM190/*RPS28B* + 3′H and grown in galactose medium. The poly(A) tail average length of *RPS28B* transcript was determined as described Figure 4B, in *CCR4* overexpressing conditions, with or without Edc3 (respectively *CCR4 over* in a wt or *edc3*Δ context). The corresponding Phosphorimager profiles are shown in the right panel.

(D) Poly(A) tails distribution of *RPS28B* transcripts when decapping is limiting. The plasmid pCM190/*RPS28B* + 3′H was introduced in a wild-type, *edc3*Δ, or *dcp2*Δ background (BMA64, LMA220, and LMA222). The poly(A) tail average length of *RPS28B* transcript was determined as described in Figure 4B. The corresponding Phosphorimager profiles are shown in the right panel.

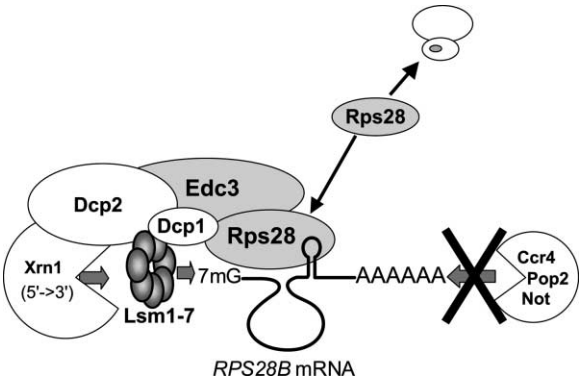


Figure 6. A Model for the Edc3-Mediated Autoregulation of the *RPS28B* mRNA

In the presence of excess Rps28, the protein binds to the conserved hairpin in the *RPS28B* mRNA 3′UTR and recruits the decapping machinery via an interaction with components of the decapping complex.

similar to those observed in the absence of Dcp2, the catalytic subunit of the decapping complex (Steiger et al., 2003; Van Dijk et al., 2002). We thus conclude that Edc3 is required for the efficient decapping and regulation of the *RPS28B* mRNA.

One intriguing question is why should only one of the two copies of the *RPS28* genes be regulated? This could be sufficient to obtain the required fine level of regulation of Rps28 expression. However, we recently made preliminary observations suggesting that the expression of the *RPS28A* gene could be also autoregulated but at another level, such as translation (Claire Torchet and A.J., unpublished data).

Several characteristics of the mechanism remain to be established. For example, it is not clear at the moment whether the Edc3-dependent degradation of the *RPS28B* mRNA occurs in the nucleus or the cytoplasm. However, since Edc3 and the components of the decapping machinery in general have been reported to be present in cytoplasmic P bodies (Huh et al., 2003; Kshirsagar and Parker, 2004; Sheth and Parker, 2003), we think it is likely that the regulation also takes place in these structures.

The mechanism that we describe here shows analogies with the deadenylation-independent nonsense-mediated decay (NMD; for review see Gonzalez et al., 2001; Wilusz et al., 2001). Previous studies suggested that wild-type transcripts (Lelivelt and Culbertson, 1999) or transcripts with extended 3′UTRs (Muhlrad and Parker, 1999) could be substrates for NMD-dependent degradation. However the *EDC3*-mediated decay of *RPS28B* mRNA does not seem to make use of the NMD factors because we could not observe any differences between the levels of the *RPS28B* mRNAs in wild-type or *upf1*Δ, *upf2*Δ, and *upf3*Δ strains (data not shown). The *EDC3*-mediated decay of the *RPS28B* mRNA thus appears independent of the NMD pathway, a conclusion consistent with the finding that *EDC3* does not affect the NMD pathway (Kshirsagar and Parker, 2004).

Another important question is whether other transcripts, in addition to the *RPS28B* mRNA, are regulated by this mode of regulation involving Edc3. The DNA-microarray experiments did not allow the identification of other transcripts regulated in an Edc3-dependent manner. However, transcripts expressed at low levels could have easily escaped our search (this should, a priori, exclude other ribosomal protein gene transcripts). Therefore, in order to address this question by other means, we performed a synthetic-lethality screen with the *edc3*Δ mutant. We found two different mutant alleles of the same gene, *NDD1*, to be synthetic lethal with *edc3*Δ (data not shown). *NDD1* encodes a transcriptional regulator involved in the G2/M checkpoint (Koranda et al., 2000). Importantly, this synthetic-lethal phenotype appeared independent of the increased level of the *RPS28B* mRNA observed in *edc3*Δ cells since the overexpression of *RPS28B* on a multicopy plasmid was not synthetic lethal with the *ndd1* mutations (data not shown). It is thus possible that at least one other transcript, involved in the same biological pathway as *NDD1*, is regulated in an Edc3-dependent manner.

Edc3 exhibits surprisingly antagonistic characteristics. On one hand, it shows features of a general decapping factor: it is found biochemically associated with the enzymatic subunit of the decaping complex Dcp2 (Gavin et al., 2002) and, as the general decapping cofactor Dcp1, is an abundant protein located within the P bodies (Ghaemmaghami et al., 2003; Huh et al., 2003; Kshirsagar and Parker, 2004; Sheth and Parker, 2003). Moreover, when decapping is compromised by mutations within *DCP1* or *DCP2*, the absence of Edc3 induces significant defects in the decapping of generic reporter transcripts (Kshirsagar and Parker, 2004). In contrast, in otherwise wild-type cells, our microarray analyses show that Edc3 does not affect the vast majority of the transcripts. It is thus likely that, if Edc3 plays a general role in decapping, this function is never rate limiting for mRNA degradation in standard growth conditions. The role of Edc3 is revealed only when it is enrolled in a mechanism that bypasses the rate limiting deadenylation step.

Experimental Procedures

Yeast Strains and Plasmids

Yeast strains used: BMA64: *MAT*α, *ura3-1*, Δ*trp1*, *ade2-1*, *leu2-3*, *112*, *his3-11,15* (from F. Lacroute, CNRS); haploid LMA36-5A, LMA36-5B, LMA36-5C, and LMA36-5D, issued from the sporulation of a heterozygous diploid deleted for *EDC3/YEL015W* obtained by transformation of genomic PCR product with D3′-YEL015wk and D5′-YEL015wk primers and pFA6a-KanMX6-pGAL1 as a template (Longtine et al., 1998) had a wild-type (LMA36-5A: *MAT*α and LMA36-5B: *MATa*) or *edc3*Δ*::KAN* deletion (LMA36-5C: *MATa* and LMA36-5D: *MAT*α) genotype; LMA203 is BMA64 containing *rps28a*Δ*::HIS3* and was obtained by transformation of a genomic PCR product with the GB157 and GB158 primers and pFA6a-*HIS3*-pGAL1 as a template (Longtine et al., 1998); LMA204 is BMA64 containing *rps28b*Δ*::TRP1* and was obtained by transformation of a genomic PCR product with the GB155 and GB156 primers and pFA6a-*TRP1*-pGAL1 as a template (Longtine et al., 1998); LMA220 was generated from LMA36-5D with the *KAN* marker replaced by the nourseothrycine resistance gene (*NAT*); LMA222 was obtained from BMA64, contains *dcp2*Δ*::KAN*, and was generated by transformation of a genomic PCR product with the GB193 and GB194 primers and genomic DNA from a *dcp2*Δ strain in BY4741 as a template; LMA225 was obtained from BMA64 and contains *rps28b*Δ*::TRP1* and *edc3*Δ*::KAN*; LMA328 and LMA329 are wild-type and *edc3*Δ strains containing the Gal promotor upstream the *CCR4* gene (*GAL::CCR4*). LMA328 and LMA329 were obtained by transformation

of BMA64 and LMA220 strains with a genomic PCR product obtained with CS47 and CS50 primers and pFA6a-*HIS3*-pGAL1 as a matrix.

Oligonucleotides used in this study are listed in Supplemental Table S1 on *Molecular Cell*'s website.

Plasmid constructions are detailed in Supplemental Data available on *Molecular Cell*'s website.

RNA Extraction and Northern Blotting

Total RNA extractions and Northern blot analysis were performed as described (Saveanu et al., 2001), except for RNAs used to perform primer extension and time course experiments that were extracted using the Epicentre kit (TEBU).

Primer Extension and Time Course

In 7 μl of buffer, 0.2 pmol of [^{32}P] 5′ end-labeled oligonucleotide GB214 complementary to *RPS28B* and *RPS28A* ORFs was annealed with 5 μg of total RNA. Extension was carried out using Superscript II reverse transcriptase (Invitrogen) for 30 min at 42°C in 1× buffer, 10 mM dithiothreitol, 0.5 μg actinomycin D, dGTP, dCTP, dTTP (0.5 mM each), and ddATP (0.04 mM). The reaction was stopped by addition of formamide-containing buffer.

RNase H Protection Assay

Deadenylation was performed as described (Muhlrad and Parker, 1992). Samples of 8 μg of RNA were annealed with 400 ng of oligonucleotide GB096 and ±400 ng oligo-dT18 at 68°C for 10 min and digested with 1.5 U RNase H at 30°C for 1 hr.

cRT-PCR

cRT-PCR was performed as described in Couttet et al. (1997), using GB196 as primer for the reverse transciption after mRNA self-ligation step and GB197 and GB203 for PCR amplification. Sequence was performed with GB197 and GB203 primers.

Acknowledgments

G.B. was supported by a CIFRE (Convention Industrielle de Formation par la Recherche en Entreprise) contract with Hybrigenics SA and the Association Nationale de la Recherche Technique (ANRT). G.B. and C.S. were supported by a grant from the European Commission (RNOMICS: QLG2-CT-2001-01554). This work was supported in part by the grant "Subventions puces Affymetrix" from the Ministry of Reaserch. We thank A. Brunet-Simon for initial work; L. Decourty for excellent technical assistance; and M. Kshirsagar and R. Parker for communicating results prior to publication.

Received: January 19, 2004
Revised: May 5, 2004
Accepted: May 11, 2004
Published: July 1, 2004

References

Bernstein, D.S., Buter, N., Stumpf, C., and Wickens, M. (2002). Analyzing mRNA-protein complexes using a yeast three-hybrid system. Methods *26*, 123–141.

Couttet, P., Fromont-Racine, M., Steel, D., Pictet, R., and Grange, T. (1997). Messenger RNA deadenylylation precedes decapping in mammalian cells. Proc. Natl. Acad. Sci. USA *94*, 5628–5633.

Duttagupta, R., Vasudevan, S., Wilusz, C.J., and Peltz, S.W. (2003). A yeast homologue of Hsp70, Ssa1p, regulates turnover of the MFA2 transcript through its AU-rich 3′ untranslated region. Mol. Cell. Biol. *23*, 2623–2632.

Eng, F.J., and Warner, J.R. (1991). Structural basis for the regulation of splicing of a yeast messenger RNA. Cell *65*, 797–804.

Fewell, S.W., and Woolford, J.L., Jr. (1999). Ribosomal protein S14 of Saccharomyces cerevisiae regulates its expression by binding to RPS14B pre-mRNA and to 18S rRNA. Mol. Cell. Biol. *19*, 826–834.

Fromont-Racine, M., Mayes, A.E., Brunet-Simon, A., Rain, J.C., Colley, A., Dix, I., Decourty, L., Joly, N., Ricard, F., Beggs, J.D., et al. (2000). Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. Yeast *17*, 95–110.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141–147.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.

Gonzalez, C.I., Bhattacharya, A., Wang, W., and Peltz, S.W. (2001). Nonsense-mediated mRNA decay in Saccharomyces cerevisiae. Gene *274*, 15–25.

Hendrick, J.L., Wilson, P.G., Edelman, I.I., Sandbaken, M.G., Ursic, D., and Culbertson, M.R. (2001). Yeast frameshift suppressor mutations in the genes coding for transcription factor Mbf1p and ribosomal protein S3: evidence for autoregulation of S3 synthesis. Genetics *157*, 1141–1158.

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686–691.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA *98*, 4569–4574.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241–254.

Koranda, M., Schleiffer, A., Endler, L., and Ammerer, G. (2000). Forkhead-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters. Nature *406*, 94–98.

Kshirsagar, M., and Parker, R. (2004). Identification of Edc3p as an enhancer of mRNA decapping in *Saccharomyces cerevisiae*. Genetics *166*, 729–739.

Lelivelt, M.J., and Culbertson, M.R. (1999). Yeast Upf proteins required for RNA surveillance affect global expression of the yeast transcriptome. Mol. Cell. Biol. *19*, 6710–6719.

Li, B., Vilardell, J., and Warner, J.R. (1996). An RNA structure involved in feedback regulation of splicing and of translation is critical for biological fitness. Proc. Natl. Acad. Sci. USA *93*, 1596–1600.

Longtine, M.S., McKenzie, A., 3rd, Demarini, D.J., Shah, N.G., Wach, A., Brachat, A., Philippsen, P., and Pringle, J.R. (1998). Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. Yeast *14*, 953–961.

McCarthy, J.E. (1998). Posttranscriptional control of gene expression in yeast. Microbiol. Mol. Biol. Rev. *62*, 1492–1553.

Meyuhas, O. (2000). Synthesis of the translational apparatus is regulated at the translational level. Eur. J. Biochem. *267*, 6321–6330.

Miller, J.H. (1972). Experiments in Molecular Genetics (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press).

Muhlrad, D., and Parker, R. (1992). Mutations affecting stability and deadenylation of the yeast MFA2 transcript. Genes Dev. *6*, 2100–2111.

Muhlrad, D., and Parker, R. (1999). Aberrant mRNAs with extended 3′ UTRs are substrates for rapid degradation by mRNA surveillance. RNA *5*, 1299–1307.

Nomura, M., and Meyuhas, O. (1999). Regulation of ribosome biosynthesis in Escherichia coli and Saccharomyces cerevisiae: diversity and common principles. J. Bacteriol. *181*, 6857–6864.

Olivas, W., and Parker, R. (2000). The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast. EMBO J. *19*, 6602–6611.

Presutti, C., Villa, T., Hall, D., Pertica, C., and Bozzoni, I. (1995). Identification of the cis-elements mediating the autogenous control of ribosomal protein L2 mRNA stability in yeast. EMBO J. *14*, 4022–4030.

Ross, J. (1996). Control of messenger RNA stability in higher eukaryotes. Trends Genet. *12*, 171–175.

Saveanu, C., Bienvenu, D., Namane, A., Gleizes, P.E., Gas, N., Jacquier, A., and Fromont-Racine, M. (2001). Nog2p, a putative GTPase

associated with pre-60S subunits and required for late 60S maturation steps. EMBO J. *20*, 6475–6484.

SenGupta, D.J., Zhang, B., Kraemer, B., Pochart, P., Fields, S., and Wickens, M. (1996). A three-hybrid system to detect RNA-protein interactions in vivo. Proc. Natl. Acad. Sci. USA *93*, 8496–8501.

Sheth, U., and Parker, R. (2003). Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. Science *300*, 805–808.

Souciet, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., de Montigny, J., Dujon, B., et al. (2000). Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. FEBS Lett. *487*, 3–12.

Steiger, M., Carr-Schmid, A., Schwartz, D.C., Kiledjian, M., and Parker, R. (2003). Analysis of recombinant yeast decapping enzyme. RNA *9*, 231–238.

Tadauchi, T., Matsumoto, K., Herskowitz, I., and Irie, K. (2001). Post-transcriptional regulation through the HO 3′-UTR by Mpt5, a yeast homolog of Pumilio and FBF. EMBO J. *20*, 552–561.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Tucker, M., and Parker, R. (2000). Mechanisms and control of mRNA decapping in Saccharomyces cerevisiae. Annu. Rev. Biochem. *69*, 571–595.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature *403*, 623–627.

Van Dijk, E., Cougot, N., Meyer, S., Babajko, S., Wahle, E., and Seraphin, B. (2002). Human Dcp2: a catalytically active mRNA decapping enzyme located in specific cytoplasmic structures. EMBO J. *21*, 6915–6924.

Vasudevan, S., and Peltz, S.W. (2001). Regulated ARE-mediated mRNA decay in *Saccharomyces cerevisiae*. Mol. Cell *7*, 1191–1200.

Vilardell, J., and Warner, J.R. (1994). Regulation of splicing at an intermediate step in the formation of the spliceosome. Genes Dev. *8*, 211–220.

Walter, A.E., Turner, D.H., Kim, J., Lyttle, M.H., Muller, P., Mathews, D.H., and Zuker, M. (1994). Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proc. Natl. Acad. Sci. USA *91*, 9218–9222.

Wang, Y., Liu, C.L., Storey, J.D., Tibshirani, R.J., Herschlag, D., and Brown, P.O. (2002). Precision and functional specificity in mRNA decay. Proc. Natl. Acad. Sci. USA *99*, 5860–5865.

Warner, J.R. (1999). The economics of ribosome biosynthesis in yeast. Trends Biochem. Sci. *24*, 437–440.

Wickens, M., Bernstein, D.S., Kimble, J., and Parker, R. (2002). A PUF family portrait: 3′ UTR regulation as a way of life. Trends Genet. *18*, 150–157.

Wilusz, C.J., Wormington, M., and Peltz, S.W. (2001). The cap-to-tail guide to mRNA turnover. Nat. Rev. Mol. Cell Biol. *2*, 237–246.

Cell, Vol. 121, 725–737, June 3, 2005, Copyright ©2005 by Elsevier Inc. DOI 10.1016/j.cell.2005.04.030

Cryptic Pol II Transcripts Are Degraded by a Nuclear Quality Control Pathway Involving a New Poly(A) Polymerase

Françoise Wyers,[4,5,7] Mathieu Rougemaille,[5,7] Gwenaël Badis,[1] Jean-Claude Rousselle,[2] Marie-Elisabeth Dufour,[4] Jocelyne Boulay,[5] Béatrice Régnault,[3] Frédéric Devaux,[6] Abdelkader Namane,[2] Bertrand Séraphin,[2,5,*] Domenico Libri,[5,*] and Alain Jacquier[1,*]

[1]Unité de Génétique des Interactions Macromoléculaires
CNRS-URA2171-Institut Pasteur
[2]PT "Proteomique"
CNRS-URA2185-Institut Pasteur
[3]PT "Puces à ADN"
Institut Pasteur
25 rue du Docteur Roux
75724 Paris Cedex 15
[4]Equipe Labelisée La Ligue
[5]Centre de Génétique Moléculaire
CNRS-UPR2167
Avenue de la Terrasse
91190 Gif sur Yvette
[6]Laboratoire de Génétique Moléculaire
Ecole Normale Supérieure
CNRS-UMR 8541
46 rue d'Ulm
75005 Paris
France

Summary

Since detection of an RNA molecule is the major criterion to define transcriptional activity, the fraction of the genome that is expressed is generally considered to parallel the complexity of the transcriptome. We show here that several supposedly silent intergenic regions in the genome of *S. cerevisiae* are actually transcribed by RNA polymerase II, suggesting that the expressed fraction of the genome is higher than anticipated. Surprisingly, however, RNAs originating from these regions are rapidly degraded by the combined action of the exosome and a new poly(A) polymerase activity that is defined by the Trf4 protein and one of two RNA binding proteins, Air1p or Air2p. We show that such a polyadenylation-assisted degradation mechanism is also responsible for the degradation of several Pol I and Pol III transcripts. Our data strongly support the existence of a posttranscriptional quality control mechanism limiting inappropriate expression of genetic information.

Introduction

Most, if not all, eukaryotic primary transcripts, whether transcribed by RNA polymerase (Pol) I, II, or III, undergo maturation, which includes endonucleolytic severing, exonucleolytic trimming, splicing, nucleotide modifications/edition, and/or capping. Interestingly, most mature 3′ ends are generated by processing. Some result from cleavage and/or trimming, while others are extended by polymerases of the β family, which add 3′ tails without the help of a DNA template: CCA for tRNA (Schurer et al., 2001) or poly(A) sequences for mRNA (Proudfoot and O'Sullivan, 2002). Addition of poly(A) tails to mRNAs occurs at an endonucleolytic cleavage site that is severed cotranscriptionally. The Pap1p protein has poly(A) polymerase activity but depends on the assembly of a large complex for its function, which also insures correct positioning and control the length of the poly(A) tail (Keller and Minvielle-Sebastia, 1997). In eucaryotes, polyadenylation of coding RNAs has at least three important functions: it is required for RNA stability, efficient nucleocytoplasmic export, and translation. Except for a limited number of exclusively nuclear species, most transcripts reach the cytoplasm, where a large majority contribute to protein synthesis.

In this canonical view of the RNA synthesis pathway, the expressed fraction of the genome in a given cell is determined by accurate promoter selection. Transcription from these landmarks generates primary transcripts that are matured into functional RNA molecules, while the remaining fragments of the primary transcripts (e.g., introns, 3′ trailers, as well as 3′ extensions and internal spacers of the pre-rRNAs and pre-tRNAs) are rapidly degraded (e.g., Kim et al. [2004], West et al. [2004]). These events may explain the high rate of degradation of a fraction of nuclear RNAs never reaching the cytoplasm (e.g., Egyhazi [1976]). In addition to degradation events targeting short-lived processing intermediates, specific nuclear RNA decay pathways also destroy aberrant pre-mRNAs or those failing to be exported (Bousquet-Antonelli et al., 2000; Das et al., 2003; Libri et al., 2002; Torchet et al., 2002). In yeast, Rat1p and the exosome are two exonucleases implicated in these nuclear degradation processes: Rat1p is a 5′–3′ exonuclease showing sequence similarity to Xrn1, the major 5′–3′ exonuclease involved in cytoplasmic mRNA decay (Johnson, 1997). Rat1p is mostly nuclear and has been implicated in the maturation of pre-rRNAs and snoRNAs (Petfalski et al., 1998) and in transcription termination (Kim et al., 2004). The exosome is a large complex of 3′–5′ nucleases that is found both in the nucleus and in the cytoplasm (Mitchell and Tollervey, 2000). The nuclear form of the complex contains two specific subunits, Rrp6p and Lrp1p, that are its only nonessential subunits. The exosome has been implicated in numerous nuclear RNA processing and degradation events including pre-rRNA and sn(o)RNA maturation (Allmang et al., 1999a; Petfalski et al., 1998) and the turnover of pre-mRNAs in processing/splicing and RNA export mutants (Bousquet-Antonelli et al., 2000; Das et al., 2003; Libri et al., 2002; Torchet et al., 2002). Presence of both Rat1p and exosome homologs in eukaryotic species suggests that the cognate decay pathways are evolutionarily conserved.

*Correspondence: bertrand.seraphin@cgm.cnrs-gif.fr (B.S.); dominico.libri@cgm.cnrs-gif.fr (D.L.); jacquier@pasteur.fr (A.J.)
[7]These authors contributed equally to this work.

To identify new targets for the nuclear exosome, we have analyzed the transcriptome of a strain lacking Rrp6p with DNA microarrays. In accordance with previous studies, polyadenylated forms of numerous Pol II and Pol III noncoding RNAs (Allmang et al., 1999a; Kadaba et al., 2004; van Hoof et al., 2000) and of transcripts derived from the rDNA locus (Kuai et al., 2004) accumulated. Surprisingly, we identified in addition new polyadenylated transcripts mapping to intergenic regions. These were characterized as novel Pol II transcription units. Accumulation of the cognate RNAs in a *rrp6* mutant results from their stabilization rather than from transcriptional activation. Interestingly, most poly(A) additions to these transcripts are not mediated by the classical polyadenylation machinery. Database searches revealed the presence of other potential poly(A) polymerases encoded by the yeast genome, including Trf4p that has recently independently been shown to be required for polyadenylation and degradation of hypomodified forms of tRNAmet (Kadaba et al., 2004). Consistently, we show that Trf4p associates with Air1p and Air2p to form a new enzyme endowed with polyadenylation activity. This complex associated with Mtr4p, a putative RNA helicase previously implicated in activation of the nuclear exosome (de la Cruz et al., 1998; Liang et al., 1996). Most importantly, the polyadenylation of most cryptic transcripts derived from intergenic regions detected in the *rrp6Δ* background was nearly completely abolished in the absence of Trf4p, leading to their further stabilization. Taken together, these results demonstrate that a novel yeast nuclear poly(A) polymerase is implicated in a quality control process targeting numerous RNA to degradation by the exosome. Notably, this mechanism appears to limit the genomic noise resulting from inappropriate transcription of intergenic regions in the genome. These observations have several evolutionary implications.

Results

New Cryptic Transcripts Accumulate in the Absence of Rrp6p

The role of the Rrp6p exonuclease in the nuclear turnover of Pol II transcripts is still unclear. To identify new Rrp6p targets, we compared the transcriptomes of an *rrp6Δ* and a wild-type strain using Affymetrix DNA microarrays spanning the entire yeast ORFeome as well as some noncoding RNAs and intergenic regions. Two microarrays were used with RNAs from a wild-type strain (BMA64; see Table S1 in the Supplemental Data available with this article online) and two with RNAs from an *rrp6Δ* strain (LMA164), and the signal intensities were compared (Figure 1A). Importantly, the fluorescent probes were generated from total RNAs using oligo(dT) as primer and were thus enriched for probes against polyadenylated RNAs. While the vast majority of the cellular ORF-containing transcripts did not differ between the two strains (only 5.1% of verified ORFs reproducibly exhibit an *rrp6Δ*/WT ratio >2), a number of signals increased significantly in the *rrp6* mutant compared to the wild-type. Signals corresponding to almost all snRNAs and snoRNAs and directly downstream sequences strongly increased in an *rrp6Δ* background (red dots, Figure 1A), presumably as a consequence of the previously reported polyadenylation of such transcripts in the absence of Rrp6p (Allmang et al. [1999a], Allmang et al. [1999b], van Hoof et al. [2000], and see below). Likewise, signals corresponding to several rRNA species increased dramatically (yellow dots, Figure 1A), consistent with the reported stabilization of polyadenylated forms of these transcripts in *rrp6* mutants (Kuai et al. [2004] and see below; van Hoof et al. [2000]). Surprisingly, a number of signals derived from intergenic regions not linked to previously reported transcripts were also specifically enhanced in the *rrp6* mutant. Many of these signals corresponded to intergenic regions containing SAGE tags (Velculescu et al., 1997) (green dots, Figure 1A). Some (20.3%) of these SAGE probes (cured for those overlapping or next to known noncoding RNAs; see Table S3) reproducibly exhibited an *rrp6Δ*/WT ratio >2 in the two independent experiments (by comparison, only 0.8% of these probes exhibited a ratio >2 in the controls where the isogenic strains, i.e., WT-1/WT-2 and *rrp6Δ-1*/*rrp6Δ-2*, were compared). Similarly, 7.7% of these probes reproducibly exhibited an *rrp6Δ*/WT ratio >3 in both independent experiments when this number was only 0.9% for the verified ORFs. This specific behavior of the SAGE probes did not result from a bias in the distribution of signal intensities between the two types of features (SAGE probes versus ORFs), since essentially identical results were obtained when comparing a subset of ORF and SAGE probes exhibiting an average signal ratio within the same intensity class (300–3000 average intensities). The peculiarity of SAGE probes was also apparent when comparing the class frequencies distribution of log2-transformed ratios between the two kind of features (Figure 1B): the *rrp6Δ* versus wild-type ratios (green curves) appear more significantly shifted toward higher values relative to the control experiments (gray curves) for the intergenic SAGEs compared to verified ORFs.

These microarray results were confirmed by real-time PCR performed on cDNAs primed with sequence specific oligonucleotides (Figure 1C), indicating that, in at least six out of eight test regions, signal increase resulted from higher transcript amounts rather than from polyadenylation of a preexisting RNA. These new regions thus differ from loci containing noncoding RNAs (snRNAs, snoRNAs, rRNAs, etc.). Oligo-directed RNase H cleavage and Northern blots performed for four of these transcripts, corresponding to Affymetrix features *NEL025c* (Figure 2A), *NBL001c*, *NPL040w*, and *NGR060w* (Figure S1), revealed that they consisted in RNAs of heterogeneous sizes (250–600 nt). The oligo-directed RNase H cleavage experiments showed that, except for *NPL040W*, these transcripts had a discrete 5′ end, and their heterogeneity thus resulted from multiple 3′ ends. We chose *NEL025C* (located on chromosome V between *RMD6* and *DLD3*) for further studies. RNaseH cleavage with oligo dT increased mobility of *NEL025c* transcripts but did not abolish size heterogeneity (Figure 2A, compare lanes 7 and 8). The polyadenylation status of these heterogeneous transcripts was further confirmed by oligo-dT affinity selection (Figure 2B). Taken together, these data indicate that these transcripts extend from a defined 5′ end to multiple, closely

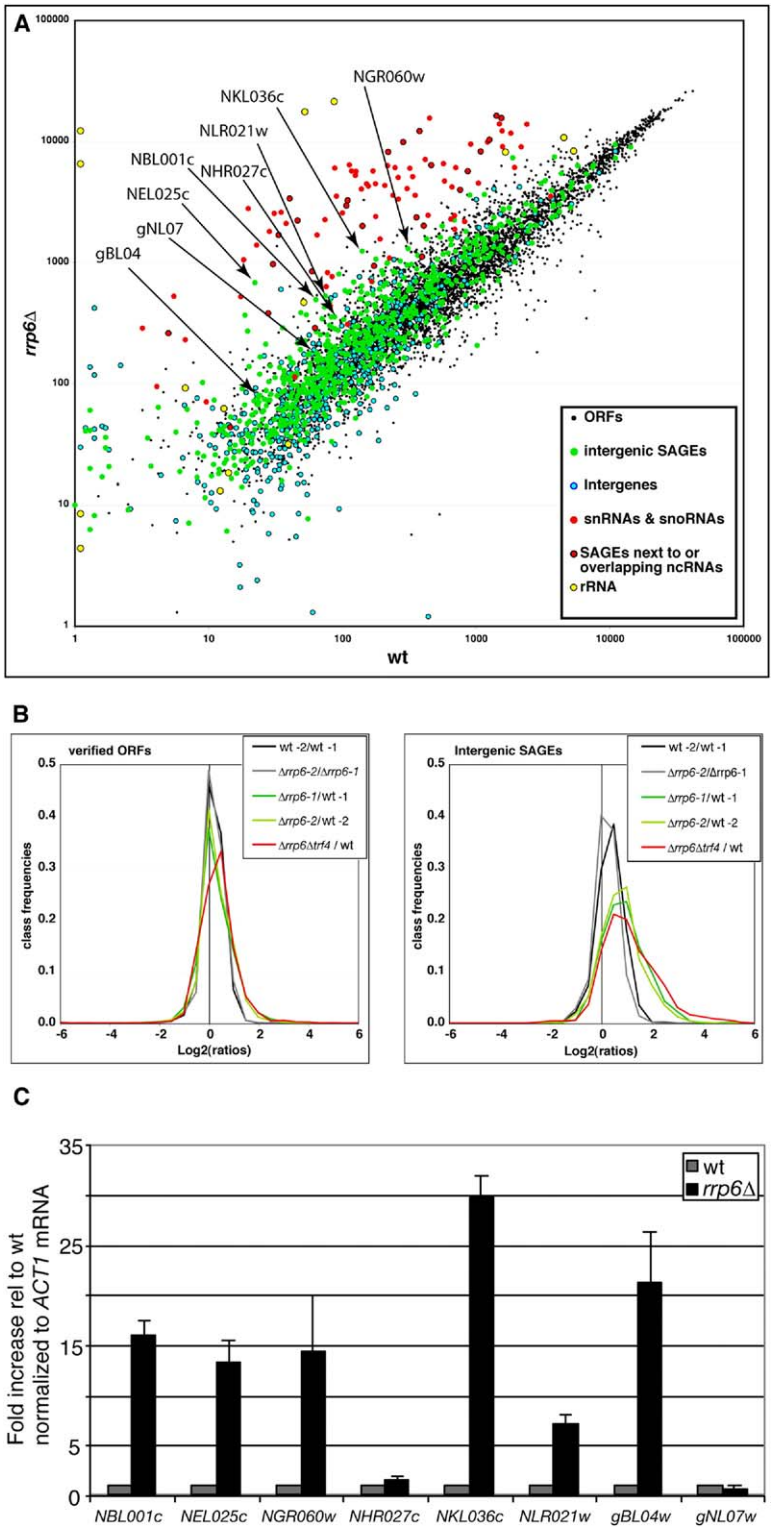


Figure 1. Genome-Wide Expression Profile of *rrp6Δ* Mutant versus Wild-Type

(A) Dot plot of signal intensities (average differences between perfect-match and mismatch oligonucleotides [Affymetrix MAS4.0 software], logarithmic scale) in wild-type (x axis) and *rrp6Δ* strain (y axis). Different classes of transcripts are color coded as indicated on the figure. The "SAGEs next to or overlapping ncRNAs" class represents features that were initially defined as intergenic SAGEs but that we found overlapping or directly juxtaposed to known noncoding RNAs—essentially snRNAs and snoRNAs; see Table S3). Arrows and labels point to dots corresponding to features that were analyzed by RT-PCR in Figure 1C.

(B) Distribution (class frequencies, one-third unit increments) of log2 transformed ratios (fold changes determined by the Affymetrix MAS4.0 software). Four microarrays were hybridized, two using RNAs from *RRP6* wild-type strains (BMA64) and two using RNAs from *rrp6Δ* strains (LMA164, see Table S1). The figure shows results obtained for the comparisons between wild-type-1 over wild-type-2 (black) or *rrp6Δ*-1 over *rrp6Δ*-2 (gray, controls) and *rrp6Δ*-1 over wild-type-1 or *rrp6Δ*-2 over wild-type-2, orange for verified ORFs and green for intergenic SAGE probes. Only verified ORFs (i.e., features defined as "ORF, verified" in the *Saccharomyces* Genome Database, www.yeastgenome.org) were taken into account in order to avoid statistical bias due to misannotated ORFs that should rather be classified as "intergenic features." Intergenic SAGE probes are as defined in the yeast S98 Affymetrix microarray and were cured for probes overlapping or directly next to known noncoding transcripts (see Table S3).

(C) The histogram shows the results of real-time PCR analysis after reverse transcription with specific oligonucleotides for eight arbitrarily chosen intergenic transcripts exhibiting a 3- to 30-fold signal increase in the *rrp6Δ* versus wild-type microarrays experiments (see Figure 1A). RNA amounts normalized to *ACT1* mRNA were expressed relative to the wild-type. Error bars were calculated from three independent experiments and represent standard deviations.

spaced 3′ ends to which poly(A) tails have been added. The oligo-dT-selected RNAs were also hybridized with a probe specific for the *NGR060W* transcripts and showed that these RNAs are also polyadenalylated in the *rrp6Δ* strain (data not shown).

Cryptic Transcripts Define New Pol II Transcription Units

Given their structure, we assessed whether *NEL025c*-derived RNAs are independent Pol II transcripts or readthrough products from neighboring genes. Immu-

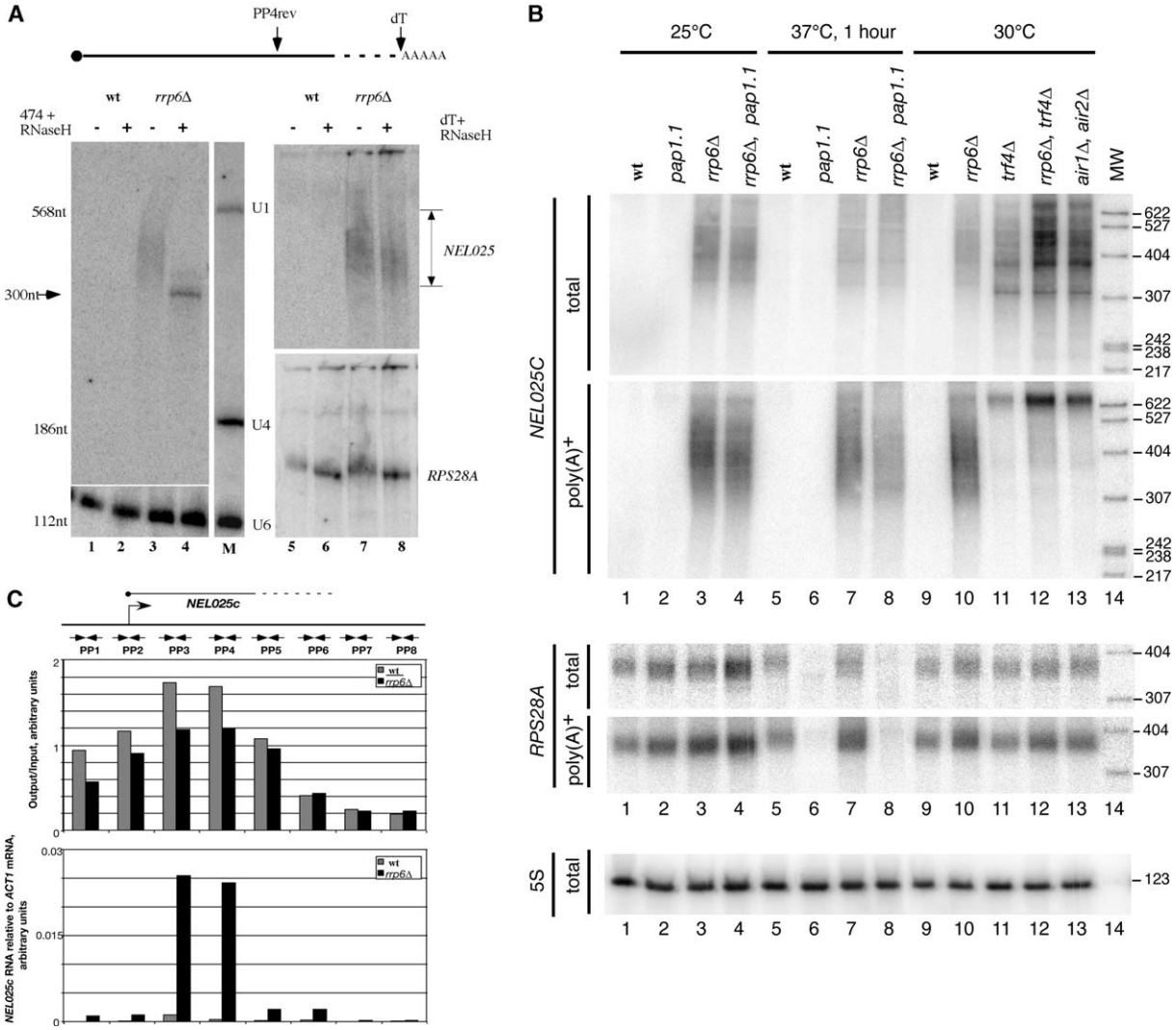


Figure 2. Characterization of the *NEL025C* Transcripts

(A) Northern blot characterization of the transcripts from the *NEL025C* region. Total RNAs from wild-type (lanes 1, 2, 5, and 6) or *rrp6Δ* (lanes 3, 4, 7, and 8) strains were separated on a 5% denaturing polyacrylamide gel. RNAs were treated with RNaseH in the presence of oligonucleotide P4rev (Table S2, lanes 2 and 4) or oligo-dT (lanes 6 and 8). The positions of the randomly primed double-strand DNA probe and oligonucleotide P4rev relative to the *NEL025c* RNA are indicated. After hybridization with the *NEL025c* probe, the filter on the right panel was stripped and rehybridized with a probe against the *RPS28A* mRNA for oligo-dT RNaseH cleavage control. Detection of U1, U4, and U6 snRNAs was used for loading control and size markers.

(B) Analysis of the polyadenylation status of *NEL025c* transcripts in different genetic backgrounds. Total RNAs (total) or oligo-dT-selected RNAs (poly[A]$^+$) were analyzed by Northern blots after separation on a 5% denaturing polyacrylamide gel and hybridization with a *NEL025c* random primed probe as in (A) (top panels). (Lanes 5–8) Cell cultures were shifted to 37°C for 1 hr prior to RNA extraction in order to inactivate Pap1p in the *rrp6Δ/pap1-1* strain. The filters used in the *NEL025c* panels were stripped and probed for *RPS28A* RNA as a control for Pap1p inactivation (*RPS28A* panels) and 5S RNA for loading control.

(C) Pol II occupancy (upper panel) in the *NEL025c* region in a wild-type (gray) and *rrp6Δ* strain (black). Chromatin immunoprecipitation (ChIP) was performed with an anti-Rpb1 antibody, and the DNA was analyzed by real-time PCR (top panel) with oligonucleotides spanning the entire region (primer pairs PP1–PP8, see Table S1) as schematized on top of the figure. RNAs were analyzed in parallel with the same primer pairs (bottom panel) by real-time RT-PCR.

noprecipitation with anti-cap antibodies (H20, kind gift of R. Lührmann) indicated that these transcripts are capped, which is a distinctive feature of Pol II transcripts and a specific mark of the transcription initiation site (Figure S2). Furthermore, inactivation of Pol II in a double mutant *rrp6Δ/rpb1-1*, expressing a thermosensitive form of the largest Pol II subunit (Nonet et al., 1987) and grown at nonpermissive temperature (37°C), resulted in the strong reduction of *NEL025c* transcript levels compared to a Pol III 5S ribosomal RNA control (Figure S3A). A similar result was also obtained for other intergenic transcripts (Figure S3B). Finally, the *NEL025c*-transcribed sequences expressed from their genomic locus but under the control of a heterologous

Pol II promoter (tetO7 operators under the control of the tetracycline repressible tTA transactivator; Gari et al. [1997]) were also strongly upregulated (~15 fold) in the absence of Rrp6p (Figure S3C). Together, these results indicate that these *rrp6Δ*-induced intergenic RNAs are independent, capped, and polyadenylated Pol II transcripts.

Cryptic Transcripts Are Unstable in Wild-Type Strains

To demonstrate that these intergenic RNAs are produced in both the wild-type and the *rrp6Δ* strain but have higher turnover rates in the former strain, we assessed Pol II occupancy at the *NEL025c* locus by chromatin immunoprecipitation (ChIP) in both strains. Real-time PCR analysis of the DNA immunoprecipitated with anti-Rpb1p antibodies was performed with primer pairs spanning the whole *NEL025c* locus. RNAs were analyzed in parallel by real-time RT-PCR with the same primer set. In striking contrast with the differences in transcript amounts, Pol II density was similar (or even slightly higher) in the wild-type strain compared to the *rrp6Δ* strain in the region tested (Figure 2C). This ChIP signal was specific since abolished by mutation of the largest Pol II subunit in an *rpb1-1* strain (Nonet et al., 1987; Schroeder et al., 2000) at the nonpermissive temperature (Figure S4). Pol II-ChIP analysis of other intergenic regions gave essentially identical results (data not shown). To further confirm that *NEL025c* transcripts are transcribed but more rapidly degraded in the presence of Rrp6p, we compared their turnover rates in a wild-type and *rrp6Δ* strains. Some *NEL025c* transcripts could be detected above background by real-time PCR in a wild-type strain, consistent with the existence of a SAGE tag in this genomic region (Velculescu et al., 1997). Use of an *rpb1-1* mutant (Nonet et al., 1987) allowed fast transcription shutoff in an otherwise wild-type or *rrp6Δ* context. As expected, the turnover rate of *NEL025c* transcripts was significantly higher in the *rpb1-1* strain compared to the *rpb1-1/rrp6Δ* strain, although the very low amount of transcripts in the *rpb1-1* strain precluded precise determination of the half-life of these RNAs (t1/2 <3 min in the *rpb1-1* strain and ~10 min in the *rpb1-1/rrp6Δ* strain); *ACT1* turnover rate was not significantly different in the two strains (Figure S5). Altogether, these data indicate that the *NEL025c* RNAs are produced in both the *rrp6Δ* and the wild-type strains, but, in the latter, the RNAs are more rapidly degraded. Degradation of these RNAs was also dependent on the integrity of the core exosome, as depletion of Rrp41p resulted in a similar stabilization of the *NEL025c* transcripts (Figure S6). Given the properties of the RNA products of these regions, revealed in the *rrp6Δ* strain, we named them CUTs for cryptic unstable transcipts.

NEL025c Transcripts Are Mainly Polyadenylated by a Pap1p-Independent Process

For most Pol II transcripts, the standard polyadenylation machinery adds poly(A) to a limited number of sites generated by cleavage. The heterogeneous 3′ ends of the CUTs were thus unexpected. To test the involvement of the standard polyadenylation machinery in *NEL025c* CUT poly(A) formation, we analyzed its polyadenylation status in an *rrp6Δ/pap1-1* double mutant shifted for 1 hr at the nonpermissive temperature (37°C). Oligo-dT-selected RNAs were analyzed by Northern blot (Figure 2B). Strikingly, Pap1p mutation did not strongly affect the amount and profile of the most abundant polyadenylated forms of these heterogeneous transcripts (300–400 nucleotides long), although the amount of the less abundant longest forms (>500 nucleotides) appear to decrease in the *rrp6Δ/pap1-1* strain compared to the *rrp6Δ* strain (Figure 2B, lanes 1–8). As a control, polyadenylation of *RPS28A* mRNA, a standard Pap1p substrate, was strongly inhibited in these conditions (Figure 2B). These observations suggested that the main polyadenylated forms of the *NEL025c* heterogeneous transcripts were polyadenylated by a machinery not involving Pap1p.

TRF4 Is the Catalytic Subunit of a Second Yeast Nuclear Poly(A) Polymerase

These results suggested the presence of at least another yeast poly(A) polymerase in addition to the classical machinery. Database searches revealed the presence of two highly related proteins, Trf4p and Trf5p, with distant similarity to Pap1p. These factors are similar to Cid1 and Cid13 from *S. pombe* and to Gld2 from *C. elegans* and related mammalian proteins that were recently described as cytoplasmic poly(A) polymerases (Kwak et al., 2004). While this work was in progress, a role for Trf4p in the polyadenylation of aberrant hypomodified $tRNA^{Met}$ was proposed (Kadaba et al., 2004).

To directly test whether Trf4p was endowed with poly(A) polymerase activity, we purified Trf4p and control factors from yeast using the TAP method (Rigaut et al., 1999) and assayed their poly(A) polymerase activity by following the incorporation of radiolabeled ATP in acid insoluble material using total yeast RNA as substrate. A strong incorporation was specifically detected with Trf4p-TAP (Figure 3A). In a similar assay, a Trf5p-TAP preparation was poorly active (data not shown). RNA polymerase activity of the Trf4p-TAP preparation is specific for ATP and could be primed by all tested substrates, including oligo(A) and tRNAs, with the exception of poly(U) (data not shown). Extension of an in vitro-transcribed RNA occurred in a time (data not shown) and Trf4-TAP concentration (Figure 3B) dependent manner in an apparent distributive reaction incorporating up to 500 residues. Mutation of two catalytic site residues (Wang et al., 2000) abolished the poly(A) polymerase activity of a Trf4-236-TAP preparation (Figure 3C). Overall, these data demonstrated the existence of a new yeast poly(A) polymerase having Trf4p as a catalytic subunit that we confirmed to be nuclear (Huh et al. [2003] and data not shown).

Air1p, or Air2p, Associates with Trf4p to Form Active Polymerases

While our results demonstrate that Trf4p is a subunit of a new poly(A) polymerase, recombinant Trf4-produced in *E. coli* was inactive in polyadenylation assays (see below), suggesting the requirement for additional factors and/or protein modification(s). Mass spectrometry

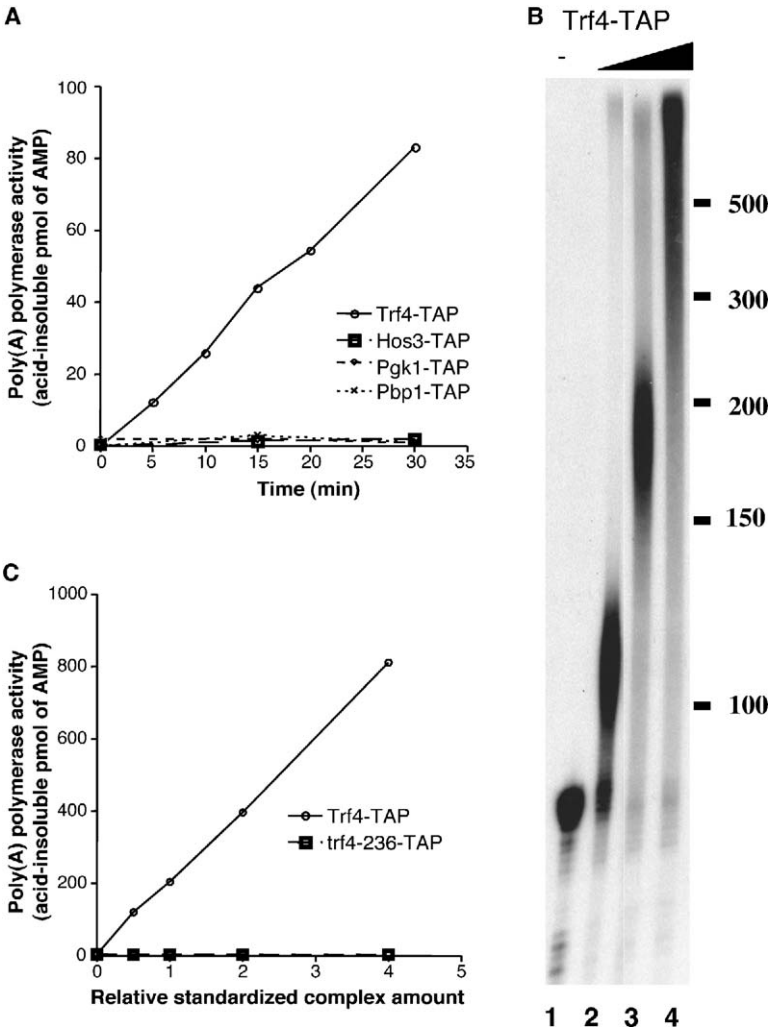


Figure 3. TRF4 Is a Subunit of a New Yeast Poly(A) Polymerase

(A) Incorporation of radioactive α32P-ATP into acid insoluble poly(A) was assayed in time course reactions using TAP-purified proteins and yeast RNA as substrate. Background activity was detected using TAP-purified Hos3, Pgk1, and Pbp1 control proteins. Identical quantities of the various proteins, as estimated from a Bradford assay, were used for each test.

(B) The polyadenylation activity associated with TRF4 and product length was tested by denaturing gel electrophoresis. An internally labeled RNA was used as substrate (lane 1) for a 30 min reaction. The product size observed at various concentrations of TAP-purified Trf4 complex was estimated by comparing with the migration of a single-stranded DNA marker (left, size in nucleotide).

(C) Mutation of the TRF4 catalytic center abolishes poly(A) polymerase activity. Incorporation of radioactive α32P-ATP into acid insoluble poly(A) in 30 min reactions was tested for TAP-purified wild-type trf4 and the trf4-236 mutant. Complex concentration was normalized by Western blotting using an antibody directed against Trf4 with concentration 0.5 corresponding to the quantity used in (A).

analysis of the purified Trf4p-TAP complex (Figure 4A) revealed the presence of additional factors, two of which that were identified as the related Air1 and Air2 proteins, which are located in the nucleus and have been previously implicated in nucleocytoplasmic mRNA transport (Inoue et al., 2000). A larger protein present at substoichiometric levels was identified as Mtr4p, a putative RNA helicase that was shown to interact functionally with the exosome (de la Cruz et al., 1998), supporting a role for Trf4p in the degradation of CUTs (see below). In addition, several ribosomal proteins were found in the purified fraction, possibly as a consequence of the implication of Trf4p in rRNA processing (see below). All these data are consistent with previous large-scale studies (Ho et al., 2002; Ito et al., 2001; Krogan et al., 2004). TAP purifications of Air1p-TAP and Air2p-TAP (Figure 4B) and the substoichiometric presence of either protein in the Trf4p-TAP preparations support the existence of two independent complexes containing either Air1p or Air2p associated with Trf4p.

Both Air1p-TAP and Air2p-TAP complexes were shown to be active in poly(A) synthesis (data not shown). The presence of either one of the two proteins is, however, required, as only in the absence of both Air1p and Air2p was the poly(A) polymerase activity abolished (Figure 4C).

Purified recombinant Air1p or Air2p failed to restore the activity of a recombinant Trf4p (Figure 4D). However, recombinant Air1p and Trf4p coexpressed in *E. coli* cells copurified with Trf4p, thus confirming a direct interaction. Most importantly, the resulting complex was active in polyadenylation (Figure 4D, a similar result was obtained for Air2-Trf4, data not shown). Thus, either Air1p or Air2p directly binds Trf4p, and these proteins are necessary and sufficient to form active polyadenylation enzymes.

Trf4 Is Required for the Polyadenylation and Degradation of the *NEL025c* Transcripts

To assess whether Trf4p plays a role in polyadenylation and/or degradation of *NEL025c* CUTs, we constructed strains deleted for *TRF4* in a wild-type or *rrp6Δ* background. The combination of the two mutations resulted in a strong synthetic growth impairment (see Figure S7), suggesting that Rrp6p and Trf4p are functionally linked.

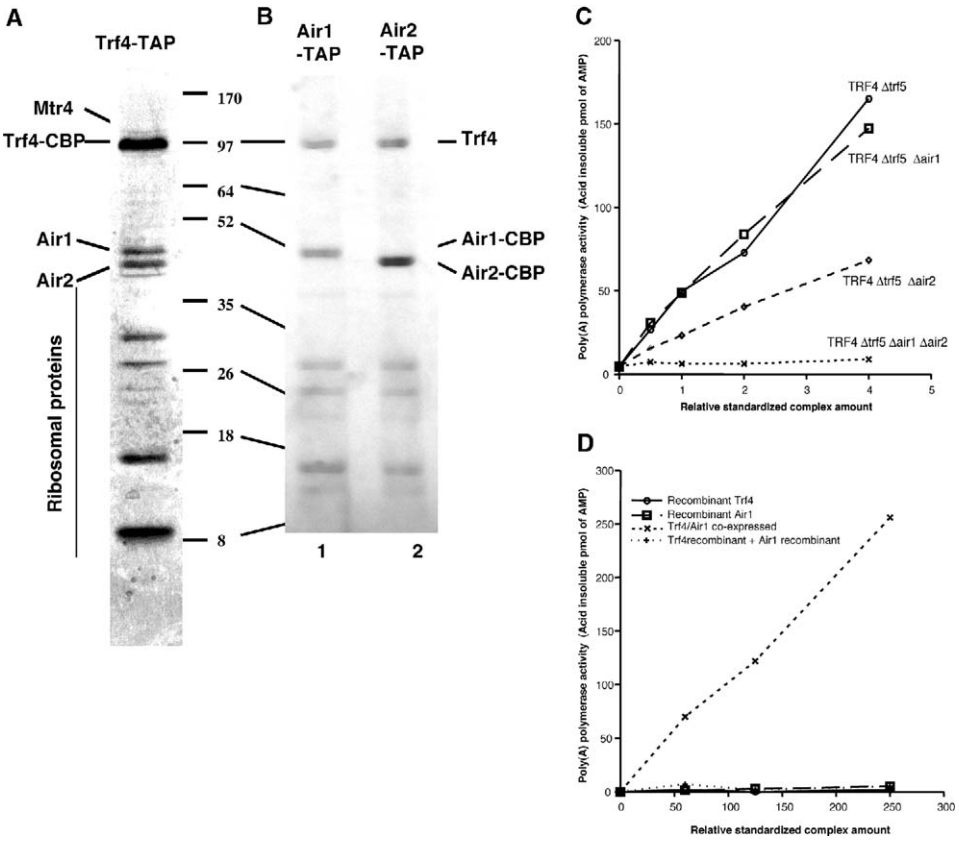


Figure 4. Trf4 Associates with Air1 or Air2 to Form an Active Poly(A) Polymerase

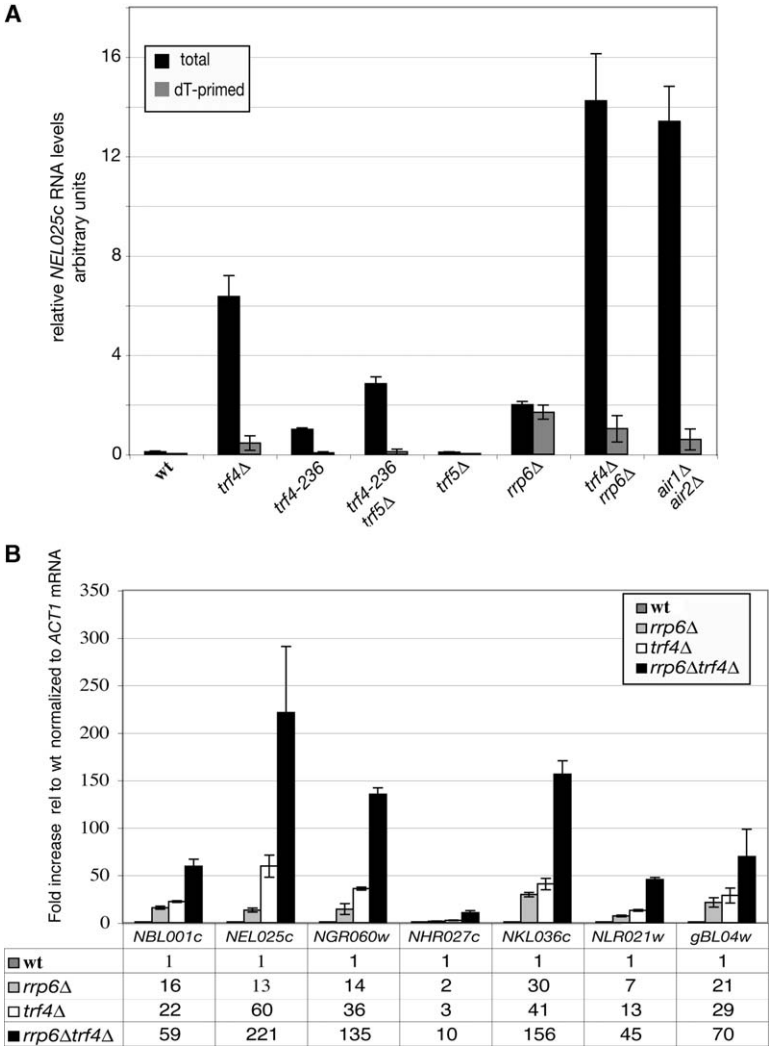
(A) A Coomassie blue-stained gel of proteins associated with TAP-purified Trf4p. Proteins identified by mass spectrometry are labeled. Position of migration of a molecular weight marker is indicated on the right.

(B) Proteins present in the TAP-purified fractions associated with Air1-TAP or Air2-TAP. Position of migration of the tagged proteins and Trf4 (identified by Western blotting) are indicated on the right, while the position of migration of a molecular weight marker is indicated on the left.

(C) Poly(A) polymerase activity requires Air1 or Air2. Concentration of the Trf4-TAP-purified complexes obtained from the strains of indicated genotypes were normalized by Western blotting using an antibody directed against Trf4. Poly(A) polymerase activity was assayed as described for Figure 3A.

(D) The poly(A) polymerase activity of recombinant Trf4, recombinant Air1, a mixture of both proteins, or a recombinant complex generated by coexpression of Trf4 and Air1 was tested. Protein concentration was normalized by Bradford assay.

To quantify the levels of both polyadenylated and nonadenylated transcripts in these strains, we first performed real-time PCR analyses after priming cDNA synthesis either with an oligonucleotide specific for *NEL025c* transcripts (total) or with oligo-dT (polyadenylated fraction) (Figure 5A). All data were normalized using *ACT1* mRNA levels. Strikingly, deletion of *TRF4* leads to stabilization of *NEL025c* transcripts (and other CUTs, Figure 5B and data not shown) to a level that is even higher than the one observed in an *rrp6*Δ strain. However, these transcripts appear to be mostly nonadenylated, in contrast to what was observed in the absence of Rrp6p (Figure 5A). Northern blot analysis confirmed that depletion of Trf4p resulted in a strong accumulation of *NEL025c* transcripts (Figure 2B, lane 11) as well as other CUTs (Figure S1). Note that, in the absence of *TRF4*, deletion of *RRP6* strongly enhances the accumulation of the *NEL025c* and other CUT transcripts, as shown both by quantitative RT-PCR and Northern blot analyses (Figures 5A, 5B, and 2B and Figure S1), suggesting that degradation of a fraction of these RNAs still occurs despite Trf4p absence. The most abundant *NEL025c* RNA species (~350 nt long) were absent from the oligo-dT selected fraction, in contrast to what was observed in the *rrp6*Δ single mutant strain, confirming that the polyadenylation of these transcripts is Trf4p dependent. In contrast, however, a larger polyadenylated product (enriched upon oligo-dT selection), of relative low abundance in the total RNA samples (Figure 2B, lane 12), was strongly stabilized in the absence of both Rrp6p and Trf4p. This polyadenylated transcript was completely absent from the oligo-dT selected fraction when Pap1p was inactivated, suggesting that it corresponds to a small fraction of *NEL025c* transcripts polyadenylated by the normal Pap1p-dependent machinery. Most interestingly, this polyadenylated RNA species accumulated only when both Rrp6p and Trf4p are absent, suggesting that, even though it might result from the normal, Pap1p-dependent, polyadenylation pathway, its precursor and/or itself are degraded by the coordinated actions of Rrp6p and Trf4p (see Discussion). In order to assess the generality of this observa-



	NBL001c	NEL025c	NGR060w	NHR027c	NKL036c	NLR021w	gBL04w
wt	1	1	1	1	1	1	1
rrp6Δ	16	13	14	2	30	7	21
trf4Δ	22	60	36	3	41	13	29
rrp6Δtrf4Δ	59	221	135	10	156	45	70

Figure 5. Real-Time RT-PCR Analysis of CUTs in Mutants of the Trf4/Exosome Degradation Pathway

(A) Real-time RT-PCR analysis of *NEL025c* transcripts in mutants of the Trf4p/exosome degradation pathways. cDNA synthesis was performed with a specific oligonucleotide (total) or oligo-dT (dT-primed) before real-time PCR analysis with primer pair PP4. These signals are proportional to the total and polyadenylated fraction, respectively. Relative normalization was performed using *ACT1* mRNA as follows: the amount of every sample was divided by a normalization index representing the ratio between the *ACT1* value in that given sample and the average value of *ACT1* mRNA in all samples. The dT-primed/total ratio for *NEL025c* transcripts in a *rrp6Δ* strain was 0.87 ± 0.09 (n = 3), which is similar (although consistently higher) than the average dT-primed/total ratio for *ACT1* mRNA (0.65 ± 0.048; n = 24).

(B) Real-time RT-PCR analysis of other CUTs (same set as in Figure 1C) in *trf4Δ* and *trf4Δ/rrp6Δ* strains. The amount of each transcript in a given mutant strain is expressed relative to the amount in a wild-type strain. Average stabilization values are indicated in the panel below the histogram. Stabilization values in an *rrp6Δ* background are reported for comparison. In (A) and (B), error bars represent standard deviations calculated from three independent experiments.

tion, we analyzed with Affymetrix microarrays (which detect mainly polyadenylated species, as they use oligo-dT primed cDNAs), the global effect of the *rrp6Δ/trf4Δ* double deletion on the stabilization of such polyadenylated forms of CUTs. Figure 1B (red curves) shows that stabilization of these minor, Trf4p-independent, polyadenylated forms of CUTs is widespread, as the signals of a large number of intergenic SAGEs were enhanced in the double mutant relative to the wild-type. Finally, the depletion of Trf5 had no marked effect the amount of CUTs (Figure 5A and data not shown) or on the profiles of the oligo-dT-selected *NEL025c* transcripts (data not shown). These data indicate that Trf4p is involved in polyadenylation of CUTs and, together with Rrp6p, in their degradation.

The Trf4-Associated Poly(A) Polymerase Activity Is Required for CUT Degradation

Because the Trf4p complex is a poly(A) polymerase in vitro and because Trf4p is involved in polyadenylation and degradation of CUTs in vivo, we assessed whether the enzymatic activity of the complex is required for CUTs degradation. We asked first whether the poly(A) polymerase catalytic site mutant (*trf4-236)* would affect CUT stability. As shown in Figure 5A, non-adenylated *NEL025c* transcripts were readily detected in a *trf4-236* strain, although they were stabilized to a lower extent than upon *TRF4* deletion. This intermediate effect was paralleled by the growth of the *trf4-236* mutant strain that was less affected than the *trf4Δ* strain (Figure S7). Interestingly, deletion of the *TRF4* paralogue *TRF5* in the *trf4-236* strain led to a stabilization of *NEL025c* transcripts that was greater than the one observed in a *trf4-236* strain, strongly suggesting a role for Trf5p in CUT degradation when Trf4p is not fully functional (Figure 5A). To further confirm that the poly(A) polymerase activity of the Trf4 complex is involved in CUT degradation, we analyzed CUT levels in a strain lacking both Air1 and Air2, as both proteins were required for poly(A) polymerase activity (see above). Real-time RT-PCR (Figure 5A, and data not shown) and Northern blot analyses (Figure 2B and Figure S1) of CUTs in this strain revealed a strong stabilization of these RNAs. For *NEL025c*, only the largest transcript accumulated in a polyadenylated form (Figure 2B). Altogether, these results strongly suggest that the

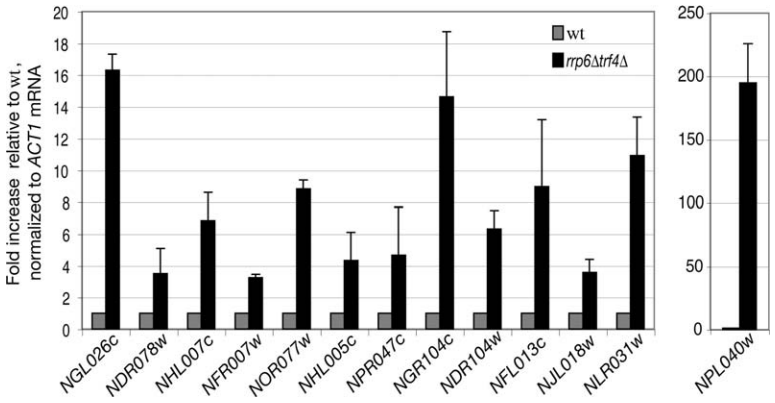


Figure 6. Most Intergenic Transcripts Are Stabilized in a *rrp6Δ/trf4Δ* Double Mutant

Real-time RT-PCR analysis in a *rrp6Δ/trf4Δ* strain of intergenic transcripts that exhibited a low *rrp6Δ*-dependent signal increase in the microarray experiments (ranging from 1.5- to 3.6-fold, 2.7-fold on average). Levels normalized to *ACT1* mRNA are expressed relative to the amount in a wild-type strain. Error bars were calculated from three independent experiments and represent standard deviations. Note that analysis for *NPL040w* is reported on a different scale, as this RNA is strongly stabilized in this strain.

poly(A) polymerase activity of Trf4p is associated with its role in CUT degradation.

Polyadenylation of rRNAs, snRNAs, and snoRNAs in an *rrp6Δ* Background Is Trf4p Dependent

Having established that Trf4p is involved in polyadenylation of CUTs, we tested whether the polyadenylated forms of the rRNAs, snRNAs, and snoRNAs observed in the *rrp6Δ* strain are also Trf4p dependent. Indeed, the presence of polyadenylated forms of U6, 5S, or 5.8S RNAs were dependent upon the presence of Trf4p (Figure 7 and data not shown). Inactivation of Pap1p had some effect on polyadenylation, in particular on the longest forms, but these effects were always weaker than the effect of *trf4Δ*. A similar observation was made for the snoRNA U18 (data not shown). Most importantly, the polyadenylated forms of these transcripts represent a small fraction of the total RNAs, and the absence of Trf4p had no strong influence on the amount of the mature forms of these transcripts (Figure 7), which is in sharp contrast with what we observed for CUTs.

Discussion

Our results support the existence of a quality control mechanism monitoring nuclear transcripts. This mechanism targets transcripts made by all three nuclear RNA polymerases. A characteristic feature of this process is the addition of poly(A) tail to the target molecules before their proper processing or degradation in an exosome-dependent manner

Numerous new RNA species accumulate in a *Δrrp6* strain. These include Pol I transcripts or derivatives thereof (e.g., 7S rRNA); Pol II transcripts, such as U18 snoRNA transcripts; and Pol III transcripts (e.g., species detected with the 5S probe). As previously reported (Allmang et al., 1999a; Kuai et al., 2004; van Hoof et al., 2000), we also found that a large fraction of these new species are polyadenylated in a *Δrrp6* strain but not in a wild-type strain. These poly(A)$^+$ species may represent normal processing intermediates with very short half-lives. Alternatively, they could represent nonfunctional transcripts targeted for degradation. Polyadenylation of such RNA species is unlikely to be restricted to *S. cerevisiae* and offers to cells a mean to control maturation or processing of these targets. Consistently, polyadenylated ribosomal RNAs were detected in the pathogenic fungi *Candida albicans* in a process controlled by the presence of serum (Fleischmann et al., 2004).

Although it has been reported (Kuai et al., 2004) that polyadenylation of several rRNA species in a *rrp6Δ* background depends on Pap1p integrity, our results only partially support this notion. In fact, Pap1p-dependent polyadenylation only accounts for a fraction of the polyadenylated rRNA species detected in *rrp6* mutants. This observation is paralleled by the analysis of snoRNA, snRNA, and CUTs. In most cases (e.g., for *NEL025c* and 5S RNAs), this fraction is minor compared to the fraction that is Trf4p dependent, and, most importantly, in no cases did mutation of Pap1p lead to stabilization of transcripts in a WT or *rrp6Δ* background. Currently, the significance of polyadenylation of these transcripts by Pap1p is unclear; it does not appear to stimulate their degradation, as shown here for Trf4p-dependent polyadenylation.

Another group of polyadenylated RNA accumulating in a *Δrrp6* strain corresponds to new cryptic Pol II transcripts. These CUT transcripts are present at extremely low concentration in wild-type cells, even though some of these transcripts were apparently detected by SAGE analyses (Velculescu et al., 1997). Nevertheless, they appear to represent bona fide transcripts generated by Pol II, containing a 5′ cap. These intergenic cryptic Pol II transcripts are usually relatively short and do not contain long or conserved reading frames. Thus, while we cannot formally exclude that they have a physiological role, their structure suggests that they result from the presence of adventitious promoters at random genomic locations. How widespread is the occurrence of cryptic intergenic transcription in the genome? We have confirmed by RT-PCR analysis that most if not all of the intergenic SAGE transcripts that exhibit an *rrp6Δ*/WT signal ratio >2 in the microrray experiments (roughly 20% of the total) are indeed responsive to mutation of the TRF4p/exosome degradation pathways. To assess whether intergenic SAGE transcripts exhibiting lower *rrp6Δ*/WT signal ratios are bona fide CUTs, we exploited the observation that, in a *trf4Δ/rrp6Δ* mutant, CUTs are stabilized to a higher level, which should improve sensitivity. Microarray analysis in this context was not informative, as stabilized CUTs are mostly non-

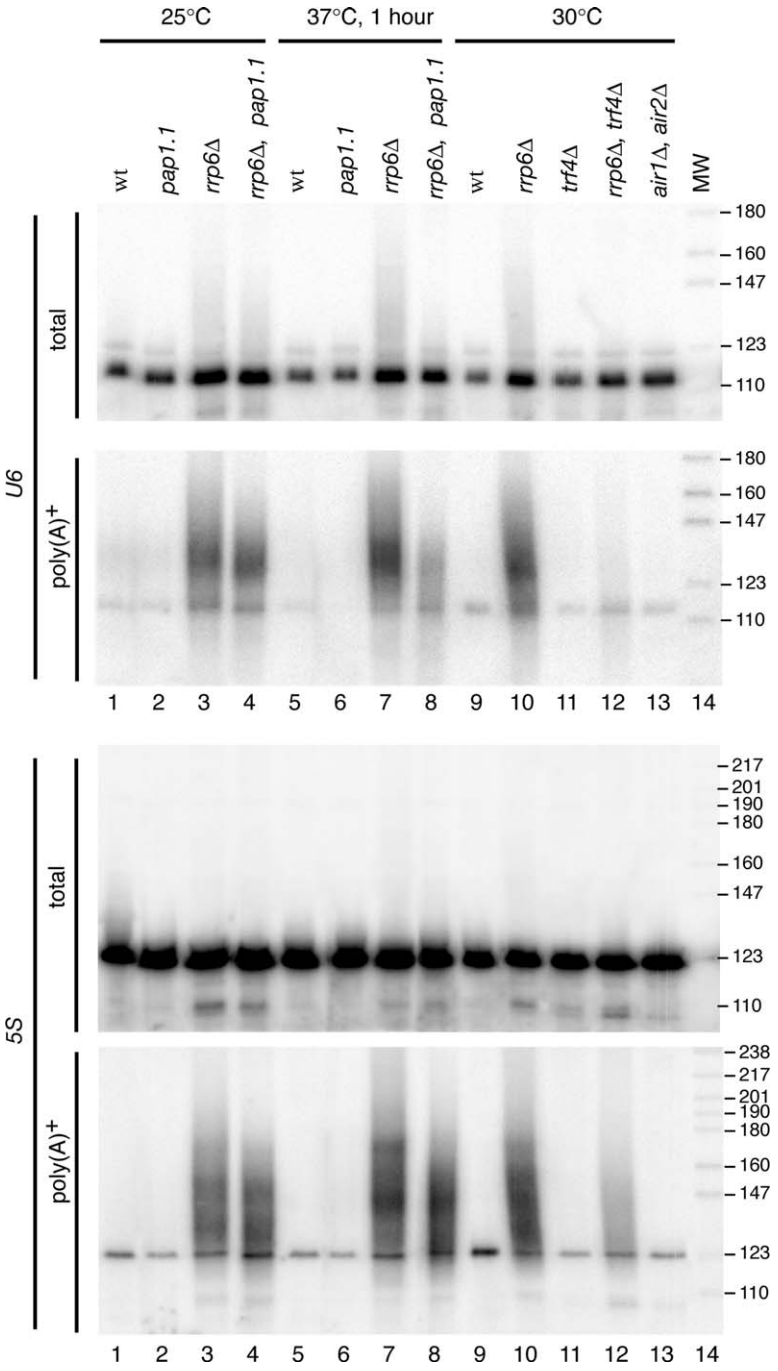


Figure 7. Analysis of the Polyadenylation Status of U6 and 5S rRNA in Different Genetic Backgrounds

As in Figure 2B, except that the filters were hybridized with [^{32}P]-labeled oligonucleotides specific for U6 snRNA (top panels) or 5S rRNA (bottom panels).

polyadenylated, while the standard Affymetrix technology only allows the detection of polyadenylated species (Figure 1B, red curves). We then extended RT-PCR analyses with sequence-specific primers to 13 additional intergenic SAGE regions that exhibited even a very modest, *rrp6Δ*-dependent signal increase in the microarray experiments (1.5- to 3.6-fold increase; 2.7-fold in average). Remarkably, all these RNAs species were responsive to the *trf4Δ/rrp6Δ* mutation (Figure 6), strongly suggesting that they are bona fide CUTs. This is consistent with the notion that a large fraction of the intergenic regions containing SAGE tag (more than 10% of the overall intergenic regions; Velculescu et al. [1997]) encode genuine transcripts that are normally targeted for degradation by the coordinated action of the nuclear exosome and the Trf4-associated complex. Consequently, as some intergenic transcripts might have escaped SAGE detection, a minimal genome-wide estimate of cryptic transcripts for all intergenic regions is likely to be more than 5%–10%. Thus, spurious intergenic transcription appears to be widely spread within the yeast genome. This is likely to be evolutionarily widespread. Indeed, microarray tiling experiments revealed the presence of numerous unsuspected tran-

scripts encoded by intergenic regions of mammalian chromosomes (Johnson et al., 2005). A relatively low specificity of promoter recognition might leave more flexibility for evolution and/or regulation. Thus, paralleling observations made with ribosome fidelity mutants (Ruusala et al., 1984), promoter recognition by the Pol II machinery may remain suboptimal. We suggest that, in addition to a chromatin-dependent repression of cryptic promoters usage, a parallel and/or overlapping strategy that involves a posttranscriptional quality control mechanism evolved to get rid of cryptic transcripts.

Our data demonstrate that Trf4 is a poly(A) polymerase. While Trf4 was previously suggested to be a DNA polymerase involved in DNA repair (Castano et al., 1996), we believe that these original data have to be reinterpreted, as its DNA polymerase activity is extremely weak compared to its poly(A) polymerase activity (Wang et al., 2000). Furthermore, Trf4 clearly affects polyadenylation in vivo, supporting the in vitro biochemical activity. Nevertheless, we cannot exclude that Trf4 has both activities. Identification of a second yeast nuclear poly(A) polymerase targeting RNA for degradation by the exosome must also be reconciled with the presence of poly(A) tails on (pre-)mRNA that are not degraded. How the cell discriminates between aberrant and functional transcripts remains unknown. Substituting the natural cryptic promoter by a heterologous one (Figure S3C) or inserting a bona fide terminator (data not shown) did not change the susceptibility of the *NEL025c* CUT toward Rrp6p.

What is the role of the Trf4p complex in recognition and degradation of unstable transcripts? Trf4p appears to have a role in CUT degradation (most likely through stimulation or targeting of the exosome) that is independent of its polyadenylation activity as the *trf4-236* point mutant, which completely lacks pol(A) polymerase activity in vitro and is much less affected than the *trf4Δ* mutant (Figure 5A and Figure S7). In at least some cases, Trf5p might substitute for Trf4p function, which is suggested by the stronger phenotype of a *trf4-236/trf5Δ* mutant compared to either single mutants (Figure 5A and Figure S7). If Air1p/Air2p were also required for Trf5p function (which is presently unclear), the stronger phenotype of *air1Δ/air2Δ* cells compared to *trf4Δ* might be explained by a concomitant impairment of both Trf4p and Trf5p activities. Finally, it is unclear whether Rrp6p and the core exosome have different roles in the degradation of CUTs. A distinct role might be consistent with the observation that the patterns and the polyadenylation status of *NEL025c* transcripts are similar but not identical in an *rrp6Δ* mutant and in the depletion of the Rrp41p core component.

Degradation of *NEL025c* transcripts might be paradigmatic for the Trf4p/exosome pathway. The long form of the *NEL025c* transcript (which might be the precursor of the shortest forms) is polyadenylated by Pap1p but degraded in a Trf4p-dependent manner (which is polyadenylation independent, as its abundance but not its polyadenylation status is affected by *TRF4* deletion). It is conceivable, for instance, that a polyadenylation-independent role of the Trf4p/Air complex in this case would be to target the exosome on CUTs or other substrates, maybe through Mtr4p, a reported constituent of the nuclear exosome that is also found associated with the Trf4/Air complex. The shorter 300–500 nt transcripts, on the contrary, would require prior polyadenylation by the Trf4/Air complex for subsequent efficient degradation. This might result from stalling of the exosome at secondary structures, which would require the secondary addition of Trf4p-dependent poly(A) tails to resume degradation. In the *rrp6Δ/trf4Δ* double mutant, the combination of compromised exosome activity and lack of the Trf4/Air complex would result in both a very inefficient targeting of the primary *NEL025c* transcript as well as inefficient removal of degradation intermediates. The role of Trf4p poly(A) polymerase activity would then be very similar to batcterial poly(A) polymerases that have been shown to facilitate mRNA degradation by the degradosome (Dreyfus and Regnier, 2002). Indeed, the group of D. Tollervey has recently found that the Trf4p-associated complex enhances the nuclear exosome activity in vitro (LaCava et al., 2005). In this vein, it is noteworthy that the degradosome is homologous to the eukaryotic exosomes (Aloy et al., 2002; Symmons et al., 2002). The existence of Trf4p and Airp homolgues in human and other species suggests that the poly(A)-stimulated 3′–5′ nuclear degradation/processing of RNA is conserved in all eucaryotes.

Experimental Procedures

Standard experimental procedures are given as Supplemental Data under the section Supplemental Experimental Procedures.

Microarray Analyses

Microarray hybridizations were performed using the Affymetrix Yeast Genome S98 Array using protocols described by Affymetrix, Inc. (Santa Clara, CA). Data were analyzed using Affymetrix Microarray Suite 4.0 software for the *rrp6Δ* results and Affymetrix Microarray Suite 5.0 for the *rrp6Δ, trf4Δ* results. Microarray data are accessible in the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE2579.

Poly(A) Polymerase Assay

Reactions (20 μl) contained 20 mM Tris-HCl (pH 7.6), 50 mM KCl, 17.5 mM MgCl2, 1 mM DTT, 0.2 mM EDTA, 100 μg/ml BSA, 10% glycerol, 0.5 mM ATP, (α32P) ATP (150–500 cpm/pmol) and 0.25 μg of substrate (total yeast RNA, poly[A] [250 nt], or oligo[A] [15 nt]). The reaction was started by the enzyme addition (2.5–50 ng of TAP-purified complexes or 50–400 ng of recombinant proteins), incubated at 30°C for 30 min, and stopped by addition of 0.5 ml of 10% TCA. The precipitate was collected on glass fiber filter, washed, and counted. Alternatively ^{32}P-labeled LucΔ RNA was used as substrate in reactions without (α32P)ATP.

Supplemental Data

Supplemental Data include seven figures, three tables, and Supplemental References and can be found with this article online at http://www.cell.com/cgi/content/full/121/5/725/DC1/.

Acknowledgments

We thank R. Lührmann for antibodies, R. Gallie for plasmid, T.H. Jensen and M. Rosbash for comments on the manuscript, David Tollervey for generously communicating unpublished results, and Kristell Wanherdrick from "Genopole Ile de France, Site de la Montagne Sainte Geneviève, Institut Curie, Paris" for performing early microarray hybridizations. Discussion and experimental input from our lab colleagues, particularly F. Lacroute, E. Kisseleva-Romanova, E. Kastenhuber, M. Minet, C. Torchet, C. Saveanu, and M. Fromont-Racine, were deeply appreciated. Work in the lab of B.S. was supported by La Ligue Contre le Cancer (Equipe Labelisée

2005), the CNRS (particularly Programme PGP 2003), and the Ministry of Research (ACI BCMS). Work in the lab of A.J. was supported by the EEC (grant RNOMICS, QLG2-CT-2001-01554) and the Ministry of Research (grant "Subventions puces Affymetrix"). Work in the lab of D.L. was supported by the CNRS. M.R. is a recipient of a fellowship from the Ministry of Research.

Received: January 10, 2005
Revised: March 25, 2005
Accepted: April 19, 2005
Published online: May 19, 2005

References

Allmang, C., Kufel, J., Chanfreau, G., Mitchell, P., Petfalski, E., and Tollervey, D. (1999a). Functions of the exosome in rRNA, snoRNA and snRNA synthesis. EMBO J. *18*, 5399–5410.

Allmang, C., Petfalski, E., Podtelejnikov, A., Mann, M., Tollervey, D., and Mitchell, P. (1999b). The yeast exosome and human PM-Scl are related complexes of 3′→5′ exonucleases. Genes Dev. *13*, 2148–2158.

Aloy, P., Ciccarelli, F.D., Leutwein, C., Gavin, A.C., Superti-Furga, G., Bork, P., Bottcher, B., and Russell, R.B. (2002). A complex prediction: three-dimensional model of the yeast exosome. EMBO Rep. *3*, 628–635.

Bousquet-Antonelli, C., Presutti, C., and Tollervey, D. (2000). Identification of a regulated pathway for nuclear pre-mRNA turnover. Cell *102*, 765–775.

Castano, I.B., Heath-Pagliuso, S., Sadoff, B.U., Fitzhugh, D.J., and Christman, M.F. (1996). A novel family of TRF (DNA topoisomerase I-related function) genes required for proper nuclear segregation. Nucleic Acids Res. *24*, 2404–2410.

Das, B., Butler, J.S., and Sherman, F. (2003). Degradation of normal mRNA in the nucleus of *Saccharomyces cerevisiae*. Mol. Cell. Biol. *23*, 5502–5515.

de la Cruz, J., Kressler, D., Tollervey, D., and Linder, P. (1998). Dob1p (Mtr4p) is a putative ATP-dependent RNA helicase required for the 3′ end formation of 5.8S rRNA in *Saccharomyces cerevisiae*. EMBO J. *17*, 1128–1140.

Dreyfus, M., and Regnier, P. (2002). The poly(A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. Cell *111*, 611–613.

Egyhazi, E. (1976). Quantitation of turnover and export to the cytoplasm of hnRNA transcribed in the Balbiani rings. Cell *7*, 507–515.

Fleischmann, J., Liu, H., and Wu, C.P. (2004). Polyadenylation of ribosomal RNA by *Candida albicans* also involves the small subunit. BMC Mol. Biol. *5*, 17.

Gari, E., Piedrafita, L., Aldea, M., and Herrero, E. (1997). A set of vectors with a tetracycline-regulatable promoter system for modulated gene expression in *Saccharomyces cerevisiae*. Yeast *13*, 837–848.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature *415*, 180–183.

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686–691.

Inoue, K., Mizuno, T., Wada, K., and Hagiwara, M. (2000). Novel RING finger proteins, Air1p and Air2p, interact with Hmt1p and inhibit the arginine methylation of Npl3p. J. Biol. Chem. *275*, 32793–32799.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA *98*, 4569–4574.

Johnson, A.W. (1997). Rat1p and Xrn1p are functionally interchangeable exoribonucleases that are restricted to and required in the nucleus and cytoplasm, respectively. Mol. Cell. Biol. *17*, 6122–6130.

Johnson, J.M., Edwards, S., Shoemaker, D., and Schadt, E.E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet. *21*, 93–102.

Kadaba, S., Krueger, A., Trice, T., Krecic, A.M., Hinnebusch, A.G., and Anderson, J. (2004). Nuclear surveillance and degradation of hypomodified initiator tRNAMet in *S. cerevisiae*. Genes Dev. *18*, 1227–1240.

Keller, W., and Minvielle-Sebastia, L. (1997). A comparison of mammalian and yeast pre-mRNA 3′-end processing. Curr. Opin. Cell Biol. *9*, 329–336.

Kim, M., Krogan, N.J., Vasiljeva, L., Rando, O.J., Nedea, E., Greenblatt, J.F., and Buratowski, S. (2004). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. Nature *432*, 517–522.

Krogan, N.J., Peng, W.T., Cagney, G., Robinson, M.D., Haw, R., Zhong, G., Guo, X., Zhang, X., Canadien, V., Richards, D.P., et al. (2004). High-definition macromolecular composition of yeast RNA-processing complexes. Mol. Cell *13*, 225–239.

Kuai, L., Fang, F., Butler, J.S., and Sherman, F. (2004). Polyadenylation of rRNA in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA *101*, 8581–8586.

Kwak, J.E., Wang, L., Ballantyne, S., Kimble, J., and Wickens, M. (2004). Mammalian GLD-2 homologs are poly(A) polymerases. Proc. Natl. Acad. Sci. USA *101*, 713–724.

LaCava, J., Houseley, J., Saveanu, C., Petfalski, E., Thompson, E., Jacquier, A., and Tollervey, D. (2005). RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. Cell *121*, this issue, 713–724.

Liang, S., Hitomi, M., Hu, Y.H., Liu, Y., and Tartakoff, A.M. (1996). A DEAD-box-family protein is required for nucleocytoplasmic transport of yeast mRNA. Mol. Cell. Biol. *16*, 5139–5146.

Libri, D., Dower, K., Boulay, J., Thomsen, R., Rosbash, M., and Jensen, T.H. (2002). Interactions between mRNA export commitment, 3′-end quality control, and nuclear degradation. Mol. Cell. Biol. *22*, 8254–8266.

Mitchell, P., and Tollervey, D. (2000). Musing on the structural organization of the exosome complex. Nat. Struct. Biol. *7*, 843–846.

Nonet, M., Scafe, C., Sexton, J., and Young, R. (1987). Eucaryotic RNA polymerase conditional mutant that rapidly ceases mRNA synthesis. Mol. Cell. Biol. *7*, 1602–1611.

Petfalski, E., Dandekar, T., Henry, Y., and Tollervey, D. (1998). Processing of the precursors to small nucleolar RNAs and rRNAs requires common components. Mol. Cell. Biol. *18*, 1181–1189.

Proudfoot, N., and O'Sullivan, J. (2002). Polyadenylation: a tail of two complexes. Curr. Biol. *12*, R855–R857.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nat. Biotechnol. *17*, 1030–1032.

Ruusala, T., Andersson, D., Ehrenberg, M., and Kurland, C.G. (1984). Hyper-accurate ribosomes inhibit growth. EMBO J. *3*, 2575–2580.

Schroeder, S.C., Schwer, B., Shuman, S., and Bentley, D. (2000). Dynamic association of capping enzymes with transcribing RNA polymerase II. Genes Dev. *14*, 2435–2440.

Schurer, H., Schiffer, S., Marchfelder, A., and Morl, M. (2001). This is the end: processing, editing and repair at the tRNA 3′-terminus. Biol. Chem. *382*, 1147–1156.

Symmons, M.F., Williams, M.G., Luisi, B.F., Jones, G.H., and Carpousis, A.J. (2002). Running rings around RNA: a superfamily of phosphate-dependent RNases. Trends Biochem. Sci. *27*, 11–18.

Torchet, C., Bousquet-Antonelli, C., Milligan, L., Thompson, E., Kufel, J., and Tollervey, D. (2002). Processing of 3′-extended read-through transcripts by the exosome can generate functional mRNAs. Mol. Cell *9*, 1285–1296.

van Hoof, A., Lennertz, P., and Parker, R. (2000). Yeast exosome

mutants accumulate 3′-extended polyadenylated forms of U4 small nuclear RNA and small nucleolar RNAs. Mol. Cell. Biol. *20*, 441–452.

Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Jr., Hieter, P., Vogelstein, B., and Kinzler, K.W. (1997). Characterization of the yeast transcriptome. Cell *88*, 243–251.

Wang, Z., Castano, I.B., De Las Penas, A., Adams, C., and Christman, M.F. (2000). Pol kappa: A DNA polymerase required for sister chromatid cohesion. Science *289*, 774–779.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5′→3′ exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522–525.

Cell

Resource

Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences

Michael F. Berger,[1,3,8] Gwenael Badis,[5,8] Andrew R. Gehrke,[1,8] Shaheynoor Talukder,[5,8] Anthony A. Philippakis,[1,3,6] Lourdes Peña-Castillo,[4] Trevis M. Alleyne,[5] Sanie Mnaimneh,[4] Olga B. Botvinnik,[1,7] Esther T. Chan,[5] Faiqua Khalid,[4] Wen Zhang,[5] Daniel Newburger,[1] Savina A. Jaeger,[1] Quaid D. Morris,[4,5] Martha L. Bulyk,[1,2,3,6,*] and Timothy R. Hughes[4,5,*]
[1]Division of Genetics, Department of Medicine
[2]Department of Pathology
Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[3]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138, USA
[4]Banting and Best Department of Medical Research
[5]Department of Molecular Genetics
University of Toronto, Toronto, ON M5S 3E1, Canada
[6]Harvard/MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115, USA
[7]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[8]These authors contributed equally to this work
*Correspondence: mlbulyk@receptor.med.harvard.edu (M.L.B.), t.hughes@utoronto.ca (T.R.H.)
DOI 10.1016/j.cell.2008.05.024

SUMMARY

Most homeodomains are unique within a genome, yet many are highly conserved across vast evolutionary distances, implying strong selection on their precise DNA-binding specificities. We determined the binding preferences of the majority (168) of mouse homeodomains to all possible 8-base sequences, revealing rich and complex patterns of sequence specificity and showing that there are at least 65 distinct homeodomain DNA-binding activities. We developed a computational system that successfully predicts binding sites for homeodomain proteins as distant from mouse as *Drosophila* and *C. elegans*, and we infer full 8-mer binding profiles for the majority of known animal homeodomains. Our results provide an unprecedented level of resolution in the analysis of this simple domain structure and suggest that variation in sequence recognition may be a factor in its functional diversity and evolutionary success.

INTRODUCTION

The approximately 60 amino acid homeobox domain or ''homeodomain'' is a conserved DNA-binding protein domain best known for its role in transcription regulation during vertebrate development. The homeodomain can both bind DNA and mediate protein-protein interactions (Wolberger, 1996); however, the precise mechanisms that dictate the physiological function and target range of individual homeodomain proteins are in general either unknown or incompletely delineated (Banerjee-Basu et al., 2003; Svingen and Tonissen, 2006). In several cases, functional specificity can be traced to the homeodomain itself (Chan and Mann, 1993; Furukubo-Tokunaga et al., 1993; Lin and McGinnis, 1992), indicating that individual homeodomains have distinct protein- and/or DNA-binding activities. Since many homeodomains have similar DNA sequence preferences, much attention has been paid to the role of protein-protein interactions in target definition (Svingen and Tonissen, 2006), despite evidence that the sequence specificity of monomers contributes to targeting specificity (Ekker et al., 1992) and that binding sequences do vary, particularly among different subtypes (Banerjee-Basu et al., 2003; Ekker et al., 1994; Sandelin et al., 2004). Indeed, it has been proposed that the DNA-binding specificity of homeodomains is determined by a combinatorial molecular code among the DNA-contacting residues (Damante et al., 1996).

Efforts to understand the physiological and biochemical functions of homeodomains have been hindered by the fact that most have only a few known binding sequences, if any. Position weight matrices (PWMs) have been compiled for 63 distinct homeodomain-containing proteins from human, mouse, *D. melanogaster*, and *S. cerevisiae* in the JASPAR (Bryne et al., 2008) and TRANSFAC (Matys et al., 2003) databases. These matrices are based on 5 to 138 individual sequences (median 18), presumably capturing only a subset of the permissible range of binding sites for these factors. Further, the accuracy of PWM models has been questioned (Benos et al., 2002), and there are many examples in which transcription factors bind sets of sequences that cannot be described in a conventional PWM representation (Blackwell et al., 1993; Chen and Schwartz, 1995; Overdier et al., 1994).

Moreover, the sequence preferences of the individual proteins can, in some cases, be altered by the binding context: For instance, the binding specificity of the complex of *Drosophila* Hox-Exd homeodomain proteins is remarkably different from

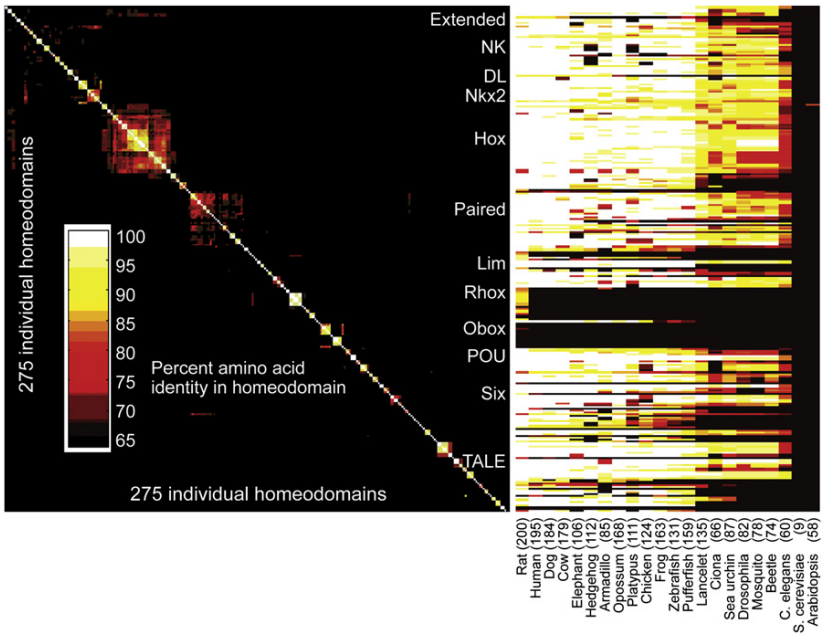


Figure 1. Conservation and Diversity of Mouse Homeodomains

Left: Heat-map showing the percent identity between different hierarchically clustered mouse homeodomains. Major homeodomain families are indicated. Right: percent identity to closest BLAST or BLAT hit in other species as indicated. The number of distinct homeodomain-containing protein counterparts in other species is given at bottom (isoforms are counted as a single entity).

that of the individual monomers (Joshi et al., 2007), raising the prospect that the monomeric binding preferences may not always be relevant to targeting in vivo. There is evidence that the sequence preferences of individual Hox proteins in *Drosophila* and mammals are significantly altered by physical interactions with protein cofactors in the PBC and Meis subfamilies, presumably through contacts to the Hox N-terminal arm that change the way the homeodomain contacts DNA (Mann and Chan, 1996; Wilson and Desplan, 1999). Other evidence, however, suggests that these examples of cofactor alterations to the monomer binding specificities are likely to be the exception rather than the rule. Carr and Biggin demonstrated that there is good correlation between monomer binding in vitro and in vivo for four fly homeodomain-containing proteins: Eve, Ftz, Bcd, and Prd (Carr and Biggin, 1999). Carroll and colleagues further showed that Ubx activity in promoting haltere development is independent of protein cofactors and that the promoters of its target genes in this pathway contain clusters of individual Ubx binding sites (Galant et al., 2002). Liberzon et al. showed not only that the specificity of the Hox-like mouse protein Pdx1 also extends beyond the TAAT core, but that the preferences at these flanking positions in vitro correlate with the ability of these sequences to stimulate transcription in vivo (Liberzon et al., 2004). In addition, for many domain classes, and in organisms ranging from yeast to human, in vivo binding sites detected by ChIP-chip typically contain sequences that reflect those preferred in vitro (Carroll et al., 2005; Harbison et al., 2004).

The mouse genome encodes a larger number of homeodomains than most vertebrates, including humans, and contains representatives of both ancient (NK, Hox) and young (Rhox, Obox) homeodomain families, encompassing striking examples of both purifying and diversifying selection (Jackson et al., 2006; Larroux et al., 2007; Rajkovic et al., 2002). The mouse homeodomain complement, estimated at 260 distinct proteins and 275 individual homeodomains (Bult et al., 2004), is broadly conserved across animals (Figure 1). For example, most mouse homeodomains (172/275, or 63%) have an identical human counterpart, and among these, most (107/172) have fewer than ten amino acid differences from their *Drosophila* counterpart. In contrast to their relative invariance over evolutionary time, however, most homeodomains within a genome are very different from other homeodomains within the same genome (Figure 1): Although there are 22 instances of mouse proteins with identical homeodomains, the median number of amino acid differences between any two mouse homeodomains is 37.

In this analysis, we sought to fully characterize the sequence preferences of mouse homeodomains in order to ask whether the binding activity is unique to each homeodomain and whether the full activity profile can be predicted from the primary amino acid sequence of the homeodomain, in a way consistent with a molecular code. We also explore the relevance of the monomeric binding preferences to binding sites in vivo. Since the mouse homeodomains exemplify the functional diversity inherited from the common ancestor of all animals, as well as the potential for homeodomain expansion and divergence, our results and conclusions are extendible across the animal kingdom.

RESULTS

Analysis of the Binding Preferences of Mouse Homeodomains to All 8-mers

Structures of homeodomains binding to DNA, as well as in vivo and in vitro selected binding sequences, are consistent with a typical binding footprint of seven or eight bases for a homeodomain monomer (Banerjee-Basu et al., 2003; Sandelin et al., 2004). To analyze the DNA-binding specificity, we used protein binding microarrays (PBMs) (Mukherjee et al., 2004) containing 41,944 60-mer probes in which all possible 10-base sequences are represented. Moreover, all nonpalindromic 8-mers occur on at least 32 spots on our microarray in different sequence contexts, thus providing a robust estimate of the binding preference of each protein to all 8-mers (Berger et al., 2006). For the facilitation of inference of wider motifs, the arrays also contain 32 instances of all gapped 8-mers up to a width of 12 bases. In total, we can reliably derive quantitative binding data for 22.3 million

gapped and contiguous 8-mers (4^8 sequence variants of 341 patterns up to 8 of 12) for any single protein. We used PBMs to analyze 194 of the 260 mouse homeodomain proteins for which we were able to produce protein as T7-driven, GST-tagged constructs by either in vitro transcription and translation or expression and purification from *E. coli*.

We systematically quantified the relative preference of each homeodomain for all possible 8-mers by several measures. These data, together with the raw microarray intensities, are in the Supplemental Data available online. The median normalized signal intensity from each 8-mer (and its Z score transform) scale almost linearly with K_a, when known (Berger et al., 2006), but may be sensitive to the amount of protein used in the assay (data not shown). We can additionally express the binding specificity of each protein as a mononucleotide PWM, or motif (contained in Table S1), but these often fail to fully capture the complete spectrum of binding activities and lack the resolution provided by individual word-by-word measurements (Benos et al., 2002; Chen et al., 2007). Here, we primarily employ a statistic we refer to as the enrichment score (E score) for each 8-mer, which is a variation on area under the ROC curve (AUC) and scales from 0.5 (highest) to −0.5 (lowest) (Berger et al., 2006). This measure is unitless and has a nonlinear scaling with intensity (there is a compression of the dynamic range among the most highly bound sequences), but on the basis of rank correlations and precision-recall analysis it is the most highly reproducible of any measure we have tested (Figure S3), and it facilitates comparison between separate experiments. On the basis of random permutations of the array data, our entire data set should contain no randomly arising E scores above 0.45. Using $E > 0.45$ for at least one 8-mer as a PBM success criterion, we obtained clear sequence preferences for 168 homeodomain proteins, including 11 different factors with identical homeodomain amino acid sequences. On average, each homeodomain had 144 such ungapped preferred 8-mers. It is possible that some proteins for which no sequence preference was obtained were improperly folded. The 26 we scored as unsuccessful, however, include seven of the nine Rhox isoforms tested, all three of the Lass isoforms tested, and both Satb isoforms tested, suggesting that these classes bind DNA nonspecifically or not at all or require modifications or cofactors not present in these experiments. This conclusion is supported by previous observations that *S*pecial *A*-*T*-rich *b*inding protein 1 (Satb1) binding preferences relate primarily to nucleotide composition and not to a specific sequence (Dickinson et al., 1992), a trend which is also present in our data (data not shown). Each of these 12 proteins exhibits a nonconsensus amino acid in at least one of the four positions conserved across nearly all homeodomains (positions 48, 49, 51, and 53 [Banerjee-Basu et al., 2003]), as do the majority of all failures that we obtained. Nonetheless, we observed sequence-specific binding for nine nonconsensus homeodomains, including Rhox6 and two novel homeodomains we have termed Dobox4 and Dobox5, indicating a potential means for acquiring additional diversity in DNA-binding specificity and function.

Comparison of PBM Data to Previously Determined Homeodomain Binding Preferences

As a first step in the analysis of our data, we compared our data to previously known binding sequences from the literature. Taking the 168 mouse proteins together with their closest ortholog in other metazoan species (regardless of the degree of similarity), the TRANSFAC and JASPAR databases contain at least one binding sequence corresponding to 97 mouse proteins or their orthologs (see the Supplemental Data for details). None of these proteins has more than 86 known binding sites, either in vitro or in vivo, in these databases. Nine of them (or an ortholog) have a PWM in the JASPAR database (derived from between 10 and 59 sequences obtained in vivo, in vitro, or both), and 58 more (or an ortholog) have a PWM in TRANSFAC (derived from between 5 and 86 binding sequences). An additional 30 of the 168 proteins we analyzed have between one and four known sites listed with a direct interaction observed in vivo or in vitro. We note that there are frequently multiple mouse homologs for each homeodomain in other species (e.g., Antp is the closest *Drosophila* homolog to the mouse Hox6, Hox7, Hox8, and Hox9 paralogs, so the Antp PWM represents the only data available for nine of the mouse homeodomains we analyzed).

Although the accuracy of the standard PWM model has been called into question, PWMs represent a straightforward means to compare binding activities on a coarse level. A visual comparison of the PWMs we derived from our data and those in the databases reveals reassuring similarities but also discrepancies with the existing literature (Table S1). For example, our PWMs for Lhx3, Meis1, Otx1/2, Nkx2-2, Pitx2, and Tgif1 are very similar to those previously determined. In some cases, however, our PWMs are somewhat different; for example, our Hmx3 PWM (resembling CAATTAA) is different from that previously determined from nine in vitro selected DNA sequences (resembling CAAGTGCGTG), although ours is very similar to those we obtained for the related proteins Hmx1 and Hmx2.

Perhaps the most obvious source of disagreement would be inconsistency in the initial data used to construct the motifs. We compared whether the individual sequences from JASPAR, which are determined by curators to be high quality, all contain 8-mers with high scores in our data. In some cases, all of the source sequences in JASPAR contain at least one 8-mer with an E score ≥ 0.45 in our data for the same protein; for example, all 41 of the human and mouse Lhx3 binding sequences meet this criterion, as do 17/18 Pbx1 binding sequences and 32/38 Nobox (Og2x) binding sequences. All of these proteins also have a PWM that is very similar to the one we derived from our data. In contrast, only one of ten in vitro selected sequences for the mouse En1 protein contains an 8-mer with $E > 0.45$ in our data, and the derived PWMs bear little resemblance (Table S1). Notably, the measured binding affinity of En1 for this one sequence was considerably higher than for any of the other nine selected sequences (Catron et al., 1993).

We conclude that our data are in most cases consistent with previous data, although in many cases there are discrepancies. We note that the previous data are also not always in agreement with each other; for example the En1 PWMs in TRANSFAC and JASPAR are quite different from each other, and also from the *Drosophila* Engrailed PWM in TRANSFAC, illustrating that motifs in databases and the literature cannot all be taken as a gold standard. We propose that heterogeneity in methods used to produce the DNA-binding data in the literature may underlie many of the differences between our results and previous findings:

Not only were the binding sites for separate proteins identified by different means, but even individual TRANSFAC matrices for single proteins are frequently derived from binding sequences compiled from multiple experimental methods. Further, these sequences often exhibit ascertainment bias reflecting which particular sequences were chosen to be examined by the investigators. In contrast, our data are homogeneous and were generated on a uniform, unbiased platform under standardized conditions, such that the binding activities of the different proteins should be directly comparable.

For 71 of the proteins we analyzed, there is no in vitro or in vivo binding site data, and for the majority, there is no PWM, in either mouse or the closest homolog in any species. To our knowledge, for several families, we describe a relatively uniform and apparently distinct binding profile for the first time. These encompass the Irx family (preferring sequences resembling TACATGTA), the Obox family (GGGGATTA), the Six family [G(G/A)TATCA], Gbx1/2 (CTAATTAG), and Pknox1/2 (CCTGTCA). Our data also include individual proteins with apparently unique sequence preferences, including Dux1 (CAATCAA), Hdx [(C/A)AATCA], Hmbox (TAACTAG), Homez (ATCGTTT), and Rhox11 [GCTGT(T/A)(T/A)]. The variety in motifs we obtained motivated us to further explore the similarities and differences among homeodomains within our data set.

Homeodomains Have Rich and Diverse Sequence Preferences

Figure 2A shows a 2D clustering analysis of the E scores of all 2585 8-mers that were bound by at least one homeodomain with $E > 0.45$. On a coarse level, the major features of the data structure correspond to the major homeodomain subclasses, and these large clusters contain sequences similar to those previously established for these subclasses, when known (Banerjee-Basu et al., 2003). For example, the largest feature (encompassing the upper-left part of Figure 2A) includes the Hox subclasses and other homeodomains that prefer a canonical TAAT core (Svingen and Tonissen, 2006). Roughly half of the homeodomains, however, have a stronger preference for other sequences, and many of the homeodomains that do bind canonical sequences also bind additional sequences (e.g., some of the Lhx classes are associated with the large TAAT binding cluster, but also have their own clusters of preferred 8-mers, boxed in Figure 2A). There are also instances of single proteins or small groups that have a distinctive 8-mer profile (Figure 2A). Indeed, when considering the top 100 highest-affinity 8-mers for each homeodomain, we identified 33 clearly separate DNA-binding activities. These binding profiles are distinguishable on the basis of limited overlap among the top 100 8-mers (among all 32,896 possible 8-mers when reverse complements are merged) for pairs of homeodomains (Figure 2B). As controls, our dataset includes 21 instances in which the same homeodomain was analyzed twice, either (1) as a freshly expressed aliquot from the same construct (three proteins) or an alternate construct (seven proteins) or (2) as a different gene with the same homeodomain sequence but different flanking residues (11 proteins). These 21 replicates invariably correlate highly: Among them, the top 100 overlap was 85 ± 8, such that proteins sharing fewer than 66 of 100 top 8-mers (99% confidence interval) were considered to have distinct binding activities. Figure 2B shows the resulting 33 specificity groups along the diagonal, accompanied by PWMs for representative members of each of the large families.

Members within each of these 33 groups, however, can be further distinguished by their lower-affinity binding sites and/or by differences in relative preference among the top 100 8-mers. For example, among the large group in the upper left of Figure 2B (bracketed) comprised of 42 proteins that are indistinguishable by the top 100 criterion, we identified 15 distinct subgroups on the basis of differences in their E score profiles over all 8-mers (Figure 3). Even though all proteins in this large group exhibit essentially the same dominant motif, clear sequence patterns are associated with the 8-mers distinctively preferred by each subgroup, and these patterns correlate with differences in their amino acid sequences (Figure 3). This is further illustrated in Figure 4. Lhx2 and Lhx4 both bind the same highest-affinity sites (8-mers containing TAATTA) but show clear, consistent preferences for different moderate- (TAATGA versus TAATCA) and lower- (TAACGA versus TAATCT) affinity sites (Figure 4A). Lhx3 and Lhx4 show greater similarity, both in binding profile and amino acid sequence, yet they have subtly different preferences for weaker 8-mers (Figure 4B). These differences only become apparent due to the richness of our dataset in capturing precise binding specificities at word-by-word resolution.

We repeated the analysis of Figure 3 for all 18 of the major groups shown along the diagonal in Figure 2B to examine whether they could be further divided by fine-grained differences in specificity (Figure S7). We considered (1) whether the motif(s) derived for any two proteins were clearly distinct and (2) whether differences in the E score profiles between proteins also contain motifs that distinguish the two binding activities. Our analysis identified a total of 65 distinct binding patterns that have a striking correlation with amino acid sequence similarity among the homeodomains (Figure S7 and see below). Although an approximation, this likely represents a lower bound on the true number of distinct patterns; for instance, our analysis places Lhx3 and Lhx4 in the same subgroup, yet we can still discern subtle differences in their 8-mer binding profiles (Figures 3 and 4).

From this analysis, we conclude that homeodomains encode distinctive DNA-binding activities and that there are often major differences between the activities of individual proteins with similar dominant sequence preferences. We also find that the dominant motif is usually unable to explain all of the data and is inferior to the full 8-mer profile in predicting the outcome of a similar experiment on an independent array (Figure S3) (Chen et al., 2007). Rather, our results are consistent with a model in which homeodomain sequence preferences may be best described as a composite of binding activities, possibly representing different binding modes with different relative affinities. This idea is supported by the report that Nkx2-5 has two distinct binding activities, one with higher affinity than the other (Chen and Schwartz, 1995); indeed, the Nkx2 group, like Lhx3 and Lhx4, is one of the 65 groups that appears as if it may be further subdivided (Figure S7).

Moreover, even the dominant motifs we obtain do not correspond perfectly with the identities of the canonical homeodomain specificity residues. The homeodomain binds DNA predominantly through interactions between helix 3 (recognition

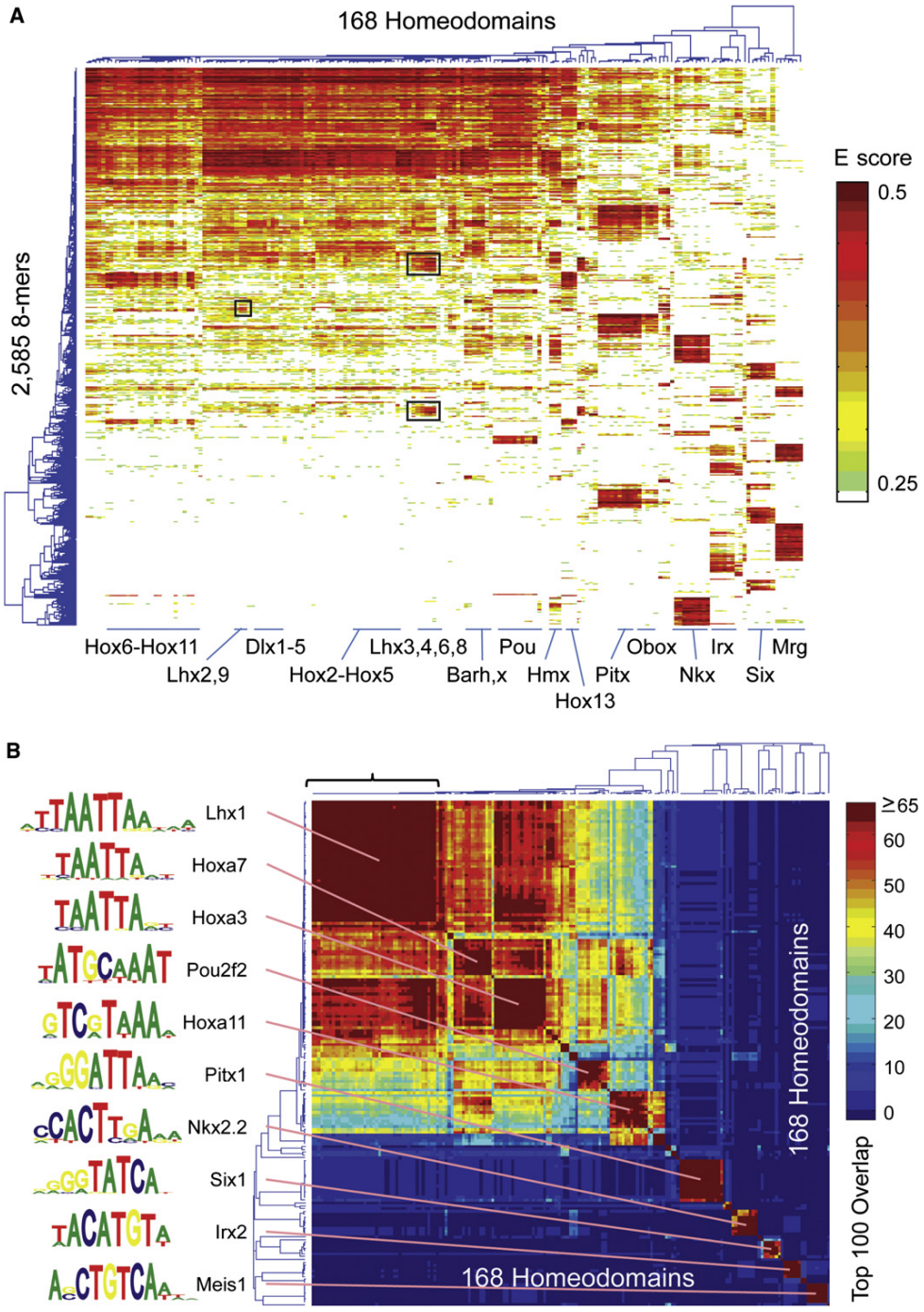


Figure 2. Overview of Homeodomains 8-mer Binding Profiles Reveals Distinct Sequence Preferences

(A) Hierarchical agglomerative clustering analysis of E score data for 2585 8-mers with E > 0.45 in at least one experiment. Boxed regions are referred to in the text. The position of exemplary homeodomain families within the dendrogram is indicated in order to highlight the diversity of overall 8-mer profiles.

(B) Clustering analysis of the matrix of overlaps in the top 100 8-mers (of all 32,896) for each pair of homeodomains. The bracket indicates the experiments analyzed in Figure 3. Logos for representative members of the major groups were determined with the Seed-and-Wobble method (Berger et al., 2006).

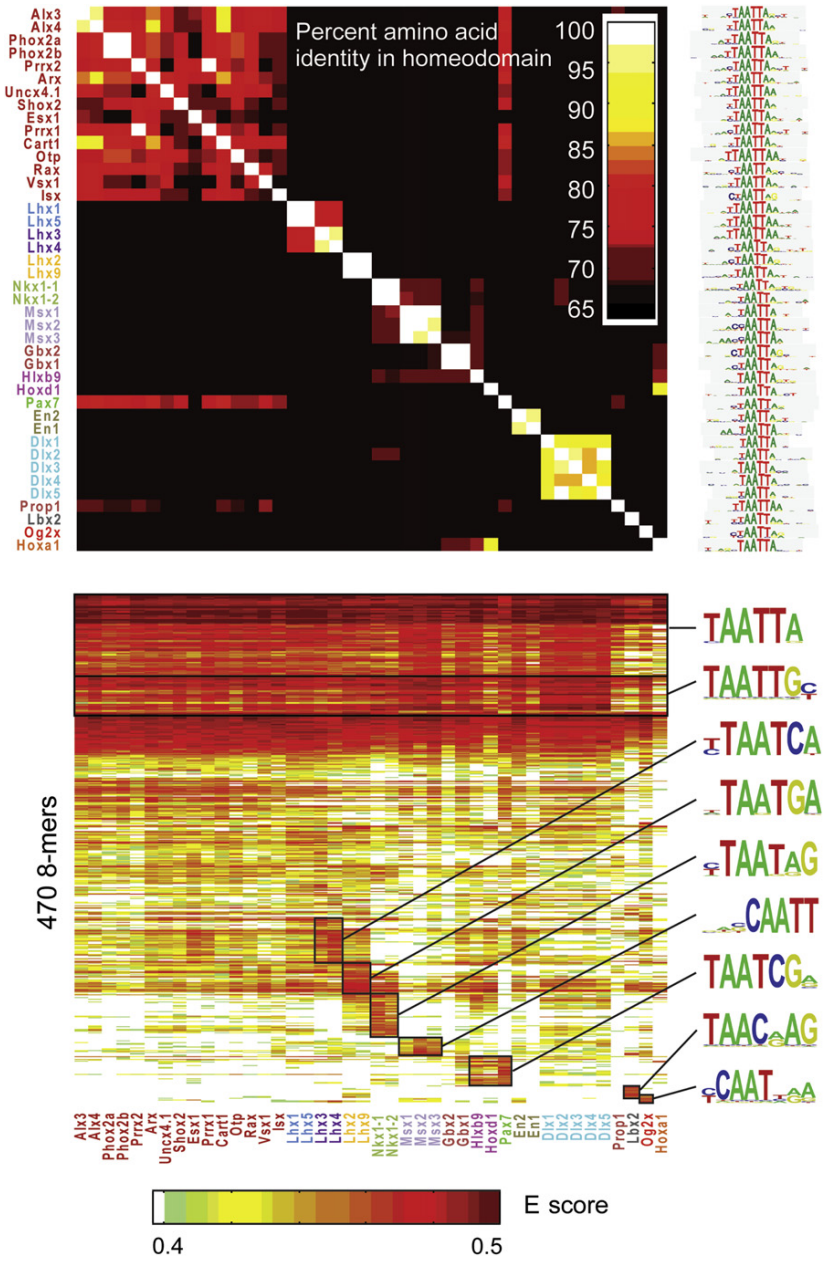


Figure 3. Homeodomains with Virtually Identical Dominant Motifs and Top 100 8-mer Preferences Have Differing Preferences for Many 8-mers

Bottom: Heat-map as in Figure 2, but restricted to the 470 8-mers with E > 0.45 in at least one of the experiments shown. Color of labels indicates groups that are distinct by our criteria. Logos were derived with ClustalW with the 8-mers in the boxed regions as inputs. Top: Amino acid similarities among these 42 homeodomains, as in Figure 1.

helix) and the major groove, and base-specific contacts made by positions 47, 50, and 54 are believed to be the main determinants of differences in binding specificity (Laughon, 1991) (Figure 5A, shown in red). Indeed, we were able to form groups harboring similar dominant motifs simply by partitioning homeodomains according to their amino acid identity at these three positions (Figure 5B). Our results are consistent with previous reports; for instance, replacement of glutamine with lysine at position 50 has been shown to dramatically alter the binding specificity through several newly formed hydrogen bonds to guanines (Tucker-Kellogg et al., 1997). These three residues alone are not sufficient to fully capture the entire binding activity, however, and in some cases, even the dominant motifs differ among proteins that have the same identity at these three residues (Figure 5B). Specific residues in the N-terminal arm have also been shown to influence binding specificities of homeodomains through minor groove interactions (Ekker et al., 1994); however, the identities at these residues (3, 6, and 7) do not correspond to the variation in Figure 5B (data not shown). Additional recognition positions must also be necessary to explain the differences in binding specificity we have observed for related homeodomains: Although we cannot exclude a molecular code controlling homeodomain DNA-binding activity (Damante et al., 1996), such a code is likely to be complex if one considers the full range of binding sequences.

Prediction of Binding Sequences across the Animal Kingdom Using Homeodomain Amino Acid Sequence Similarity

To more systematically and thoroughly approach the problem of identifying determinants of homeodomain sequence preferences, we tested the efficacy of a variety of methods to predict the full 8-mer binding profiles by using only the amino acid sequences as inputs (see the Supplemental Data for details). We evaluated each approach using leave-one-out crossvalidation (in which each homeodomain in turn was "held out" and its full 8-mer binding profile was predicted) to test our success at reproducing the 8-mer data for each of the 157 non-identical homeodomains, using Spearman correlation, top 100 overlap, and root mean squared error as success criteria in predicting the 8-mer profile. The most effective overall approach was a nearest-neighbor method, in which the 8-mer data were transferred from the homeodomain with the fewest number of mismatches over a set of 15 DNA-contacting amino acids

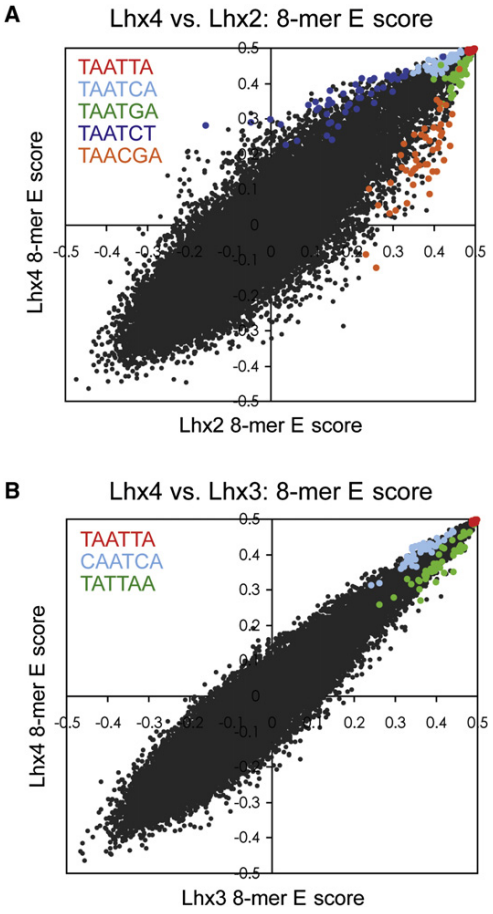


Figure 4. Scatter Plots Showing Differences in E Scores for Individual 8-mers between Lhx Family Members

(A) Comparison of Lhx2 and Lhx4.

(B) Comparison of Lhx3 and Lhx4.

8-mers containing each 6-mer sequence (inset) are highlighted, revealing clear systematic differences between sequence preferences despite essentially identical dominant motifs and sets of top 100 8-mers for these homeodomains.

(averaging the E scores in the case of ties). These 15 residues (3, 5, 6, 25, 31, 44, 46, 47, 48, 50, 51, 53, 54, 55, and 57; Figure 5A) account for all specific base-pair and phosphate backbone contacts in crystal structures for the *Engrailed* homeodomain (Fraenkel et al., 1998; Kissinger et al., 1990). The number of overlaps between the measured and predicted top 100 8-mers correlates with the distance to the closest example in the data, with zero, one, or two mismatches typically yielding predictions that are as close as an experimental replicate (Figure 6A). This result is consistent with our previous assessment of homeodomain DNA-binding activity subclassifications because there are more than 65 different naturally occurring variants among these 15 residues, groupings of which closely correspond to those obtained from the 8-mer profiles (see the Supplemental Data for details).

Consistent with the fact that much of the amino acid sequence variation among animal homeodomains is found in the mouse (Figure 1), the number of mismatches among these 15 amino acids from most mouse homeodomains to their homologs in species as distant as *Drosophila* is zero (Figures 1 and 6A and the Supplemental Data). We therefore applied the nearest-neighbor approach to project high-confidence 8-mer binding profiles for homeodomain proteins in 24 species (Supplemental Data). We found that in many cases, the predicted data were consistent with known motifs and binding sequences, even when the remainder of the homeodomain sequence had diverged considerably. We experimentally determined 8-mer E scores for the *C.elegans* homeodomain protein Ceh-22 by PBM and observed striking correlation with its predicted profile (Pearson correlation = 0.93, 78 of the top 100 overlap; Figure 6B) despite an overall difference of 11 amino acids within the homeodomain to the most similar mouse protein. Our inferred 8-mer profiles closely mirror quantitative in vitro measurements for the *Drosophila Engrailed* homeodomain, as well (Figure S8) (Damante et al., 1996).

Sequences Preferred by Homeodomains In Vitro Correspond to Sites Preferentially Bound In Vivo

Finally, we asked whether the homeodomain monomer binding preferences we identified in vitro reflect sequences preferred in vivo. Anecdotally, our highest predicted binding sequences do correspond to known in vivo binding sites. For example, in the predicted 8-mer profile for sea urchin Otx, a previously identified in vivo binding sequence (TAATCC, from the Spec2a RSR enhancer) (Mao et al., 1994), is contained in our top predicted 8-mer sequence, and, more strikingly, it is embedded in our fifth-highest predicted 8-mer sequence (TTAATCCT). At greater evolutionary distance, three of the four *Drosophila* Tinman binding sites in the minimal Hand cardiac and hematopoietic (HCH) enhancer (Han and Olson, 2005) are contained within the second (TCAAGTGG), fifth (ACCACTTA), and ninth (GCACTTAA) ranked 8-mers (the fourth overlaps the 428th ranked 8-mer [CAATTGAG], but also overlaps with a GATA binding site (Han and Olson, 2005) and may have constraints on its sequence in addition to binding Tinman).

To ask more generally whether occupied sites in vivo contain sequences preferred in vitro, we examined six ChIP-chip or ChIP-seq data sets in the literature that involved immunoprecipitation of homeodomain proteins that we analyzed, or homologs of proteins we analyzed that shared at least 14 of the 15 DNA-contacting amino acids. In all cases, we observed enrichment for monomer binding sites in the neighborhood of the bound fragments, with a peak at the center (Figure 7 and Figure S9). Figures 7A and 7B show two examples, *Drosophila* Caudal (Li et al., 2008) and human Tcf1/Hnf1 (Odom et al., 2006). For Caudal, the size of this ratio peak increased dramatically with E score cutoff, indicating that the most preferred in vitro monomer binding sequences correspond to the most enriched in vivo binding sites (cutoff E > 0.49) (Figure 7D) (51% of bound fragments have such an 8-mer, versus 17% in randomly selected fragments). For Tcf1/Hnf1, however, the majority of sequences bound in vivo do not contain the best in vitro binding sequences (E > 0.49), although most do contain at least one 8-mer with E > 0.45 (Figure 7C) (53%, versus 27% in random fragments), suggesting utilization of weaker binding sites. Similar results were obtained with PWMs (data not shown). Thus, the requirement for

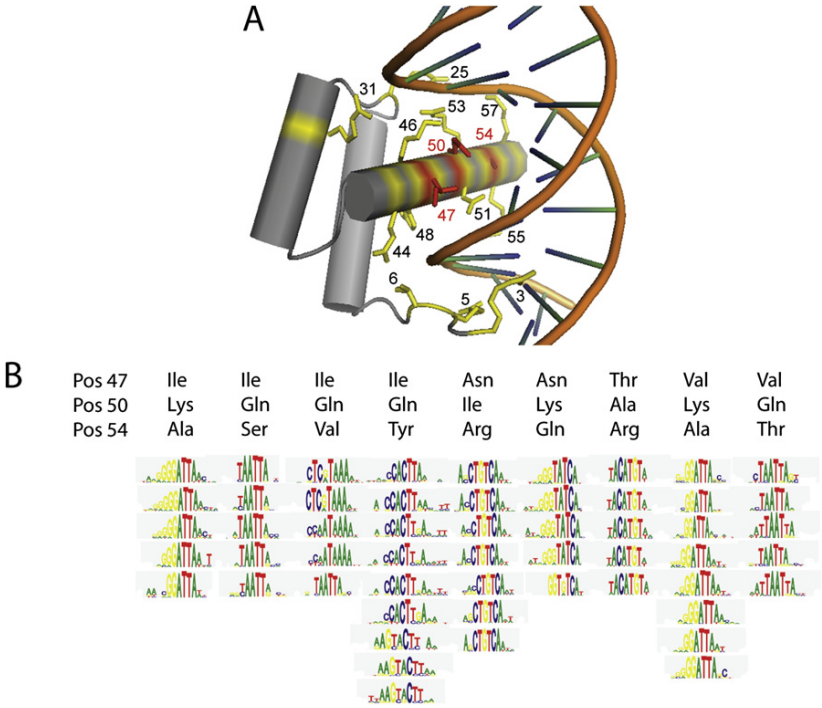


Figure 5. Correspondence between Canonical Homeodomain Amino Acid Sequence Specificity Residues and Dominant Motifs

(A) Protein-DNA interface for the *Drosophila* Engrailed protein (Kissinger et al., 1990). The three primary specificity residues discussed in the text are shown in red. The remaining residues considered in our nearest-neighbor analysis are in yellow.

(B) Motifs for all homeodomains in our dataset containing each of the displayed combinations of residues. For clarity, only those combinations occurring between five and ten times are shown. Logos represent PWMs determined with the Seed-and-Wobble method (Berger et al., 2006).

highest-affinity binding sequences may vary among homeodomain proteins, species, or under different physiological contexts. Nonetheless, a large proportion of the in vivo binding events apparently involve the monomeric homeodomain sequence preferences, which can be derived in vitro.

DISCUSSION

Our data provide a new level of resolution in the analysis of homeodomain sequence specificity. Our analyses show that homeodomains have distinctive sequence preferences, which may contribute to the strong selective pressure on their amino acid sequences, as well as to the biological specificity in target genes and diversity in function among the homeodomain proteins. Our findings should provide a fertile basis for future study of homeodomain function and evolution and may influence our understanding of evolved diversity in other transcription factor families.

One of the long-standing goals in the study of DNA-protein interactions has been to elucidate the relationships between amino acid residues and base preferences. Although it is clear that key residues can exert a strong influence, with others held constant (Hanes and Brent, 1989; Treisman et al., 1989), there is also evidence that alterations in the overall structure of DNA-binding domains can influence the DNA sequence preferences in unexpected ways (Miller and Pabo, 2001; Wolfe et al., 2001). Interactions among residues in the PWM (Benos et al., 2002) further complicate derivation of a deterministic recognition code. Full 8-mer profiles provide a new way to approach this problem. Although there is a correspondence between the canonical homeodomain DNA-binding specificity residues and the dominant motif, the correspondence is imperfect, and the dominant motif does not fully describe the complete binding profile, consistent with a model in which homeodomains have multiple binding modes. Perhaps as a consequence, our analyses suggest that categorization of the 8-mer profile on the basis of the full suite of DNA-contacting residues may be a more appropriate and practical paradigm for homeodomain sequence recognition than a molecular encoding of a PWM.

This idea is supported by our accurate prediction of full binding profiles over vast evolutionary differences. In fact, it is striking how little the entire homeodomain family has evolved at DNA-contacting residues since the common ancestor of all animals, considering that the potential for diversity in homeodomain DNA-binding activity seems well suited for duplication and divergence. Although newer binding activities (e.g., those of the Oboxes, Dobox4, Dobox5, Rhox6, and Rhox11) have apparently arisen since the divergence of mice and humans (there is no apparent homolog of these homeodomains in any species more distant than rat), the range of possible configurations even at the three canonical specificity residues (47, 50, and 54) appears to be sparsely populated in nature.

In all cases we tested, including predicted profiles for *Drosophila* homeodomains, the preferred monomer binding 8-mer sequences we obtained in vitro are enriched at the center of genomic fragments bound by the same protein in vivo. From this, we conclude that monomer binding preferences are likely to be a component of targeting mechanisms in general. Other factors (e.g., the chromatin landscape and protein-protein interactions) must also play a role in targeting because only a small fraction of all possible binding sites are occupied. We cannot exclude the possibility that the homeodomains we analyzed can undergo a radical change in binding specificity when they form complexes and that they rely on this or other mechanisms for a subset of in vivo binding events. Nonetheless, our demonstration that there are strong relationships between in vitro sequence preferences and in vivo binding sites supports the biological relevance of binding preferences of homeodomain monomers and indicates that our data should be of widespread use for identifying regulatory sites in vivo.

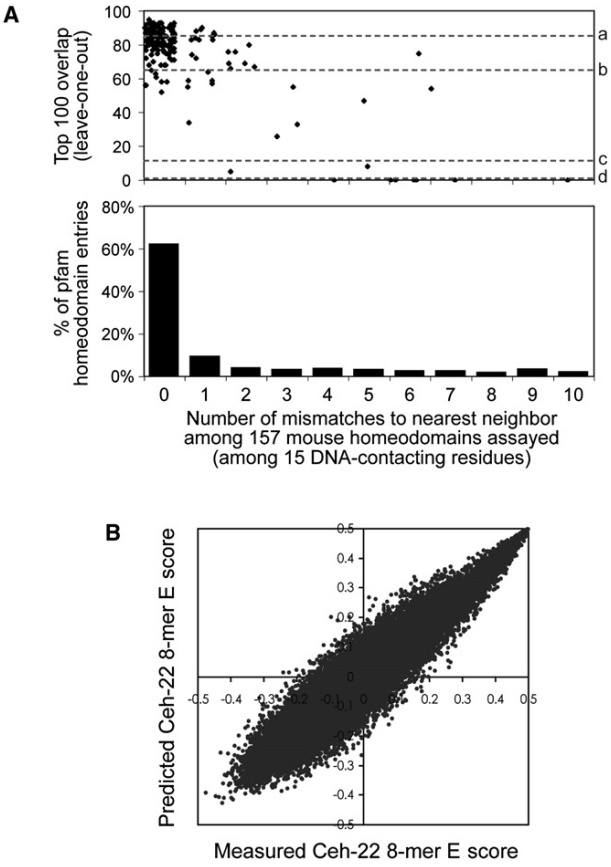


Figure 6. Correspondence between Homeodomain DNA-Contacting Amino Acid Sequence Residues and 8-mer DNA-Binding Profiles

(A) Top: Scatter plot showing the top 100 overlap between real and predicted 8-mer binding profiles from leave-one-out crossvalidation for our nearest-neighbor approach. Dashed lines indicate the following benchmarks: median, experimental replicates (a), 99% confidence, experimental replicates (b), median, randomized homeodomain labels (c) and median, randomized 8-mer labels (d). Within each bin, the x axis values have been nudged randomly for visualization. Bottom, the proportion of 3693 Pfam entries with the indicated identity to at least one mouse homeodomain analyzed.

(B) Predicted versus measured 8-mer E scores for *C. elegans* Ceh-22.

EXPERIMENTAL PROCEDURES

Cloning, Expressing, and Purifying Homeodomains

Homeodomain open reading frames, consisting of the Pfam-defined homeodomain and 15 amino acids of flanking sequence (or to the end of the full open reading frame) were cloned into pMAGIC1 (Li and Elledge, 2005) by either RT-PCR from pooled mouse mRNA or by gene synthesis (DNA 2.0). All clones were sequence verified (supplementary file "Protein and DNA sequence," available at http://hugheslab.ccbr.utoronto.ca/supplementary-data/homeodomains1/). We transferred the inserts into a T7-GST-tagged variant of pML280 following Li and Elledge (2005). We expressed proteins by either (1) purification from *E. coli* C41 DE3 cells (Lucigen) or (2) in vitro translation reactions (Ambion ActivePro Kit) without purification. Essentially identical results were obtained by either method (Figure S1).

Microarray Design and Use

The construction of "all 10-mer" universal PBMs with a de Bruijn sequence of order 10 has already been described (Berger et al., 2006) and is described in more detail in conference proceedings posted at http://thebrain.bwh.harvard.edu/RECOMB2007.pdf (Philippakis et al., 2008). For this study, we further optimized our design to achieve greater coverage of gapped *k*-mers (see the Supplemental Data for details). PBM assays were performed essentially as described previously (Berger et al., 2006), except that four proteins were simultaneously assayed in separate sectors of a single microarray and scanned with at least three different laser power settings to best capture a broad range of signal intensities and ensure signal intensities below saturation for all spots. Images were analyzed with GenePix Pro version 6.0 software (Molecular Devices), bad spots were manually flagged and removed, and data from multiple Alexa Fluor 488 scans of the same slide were combined with "masliner" software (Dudley et al., 2002) and normalized as described previously (Berger et al., 2006).

Sequence Analysis and Motif Construction

We provide several scores for each 8-mer in each experiment: (1) median intensity, (2) Z score, (3) enrichment score (E score), and (4) false discovery rate Q value for the E score. The median intensity and Z score measures follow standard statistical procedures. The E score has already been described in detail (Berger et al., 2006). In brief, for each 8-mer (contiguous or gapped), we consider the collection of all probes harboring a match as the "foreground" feature set and the remaining probes as a "background" feature set. We compare the ranks of the top half of the foreground with the ranks of the top half of the background by computing a modified form of the Wilcoxon-Mann-Whitney (WMW) statistic scaled to be invariant of foreground and background sample sizes. The E Score ranges from +0.5 (most favored) to −0.5 (most disfavored). We compute a false discovery rate Q value for the E score by comparing it to the null distribution of E scores (over 32,896 8-mers) calculated by randomly shuffling the mapping among the 41,944 probe sequences and intensities (repeated 20 times) (Subramanian et al., 2005). In computing all of the above scores, we do not consider probes for which the 8-mer occupies the most distal position on the probe (5′ with respect to the template strand) or for which the 8-mer overlaps the 24 nt primer region. We derive PWMs with the "Seed-and-Wobble" algorithm (Berger et al., 2006).

Predicting 8-mer Profiles and Scoring the Predictions

We considered two general methods for predicting 8-mer binding profiles on the basis of the primary amino acid sequence: nearest neighbor and regression. In the nearest neighbor (NN) approach, we predicted the 8-mer profile of any given homeodomain protein by taking the 8-mer profile(s) of its nearest neighbor(s) (averaging in the case of a tie). For regression, we converted the homeodomain amino acid sequence alignment to a binary representation by replacing all 20 standard amino acids in any of the canonical residue positions with unique 20 bit binary flags, the dimensionality was reduced by Principal Components Analysis (PCA), and a distinct model was learned for each 8-mer and for each homeodomain (i.e., a separate model for all 157 × 32,896 entries in the data table). We considered several variations of the distance metric used (e.g., number of mismatches versus amino acid similarity scores) and/or the residues considered (all 57 residues, 15 DNA-contacting residues, or five known specificity residues).

Chromatin Immunoprecipitation Analyses

We obtained 1331 bound sequences in the Caudal data set by selecting those in the 1% false discovery rate set where a peak was also reported (Li et al., 2008). We obtained 427 bound sequences in the Tcf1/Hnf1 data set (Odom et al., 2006) by implementing a program to perform the procedure described at http://jura.wi.mit.edu/young_public/hESregulation/Regions.html to the raw data. To create Figures 7A and 7B, we added 1 kb to either side of the ChIP-chip peak (for Caudal) or the center of the identified bound sequence (for Tcf1/Hnf1) and determined the relative enrichment in overlapping 500-base windows, using a 10-fold excess of 2 kb random genomic regions taken from the *Drosophila* genome (for Caudal) or the human genome (for Tcf1/Hnf1) as a background set.

Data Availability

Supplementary data, all original data files and array probe sequences are online at http://hugheslab.ccbr.utoronto.ca/supplementary-data/homeodomains1/ and http://the_brain.bwh.harvard.edu/pbms/webworks2/.

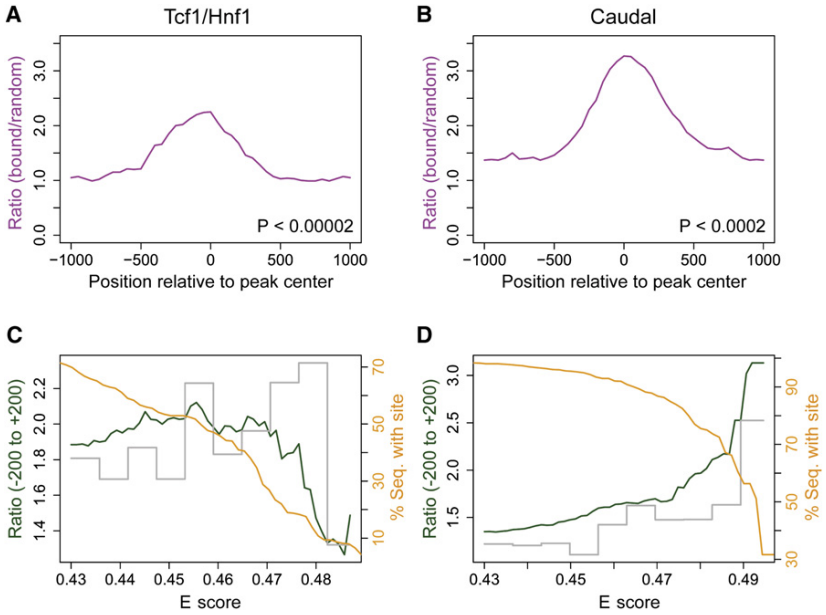


Figure 7. Enrichment of Sequences Preferred In Vitro within Genomic Sequences Bound In Vivo by the Same Protein

(A) Comparison of bound to randomly selected sequences for human Tcf1/Hnf1 (Odom et al., 2006), showing the relative enrichment of our 8-mers (at 0.456 cutoff). p value was calculated for the interval (−200 to +200) by the Wilcoxon-Mann-Whitney rank sum test, comparing the number of occurrences per sequence in the bound set versus the background set.

(B) Same as (A), but for *Drosophila* Caudal (Li et al., 2008) (at 0.493 cutoff).

(C) Relative enrichment (green line) in the −200 to +200 window as a function of the E score cutoff for Tcf1/Hnf1. The orange line shows the proportion of bound fragments with at least one such sequence in the same interval. The gray bars show the relative enrichment of 8-mers within each interval of 0.006, e.g., only 0.43–0.436 for the first interval.

(D) Same as (C), but for Caudal.

ACCESSION NUMBERS

Microarray data reported in this paper have been deposited in the NCBI GEO database with Platform ID GPL6760 and Series ID GSE11239.

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Experimental Procedures, Results, and Discussion, nine figures, four tables, and Supplemental References and can be found with this article online at http://www.cell.com/cgi/content/full/133/7/1266/DC1/.

ACKNOWLEDGMENTS

This project was supported by funding from the Canadian Institutes of Health Research (MOP-77721), Genome Canada through the Ontario Genomics Institute, the Ontario Research Fund, and the Canadian Institute for Advanced Research to T.R.H., M.L.B. and G.B., and by grant R01 HG003985 from NIH/NHGRI to M.L.B. M.F.B. was supported in part by a U.S. National Science Foundation Graduate Research Fellowship. A.A.P. was supported in part by a National Defense Science and Engineering Graduate Fellowship from the U.S. Department of Defense and an Athinoula Martinos Fellowship from HST. S.A.J. was supported in part by a U.S. National Science Foundation fellowship Postdoctoral Research Fellowship in Biological Informatics. We thank Genita Metzler, Hanna Kuznetsov, Chi-Fong Wang, Anastasia Vedenko, Frédéric Bréard, David Coburn, Dimitri Terterov, Ally Yang, Harm van Bakel, Wing Chang, and John Calarco for technical assistance and Shoshana Wodak, Fritz Roth, Charlie Boone, Jack Greenblatt, Ben Blencowe, Bill Stanford, and Trevor Siggers for helpful discussions and critical evaluation of the manuscript.

Received: November 21, 2007
Revised: March 10, 2008
Accepted: May 12, 2008
Published: June 26, 2008

REFERENCES

Banerjee-Basu, S., Moreland, T., Hsu, B.J., Trout, K.L., and Baxevanis, A.D. (2003). The Homeodomain Resource: 2003 update. Nucleic Acids Res. *31*, 304–306.

Benos, P.V., Bulyk, M.L., and Stormo, G.D. (2002). Additivity in protein-DNA interactions:How good an approximation is it? Nucleic Acids Res. *30*, 4442–4451.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. *24*, 1429–1435.

Blackwell, T.K., Huang, J., Ma, A., Kretzner, L., Alt, F.W., Eisenman, R.N., and Weintraub, H. (1993). Binding of myc proteins to canonical and noncanonical DNA sequences. Mol. Cell. Biol. *13*, 5216–5224.

Bryne, J.C., Valen, E., Tang, M.H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. Nucleic Acids Res. *36*, D102–D106.

Bult, C.J., Blake, J.A., Richardson, J.E., Kadin, J.A., Eppig, J.T., Baldarelli, R.M., Barsanti, K., Baya, M., Beal, J.S., Boddy, W.J., et al. (2004). The Mouse Genome Database (MGD): Integrating biology with the genome. Nucleic Acids Res. *32*, D476–D481.

Carr, A., and Biggin, M.D. (1999). A comparison of in vivo and in vitro DNA-binding specificities suggests a new model for homeoprotein DNA binding in Drosophila embryos. EMBO J. *18*, 1598–1608.

Carroll, J.S., Liu, X.S., Brodsky, A.S., Li, W., Meyer, C.A., Szary, A.J., Eeckhoute, J., Shao, W., Hestermann, E.V., Geistlinger, T.R., et al. (2005). Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell *122*, 33–43.

Catron, K.M., Iler, N., and Abate, C. (1993). Nucleotides flanking a conserved TAAT core dictate the DNA binding specificity of three murine homeodomain proteins. Mol. Cell. Biol. *13*, 2354–2365.

Chan, S.K., and Mann, R.S. (1993). The segment identity functions of Ultrabithorax are contained within its homeo domain and carboxy-terminal sequences. Genes Dev. *7*, 796–811.

Chen, C.Y., and Schwartz, R.J. (1995). Identification of novel DNA binding targets and regulatory domains of a murine tinman homeodomain factor, nkx-2.5. J. Biol. Chem. *270*, 15628–15633.

Chen, X., Hughes, T.R., and Morris, Q. (2007). RankMotif++: A motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. Bioinformatics *23*, i72–i79.

Damante, G., Pellizzari, L., Esposito, G., Fogolari, F., Viglino, P., Fabbro, D., Tell, G., Formisano, S., and Di Lauro, R. (1996). A molecular code dictates sequence-specific DNA recognition by homeodomains. EMBO J. *15*, 4992–5000.

Dickinson, L.A., Joh, T., Kohwi, Y., and Kohwi-Shigematsu, T. (1992). A tissue-specific MAR/SAR DNA-binding protein with unusual binding site recognition. Cell *70*, 631–645.

Dudley, A.M., Aach, J., Steffen, M.A., and Church, G.M. (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc. Natl. Acad. Sci. USA *99*, 7554–7559.

Ekker, S.C., von Kessler, D.P., and Beachy, P.A. (1992). Differential DNA sequence recognition is a determinant of specificity in homeotic gene action. EMBO J. *11*, 4059–4072.

Ekker, S.C., Jackson, D.G., von Kessler, D.P., Sun, B.I., Young, K.E., and Beachy, P.A. (1994). The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins. EMBO J. *13*, 3551–3560.

Fraenkel, E., Rould, M.A., Chambers, K.A., and Pabo, C.O. (1998). Engrailed homeodomain-DNA complex at 2.2 A resolution: A detailed view of the interface and comparison with other engrailed structures. J. Mol. Biol. *284*, 351–361.

Furukubo-Tokunaga, K., Flister, S., and Gehring, W.J. (1993). Functional specificity of the Antennapedia homeodomain. Proc. Natl. Acad. Sci. USA *90*, 6360–6364.

Galant, R., Walsh, C.M., and Carroll, S.B. (2002). Hox repression of a target gene: Extradenticle-independent, additive action through multiple monomer binding sites. Development *129*, 3115–3126.

Han, Z., and Olson, E.N. (2005). Hand is a direct target of Tinman and GATA factors during Drosophila cardiogenesis and hematopoiesis. Development *132*, 3525–3536.

Hanes, S.D., and Brent, R. (1989). DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. Cell *57*, 1275–1283.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99–104.

Jackson, M., Watt, A.J., Gautier, P., Gilchrist, D., Driehaus, J., Graham, G.J., Keebler, J., Prugnolle, F., Awadalla, P., and Forrester, L.M. (2006). A murine specific expansion of the Rhox cluster involved in embryonic stem cell biology is under natural selection. BMC Genomics *7*, 212.

Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B., and Mann, R.S. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell *131*, 530–543.

Kissinger, C.R., Liu, B.S., Martin-Blanco, E., Kornberg, T.B., and Pabo, C.O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: A framework for understanding homeodomain-DNA interactions. Cell *63*, 579–590.

Larroux, C., Fahey, B., Degnan, S.M., Adamski, M., Rokhsar, D.S., and Degnan, B.M. (2007). The NK homeobox gene cluster predates the origin of Hox genes. Curr. Biol. *17*, 706–710.

Laughon, A. (1991). DNA binding specificity of homeodomains. Biochemistry *30*, 11357–11367.

Li, M.Z., and Elledge, S.J. (2005). MAGIC, an in vivo genetic method for the rapid construction of recombinant DNA molecules. Nat. Genet. *37*, 311–319.

Li, X.Y., Macarthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C.L., et al. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. *6*, e27.

Liberzon, A., Ridner, G., and Walker, M.D. (2004). Role of intrinsic DNA binding specificity in defining target genes of the mammalian transcription factor PDX1. Nucleic Acids Res. *32*, 54–64.

Lin, L., and McGinnis, W. (1992). Mapping functional specificity in the Dfd and Ubx homeo domains. Genes Dev. *6*, 1071–1081.

Mann, R.S., and Chan, S.K. (1996). Extra specificity from extradenticle: the partnership between HOX and PBX/EXD homeodomain proteins. Trends Genet. *12*, 258–262.

Mao, C.A., Gan, L., and Klein, W.H. (1994). Multiple Otx binding sites required for expression of the Strongylocentrotus purpuratus Spec2a gene. Dev. Biol. *165*, 229–242.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. (2003). TRANSFAC: Transcriptional regulation, from patterns to profiles. Nucleic Acids Res. *31*, 374–378.

Miller, J.C., and Pabo, C.O. (2001). Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. J. Mol. Biol. *313*, 309–315.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat. Genet. *36*, 1331–1339.

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Nekludova, L., Rolfe, P.A., Danford, T.W., Gifford, D.K., Fraenkel, E., Bell, G.I., and Young, R.A. (2006). Core transcriptional regulatory circuitry in human hepatocytes. Mol. Syst. Biol. *2*, 2006.0017.

Overdier, D.G., Porcella, A., and Costa, R.H. (1994). The DNA-binding specificity of the hepatocyte nuclear factor 3/forkhead domain is influenced by amino-acid residues adjacent to the recognition helix. Mol. Cell. Biol. *14*, 2755–2766.

Philippakis, A.A., Qureshi, A., Berger, M.F., and Bulyk, M.L. (2008). Design of compact, universal DNA microarrays for protein binding microarray experiments. J. Comput. Biol., in press.

Rajkovic, A., Yan, C., Yan, W., Klysik, M., and Matzuk, M.M. (2002). Obox, a family of homeobox genes preferentially expressed in germ cells. Genomics *79*, 711–717.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: An open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. *32*, D91–D94.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Svingen, T., and Tonissen, K.F. (2006). Hox transcription factors and their elusive mammalian gene targets. Heredity *97*, 88–96.

Treisman, J., Gonczy, P., Vashishtha, M., Harris, E., and Desplan, C. (1989). A single amino acid can determine the DNA binding specificity of homeodomain proteins. Cell *59*, 553–562.

Tucker-Kellogg, L., Rould, M.A., Chambers, K.A., Ades, S.E., Sauer, R.T., and Pabo, C.O. (1997). Engrailed (Gln50→Lys) homeodomain-DNA complex at 1.9 A resolution: Structural basis for enhanced affinity and altered specificity. Structure *5*, 1047–1054.

Wilson, D.S., and Desplan, C. (1999). Structural basis of Hox specificity. Nat. Struct. Biol. *6*, 297–300.

Wolberger, C. (1996). Homeodomain interactions. Curr. Opin. Struct. Biol. *6*, 62–68.

Wolfe, S.A., Grant, R.A., Elrod-Erickson, M., and Pabo, C.O. (2001). Beyond the "recognition code": Structures of two Cys2His2 zinc finger/TATA box complexes. Structure *9*, 717–723.

Molecular Cell
Resource

A Library of Yeast Transcription Factor Motifs Reveals a Widespread Function for Rsc3 in Targeting Nucleosome Exclusion at Promoters

Gwenael Badis,[1] Esther T. Chan,[2] Harm van Bakel,[1] Lourdes Pena-Castillo,[1] Desiree Tillo,[2] Kyle Tsui,[3] Clayton D. Carlson,[4] Andrea J. Gossett,[6] Michael J. Hasinoff,[4] Christopher L. Warren,[4] Marinella Gebbia,[1] Shaheynoor Talukder,[1] Ally Yang,[1] Sanie Mnaimneh,[1] Dimitri Terterov,[1] David Coburn,[1] Ai Li Yeo,[7] Zhen Xuan Yeo,[7] Neil D. Clarke,[7] Jason D. Lieb,[6] Aseem Z. Ansari,[4,5] Corey Nislow,[1,2] and Timothy R. Hughes[1,2,*]
[1]Banting and Best Department of Medical Research
[2]Department of Molecular Genetics
[3]Department of Pharmaceutical Sciences
University of Toronto, Toronto, ON M5S 3E1, Canada
[4]Department of Biochemistry
[5]The Genome Center
University of Wisconsin-Madison, Madison, WI 53706, USA
[6]Department of Biology, Carolina Center for Genome Sciences and Lineberger Comprehensive Cancer Center, CB# 3280, 408 Fordham Hall, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3280, USA
[7]Computational and Systems Biology, Genome Institute of Singapore, 60 Biopolis Street, Singapore, 138672, Republic of Singapore
*Correspondence: t.hughes@utoronto.ca
DOI 10.1016/j.molcel.2008.11.020

SUMMARY

The sequence specificity of DNA-binding proteins is the primary mechanism by which the cell recognizes genomic features. Here, we describe systematic determination of yeast transcription factor DNA-binding specificities. We obtained binding specificities for 112 DNA-binding proteins representing 19 distinct structural classes. One-third of the binding specificities have not been previously reported. Several binding sequences have striking genomic distributions relative to transcription start sites, supporting their biological relevance and suggesting a role in promoter architecture. Among these are Rsc3 binding sequences, containing the core CGCG, which are found preferentially ~100 bp upstream of transcription start sites. Mutation of *RSC3* results in a dramatic increase in nucleosome occupancy in hundreds of proximal promoters containing a Rsc3 binding element, but has little impact on promoters lacking Rsc3 binding sequences, indicating that Rsc3 plays a broad role in targeting nucleosome exclusion at yeast promoters.

INTRODUCTION

The targeting of a transcription factor (TF) to specific genomic loci is determined by its DNA-binding activity, which is typically encoded by a conserved DNA-binding domain (DBD), together with cofactor interactions and the chromatin state of potential targets (Barrera and Ren, 2006). A foundation of any complete and accurate model of transcriptional regulation will be knowledge of the sequence specificities of DNA-binding proteins (Beer and Tavazoie, 2004; Segal et al., 2008). Despite intense study, there is currently no organism for which a complete encyclopedia of such TF sequence specificities exists. Even in the well-studied yeast *S. cerevisiae*, prior to this study, binding sequences were understood with confidence for only about half of its ~200 TFs. The majority of yeast TFs have been analyzed by ChIP-chip, but even when assayed under several different growth conditions (Harbison et al., 2004), these experiments often fail to identify either significant binding events or associated motifs, presumably because the TF is not binding DNA under the assay conditions. Further complicating de novo motif identification is the possibility that ChIP-chip and related techniques (e.g., ChIP-seq) may identify binding sequences for cofactors rather than the intended TF (Carroll et al., 2005). In some cases, it may be possible to infer TF sequence preferences on the basis of similarity among DBDs or identities of DNA-contacting residues (Berger et al., 2008; Wolfe et al., 2000), but for no DBD class is there a complete and accurate combinatorial code that dictates sequence specificity.

Incomplete knowledge of TF binding specificities hinders our understanding of basic mechanisms of transcription and nuclear organization. For example, RSC (*r*emodel the *s*tructure of *c*hromatin) is an abundant nuclear protein complex with a role in nucleosome organization at many yeast promoters (Cairns et al., 1996; Ng et al., 2002; Parnell et al., 2008). RSC contains two Gal4-class transcription factor-like proteins (Rsc3 and Rsc30) with very similar amino acid (AA) sequences but apparently different cellular functions (Angus-Hill et al., 2001; Wilson et al., 2006). Neither Rsc3 nor Rsc30 has known sequence specificity, and the mechanisms that target RSC to individual loci remain poorly defined.

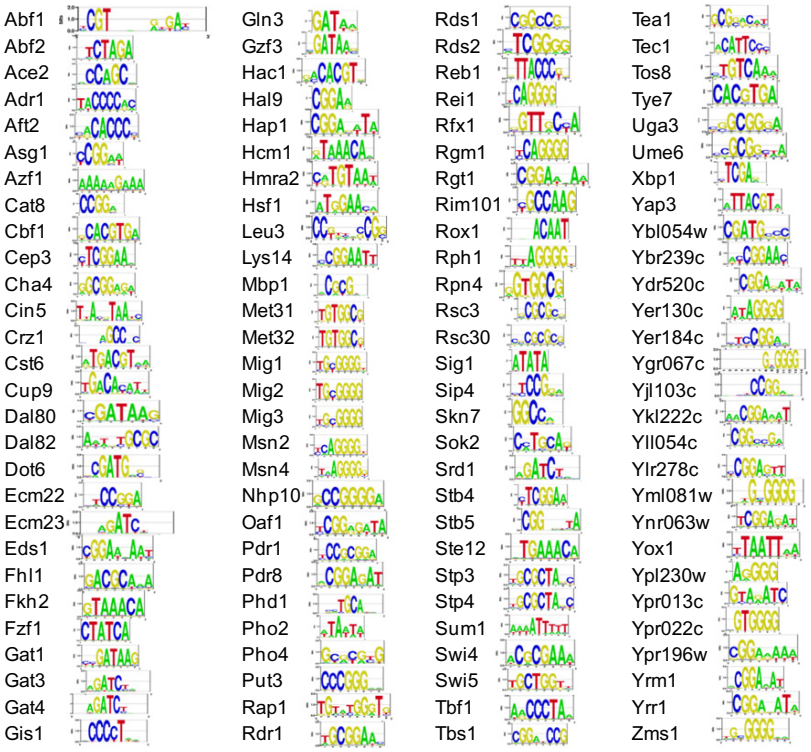


Figure 1. Motifs Identified in Our Study
Motifs represent Position Weight Matrices (PWMs), following Granek and Clarke (2005).

More generally, the mechanisms responsible for nucleosome-free regions (NFRs) in yeast promoters are incompletely understood. Current models of intrinsic nucleosome-DNA preference do not explain all of the observed nucleosome positioning and occupancy (Lee et al., 2007; Segal et al., 2006; Yuan and Liu, 2008). TF binding sequences are often enriched in NFRs (Lee et al., 2007; Liu et al., 2006), and in at least some cases, TFs make strong contributions to the local chromatin landscape. For example, Abf1, Reb1, and Rap1 are found frequently in yeast promoters and are able to define chromatin domains and enable activation or repression by other TFs in diverse pathways (Chasman et al., 1990; Elemento and Tavazoie, 2005; Fourel et al., 2002; Planta et al., 1995). Abf1, Reb1, or Rap1 binding sites are found in only a minority of promoters, however (Harbison et al., 2004), highlighting the probability that additional nucleosome-displacing factors, or combinations of factors, remain to be identified.

Here, we have measured the sequence preferences of the majority of yeast TF DBDs, using a combination of systematic microarray-based approaches. These data provide a resource for genomic analyses and for the study of the evolution of both the genome and the TFs themselves. Our data include binding preferences for 38 proteins, for which to the best of our knowledge, there was previously no reported binding specificity information, and provide independent support for many more that were previously inferred from ChIP-chip or identified on the basis of one or a few binding sequences. Among the proteins for which we have defined specificities for the first time are Rsc3 and Rsc30. Binding sequences for these proteins occur preferentially between 125 and 75 bp upstream of TSS (transcription start site), and Rsc3 is essential for the maintenance of a nucleosome-free region in hundreds of yeast promoters as well as transcript abundance from these promoters.

RESULTS

Creation of a Library of Sequence Specificities for 112 Yeast TFs

We began by creating a list of 218 yeast proteins that either contain a TF DBD or are known to bind to specific DNA sequences and regulate transcription (Table S1 available online). We were able to clone 207 of the 218 DBDs (or full-length proteins in the event that the DBD is unknown) as GST and/or MBP fusion proteins and, upon expression, obtained a protein for 195. We analyzed the sequence specificities of these 195 using at least one of three methods: (1) Protein Binding Microarrays (PBMs), in which the proteins are applied to an Agilent microarray consisting of 40,330 double-stranded 60-mers, each containing a unique 35-mer, such that all 10-mers are represented once and only once (Berger et al., 2006; Mintseris and Eisen, 2006); (2) Cognate Site Identifier (CSI) (Warren et al., 2006), in which proteins are applied to a NimbleGen array of 262,148 DNA hairpins each containing an 11 bp randomized region permitting display of all possible 10-mers; and/or (3) DNA immunoprecipitation-chip (DIP-chip) (Liu et al., 2005), in which a purified transcription factor, bound to yeast genomic DNA, is immunoprecipitated in vitro and analyzed using microarrays.

Table S1 and our project website contain a summary of which proteins were analyzed by each method and details on motif derivation. The majority of data produced resulted from PBMs (Berger et al., 2006). To discover the motifs preferentially bound by each protein in the PBM experiments, we first took the median signal intensity across the array from the 32 spots containing each 8-mer and expressed this as a Z score (Berger et al., 2006). We then sought DNA sequence motifs (Position Weight Matrices or PWMs) that produced predicted binding scores (Granek and Clarke, 2005) that correlated with the 8-mer-based Z scores for each factor (see Experimental Procedures for details). The 112 resulting motifs identified are shown in Figure 1. Figure 2A illustrates how the PWM-derived scores correlate with the 8-mer Z score data for Gzf3. Figure 2B, which shows a comparison of 8-mer Z scores obtained for Gzf3 using either PBM or CSI, demonstrates that the imperfect correlation cannot be attributed primarily to measurement noise in the assay or the array platform, because the 8-mer profile is consistent between these two different experiment types, even among less-preferred 8-mers. This observation may reflect shortcomings in

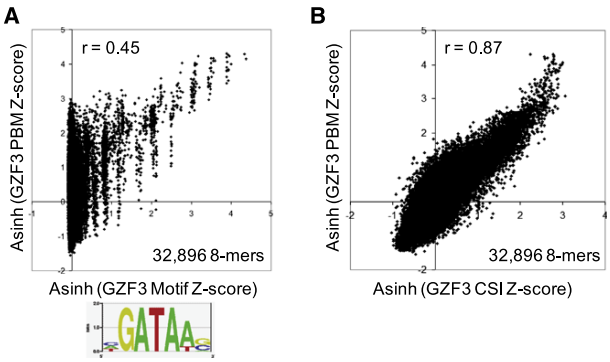


Figure 2. Comparison of Motif Representation and Reproducibility of 8-mer Profiles across Platforms

(A) PWM scores (Granek and Clarke, 2005) for all possible 8-mers for the single motif with highest Pearson correlation to the PBM 8-mers, plotted against the Z scores from the PBM. Data are plotted as asinh values, which are similar to natural log, but return real values for negative numbers (by definition, half of all Z scores are negative).

(B) CSI Z scores (combined from up to four array spots containing the 8-mer) versus Z scores from PBM.

PWM and consensus models (Benos et al., 2002). PWMs do, however, identify the best binding sequences in all of our experiments, and since they are compact, intuitive, and compatible with existing analysis techniques, we used PWMs for the remainder of our analyses.

63 of the 112 Motifs in our Library Correspond to Known Motifs

We next asked if the 112 motifs we obtained agree with those previously identified for the same proteins, from either global ChIP-chip analysis (Harbison et al., 2004; MacIsaac et al., 2006), or individual studies in the literature (Nash et al. [2007] and others), by manual comparison of logos, consensus sequences, and individual binding sites (Table S1). Sixty-three of our motifs bear an obvious correspondence to previous information (although not always all previous information), while 11 are inconsistent. For the remaining 38, we did not find any previously known motifs, although most of these motifs we obtained are consistent with expectation in some way (see below).

In Cases of Discrepancies with Existing Data, Evidence Supports the Newly Discovered Motifs

For some of the 11 discrepancies, additional evidence suggests that our measurements are likely to represent at least a correct in vitro monomeric binding sequence (Table S2). For example, our Fhl1 motif is a close match to that of its human homolog, FoxN1 (Schlake et al., 1997). Our motifs for Stp4 and Yml081w are very similar to those we obtained from Stp3 and Zms1, respectively, their corresponding yeast paralogs that arose from an ancient whole genome duplication (WGD) (Kellis et al., 2004). We verified by Electrophoretic Mobility Shift Assay (EMSA) that Stp3 and Yml081w bind to DNA sequences matching our motifs and not those previously described (Figure S1).

A few other discrepancies can be explained by the methodology we employed. For example, the A/T-rich motif we obtained for Sum1 is different from the published motif because when cloning DBDs, we selected the N-terminal AT hook domain rather than the C-terminal fragment that binds the established Sum1 motif but does not contain a known conserved domain (Pierce et al., 2003). Despite this discrepancy, promoter scans with our Sum1 motif do have a high correspondence to ChIP-chip results, suggesting that this additional DNA-binding activity of Sum1 may contribute to targeting in vivo (Spearman correlation $p < 10^{-92}$; Wilcoxon Rank Sum $p < 0.000011$ with 61 targets defined by Harbison et al. [2004] at $p < 0.001$).

Other variations from the literature are likely reproducible in vitro phenomena that are characteristic of members of a structural class. Four of the eight GATA-class proteins we analyzed (Ecm23, Srd1, Gat3, and Gat4) bound unexpectedly to sequences resembling the palindrome AGATCT. No binding sequences have been described for three of these four proteins, Ecm23, Srd1, or Gat4, and we know of no other in vitro or in vivo data that confirms or refutes our observations. A noncanonical motif different from AGATCT was derived for the fourth protein, Gat3, on the basis of ChIP-chip and sequence conservation of putative target sites (MacIsaac et al., 2006), and has not been experimentally pursued to our knowledge. Our motif does not correlate with the ChIP-chip data, which is highly enriched for subtelomeric loci. However, we confirmed by EMSA that Gat3 binds the sequence we identified more strongly than the sequence identified by ChIP-chip and that Ecm23 binds to the newly-identified motif (Figure S1).

Three of the discrepancies (Ecm22, Put3, and Ume6) are for Gal4-class proteins, which also have characteristic behavior in our analyses. It appears that our data largely capture monomeric specificities rather than the dimeric motifs typically associated with proteins in this class (MacPherson et al., 2006) (for all DBD classes, we counted correct monomeric specificities as consistent with previous information for dimeric proteins). Still, all but two of the motifs we obtained for Gal4-class proteins do contain the expected CGG core sequence (MacPherson et al., 2006), which is not always the case for the motifs derived from other studies. The capture of monomeric specificities could be a consequence of the domain definitions used for expression or the epitope-tagging strategy. In order to include dimerization contacts, our Gal4-class contacts included 50 AAs of flanking sequence beyond the boundaries of the DBD (or to the end of the protein if within 50 AAs). The choice of flanking sequence length was based on inspection of crystal structures of Gal4-class dimers binding to the DNA. However, the family is structurally diverse in the way the DBD dimerizes, and it may be that for some members of the family, the flanking sequence that was included was insufficient to mediate dimerization. In addition, our constructs are N-terminal GST fusions; Gal4-class DBDs are typically found at the N-terminus of yeast proteins and either dimerization or DNA binding by dimers may be influenced by N-terminal GST tags. The array designs we used may also fail to detect long motifs because the arrays are designed primarily to detect sequences up to ~10 bases (for PBM and CSI). Nonetheless, Gal4-class proteins do sometimes function in vivo as monomers (Kim et al., 2003; Larochelle et al., 2006; Vik and Rine, 2001) and several of our monomeric motifs are enriched in the promoters of functionally-related genes and at specific promoter positions (see below).

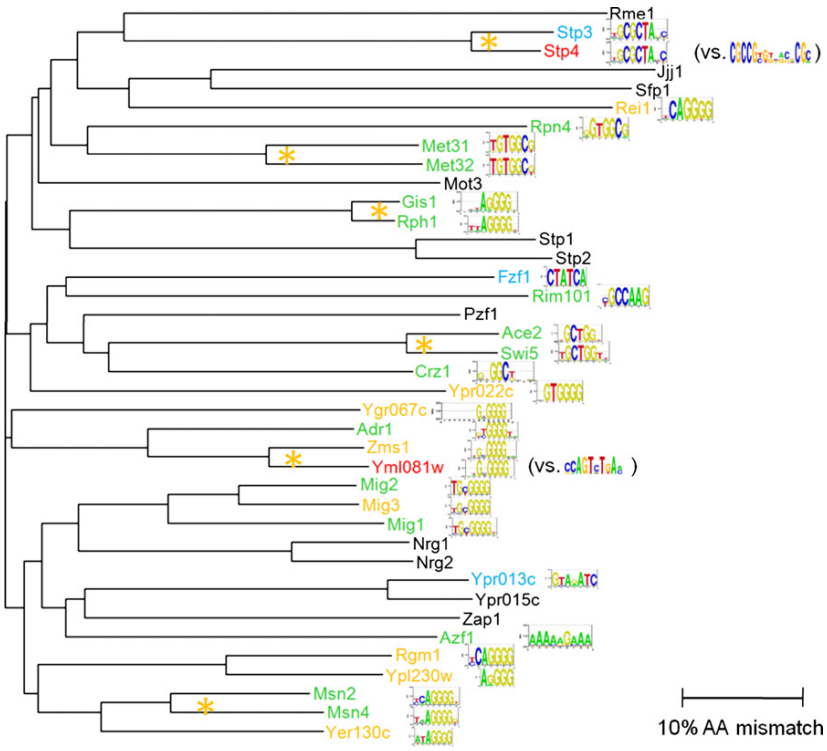


Figure 3. Similarity among C2H2 Zinc Finger Motifs Reflects DNA-Binding Domain Sequence Similarity

The phylogram tree was created based on AA similarity in DBDs using online EBI ClustalW with default settings. Motifs discovered in this study are shown next to protein names; inconsistent motifs from (MacIsaac et al., 2006) are shown for Stp4 and Yml081w. Yellow asterisks represent pairs arising from the WGD. Colors of protein names reflect our classifications of consistency with prior data: green, known motif obtained; red, discrepancy between our motif and that previously reported; yellow, new motif but consistent with expectation based on homology; blue, new unexpected motif.

Correspondence between Amino Acid Sequence Similarity and DNA-Binding Specificities Supports New Motifs

Based on our examination of literature and databases (described above and in Table S1), we classified 38 of the proteins for which we obtained a motif as having no previously established binding sequences. However, most of the 38 are members of structural classes that have characteristic binding site properties, and many are members of gene families that might be expected to share related sequence specificities. Indeed, most of our new motifs conform to expectation. The C2H2 zinc finger family provides several such examples (Figure 3). All three Mig proteins share virtually identical DNA-binding activities, as expected (Lutfiyya et al., 1998), as do Stp3 and Stp4, as described above. In contrast, C2H2 zinc finger proteins with unique motifs (Azf1, Crz1, Fzf1, Rpn4, Rei1, and Rim101) all have less than 60% identity to any other yeast protein in the DBD. ClustalW-derived phylograms similar to Figure 3 are given for all other structural classes in Figure S2. Three major observations include (1) Two Gal4-class proteins with related DBD sequences, Rsc3 and Rsc30, prefer sites that contain CGCG rather than the CGG typical of this class of proteins. Not coincidentally, perhaps, these two proteins are also unusual in having glycine at a position that is almost always lysine or arginine (corresponding to K20 in the Gal4 DBD). The lysine or arginine normally found at this position is in close proximity to the phosphate backbone in crystal structures of protein-DNA complexes (Figure S3). It is also just two positions C terminal to the residue that makes base-specific contacts to the usual CGG half-site. Thus, the unusual glycine at this position in Rsc3 and Rsc30 may affect the orientation of the domain with respect to DNA, resulting in the unusual DNA-binding specificity discovered here. (2) Dot6 and Ybl054w, a pair of related SANT domain proteins originating from the WGD (Kellis et al., 2004), both bound to sequences containing the core CGATG, which resembles the PAC (Polymerase A and C) motif (Dequard-Chablat et al., 1991). However, we found no evidence indicating that they bind to the promoters of genes containing these motifs (Harbison et al., 2004). (3) We obtained similar motifs containing the core TGTCA for Tos8 and Cup9, a pair of homeodomain proteins originating from the WGD. Neither protein has previously established binding specificity.

Many Motifs Are Enriched Upstream of Functionally Related Genes

We next scanned the yeast genome with the motifs and asked if the potential binding sites for each TF are associated with genes that share functional classes. Twenty-seven of the 112 motifs had a hypergeometric p value of $< 5 \times 10^{-6}$ (corresponding to a Bonferroni-corrected p value of 0.01) for enrichment of at least one GO Biological Process category among the top 100 promoter/motif hits. Expected enrichments include Ste12 (Sterile 12), with "cell-cell fusion" ($p < 2.2 \times 10^{-14}$), and Pdr1 (Pleiotropic Drug Resistance), with "response to drug" ($p < 1 \times 10^{-6}$). Our analysis is consistent with the function of Rgt1 (Restores Glucose Transport 1) as a Gal4-class TF that binds DNA as a monomer in vivo (Kim et al., 2003), since our monomeric motif is associated with "hexose transport" ($p < 6.1 \times 10^{-10}$). Ypr196w and Ydr520c binding sequences were also enriched in the promoters of hexose transporters ($p < 2.4 \times 10^{-8}$; 6.35×10^{-7}); the motifs for these proteins are related to that of Rgt1 and the top promoter/motif matches are found in an overlapping, but not identical, set of transporters, suggesting a more complex regulatory network of sugar utilization than is currently known. We were also intrigued to find that the monomeric motif we obtained for Lys14 has the same enrichment in promoters of lysine biosynthesis genes as the established dimeric motif ($p < 3.8 \times 10^{-6}$

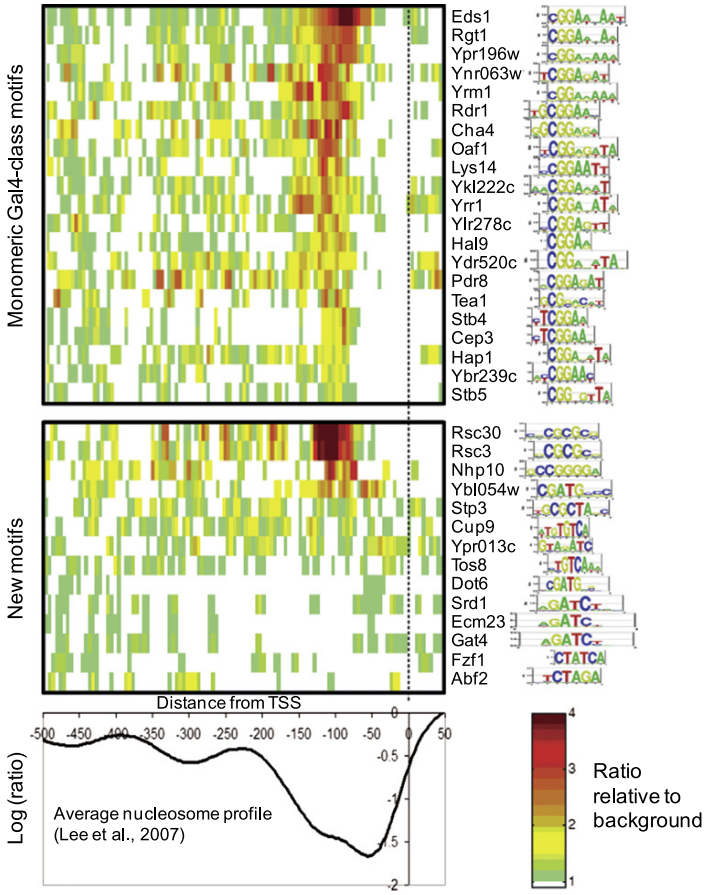


Figure 4. Bias in the Position of TF Binding Sequences

Bias in the position of TF binding sequences in 5,015 promoters with well-defined TSS (Lee et al., 2007). Motif scores (Granek and Clarke, 2005) were calculated for 8 bp windows, and high-scoring 8-mers were tallied along equivalent positions of all of the yeast promoter sequences using a cutoff selected to capture only the linear range of 8-mer binding Z score versus PWM score in PBM experiments (cutoff values are given in the Supplemental Data). Background was calculated from the first 100 bases of yeast ORFs. TFs are sorted by relative enrichment between −125 and −75.

for both), suggesting that both binding modes may be used in vivo.

Many New Motifs Are Preferentially Found in the NFR

We next examined how the occurrences of the motifs we discovered were distributed within promoters. Figure 4 (top) shows that most of our 21 monomeric Gal4 motifs occur preferentially in the position of the NFR (approximately −130 to −50, relative to TSS), providing support for their widespread in vivo relevance. Figure 4 (middle) shows 14 motifs we classified as new and unexpected; several of these are also located preferentially in the NFR. The most striking examples are Rsc3 and Rsc30, which share very similar binding preferences to sequences containing CGCG. At a stringent motif score threshold, these sequences are 16-fold more likely to occur in the position of the NFR than they are within genes. Only a handful of other TFs have this extreme bias (Lee et al., 2007), most notably Abf1 and Reb1, which are capable of remodeling chromatin in the vicinity of their binding sites. At a more liberal PWM score threshold, 708 yeast genes contain a potential Rsc3 binding sequence in the NFR region (−130 to −75), compared to only 146 found in an identical amount of ORF sequence. These 708 genes represent a broad spectrum of functional classes, including 169 (of 1101) that are essential for cell viability (hypergeometric $p < 2.2 \times 10^{-6}$). Given that RSC is an abundant protein complex that repositions nucleosomes (Angus-Hill et al., 2001; Cairns et al., 1996; Parnell et al., 2008), we reasoned that Rsc3 and Rsc30 may play a broad role in directing the establishment or maintenance of nucleosome-free regions in promoters. We focused on Rsc3 because it is essential, and therefore, it must be active under typical laboratory growth conditions.

Promoters Containing Rsc3 Binding Sequences Are Likely to Be Bound by RSC

Three previous studies have analyzed RSC binding sites in the yeast genome using ChIP-chip (Damelin et al., 2002; Ng et al., 2002; Parnell et al., 2008), two involving Rsc3. Promoters containing the Rsc3 motif displayed a statistically significant correspondence to overall RSC occupancy in these previous studies: among 5,015 (4,947 with ChIP-chip data) yeast genes with well-defined TSS (Lee et al., 2007), 2,325 (2,296 with ChIP-chip data) have a match to our Rsc3 motif (using our most liberal cutoff). Among these are 416 of 667 RSC targets defined in Ng et al., using a combined p value cutoff of < 0.01 (the p value of this overlap among 4,947 genes is $p < 4.36 \times 10^{-19}$). The correspondence to Rsc3 ChIP-chip occupancy (defined in Ng et al. [2002] using a p value cutoff < 0.01) is lower, although still significant (162 out of 293 targets; $p < 0.0011$). We note, however, along with others (Parnell et al., 2008), that ChIP-chip experiments with RSC subunits, particularly Rsc3, tend to have very low enrichment ratios. One possible explanation, consistent with the activity of RSC as an enzyme that displaces nucleosomes, may be that the association of RSC with target promoters is transient, as may be the case for the DNA-binding TFIIIC module, which also has relatively low ChIP-chip enrichments (Roberts et al., 2003; Soragni and Kassavetis, 2008). We therefore sought an alternative functional assay to ask if Rsc3 binding sites in promoters influence nucleosome occupancy.

RSC3 Is Required for the Formation of Nucleosome-free Regions at Promoters Containing Rsc3 Binding Sites

We assayed nucleosome occupancy in the *rsc3-1* mutant (Angus-Hill et al., 2001) using MNaseI digestion mapping and full-genome tiling arrays with 4 nt resolution (Lee et al., 2007). The biochemical defect of *rsc3-1* is unknown, but the mutations (M709I and L828S) are outside the DBD (AA1-37). We compared nucleosomal DNA enrichment (i.e., ratio of nucleosomal DNA

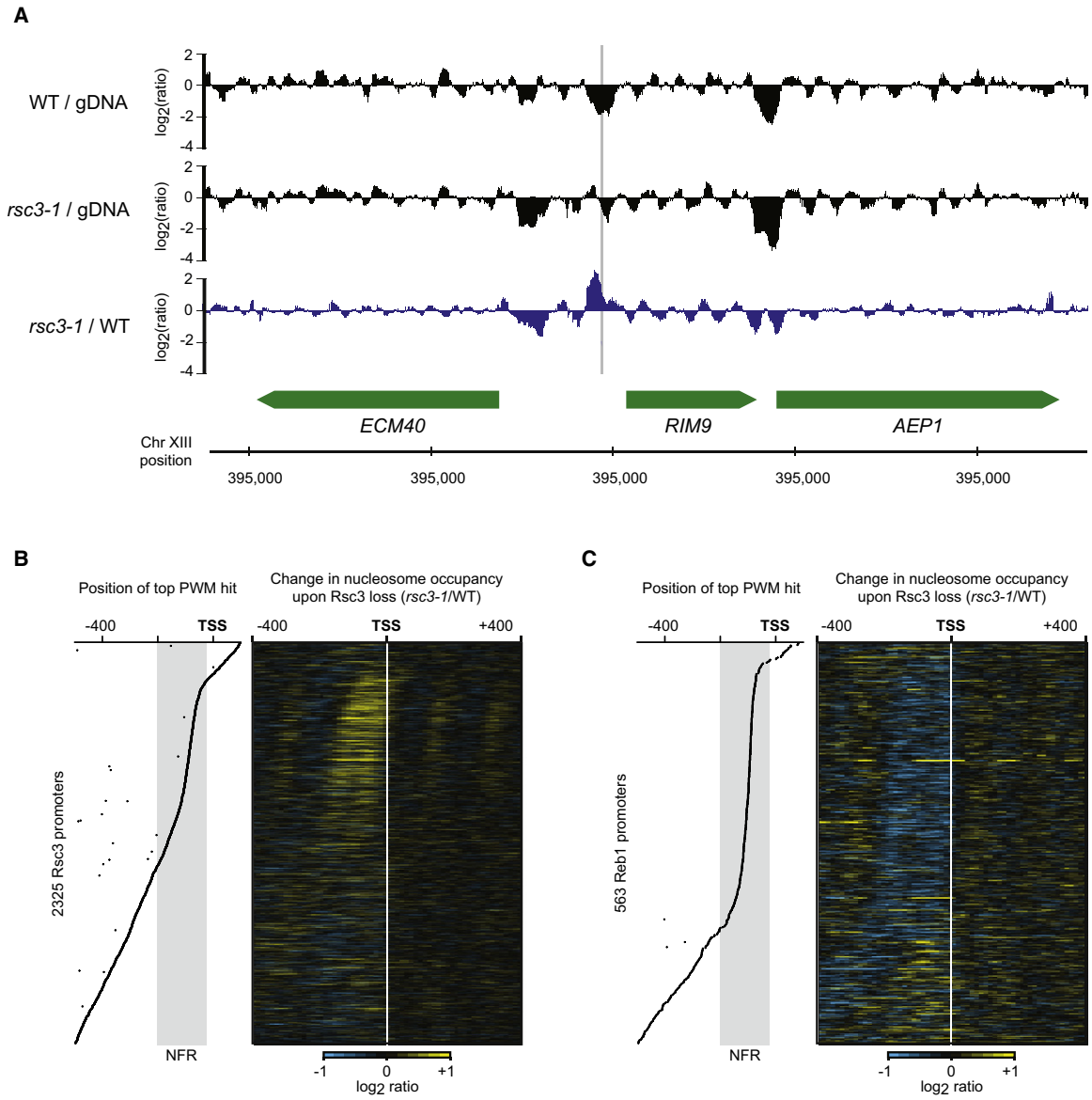


Figure 5. Rsc3 Influences Nucleosome Occupancy at Proximal Promoters Containing Rsc3 Binding Sequences

(A) A segment of Chromosome XIII with a Rsc3 binding sequence (gray vertical line) that is depleted in wild-type but occupied in the *rsc3-1* mutant.

(B and C) Changes in promoter nucleosome occupancy profiles between *rsc3-1* and a wild-type control for promoters containing Rsc3 binding sequences (B) or containing Reb1 binding sequences, but not Rsc3 binding sequences (C). Promoters are sorted by the position of the highest scoring Rsc3 or Reb1 binding sequence location in the promoter, which is shown at left in (B) and (C). Additional sites of equivalent PWM score are also indicated.

versus total genomic DNA) in the *rsc3-1* mutant to that in an isogenic wild-type control grown at the same temperature (37 degrees for 6 hr). Figure 5A shows an example locus in which nucleosome depletion over a Rsc3 binding sequence in a promoter region is dependent on *RSC3*. Figure 5B shows that this phenomenon occurs at many yeast promoters, with a clear preference for the affected region to be located near −100 from TSS. Moreover, the location of the increase in nucleosome occupancy (and the position of the NFR itself) tracks with the Rsc3 binding sequence across hundreds of promoters. Such changes are not observed at promoters that do not contain Rsc3 binding sequences (Figure 5C); in fact, nucleosome occupancy appears to decrease in these promoters, perhaps as a consequence of microarray signal normalization or redistribution of nucleosomes in vivo. This observation illustrates specificity of this phenomenon for Rsc3 binding sequences and not just NFRs in general. Unlike a previous study that used a greater tiling interval on selected promoters to examine the effects of mutating another RSC

subunit (Parnell et al., 2008), we saw little or no effect on nucleosome positioning or occupancy at tRNA genes (Figure S4), indicating that the effects we observed are distinct from a general loss of RSC activity. We also surveyed RNA abundance in the *rsc3-1* strain using the same arrays and observed a clear trend in which the Pol II promoters with an increase in nucleosome occupancy tend to exhibit lower RNA abundance (Figure 6). Overall, our results are consistent with a function for Rsc3 in nucleosome removal and promoting transcription from Pol II promoters that contain Rsc3 binding sequences in the NFR region.

In order to ask whether the effect of Rsc3 is mediated by RSC, we compared the relative occupancy of Rsc8 in wild-type and *rsc3-1* strains using ChIP-chip. In previous studies (Damelin et al., 2002; Ng et al., 2002; Parnell et al., 2008), Rsc8 has the highest occupancy ratios of any RSC subunit with up to 6-fold enrichment at tRNAs. In our wild-type strain, Rsc8 occupancy ratios are also highest at tRNAs (maximum enrichment 8.5-fold in our analysis, Figure S4), and at Pol II promoters, there is a significant correspondence between Rsc8 occupancy and the Rsc3 motif score (Spearman rank correlation $p < 1.3 \times 10^{-9}$). We found that occupancy at tRNAs is not affected by *rsc3-1* (Figure S4), suggesting that RSC is targeted to Pol III transcripts by a *RSC3*-independent mechanism. Surprisingly, in *rsc3-1*, we saw a global (albeit modest) increase in occupancy of Rsc8 at Pol II promoters (Figure 6), which could be an indirect effect of the fitness defects seen in *rsc3-1* mutant cells (Angus-Hill et al., 2001), and/or the dramatic alterations we observed in chromatin organization and transcript profiles. Nonetheless, the increase is clearly smaller for promoters in which nucleosome occupancy increases in response to *rsc3-1* (Figure 6), and it is also smaller for those promoters carrying a Rsc3 sequence (Wilcoxon rank sum test $p < 2.7 \times 10^{-5}$ among Rsc8-bound promoters, with Rsc3 positives defined as genes with a Rsc3 site in the NFR [−150 to −70]). Together, these observations suggest that Rsc3 may function by targeting RSC but do not rule out the possibility that Rsc3 acts by other mechanisms.

Other TFs Contribute to Nucleosome Occupancy at Promoters Containing Their Cognate Binding Sequences

Finally, we asked whether other TFs have an impact on nucleosome occupancy and transcription similar to that observed for Rsc3. Indeed, the correspondence between Rsc3 binding sequences and the impact of the *rsc3-1* mutant on nucleosome occupancy in promoters and transcript levels from the corresponding gene is similar to that seen with Abf1 and Reb1 (Figure 6 and Figure S5). Binding sequences for these TFs are found in the proximal promoter of hundreds of yeast genes, and as predicted from their known roles as chromatin modifiers, mutation of each TF results in a specific increase in the occupancy of nucleosomes over the potential binding site (Figure 6), with the most affected NFRs in the mutants typically containing the TF binding sequence. We also analyzed nucleosome occupancy in mutants in the essential DNA-binding proteins Tbf1, Rap1, and Mcm1; all three appear to influence nucleosome occupancy at promoters containing their cognate binding sequences, although the number of promoters affected is smaller than for Rsc3, Abf1, and Reb1 (Figures S5 and S6). By way of comparison, there is no relationship between binding sequences for Cep3, a centromere-binding protein, and nucleosome occupancy at Pol II promoters in a *cep3* mutant (Figure 6 and Figure S5). There is, however, a match to the Cep3 motif in all sixteen yeast centromeres, and the array signal in our nucleosome preparations at each centromere is depleted in the *cep3* mutant (Figure S7).

DISCUSSION

Our in vitro survey of yeast TF-DBD sequence specificities raises the number of yeast TFs with known sequence preference to 174, or ~80% (Table S1). This expanded index of sequence preferences provides a resource for exploration of the function and evolution of gene regulatory networks. Our comparison of predicted promoter preferences to GO categories represents only one possible exploratory approach; by examining correlations between theoretical promoter affinity for TFs (Granek and Clarke, 2005) and relative induction or repression in individual microarray experiments, we have identified many statistically significant associations (A.L.Y., Z.X.Y., N.D.C., and T.R.H., unpublished data). In addition, because motif representations almost certainly do not fully describe in vitro TF binding preferences (e.g., see Figure 2) and because previous studies have concluded that weak and/or noncanonical binding sites are likely to be functional in some instances (Blackwell et al., 1993; Buck and Lieb, 2006; Tanay, 2006), in the future it may be useful to scan the genome with indices of relative affinity to individual sequences rather than positional models of specificity.

One aspect of global gene expression and regulation that has been difficult to model is precisely how factors within cells assemble at promoters rather than other genomic locations with similar sequence characteristics. In our study, Rsc3 emerged as a major player in NFR formation/maintenance and promoter function for hundreds of yeast genes. Our data are consistent with prior conjecture that Rsc3 uses its sequence-specific binding activity to target RSC to promoters and creating the NFR (Angus-Hill et al., 2001; Parnell et al., 2008; Wilson et al., 2006). Our data are also consistent with previous ChIP-chip analyses of RSC because promoters containing Rsc3 binding site are enriched in RSC immunoprecipitates. Rsc3 itself is frustratingly refractory to study by ChIP-chip (Parnell et al., 2008); although there is a significant enrichment of Rsc3 binding sites among ChIP-chip targets, the enrichment ratios, the overlap with Rsc3 binding sequences, and the resolution of published ChIP-chip data (Damelin et al., 2002; Ng et al., 2002; Parnell et al., 2008) are all too low to specify exact target interactions. Therefore, we cannot rule out that the effects of Rsc3 on occupancy of many promoters are indirect, although we have no other explanation for the extremely strong association between Rsc3 binding sequences and the promoter nucleosome occupancy changes in the *rsc3-1* mutant (Figures 5 and 6). Several other TFs bind to sequences containing CGCG (e.g., Mbp1, Swi6, Dal82, and Rsc30), but no other known TF binding site (Harbison et al., 2004) or binding sequence (MacIsaac et al., 2006 and this study) correlates as powerfully with the *rsc3-1* data as does that of our Rsc3 PWM (Spearman rank correlation $p < 4.4 \times 10^{-43}$ between the Rsc3 PWM score and the relative change in the NFR in *rsc3-1* shown in Figure 6). Moreover, motif searches in

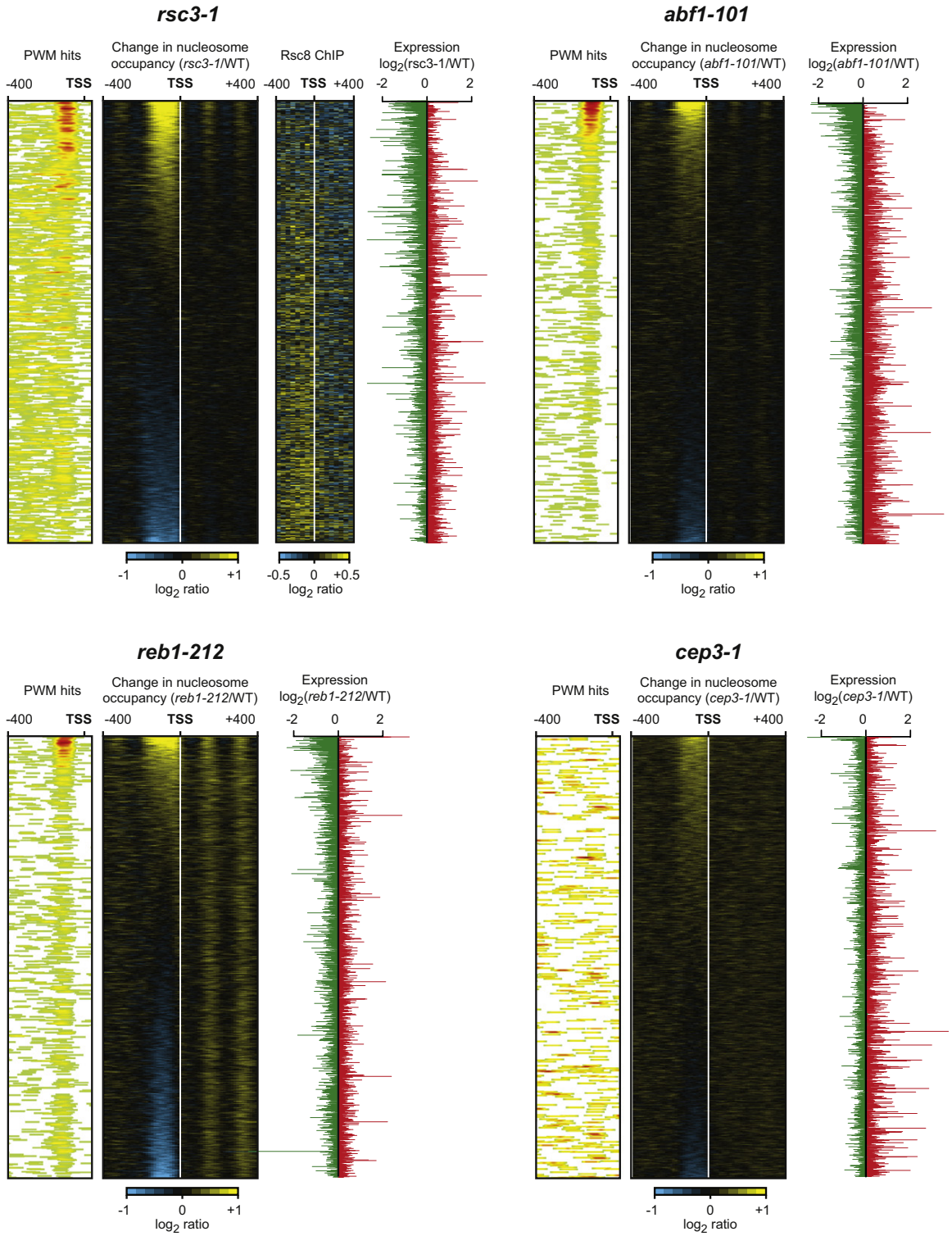


Figure 6. Comparison of the Effects of Mutations in Essential DNA-Binding Proteins on Nucleosome Profiles at all Promoters

Within each panel, promoters are sorted by change in occupancy in the NFR. Locations of binding sequences for the mutated factor are illustrated at left in tiling intervals matching those of the array and shown as heat-maps. The change in nucleosome occupancy in the mutant is shown in the middle. Relative transcript levels are illustrated at right. The *rsc3-1* panel (upper left) also shows the change in relative enrichment in Rsc8-TAP ChIP-chip between the *rsc3-1* and wild-type strains.

the promoters most affected in *rsc3-1* yield CGCG-containing motifs (data not shown).

Promoters in diverse organisms are enriched for both characteristic DNA structural features and binding sites for specific proteins (Lee et al., 2007). Our analyses extend these observations and, furthermore, demonstrate that many TFs contribute globally to either establishment or maintenance of the NFR (Figures 5 and 6 and Figures S3 and S4). Our data also link NFR formation to promoter function, since in all of the TF mutants we analyzed, an increase in nucleosome occupancy in the NFR generally corresponds to a decrease in transcript levels (Figure 6 and Figure S4). However, it is also true that correlation between binding sequences and effect of mutation is imperfect in all of the TF mutants we analyzed, supporting the notion that NFRs, and promoters, are created by a combination of factors, likely including both DNA structural features and specific TF recognition sites. It is curious and somewhat unexpected that the TFs that play key roles in NFR formation in yeast are not highly Fconserved proteins: obvious orthologs of Reb1, Abf1, and Rsc3 are not found outside of fungi (Wilson et al., 2006). Possibly, TFs involved in promoter establishment evolve with gene architecture, chromosome structure, and nuclear organization. If this is the case, then large-scale study of TF binding specificities in other organisms may be needed as much to understand how the cell identifies genomic landmarks as to map regulatory pathways.

EXPERIMENTAL PROCEDURES

Additional details and data are found in the Supplemental Experimental Procedures (see below) and on our project web site (http://hugheslab.ccbr.utoronto.ca/supplementary-data/yeastDBD/).

Cloning and Protein Expression

We cloned PCR amplicons (pfam-defined DBDs plus 50 flanking residues) into pMAGIC (Li and Elledge, 2005). Resulting inserts were transferred into pTH1137, a T7-GST-tagged variant of pML280 (Berger et al., 2008). We obtained proteins by either purification from *E. coli* C41 DE3 cells (Lucigen), or in vitro transcription/translation reactions (Ambion ActivePro Kit) without purification, as indicated on our project web site.

Microarray Analysis of TF Binding Specificities

The Supplemental Experimental Procedures contain a detailed description of microarray analyses and motif derivation methods. PBM arrays and assays were as described (Berger et al., 2006). CSI methods essentially followed (Warren et al., 2006). DIP-chip was carried out as described previously (Liu et al., 2005), and the resulting DNA was hybridized to NimbleGen microarrays covering the yeast genome at 32 bp resolution.

Nucleosome and Expression Analyses Using Tiling Arrays

Extraction of nucleosomal DNA from the samples and hybridization onto the yeast tiling array was performed according to Lee et al. (2007). Isolation of total RNA and hybridization onto the tiling arrays followed (Juneau et al., 2007), except that Actinomycin D was added in a final concentration of 6 μg/ml during cDNA synthesis to prevent antisense artifacts (Perocchi et al., 2007).

ChIP-chip

We grew isogenic wild-type and *rsc3-1* strains, each carrying Rsc8-TAP, in parallel under *rsc3-1*-restrictive growth conditions. After formaldehyde crosslinking and chromatin extraction, we performed a single pull-down with IgG sepharose. Following decrosslinking, we analyzed these samples on Nimblegen tiling arrays using a two-color procedure, comparing the pulled-down DNA to genomic DNA. We then compared relative enrichment between wild-type and *rsc3-1*.

Scoring Promoter Sequences and GO Enrichment

The probability of a transcription factor binding somewhere within a promoter was estimated using PWMs obtained in this study and the program GOMER (Granek and Clarke, 2005), run with default parameters, with promoters defined as the 600 bp region 5′ to the ORF. The top 100 hits were input into FunSpec (Robinson et al., 2002).

Additional Information

Additional information including clone sequences and 8-mer scores and motifs for all TFs can be found in Tables S4–S7.

ACCESSION NUMBERS

Affymetrix tiling array data are available at ArrayExpress (record E-MEXP-1754); all other microarray data are available at GEO (record GSE12349).

SUPPLEMENTAL DATA

The Supplemental Data include eight tables, seven figures, Supplemental Experimental Procedures, and Supplemental References and can be found with this article online at http://www.cell.com/molecular-cell/supplemental/S1097-2765(08)00842-3.

ACKNOWLEDGMENTS

This work was supported by Genome Canada through the Ontario Genomics Institute, the Ontario Research Fund, a grant from the CIHR to C.N. and T.R.H. (MOP 86705), and grants from NIH (GM069420) and USDA/Hatch to A.Z.A. G.B.B. was supported by a CIHR postdoctoral fellowship, H.v.B. by the Netherlands Organization for Scientific Research (825.06.033), C.D.C. by American Heart Association Predoctoral Fellowship No. 0615615Z, and C.L.W. by Computation and Informatics in Biology and Medicine Training Grant T15LM007359. A.Z.A. is a Shaw Scholar. J.D.L. and A.J.G. are supported by NIH R01-GM072518. We thank Brenda Andrews, Charlie Boone, Li Zhijang, Zhaolei Zhang, Quaid Morris, Larry Hiesler, Martha Bulyk, Mike Berger, Cong Zhu, and Andrew Gehrke for assistance and helpful discussions.

Received: August 13, 2008
Revised: November 5, 2008
Accepted: November 26, 2008
Published: December 24, 2008

REFERENCES

Angus-Hill, M.L., Schlichter, A., Roberts, D., Erdjument-Bromage, H., Tempst, P., and Cairns, B.R. (2001). A Rsc3/Rsc30 zinc cluster dimer reveals novel roles for the chromatin remodeler RSC in gene expression and cell cycle control. Mol. Cell *7*, 741–751.

Barrera, L.O., and Ren, B. (2006). The transcriptional regulatory code of eukaryotic cells–insights from genome-wide analysis of chromatin organization and transcription factor binding. Curr. Opin. Cell Biol. *18*, 291–298.

Beer, M.A., and Tavazoie, S. (2004). Predicting gene expression from sequence. Cell *117*, 185–198.

Benos, P.V., Bulyk, M.L., and Stormo, G.D. (2002). Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res. *30*, 4442–4451.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266–1276.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., III, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. *24*, 1429–1435.

Blackwell, T.K., Huang, J., Ma, A., Kretzner, L., Alt, F.W., Eisenman, R.N., and Weintraub, H. (1993). Binding of myc proteins to canonical and noncanonical DNA sequences. Mol. Cell. Biol. *13*, 5216–5224.

Buck, M.J., and Lieb, J.D. (2006). A chromatin-mediated mechanism for specification of conditional transcription factor targets. Nat. Genet. *38*, 1446–1451.

Cairns, B.R., Lorch, Y., Li, Y., Zhang, M., Lacomis, L., Erdjument-Bromage, H., Tempst, P., Du, J., Laurent, B., and Kornberg, R.D. (1996). RSC, an essential, abundant chromatin-remodeling complex. Cell *87*, 1249–1260.

Carroll, J.S., Liu, X.S., Brodsky, A.S., Li, W., Meyer, C.A., Szary, A.J., Eeckhoute, J., Shao, W., Hestermann, E.V., Geistlinger, T.R., et al. (2005). Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell *122*, 33–43.

Chasman, D.I., Lue, N.F., Buchman, A.R., LaPointe, J.W., Lorch, Y., and Kornberg, R.D. (1990). A yeast protein that influences the chromatin structure of UASG and functions as a powerful auxiliary gene activator. Genes Dev. *4*, 503–514.

Damelin, M., Simon, I., Moy, T.I., Wilson, B., Komili, S., Tempst, P., Roth, F.P., Young, R.A., Cairns, B.R., and Silver, P.A. (2002). The genome-wide localization of Rsc9, a component of the RSC chromatin-remodeling complex, changes in response to stress. Mol. Cell *9*, 563–573.

Dequard-Chablat, M., Riva, M., Carles, C., and Sentenac, A. (1991). RPC19, the gene for a subunit common to yeast RNA polymerases A (I) and C (III). J. Biol. Chem. *266*, 15300–15307.

Elemento, O., and Tavazoie, S. (2005). Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol. *6*, R18.

Fourel, G., Miyake, T., Defossez, P.A., Li, R., and Gilson, E. (2002). General regulatory factors (GRFs) as genome partitioners. J. Biol. Chem. *277*, 41736–41743.

Granek, J.A., and Clarke, N.D. (2005). Explicit equilibrium modeling of transcription-factor binding and gene regulation. Genome Biol. *6*, R87.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99–104.

Juneau, K., Palm, C., Miranda, M., and Davis, R.W. (2007). High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. Proc. Natl. Acad. Sci. USA *104*, 1522–1527.

Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature *428*, 617–624.

Kim, J.H., Polish, J., and Johnston, M. (2003). Specificity and regulation of DNA binding by the yeast glucose transporter gene repressor Rgt1. Mol. Cell. Biol. *23*, 5208–5216.

Larochelle, M., Drouin, S., Robert, F., and Turcotte, B. (2006). Oxidative stress-activated zinc cluster protein Stb5 has dual activator/repressor functions required for pentose phosphate pathway regulation and NADPH production. Mol. Cell. Biol. *26*, 6690–6701.

Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R., and Nislow, C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. Nat. Genet. *39*, 1235–1244.

Li, M.Z., and Elledge, S.J. (2005). MAGIC, an in vivo genetic method for the rapid construction of recombinant DNA molecules. Nat. Genet. *37*, 311–319.

Liu, X., Lee, C.K., Granek, J.A., Clarke, N.D., and Lieb, J.D. (2006). Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. Genome Res. *16*, 1517–1528.

Liu, X., Noll, D.M., Lieb, J.D., and Clarke, N.D. (2005). DIP-chip: rapid and accurate determination of DNA-binding specificity. Genome Res. *15*, 421–427.

Lutfiyya, L.L., Iyer, V.R., DeRisi, J., DeVit, M.J., Brown, P.O., and Johnston, M. (1998). Characterization of three related glucose repressors and genes they regulate in Saccharomyces cerevisiae. Genetics *150*, 1377–1391.

MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., Stormo, G.D., and Fraenkel, E. (2006). An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics *7*, 113.

MacPherson, S., Larochelle, M., and Turcotte, B. (2006). A fungal family of transcriptional regulators: the zinc cluster proteins. Microbiol. Mol. Biol. Rev. *70*, 583–604.

Mintseris, J., and Eisen, M.B. (2006). Design of a combinatorial DNA microarray for protein-DNA interaction studies. BMC Bioinformatics *7*, 429.

Nash, R., Weng, S., Hitz, B., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., et al. (2007). Expanded protein information at SGD: new pages and proteome browser. Nucleic Acids Res. *35*, D468–D471.

Ng, H.H., Robert, F., Young, R.A., and Struhl, K. (2002). Genome-wide location and regulated recruitment of the RSC nucleosome-remodeling complex. Genes Dev. *16*, 806–819.

Parnell, T.J., Huff, J.T., and Cairns, B.R. (2008). RSC regulates nucleosome positioning at Pol II genes and density at Pol III genes. EMBO J. *27*, 100–110.

Perocchi, F., Xu, Z., Clauder-Munster, S., and Steninmetz, L.M. (2007). Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Nucleic Acids Res. *35*, e128.

Pierce, M., Benjamin, K.R., Montano, S.P., Georgiadis, M.M., Winter, E., and Vershon, A.K. (2003). Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression. Mol. Cell. Biol. *23*, 4814–4825.

Planta, R.J., Goncalves, P.M., and Mager, W.H. (1995). Global regulators of ribosome biosynthesis in yeast. Biochem. Cell Biol. *73*, 825–834.

Roberts, D.N., Stewart, A.J., Huff, J.T., and Cairns, B.R. (2003). The RNA polymerase III transcriptome revealed by genome-wide localization and activity-occupancy relationships. Proc. Natl. Acad. Sci. USA *100*, 14695–14700.

Robinson, M.D., Grigull, J., Mohammad, N., and Hughes, T.R. (2002). FunSpec: a web-based cluster interpreter for yeast. BMC Bioinformatics *3*, 35.

Schlake, T., Schorpp, M., Nehls, M., and Boehm, T. (1997). The nude gene encodes a sequence-specific DNA binding protein with homologs in organisms that lack an anticipatory immune system. Proc. Natl. Acad. Sci. USA *94*, 3842–3847.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. Nature *451*, 535–540.

Soragni, E., and Kassavetis, G.A. (2008). Absolute gene occupancies by RNA polymerase III, TFIIIB and TFIIIC in Saccharomyces cerevisiae. J. Biol. Chem. *283*, 26568–26576.

Tanay, A. (2006). Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res. *16*, 962–972.

Vik, A., and Rine, J. (2001). Upc2p and Ecm22p, dual regulators of sterol biosynthesis in Saccharomyces cerevisiae. Mol. Cell. Biol. *21*, 6395–6405.

Warren, C.L., Kratochvil, N.C., Hauschild, K.E., Foister, S., Brezinski, M.L., Dervan, P.B., Phillips, G.N., Jr., and Ansari, A.Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. Proc. Natl. Acad. Sci. USA *103*, 867–872.

Wilson, B., Erdjument-Bromage, H., Tempst, P., and Cairns, B.R. (2006). The RSC chromatin remodeling complex bears an essential fungal-specific protein module with broad functional roles. Genetics *172*, 795–809.

Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000). DNA recognition by Cys2His2 zinc finger proteins. Annu. Rev. Biophys. Biomol. Struct. *29*, 183–212.

Yuan, G.C., and Liu, J.S. (2008). Genomic sequence is highly predictive of local nucleosome depletion. PLoS Comput. Biol. *4*, e13. 10.1371/journal.pcbi.0040013.

Diversity and Complexity in DNA Recognition by Transcription Factors

Gwenael Badis,[1]* Michael F. Berger,[2,3]* Anthony A. Philippakis,[2,3,4]* Shaheynoor Talukder,[1,5]* Andrew R. Gehrke,[2]* Savina A. Jaeger,[2]* Esther T. Chan,[5]* Genita Metzler,[6] Anastasia Vedenko,[7] Xiaoyu Chen,[1] Hanna Kuznetsov,[6] Chi-Fong Wang,[8] David Coburn,[1] Daniel E. Newburger,[2] Quaid Morris,[1,5,9,10] Timothy R. Hughes,[1,5,10]† Martha L. Bulyk[2,3,4,11]†

Sequence preferences of DNA binding proteins are a primary mechanism by which cells interpret the genome. Despite the central importance of these proteins in physiology, development, and evolution, comprehensive DNA binding specificities have been determined experimentally for only a few proteins. Here, we used microarrays containing all 10–base pair sequences to examine the binding specificities of 104 distinct mouse DNA binding proteins representing 22 structural classes. Our results reveal a complex landscape of binding, with virtually every protein analyzed possessing unique preferences. Roughly half of the proteins each recognized multiple distinctly different sequence motifs, challenging our molecular understanding of how proteins interact with their DNA binding sites. This complexity in DNA recognition may be important in gene regulation and in the evolution of transcriptional regulatory networks.

The interactions between transcription factors (TFs) and their DNA binding sites are an integral part of the gene regulatory networks that control development, core cellular processes, and responses to environmental perturbations. However, only a handful of sequence-specific TFs have been characterized well enough to identify all the sequences that they can and, just as importantly, cannot bind. Computational analysis of microarray readout of chromatin immunoprecipitation experiments (ChIP-chip) suggests extensive use of low-affinity binding sites in yeast (*1*), and computational models of gene expression during fly embryonic development suggest that low-affinity binding sites contribute as much as high-affinity sites (*2*).

The availability of TF binding data spanning the full affinity range would improve our understanding of the biophysical phenomena underlying protein-DNA recognition and would also improve accuracy in analyzing cis regulatory elements. Here we report the comprehensive determination of the DNA binding specificities of 104 known and predicted mouse TFs with the use of the universal protein binding microarray (PBM) technology (*3*). These TFs represent 22 different DNA binding domain (DBD) structural classes that are the major DBD classes found in metazoan TFs.

We created N-terminal glutathione *S*-transferase (GST) fusion constructs of the DBDs of 104 known and predicted mouse TFs (fig. S1 and table S1) (*4*). Five of these proteins—Max, Bhlhb2, Gata3, Rfx3, and Sox7—were also represented as full-length fusions to N-terminal GST, yielding a total set of 109 nonredundant proteins represented by 115 samples (*5*). Each protein was used in two PBM experiments (*6*, *7*) (figs. S2 to S4 and table S2). DNA binding site motifs were initially derived by the Seed-and-Wobble algorithm (*3*, *8*); Seed-and-Wobble first identifies the single 8-mer (ungapped or gapped) with the greatest PBM enrichment score (E score) (*3*) and then systematically tests the relative preference of each nucleotide variant at each position, both within and outside the seed (*5*). Later analyses incorporated additional motif-finding algorithms, including RankMotif++ (*9*) and Kafal (*5*).

Beyond simply providing a DNA binding site motif, these data provide a rank-ordered listing of the preference of a protein for every gapped and ungapped *k*-mer "word," where *k* is the number of informative nucleotide positions in the binding site. This data set consists of 9.6 million measurements, from which we can derive binding data for 22.3 million ungapped and gapped 8-mers (up to 12 positions) for each protein. For each of the 8-mers for each protein, we report its E score, median signal intensity Z score, and false discovery rate *Q* value (*5*). We found that the average number of ungapped 8-mers considered "bound" at a *Q* value threshold of 0.001 varied across classes, ranging from 65 for the MADS class factor SRF to 871 for the E2F class.

For TFs that had previously known binding site motifs, we observed general agreement with earlier motif data (fig. S5 and table S3) (*5*). Comparisons to dissociation constant data (*10*) for Max and for the yeast TF Cbf1 (*3*) indicate that words with higher E scores are generally bound with higher affinity (*3*) (fig. S6). Confirmation by electrophoretic mobility shift assays (EMSAs) for three newly characterized proteins and one recently characterized protein (*11*) —Zfp740, Osr2, Sp100, and Zfp161 (ZF5) (*12*), respectively—is shown in fig. S7.

To examine correlations among the proteins' DNA binding specificities and to identify DNA sequences that distinguish the binding profiles of different TF families, we hierarchically clustered the *k*-mers that met a stringent binding threshold (E score $\geq$ 0.45) for at least one of the proteins. We used E scores because they are robust to differences in protein concentration and thus facilitate comparison of *k*-mer data across arrays (*3*); we consider them as a proxy for relative affinities. Different DBD classes generally recognize distinct portions of sequence space (Fig. 1A and fig. S8). However, even proteins with up to 67% amino acid sequence identity exhibited distinct DNA binding profiles. For example, although Irf4 and Irf5 both bind the same highest-affinity sites (8-mers containing CGAAAC), they prefer different lower-affinity sites (TGAAAG versus CGAGAC) (Fig. 1B). We verified for five TFs that the full-length protein displays a virtually identical spectrum of 8-mer preferences to that of the DBD and that the spectrum is distinct from other proteins of the same structural class (figs. S2 and S9).

Our data set includes most members of three TF structural classes in mouse: (i) Sox and Sox-related, (ii) IRF, and (iii) AP-2. In an extreme case, we find no evidence that the binding profiles of the AP-2 class members are different from each other (fig. S9B), consistent with reports that the human counterparts of AP-2α, AP-2β, and AP-2γ all bind GCCNNNGGC (*13*). In contrast, members of the IRF class all appeared to have different binding profiles (fig. S9L).

The Sox and Sox-related family presents an intriguing instance of highly conserved DBDs with closely related but distinct binding preferences. We found marked differences in the binding specificities of the Sox (*14*), Tcf/Lef (*15*, *16*), and Hbp1/Bbx (*17*) families (Fig. 1C). In most cases, our data are roughly consistent with known binding sequences (Fig. 1C), although there are also clear differences: Hpb1 and Bbx have been described as preferring WRAATGGG (*17*), whereas in our data, Hbp1 and Bbx prefer TGAATG and have lesser preference for AATGGG. Our data confirm that there are at least four different varieties of Sox and Sox-related DNA binding specificity (Fig. 1C) and suggest that there are subtle variations among Sox proteins (Fig. 1B).

Several TFs had two distinct sets of high-scoring *k*-mers. For example, the nuclear receptor

[1]Banting and Best Department of Medical Research, University of Toronto, Toronto, ON M5S 3E1, Canada. [2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA. [3]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138, USA. [4]Harvard–Massachusetts Institute of Technology (MIT) Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA. [5]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada. [6]Department of Biology, MIT, Cambridge, MA 02139, USA. [7]Department of Biology, Wellesley College, Wellesley, MA 02481, USA. [8]Department of Physics, MIT, Cambridge, MA 02139, USA. [9]Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada. [10]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON M5S 3E1, Canada. [11]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: t.hughes@utoronto.ca (T.R.H.); mlbulyk@receptor.med.harvard.edu (M.L.B.)

hepatic nuclear factor 4 alpha [Hnf4a; C4 zinc finger (ZnF) DBD] exhibits strong binding to both sequences containing GGTCCA and sequences containing GGTCCA (Fig. 2A), whereas all four other C4 ZnF TFs that we examined bind only to GGTCA. We confirmed binding of Hnf4a to both variants by EMSA (fig. S10). TFs that can recognize two distinctly different DNA sequences have been noted before (*18*). We

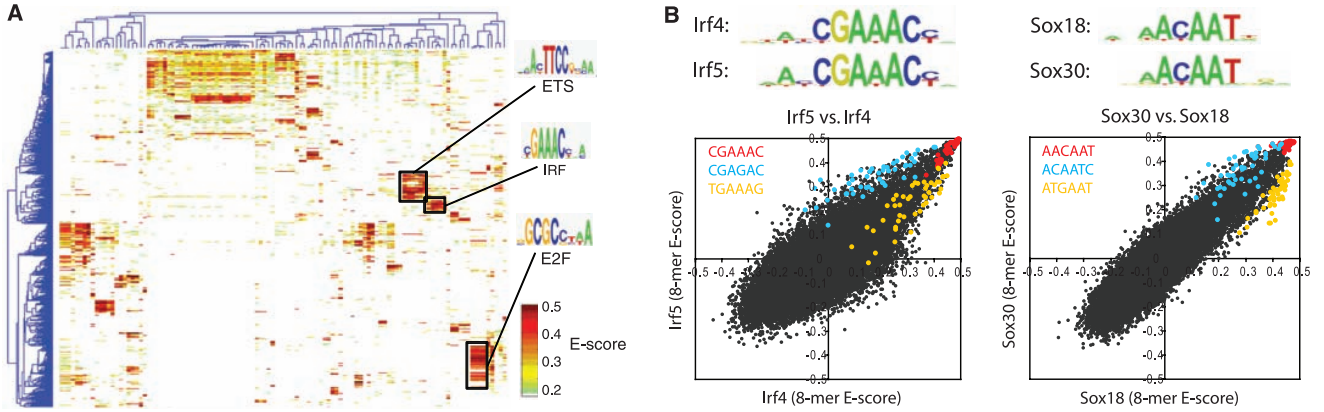


Fig. 1. High-resolution PBM *k*-mer data. (**A**) Heat map of two-dimensional hierarchical agglomerative clustering analysis of 4740 ungapped 8-mers (rows) over 104 nonredundant TFs (columns), with both 8-mers and proteins clustered using averaged E score from the two different array designs. The 4740 8-mers were selected because they have an E score of 0.45 or greater for at least one of the proteins. A motif representative of the 8-mers contained in each of the indicated clusters is shown, derived from running the 8-mers on ClustalW (*32*) and entering groups of related aligned sequences into WebLogo (*33*). (**B**) Scatter plots comparing 8-mer scores for each pair of TFs, whose primary Seed-and-Wobble logos are shown above the plots. 8-mers containing each 6-mer sequence (inset) are highlighted, revealing consistent differences between sequence preferences among lower-affinity 8-mers, despite identical preferences for the same highest-affinity 8-mers. (Left) Irf5 versus Irf4; (right) Sox30 versus Sox18. (**C**) Clustergram of *k*-mers for the Sox family of TFs. 310 8-mers with E score ≥ 0.45 for at least one of the 21 Sox and Sox-related TFs were hierarchically clustered according to their relative ranks for each TF, and then the rows, corresponding to *k*-mers, were rearranged to group together 8-mers with shared sequence patterns.

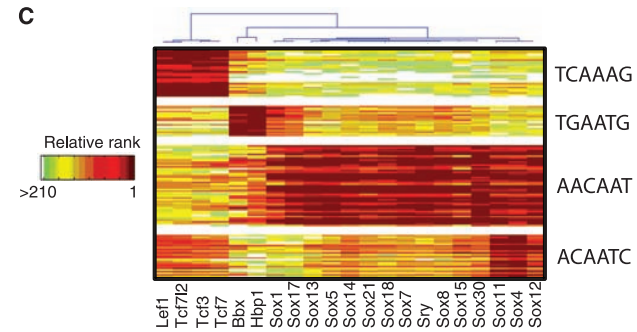
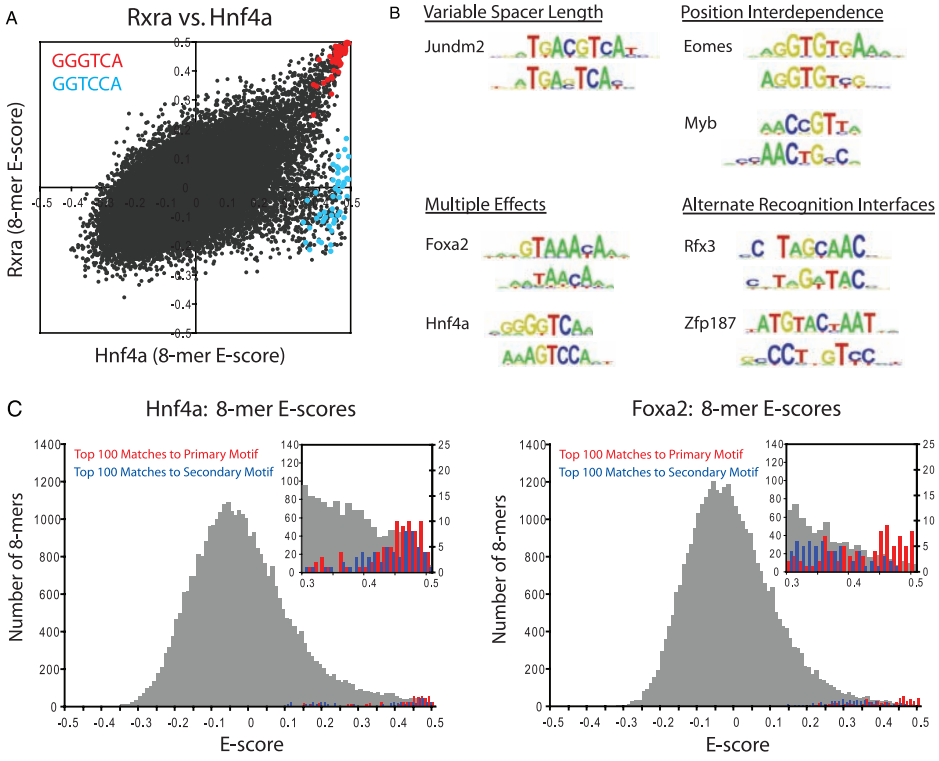


Fig. 2. TF binding site secondary motifs. (**A**) Scatter plot comparing 8-mer E scores for closely related TFs. Hnf4a and Rxra (two C4 zinc finger DBDs) both exhibit strong binding to 8-mers containing GGGTCA (red), whereas Hnf4a shows specific binding to an additional set of 8-mers containing GGTCCA (blue). (**B**) Examples of motifs from different categories of secondary motifs. (**C**) Histograms of E scores for all 8-mers (gray), the top 100 8-mer matches to the primary motif (red), and the top 100 8-mer matches to the secondary motif (blue). 8-mers were scored for matches to PWMs according to the GOMER (*27*) scoring framework. Insets provide a magnified display of the tails of the distributions; *y*-axis labels along the right of each inset refer to the red and blue bars. On the basis of the 8-mer scores, the primary and secondary Hnf4a motifs are essentially interchangeable (left), whereas Foxa2 shows a clear preference for 8-mers corresponding to its primary motif (right).



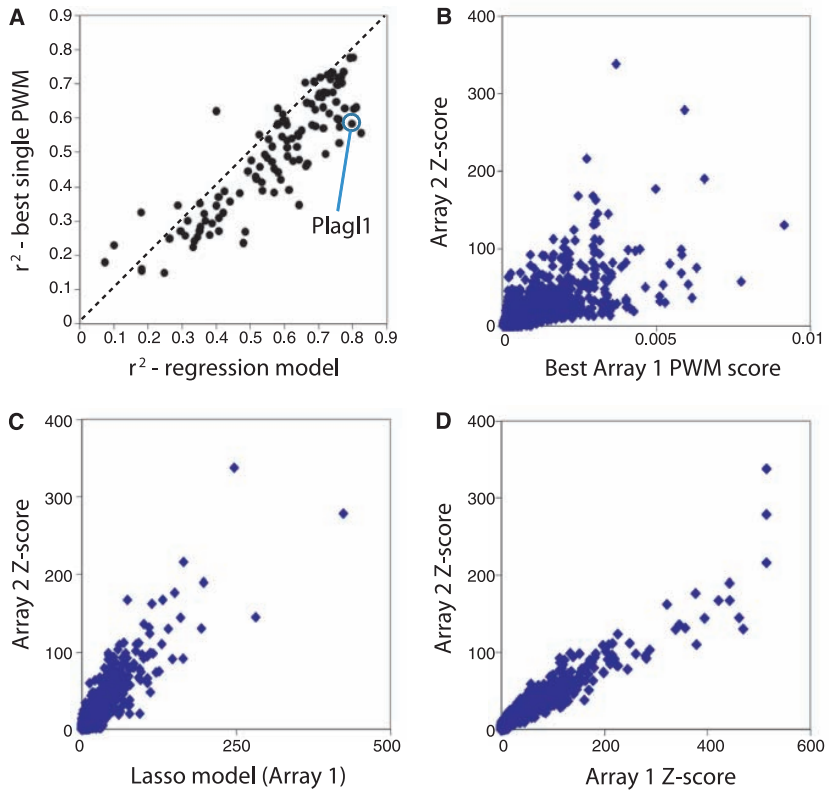


Fig. 3. Multiple-motif models typically better represent the binding profiles than do single-motif models. (**A**) Considering all TFs in this study, in general, multiple-motif models are a better representation of the data than are single-motif models. Variance in 8-mer median intensity (Z score) on Array 2 explained by our PWM regression model (*x* axis) compared to GOMER (*27*) scores for the single best PWM model obtained (best is defined as highest variance explained) over all 8-mers, with models derived from Array 1; the GOMER scoring framework calculates binding probabilities over the 8-mers according to PWMs (*27*). Each point represents one of the TFs analyzed. (**B**) The GOMER score for the best PWM derived from Array 1 is compared to the Z scores from Array 2, for Plagl1 as a case example. Each point is a single 8-mer; all 32,896 8-mers are shown. (**C**) Same as (B), except that the Array 1 regression model scores [which are a linear combination, built by using the least absolute shrinkage and selection operator (Lasso) algorithm (*34*), of GOMER scores from individual motifs] are compared to the Z scores from Array 2. (**D**) 8-mer Z scores for Plagl1 derived from Array 1 compared to the Z scores from Array 2. Each point is a single 8-mer; all 32,896 8-mers are shown.

hypothesized that the existence of secondary motifs may be a general phenomenon, and therefore, we searched for alternate binding preferences throughout our entire data set.

To aid in the identification of secondary binding preferences, we further developed our Seed-and-Wobble algorithm to search specifically for motifs that represent the *k*-mers of high signal intensity that are not explained well by the primary motif; we refer to these as the secondary motifs. A further iteration can be employed to search for a tertiary motif. As an initial test case, we examined PBM data for the human TF Oct-1 (*3*); the PBM-derived Oct-1 primary motif corresponded to the full Oct-1 DNA binding site motif, whereas the secondary and tertiary motifs corresponded to the binding site motifs of the POU_{HD} and POU_S domains (*19*), respectively (fig. S11). Analysis of 100 simulated long, 14–base pair (bp) motifs (*5*) indicated that Seed-and-Wobble was highly successful in identifying the simulated motifs and that essentially all of the secondary motifs we found in analyzing the real PBM data were unlikely to be attributable to a motif-finding artifact due to long motifs (*5*).

We observed clear secondary DNA binding preferences for nearly half of our 104 mouse TFs. Their secondary motifs fell into four different categories (Fig. 2B and supporting online material text), which we annotated manually. We confirmed binding to the secondary motifs by 6 TFs—Hnf4a, Nkx3.1, Myb, Mybl1, Foxj3, and Rfxdc2—by EMSAs (fig. S10).

We found 19 clear cases of "position interdependence" TFs, which exhibited strong interdependence (*20*) among the nucleotide positions of their binding sites. Position interdependencies frequently spanned more than just dinucleotides; for example, estrogen related receptor alpha has a strong preference for binding either CAAGGTCA or AGGGGTCA, but not CAGGGTCA or CGGGGTCA. Interdependent nucleotide positions were not always adjacent to each other; for example, Myb (fig. S10) exhibited strong interdependence at positions separated by 1 nucleotide,

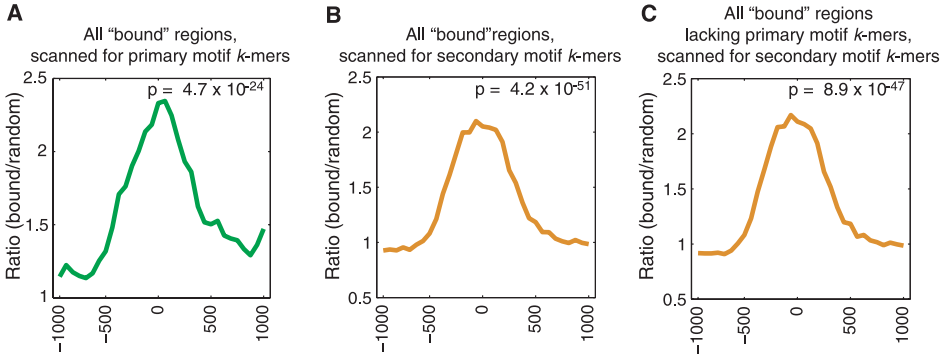


Fig. 4. Enrichment of primary versus secondary motif sequences bound in vitro within genomic regions bound in vivo. Relative enrichment of *k*-mers corresponding to the primary versus secondary Seed-and-Wobble motifs within (**A**) and (**B**) all bound genomic regions in ChIP-chip data or (**C**) those bound regions lacking primary motif *k*-mers, as compared to randomly selected sequences, was calculated (*5*) for Hnf4a (Gene Expression Omnibus accession number GSE7745). ChIP-chip "bound" peaks were identified according to the criteria of that study (*28*). A window size of 500 bp with a step size of 100 bp was used. The GOMER thresholds used are 2.958×10^{-7} and 8.419×10^{-7}, corresponding to 9 primary and 20 secondary 8-mers scanned, respectively, for Hnf4a. *P* values for enrichment of 8-mers within the bound genomic regions shown in each panel were calculated for the interval from –250 to +250 by the Wilcoxon-Mann-Whitney rank sum test, comparing the number of occurrences per sequence in the bound set versus the background set.

with preference for binding either AACCGTCA or AACTGCCA. Although position interdependence has been observed (*21*–*25*), that this phenomenon occurs on such a broad scale was not known and has important implications because commonly used TF binding site models assume mononucleotide independence.

One protein, the mouse transcriptional regulator Jundm2, which is a member of the basic leucine zipper structural class, bound to a "variable spacer length" motif (fig. S12). "Multiple effects" motifs appeared to display a combination of position interdependence and variable distances separating different parts of their motifs; at least 16 TFs fell into this category.

Finally, at least five secondary motifs in the "alternate recognition interfaces" category were not readily explainable by either a variable spacer length or position interdependence. This category is the most intriguing, as it suggests that some TFs recognize their DNA binding sites through multiple, completely different interaction modes, either through alternate structural features or by switching between alternate conformations. Support for this hypothesis comes from the co-crystal structure of human Rfx1 bound to DNA, which indicated that Rfx1 uses β strands and a connecting loop to interact with the major groove of one half-site and an α helix to interact with the minor groove of the other half-site (*26*). It is likely that Rfx3, Rfx4, and Rfxdc2 use this same mechanism of alternative DNA recognition modes (fig. S13).

For several TFs, the secondary motifs were bound nearly as well as the primary motifs, whereas in most cases, the motifs represented different affinity classes. For example, the top 20 8-mers that matched Hnf4a's primary motif were fairly evenly intermingled [$P = 0.037$ by Wilcoxon-Mann-Whitney U test, using GOMER (generalizable occupancy model of expression regulation) (*27*) scoring of motifs] with those that matched its secondary motif (Fig. 2C, left). In contrast, for Foxa2, the secondary motif represented lower-affinity binding sequences ($P = 1.94 \times 10^{-6}$) (Fig. 2C, right).

We further considered the possibility that some proteins' DNA binding specificities might be represented best by multiple motifs. We applied a linear regression approach (*5*) to learn weighted combinations of position weight matrices (PWMs) generated from several different motif-finding algorithms. We found that the binding profiles for all but 15 proteins were represented best by more than one motif (Fig. 3 and fig. S14). Some of these multiple motifs did not appear to represent different protein-DNA interaction properties described above, but nevertheless, they captured different subsets of the *k*-mer data.

We explored the in vivo usage of the secondary motifs by considering their TF occupancy. We calculated the relative enrichment of 8-mers corresponding to the primary versus secondary Seed-and-Wobble motifs within genomic regions bound in ChIP-chip data, as compared with randomly selected sequences (*5*) for Hnf4a (Fig. 4 and fig. S15, A, C, and D). As expected, Hnf4a-bound regions are enriched for matches to 8-mers corresponding to the primary motif for Hnf4a PBM data, with the greatest enrichment toward the centers of the bound regions (Fig. 4A). Hnf4a-bound regions are also enriched for matches to 8-mers corresponding to the secondary motif (Fig. 4B). Hnf4a secondary motif 8-mers are enriched even among those Hnf4a-bound regions that lack primary motif 8-mers (Fig. 4C), suggesting that the secondary motif can recruit Hnf4a to genomic loci independently of the primary motif. We observed similar results for Bcl6 (*28*) (fig. S15).

Our characterization of 104 TFs from 22 different structural classes revealed a prevalence of complexity and richness in DNA binding preferences, both across and within classes. The breadth of the observed "secondary motif" phenomenon had not been described before, and it has important implications for understanding how proteins interact with their DNA binding sites and for genome analysis.

Further experiments and analyses are needed to determine whether the same TF exerts different gene regulatory effects through distinct sequence motifs, as well as to determine whether TF-specific differences among members of a TF family (*29*) contribute to differences in binding in vivo and to distinct physiological functions. Although TFs bind a rich spectrum of *k*-mers not fully captured even by multiple PWMs, using a multiple-motif model is of practical consequence because most genome analysis tools employ PWMs. Algorithms that consider the quantitative nature of *k*-mer binding data in scoring candidate regulatory elements need to be developed.

Finally, these PBM data are likely to be highly informative for well-conserved homologs in other organisms. Generating [or inferring (*29*)] PBM data for all regulatory factors in all major model organisms is an important goal, as such *k*-mer data probably will be useful for improved prediction and analysis of regulatory elements, including the identification of direct versus indirect TF binding sites from ChIP data (*30*). Moreover, such data would aid in understanding the evolution of cis regulatory elements and transcriptional regulatory networks.

References and Notes

1. A. Tanay, *Genome Res.* **16**, 962 (2006).
2. E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, U. Gaul, *Nature* **451**, 535 (2008).
3. M. F. Berger *et al.*, *Nat. Biotechnol.* **24**, 1429 (2006).
4. M. Z. Li, S. J. Elledge, *Nat. Genet.* **37**, 311 (2005).
5. Materials and methods are available as supporting material on *Science* Online.
6. M. L. Bulyk, X. Huang, Y. Choo, G. M. Church, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 7158 (2001).
7. S. Mukherjee *et al.*, *Nat. Genet.* **36**, 1331 (2004).
8. M. F. Berger, M. L. Bulyk, *Nat. Protocols* **4**, 393 (2009).
9. X. Chen, T. R. Hughes, Q. Morris, *Bioinformatics* **23**, i72 (2007).
10. S. J. Maerkl, S. R. Quake, *Science* **315**, 233 (2007).
11. V. Matys *et al.*, *Nucleic Acids Res.* **34**, D108 (2006).
12. S. V. Orlov *et al.*, *FEBS J.* **274**, 4848 (2007).
13. J. M. Bosher, N. F. Totty, J. J. Hsuan, T. Williams, H. C. Hurst, *Oncogene* **13**, 1701 (1996).
14. S. Mertin, S. G. McDowall, V. R. Harley, *Nucleic Acids Res.* **27**, 1359 (1999).
15. M. van de Wetering, M. Oosterwegel, D. Dooijes, H. Clevers, *EMBO J.* **10**, 123 (1991).
16. A. Travis, A. Amsterdam, C. Belanger, R. Grosschedl, *Genes Dev.* **5**, 880 (1991).
17. S. G. Tevosian *et al.*, *Genes Dev.* **11**, 383 (1997).
18. K. Pfeifer, T. Prezant, L. Guarente, *Cell* **49**, 19 (1987).
19. J. D. Klemm, M. A. Rould, R. Aurora, W. Herr, C. O. Pabo, *Cell* **77**, 21 (1994).
20. P. V. Benos, M. L. Bulyk, G. D. Stormo, *Nucleic Acids Res.* **30**, 4442 (2002).
21. P. V. Benos, A. S. Lapedes, G. D. Stormo, *Bioessays* **24**, 466 (2002).
22. M. L. Bulyk, P. L. Johnson, G. M. Church, *Nucleic Acids Res.* **30**, 1255 (2002).
23. M.-L. Lee, M. Bulyk, G. Whitmore, G. Church, *Biometrics* **58**, 981 (2002).
24. T. K. Man, G. D. Stormo, *Nucleic Acids Res.* **29**, 2471 (2001).
25. Y. Barash, G. Elidan, N. Friedman, T. Kaplan, paper presented at the Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB), Berlin, 10 to 13 April 2003.
26. K. S. Gajiwala *et al.*, *Nature* **403**, 916 (2000).
27. J. A. Granek, N. D. Clarke, *Genome Biol.* **6**, R87 (2005).
28. S. M. Ranuncolo *et al.*, *Nat. Immunol.* **8**, 705 (2007).
29. M. F. Berger *et al.*, *Cell* **133**, 1266 (2008).
30. C. Zhu *et al.*, *Genome Res.* **19**, 556 (2009).
31. D. E. Newburger, M. L. Bulyk, *Nucleic Acids Res.* **37**, D77 (2009).
32. R. Chenna *et al.*, *Nucleic Acids Res.* **31**, 3497 (2003).
33. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, *Genome Res.* **14**, 1188 (2004).
34. R. Tibshirani, *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267 (1996).

35. This project was supported by funding from the Canadian Institutes of Health Research (MOP-77721 and postdoctoral fellowship to G.B.); Genome Canada through the Ontario Genomics Institute, the Ontario Research Fund, and the Canadian Institute for Advanced Research (to T.R.H.); NSF (to M.F.B.); the Canadian Foundation for Innovation and Ontario Research Fund (to Q.M.); and grant R01 HG003985 from NIH/National Human Genome Research Institute (to M.L.B.). We thank L. Peña-Castillo, A. Cheung, M. Chan, S. Bhinder, F. Bréard, P. Qureshi, S. Mnaimneh, M. Kekis, F. Khalid, J. Holroyd, D. Terterov, and K. Robasky for technical assistance and S. Gisselbrecht, K. Struhl, and S. Sunyaev for critical reading of the manuscript. PBM data are available at http://the_brain.bwh.harvard.edu/pbms/webworks/ and also via the publicly available UniPROBE database (*31*).

Supporting Online Material

www.sciencemag.org/cgi/content/full/1162327/DC1
Materials and Methods
SOM Text
Figs. S1 to S15
Tables S1 to S3
References

25 June 2008; accepted 1 May 2009
Published online 14 May 2009;
10.1126/science.1162327
Include this information when citing this paper.

Published online 13 August 2013

Nucleic Acids Research, 2013, Vol. 41, No. 20 ***9461–9470***
doi:10.1093/nar/gkt726

Yeast ribosomal protein L7 and its homologue Rlp7 are simultaneously present at distinct sites on pre-60S ribosomal particles

Reyes Babiano[1], Gwenael Badis[2], Cosmin Saveanu[2], Abdelkader Namane[2], Antonia Doyen[2], Antonio Díaz-Quintana[3], Alain Jacquier[2], Micheline Fromont-Racine[2,*] and Jesús de la Cruz[1,*]

[1]Departamento de Genética, Universidad de Sevilla, E-41012 Seville, Spain, [2]Institut Pasteur, Génétique des Interactions Macromoléculaires, CNRS UMR-3525, Paris, France and [3]Instituto de Bioquímica Vegetal y Fotosíntesis, Universidad de Sevilla-CSIC, Seville, Spain

Received June 18, 2013; Revised July 22, 2013; Accepted July 24, 2013

ABSTRACT

Ribosome biogenesis requires >300 assembly factors in *Saccharomyces cerevisiae*. Ribosome assembly factors Imp3, Mrt4, Rlp7 and Rlp24 have sequence similarity to ribosomal proteins S9, P0, L7 and L24, suggesting that these pre-ribosomal factors could be placeholders that prevent premature assembly of the corresponding ribosomal proteins to nascent ribosomes. However, we found L7 to be a highly specific component of Rlp7-associated complexes, revealing that the two proteins can bind simultaneously to pre-ribosomal particles. Cross-linking and cDNA analysis experiments showed that Rlp7 binds to the ITS2 region of 27S pre-rRNAs, at two sites, in helix III and in a region adjacent to the pre-rRNA processing sites C_1 and E. However, L7 binds to mature 25S and 5S rRNAs and cross-linked predominantly to helix $ES7^L$b within 25S rRNA. Thus, despite their predicted structural similarity, our data show that Rlp7 and L7 clearly bind at different positions on the same pre-60S particles. Our results also suggest that Rlp7 facilitates the formation of the hairpin structure of ITS2 during 60S ribosomal subunit maturation.

INTRODUCTION

In eukaryotes, ribosome biogenesis is a complex multi-step and multi-component process, which occurs primarily in the nucleolus, although late steps occur in the nucleoplasm and in the cytoplasm [for reviews, see (1,2)] Most of our knowledge concerning ribosome biogenesis in eukaryotes derives from studies with *Saccharomyces cerevisiae*. In the yeast nucleolus, the mature 18S, 5.8S and 25S rRNAs are co-transcribed by RNA polymerase I as a single large precursor rRNA (pre-rRNA) that undergoes co- or post-transcriptional processing, whereas the pre-5S is independently transcribed by RNA polymerase III (Supplementary Figure S1). Pre-rRNA processing occurs concomitantly to most rRNA modification reactions, folding of pre-rRNAs and assembly of most ribosomal proteins (r-proteins) to form pre-ribosomal particles (Supplementary Figure S2). Pre-ribosomal particles contain, in addition to pre-rRNAs and r-proteins, non-ribosomal *trans*-acting factors (1).

In yeast, roughly 300 protein *trans*-acting factors, involved in ribosome biogenesis, have been identified (3,4). These factors likely confer speed, accuracy and directionality to the ribosome synthesis process. The precise mechanisms by which protein *trans*-acting factors operate are still largely unknown. The use of affinity purification combined with quantitative mass spectrometry techniques like isobaric Tag for Relative and Absolute Quantification or SILAC (Stable Isotope Labelling with Amino-acids in Cell culture) allow to measure the timing of binding to and dissociation from pre-ribosomal particles for many protein *trans*-acting factors [e.g., (5,6)]. To better understand the role of these factors in ribosome biogenesis, other experimental approaches have been developed, among them, *in vivo* cross-linking and cDNA analysis (CRAC). This technique allows the identification of the interaction sites between several protein *trans*-acting factors and pre-rRNAs or snoRNAs [e.g. (7–11)].

*To whom correspondence should be addressed. Tel: +34 954 5577106; Fax: +34 954 557104; Email: jdlcd@us.es
Correspondence may also be addressed to Micheline Fromont-Racine. Tel: +33 1 4061 3205; Fax: +33 1 4568 3456; Email: mfromont@pasteur.fr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Ribosomal proteins are active players in the maturation and the nucleo-cytoplasmic transport of pre-ribosomal particles. As for *trans*-acting factors, functional analyses have revealed how loss-of-function mutations in r-protein genes negatively impact on pre-rRNA processing and pre-ribosomal particles transport [e.g. (12–15)]. Little is known about the specific role of r-proteins in driving formation or re-arrangement of structures within pre-ribosomal particles [e.g. (6)]. Moreover, the course of assembly of the r-proteins remains unclear, especially for r-proteins of the large r-subunit [(16–18) and references therein].

A set of *trans*-acting factors, similar to selected r-proteins throughout their entire primary sequence, provides interesting insight into the assembly process and especially into the evolution of ribosome assembly factors and r-proteins. Amongst them are Rlp7, which is paralogous to L7 [L30 in the Yusupov's nomenclature (19)], as is Rlp24 to L24 (L24e), Mrt4 to P0 and Imp3 to S9 (S4) (20–24). Considering the high degree of homology found, it has been proposed that these factors and their r-protein counterpart successively bind the same rRNA structure, but although the factor binds to the rRNA site within a pre-ribosomal particle, the r-protein binds the same site within the mature r-subunit [discussed in (20,25)]. However, although we have experimentally demonstrated this 'placeholder hypothesis' for the relationship between Mrt4 and P0 (23), no validation has been shown for other paralogous pairs. In this work, we have studied the relationship between Rlp7 and L7 at different levels. We report distinct rRNA binding sites for Rlp7 and L7 and co-existence of the two proteins on the same pre-ribosomal particles. Our findings clearly show that the placeholder hypothesis is far from being a rule in ribosome biogenesis and provides insights into the molecular role of Rlp7 during 60S r-subunit assembly.

MATERIALS AND METHODS

Strains and microbiological methods

The yeast strains used in this study, which were derivatives of BY4741 or BMA64-1B, are listed in Supplementary Table S1. Most strains were generated by standard recombination techniques or by genetic crosses followed by sporulation, tetrad dissection and phenotypic analysis. All strains were checked by PCR and, when possible, by western blotting. Growth and handling of yeast and standard media were performed by established procedures (26). Yeast cells were grown at 30°C in rich or minimal medium containing either 2% galactose (YPGal, SGal), 2% glucose (YPD, SD) or in minimal medium containing 2% raffinose (SRaf).

Plasmid constructions

Plasmids are listed in Supplementary Table S2. To generate YCplac111-RLP7-HA, a 1.6 kb PCR product containing the *RLP7* ORF lacking the termination codon and an additional 1 kb upstream the ORF was cloned into pHAC111 (27). The structure of the resulting plasmid was verified by DNA sequencing. This construct complemented the growth of strains harbouring the *GAL::RLP7* allele to the wild-type extent in glucose-containing media.

Purification of complexes for SILAC quantification and SILAC data analysis

Cells from untagged and Rlp7-TAP tagged strains were grown in minimal medium in presence of either labelled L-lysine-$^{13}C_6$,$^{15}N_2$ (Sigma-Aldrich) or regular L-lysine, respectively (50 mg/l). Cell pellets from 1 l of each culture at 1.5 OD_{600} per ml were suspended in lysis buffer, mixed and broken with a French Press. One-step purification for SILAC experiments was performed using magnetic beads coated with immunoglobulin G (IgG) (Life Technologies) as described in (28,29). Proteins were identified by LC-MS/MS on a LTQ-Orbitrap velos instrument (Thermo Fisher Scientific, Bremen) as described in (28). Briefly, protein samples were treated with Endoprotease Lys-C and Trypsin. Digested peptides were desalted and then analysed by LC-MS/MS on a LTQ-Orbitrap velos instrument (Thermo Fisher Scientific, Bremen) (28). Raw MS data from the LTQ-Orbitrap were analysed using the MaxQuant software version 1.3.0.5 (30,31), supported by Andromeda (32) applying a false-discovery rate for both peptide and protein identification at $P \leq 0.01$. MS/MS spectra were searched against a concatenated *S. cerevisiae* decoy database from UniprotKB. Two missed cleavage were allowed, and only peptides with a minimum of seven amino acids were considered for identification. After data processing, SILAC quantification (H/L ratios) values from the 'proteinGroups.txt' output file of MaxQuant were taken for further analysis. The protein list was filtered to remove contaminants and reversed sequences. Proteins with a minimum of two measurements (ratio count ≥ 2) were selected for further analysis, and the H/L ratios were $\log_2$ transformed. The data are shown as a Supplementary Excel Table (Supplementary Data Set S1).

Affinity purifications

One-step purification of HTP- and TAP-tagged or untagged cells was performed with IgG-Sepharose beads. About 100 ml of HTP/TAP-tagged or untagged negative control cells were grown to an OD_{600} of 0.8, washed with cold water, harvested and concentrated in 500 µl of ice-cold TNM150 lysis buffer [50 mM Tris (pH 7.8), 1.5 mM $MgCl_2$, 150 mM NaCl, 0.1% NP-40, 5 mM β-mercaptoethanol] containing a protease inhibitor cocktail (Complete, Roche). Cells were disrupted by vigorous shaking with glass beads in a Fastprep®-24 (MP Biomedicals) at 4°C, and total cell extracts were obtained by centrifugation in a microcentrifuge at the maximun speed (ca. 16 100× *g*) for 15 min at 4°C. Each supernatant obtained was mixed with 50 µl of IgG-Sepharose beads (GE-Healthcare), previously equilibrated with the TNM150 buffer, and incubated for 2 h at 4°C with end-over-end tube rotation. After incubation, the beads were extensively washed 10 times with 1 ml of the same buffer at 4°C and finally collected. Protein was extracted with Laemmli buffer from both whole-cell extracts and 1/10th of the beads. Proteins were analysed by

western blotting using Peroxidase anti-peroxidase soluble complex (Sigma). RNA was extracted from total cell extracts, and the rest of the beads as described in (33,34) and analysed by northern blotting. The oligonucleotides used for northern blot hybridizations are described in the Supplementary Table S3.

Western blotting analyses and antibodies

Proteins were separated by SDS–PAGE and transferred onto nitrocellulose membranes by standard procedures. The following primary antibodies were used: Peroxidase anti-peroxidase at a dilution of 1:10000, rabbit polyclonal anti-L1 (1:10000; gift from F. Lacroute) (35), rabbit polyclonal anti-L35 (1:5000; gift from M. Seedorf) (36), rabbit polyclonal anti-Has1 (1:5000) (37), mouse monoclonal anti-Nop1 (1:5000; MCA28F2, EnCor Biotechnology) and mouse monoclonal anti-HA (1:5000, Roche). Secondary goat anti-rabbit or anti-mouse horseradish peroxidase-conjugated antibodies (Bio-Rad) were used at a dilution of 1:5000. Proteins were detected using an enhanced chemiluminescence detection kit (Super-Signal West Pico, Pierce).

CRAC and sequence analysis

In vivo CRAC experiments were performed as described previously (9). Briefly, cells expressing HTP-tagged Rlp7 and a non-tagged negative control strain were ultraviolet irradiated; cell extracts were then performed and subjected to a first affinity purification on IgG-Sepharose beads. Purified complexes were partially RNase-digested and subjected to a second affinity purification step on a nickel column under denaturing conditions. RNA molecules cross-linked to Rlp7-HTP were ligated to linkers, amplify by RT-PCR and subjected to Solexa sequencing to identify the relative location along the rDNA of the recovered RNA pieces. Similar experiments were performed with HTP-tagged L7 strains. Data analyses were done with the Integrative Genomics Viewer software (38).

In silico analysis of ribosome structure

Atomic coordinates of the yeast 60S r-subunit were retrieved from the Protein Data Bank (PDB; www.rcsb.org) with the accesion number 3U5D and 3U5E (19). L7 structure was substracted from the 3U5E file (chain F). Rlp7 structure was built with the Modeller 9v2 program (39) using as templates the coordinates provided in pdb files 2ZKR, 3IZR, 3JYW, 3U5E, 3U5E, 3U5I, 3ZF7, 4A1C, 4A1E and 4B6A. Sequence identities percentages of the target versus the templates ranged from 42 (2ZKR) to 48% (3IZR). All models were visualized with the UCSF Chimera program (40). Secondary structure of pre- and mature rRNAs were taken from Granneman *et al.* (7) and The Comparative RNA Web Site (http://www.rna.icmb.utexas.edu/), respectively.

RESULTS

Overlapping pattern of Rlp7 and L7 association with pre-ribosomal particles

Yeast r-protein L7 shares a notable sequence similarity with the *trans*-acting factor Rlp7 throughout its entire primary sequence (41) (Supplementary Figure S3). Given this close homology, we could even model the predicted structure of most of the Rlp7 protein (from Lys_{84} to Asn_{322}) on the basis of the structure of L7 in the crystal of yeast 60S r-subunit (19). The N-terminal part of Rlp7 (from Met_1 to Asp_{83}) shows a high probability of being intrinsically disordered. This is also the case for the N-terminal region of L7 r-protein (from Met_1 to Lys_{21}) (A.D.-Q., unpublished results). As expected, both core structures satisfactorily superimposed (Supplementary Figure S4). Taking into account the degree of homology and the predicted structural resemblance, we aimed to test whether Rlp7 and L7 could exchange at a particular stage of the 60S r-subunit maturation pathway.

To determine whether Rlp7 and L7 co-exist in the same pre-ribosomal particles, we performed a SILAC experiment combined with LC-MS/MS mass spectrometry to monitor the relative amount of L7 present in affinity purified pre-ribosomal complexes from a strain expressing TAP-tagged Rlp7. We have identified 60 pre-60S factors and most r-proteins from 60S (Rpl) and 40S (Rps) r-subunits (Figure 1 and Supplementary Data Set S1). Proteins clearly arranged in two groups. Group 1 (specifically enriched proteins; $\log_2$ of SILAC ratio between -6 and -1) comprises most Rpl proteins (including L7) and 31 pre-60S factors, in addition to Rlp7, which were strongly enriched in the Rlp7-TAP associated sample; among them, early-acting pre-60S factors including several A_3 factors (Erb1, Nop7, Nsa3, Ytm1 and Nop15) (6) were specifically abundant as judged from the total MS signal intensity. Group 2 (non-specific proteins; $\log_2$ of SILAC ratio between -1 and 0) includes 28 pre-60S factors, among them late-acting pre-60S factors (e.g. Arb1, Bud20, Arx1 and Alb1) and components of snoRNPs (e.g. Nop56, Cbf5, Nop1, Nhp2, Nop58, Gar1 and Snu13). Interestingly, all detected Rps proteins, the Rpl proteins from the r-stalk (P0, P1 and P2) and L10 appeared at values considered as contamination ($\log_2$ of SILAC ratio > 0). These results are consistent with the fact that the assembly of the r-stalk proteins and L10 into 60S r-subunits occurs predominantly in the cytoplasm (42,43). These results, confirmed by a second independent SILAC experiment, clearly identify Rlp7 as an early pre-60S assembly factor and strongly suggest that Rlp7 and L7 bind to the same pre-ribosomal particles (Supplementary Data Set S1).

To confirm these results, we performed IgG-Sepharose purification with extracts of cells expressing both C-terminal TAP-tagged L7B and HA-tagged Rlp7 proteins. Analysis of purified complexes by SDS–PAGE and western blotting demonstrated the association of Rlp7-HA with L7B-TAP, the r-proteins L1 and L35 and the *trans*-acting factor Has1 (Figure 2). This association seems to be specific, as no co-purification was observed

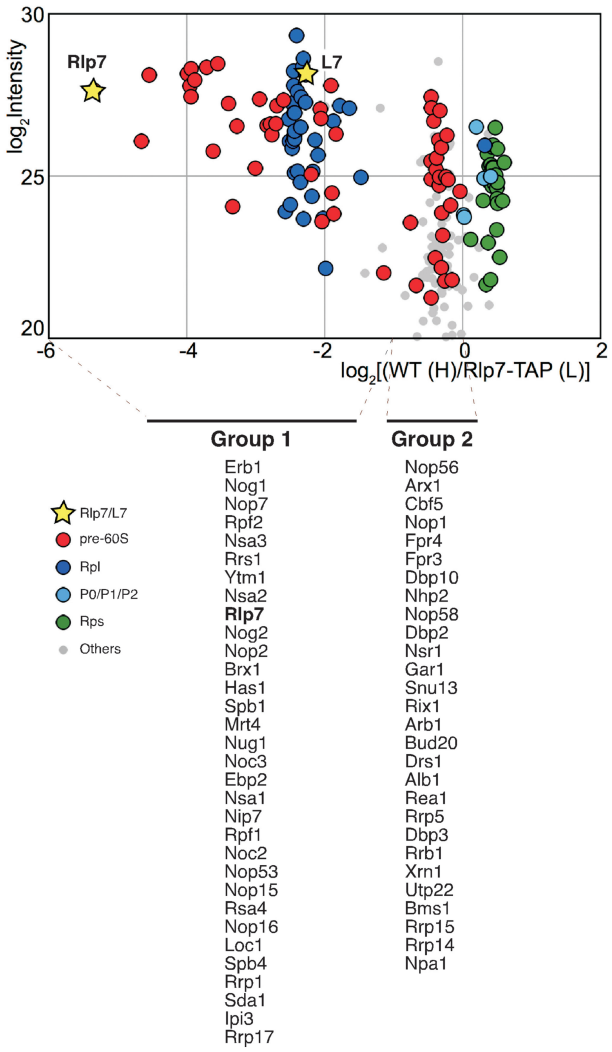


Figure 1. Rlp7-associated pre-ribosomal particles contain the L7 r-protein. Wild-type and Rlp7-TAP cells were mixed in equal proportions prior to complex purification. The $\log_2$ of SILAC ratios (median value of Wild-type/Rlp7-TAP peptide ratio) were plotted against the sum of the intensity of all the peptides for each protein. Dots are coloured according to protein function: pre-60S factors (red), 60S r-proteins (Rpl, blue), r-stalk proteins (P0/P1/P2, light blue), 40S r-proteins (Rps, green) and proteins of other different functions (grey). Yellow stars indicate the Rlp7 and L7 values. The identity of pre-60S factors specifically enriched (Group 1) or not (Group 2) is indicated below the graph. These factors are listed from their highest to lowest intensity values (see Supplementary Data Set S1).

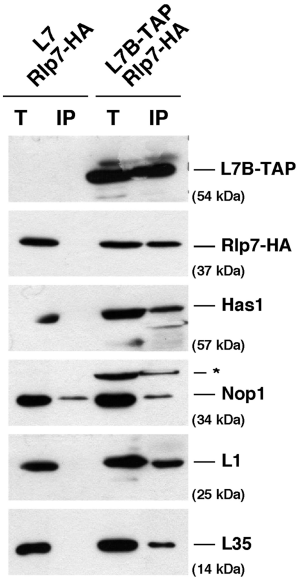


Figure 2. Rlp7 and L7 association with pre-ribosomal particles is not mutually exclusive. Extracts were prepared from cells co-expressing Rlp7-HA and L7B-TAP or, as a control, Rlp7-HA and untagged L7B. Total extracts (T) and L7B-TAP affinity-purified samples (IP) were analysed by western blotting. Co-precipitation of Rlp7-HA, Has1, Nop1 and r-proteins L1 and L35 was tested with specific antibodies. The asterisk corresponds to a L7B-TAP cross-reaction.

when a strain harbouring a non-tagged L7B was used as a control. In contrast, the *trans*-acting factor Nop1 did not associate with L7B-TAP (Figure 2). This result correlates well with the identification of Has1 in Group 1 and Nop1 in Group 2 of Rlp7-TAP associated proteins. We conclude that both Rlp7 and L7 are able to simultaneously bind to the same particles.

Rlp7 has been shown to localize in the nucleolus (20,21) and associate to pre-60S r-particles (6,21). To further characterize Rlp7-containing pre-ribosomal particles, we performed a one-step IgG-Sepharose purification with lysates from a C-terminal Rlp7-HTP (His_6-TEV-Protein A)-tagged strain and a non-tagged negative control and identified the associated pre-rRNA species. As shown in Figure 3, there was a significant enrichment for the $27SA_2$, 27SB and 7S pre-rRNAs and a modest enrichment for 35S pre-rRNA for the Rlp7-HTP-precipitated preparations. No enrichment over the background levels was detected for mature rRNAs. These results indicate that Rlp7 is a stable component of early and medium pre-60S particles that dissociates from particles following 7S pre-rRNA processing. To explore L7 timing of assembly, we performed the reverse experiment by affinity purification of L7B-HTP containing complexes from C-terminal HTP-tagged L7B strain. As shown in Figure 3, and as expected for a 60S r-protein [e.g. (14,18)], there was significant co-purification of mature rRNAs with L7B-HTP. Precursors 27S and 7S pre-rRNAs were clearly also detected, in contrast to 35S and 20S pre-rRNAs that were found at the background level (Figure 3). The contribution of the *RPL7A* gene to growth and ribosome biogenesis is more important than that of *RPL7B* (15) (Supplementary Figure S5); thus, to improve the pre-rRNA co-precipitation by L7B-HTP, we made use of an isogenic HTP-tagged strain disrupted for the *RPL7A* gene. As expected, this strain displayed a slow-growth phenotype (Supplementary Figure S5). As also shown in Figure 3, L7B-HTP more efficiently co-precipitated pre- and mature rRNAs, especially 7S pre-rRNA, in this genetic background. Thus, L7 stably assembles into

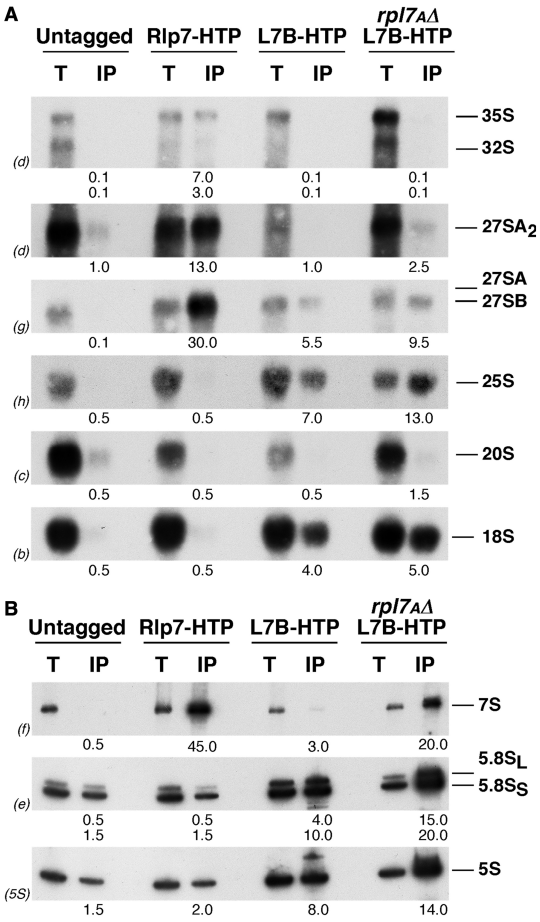


Figure 3. Association of Rlp7 and L7 with r-particles. HTP-tagged Rlp7 and L7 were affinity-purified from extracts of the indicated strains. RNA was isolated from total extract (T) and purified samples (IP) and analysed by northern blotting. (**A**) Large pre- and mature rRNAs. (**B**) Small pre- and mature rRNAs. Probes (in parentheses) are described in Supplementary Figure S1. Signal intensity was measured by phosphorimager scanning; values (below each IP lane) refer to the percentage of each RNA recovered after purification.

early and medium pre-60S particles. Consistently, when we monitor the localization of a functional L7B-GFP construct on induction of the dominant negative *NMD3Δ100* allele, which lead to the retention of pre-60S r-particles in the nucle(ol)us (44), this accumulates in the nucle(ol)us from most cells examined (Supplementary Figure S6), similarly as does a L3-GFP reporter used as positive control for nucle(ol)ar assembly.

Taken together, these data demonstrate that Rlp7 and L7 are present in similar pre-ribosomal complexes at the same time in the nuclear stages of 60S r-subunit assembly. This conclusion is outlined in Supplementary Figure S2.

Rlp7 and L7 bind distinct sites within pre-ribosomal particles

L7 is a RNA binding protein (19) and due to its homology with L7, Rlp7 is also predicted to bind RNA. To find out whether Rlp7 and L7 share the same binding site on pre-60S ribosome, we attempted to identify their *in vivo* binding sites by using the CRAC method (9). The Rlp7-HTP strain did not show any growth phenotype at 30°C (Supplementary Figure S5). We also analysed L7B-HTP tagged strains, with or without the *RPL7A* endogenous copy, both presenting a phenotype consistent with the non-tagged corresponding strain (Supplementary Figure S5). We found that Rlp7-HTP directly and specifically contacts two regions in ITS2 (Figures 4 and 5, Supplementary Figure S7A, S8 and S9), whereas the non-ultraviolet cross-linked Rlp7-HTP protein did not significantly cross-link detectable rRNA (Supplementary Figure S7A). These two regions of ITS2 overlap the boundaries of 25S 5′ end and 5.8S 3′ end. Previous CRAC experiments with other A_3 factors, Erb1, Nop7, Nop15 and Nsa3 revealed binding sites close and even overlapping these boundaries, consistent with the described collective role of these factors for processing of the $27SA_3$ pre-rRNA (6,7) (Figure 5 and Supplementary Figure S7A). Indeed, loss-of-function of the A_3 factors (including Rlp7) leads to the accumulation of the $27SA_3$ pre-rRNA, reduced formation of $27SB_S$ relative to $27SB_L$ pre-rRNA and loss of cleavage at site C_2 in ITS2 (6,7,20,21). Interestingly, the nucleotide substitutions analysis at specific positions in the sequence reads allowed us to precisely identify cross-link sites of Rlp7-HTP as nucleotides distributed in two groups, one adjacent to the pre-rRNA processing sites C_1 and E that define the 5′ end and 3′ end of mature 25S and 5.8S rRNAs, respectively, and another at helix III at the 3′ end of ITS2, between nucleotides 200 and 225 (Figure 5 and Supplementary Figure S8).

We also performed CRAC analyses with strains expressing HTP-tagged L7B either harbouring the wild-type *RPL7A* (LMA1551 strain) or the null *rpl7AΔ* allele (LMA1730 strain). The LMA1551 strain did not give a sufficient signal to get robust identification of binding sites; thus, we pursued the experiment only with the LMA1730 strain (Supplementary Figure S7B). We found that L7 mainly cross-linked to helix $ES7^L$b in 25S rRNA, which corresponds to an expansion segment exclusive of eukaryotic 25S rRNA (19), and a long region of 5S rRNA over H2, H4 and H5 (Figures 4 and 6, Supplementary Figures S8 and S9). Base substitutions at positions U_{508}, U_{520} and G_{579} in the 25S rRNA identified positions where L7 intimately contact 25S rRNA (Supplementary Figure S8). These results are in full agreement with the L7 interaction sites on domain II of 25S rRNA deduced from the large r-subunit crystal structure analysis (19). Additionally, minor cross-linking regions were identified along 25S rRNA and even at 18S rRNA that are at the threshold of what could be considered background (Figure 4).

Altogether, these findings clearly indicate that the Rlp7 and L7 binding sites are distinct in pre- and/or mature rRNA and are distant from each other, consistent with simultaneous binding of Rlp7 and L7 to pre-60S ribosomal particles without steric conflict.

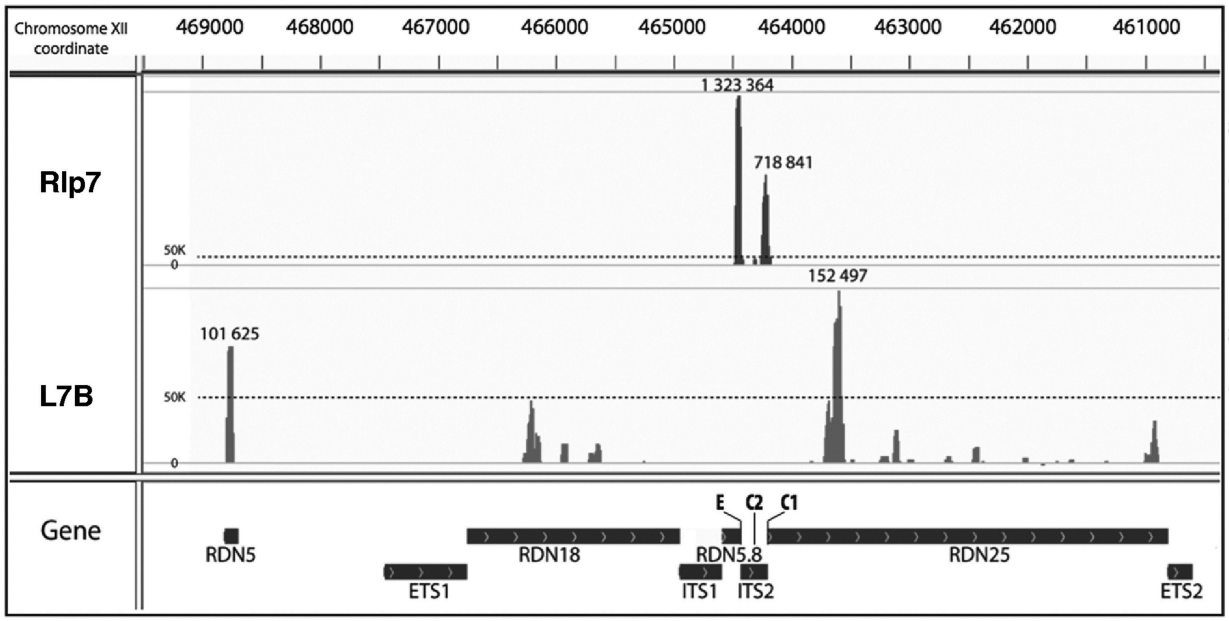


Figure 4. Identification of Rlp7 and L7 binding sites on pre- and mature rRNAs. The histograms, plotted using the Integrative Genomics Viewer software, display the sequences identified on CRAC analysis and the number of hits mapped to the rDNA. CRAC was performed with Rlp7-HTP (Rlp7) and L7B-HTP *rpl7AΔ* (L7B) cells. The maximum number of hits in the main peaks is shown.

DISCUSSION

In this work, we report the unexpected discovery that both r-protein L7 and its related pre-60S factor Rlp7 bind to the same nuclear 60S precursors. Moreover, we show that despite their sequence and predicted structural similarity, the proteins bind to distinct sites on pre-rRNA.

Our SILAC and immunoprecipitation experiments indicate that Rlp7 mainly associates with early and medium nucle(ol)ar pre-60S r-particles and likely dissociates after cleavage of 27SB pre-rRNA, following exonucleolytic 7S pre-rRNA processing. These results complement those previously obtained by Woolford and co-workers (6). Strikingly, the most enriched and abundant pre-60S factors associated to Rlp7-TAP (Erb1, Nop7, Nsa3, Ytm1 and Nop15) are components of the A_3 factors cluster whose loss-of-function leads to processing defects at A_3 site with the subsequent accumulation of the $27SA_3$ precursor and depletion of its immediate $27SB_S$ and 7S pre-rRNA products [(6,7) and references therein]. The A_3 factors are also intriguing because their association with pre-ribosomal particles appears to be interdependent (6). Moreover, A_3 factors are required for proper assembly of four r-proteins (L17, L26, L35 and L37) that predominantly bind to 5.8S/25S rRNA domain I, which in turn enable cleavage of ITS2 at site C_2 (6,14,18,47). Consistently, our CRAC analyses show that Rlp7 binds to ITS2 at a position adjacent with that of the 3′ end of mature 5.8S rRNA (site E) and the 5′ end of mature 25S rRNA (site C_1). The Rlp7 binding sites on pre-rRNA partially overlap with those previously reported for the A_3 factor Nsa3 and are close to those of other A_3 factors (Nop12, Nop15, Erb1 and Nop7) (7). It has been shown that binding of Nsa3 to 27S pre-rRNAs is required to maintain a flexible and an open structure in ITS2 (the so-called ring conformation), which prevents the formation of the mutually exclusive hairpin structure of ITS2 (7). However, assuming that (i) Rlp7 contacts rRNA in a similar manner as L7 (see Supplementary Figure S9) and that (ii) one Rlp7 molecule simultaneously binds to the two Rlp7 binding sites, we propose that Rlp7 might preferentially interact with the hairpin structure of ITS2, thereby promoting the transition from the ring to the hairpin conformation of ITS2 that it is essentially required for ITS2 processing at site C_2 (48). Consistently with this hypothesis, we identified the precise Rlp7 cross-linking sites by the presence of point mutations exactly at the site E and one nucleotide after the site C_1. Sites E and C_1 represent, respectively, precise positions where the exonuclease activities of the exosome and Rat1-Xrn1-Rrp17 stop at ITS2 (49–51). Whether Rlp7 acts blocking progression of these exonucleases beyond the processing sites is an attractive suggestion, which fits well with our results, but it needs further experimental evidence to be proven.

Our results also clearly indicate the early nucle(ol)ar assembly of L7 to ribosome precursors containing 27S pre-rRNAs. RNA and protein precipitation experiments strongly suggest that both Rlp7 and L7 co-exist in the same pre-ribosomal particles. Moreover, SILAC shows that L7 is as abundant as any other 60S r-protein in an Rlp7-TAP purified fraction. CRAC examination of the RNA binding sites of L7 in r-particles showed specific cross-links to helix $ES7^L$b and 5S rRNA. These sites are far from the CRAC identified Rlp7 binding sites (see Supplementary Figure S9), therefore, suggesting mutually

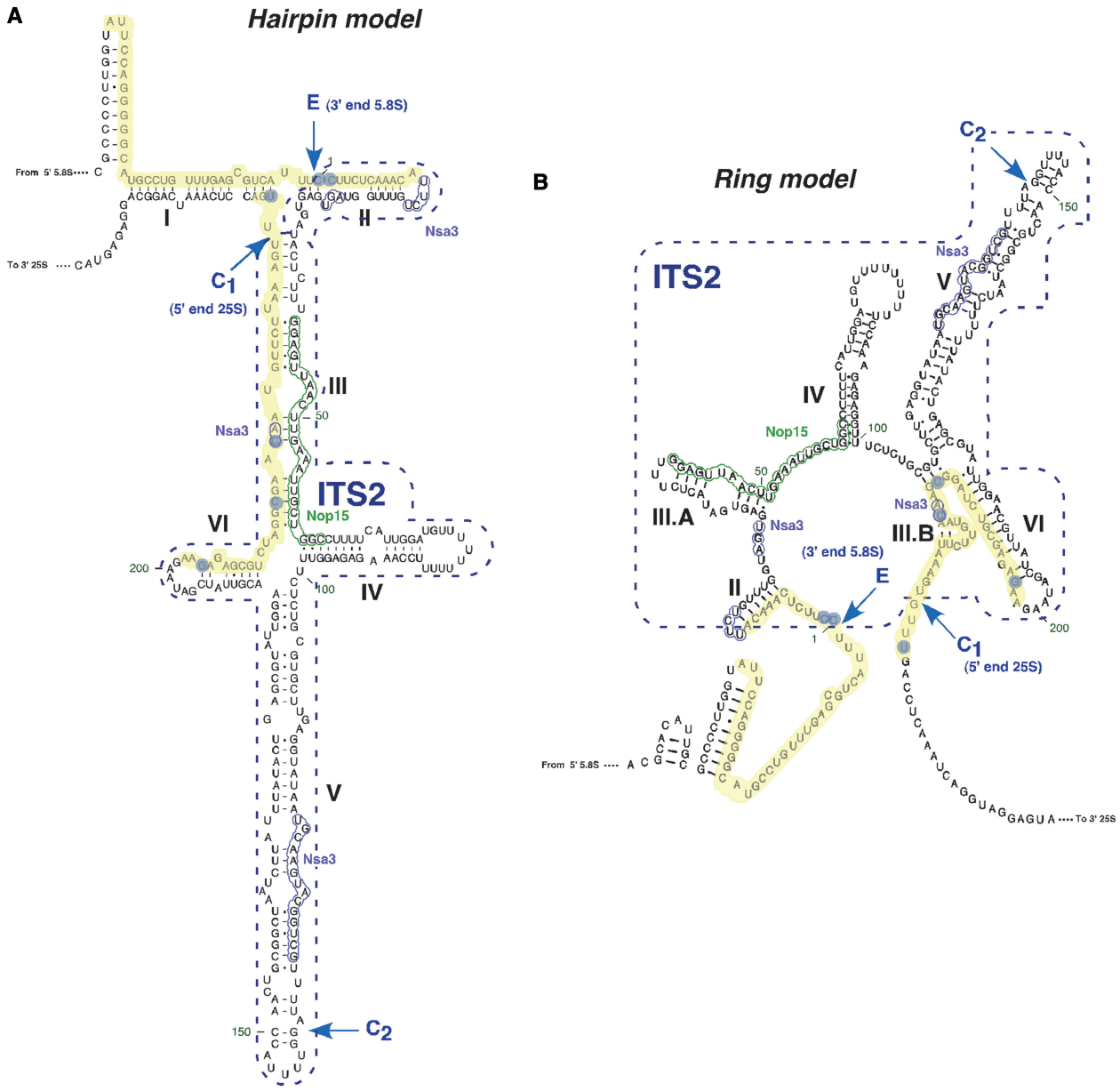


Figure 5. Localization of the CRAC interaction sites of Rlp7 with pre-rRNA sequences displayed on the 'hairpin model' (**A**) and on the 'ring model' (**B**) for yeast ITS2 secondary structure [for a reference, see (45)]. The CRAC sites are highlighted in yellow; blue circles indicate frequently mutated residues found in the experiments (see Supplementary Figure S8). The Nsa3 and Nop15 CRAC sites, as described in (7), are represented as purple and green, respectively. The location of the CRAC sites cleavage sites C_1, C_2 and E are also indicated.

independent binding of both proteins with pre-ribosomal particles. Consistent with this, Woolford Jr and co-workers have demonstrated that levels of L7 in pre-60S r-particles were practically unaffected on depletion of Rlp7 (6). In the opposite experiment, depletion of L7 only slightly diminished levels of Rlp7 in pre-60S r-particles (15). Thus, Rlp7 is likely not the placeholder for the assembly of L7 r-protein; thus, L7 does not exchange with Rlp7 during 60S r-subunit biogenesis. This scenario is the opposite to that we have previously reported for the Mrt4-P0 pair of paralogues (23,42). In addition to Mrt4-P0 and Rlp7-L7, there are at least two other pairs of paralogues comprised by an r-like assembly factor and an r-protein in yeast, Imp3-S9 and Rlp24-L24. Whether the dynamics of these pairs during ribosome biogenesis resembles that of Mrt4-P0 or that of Rlp7-L7 evidently needs further investigation.

In conclusion, the hypothesis that a distinct *trans*-acting factor serves as a placeholder for its homologous r-protein is not applicable in all circumstances and, indeed, paralogues can co-exist within the same pre-ribosomal

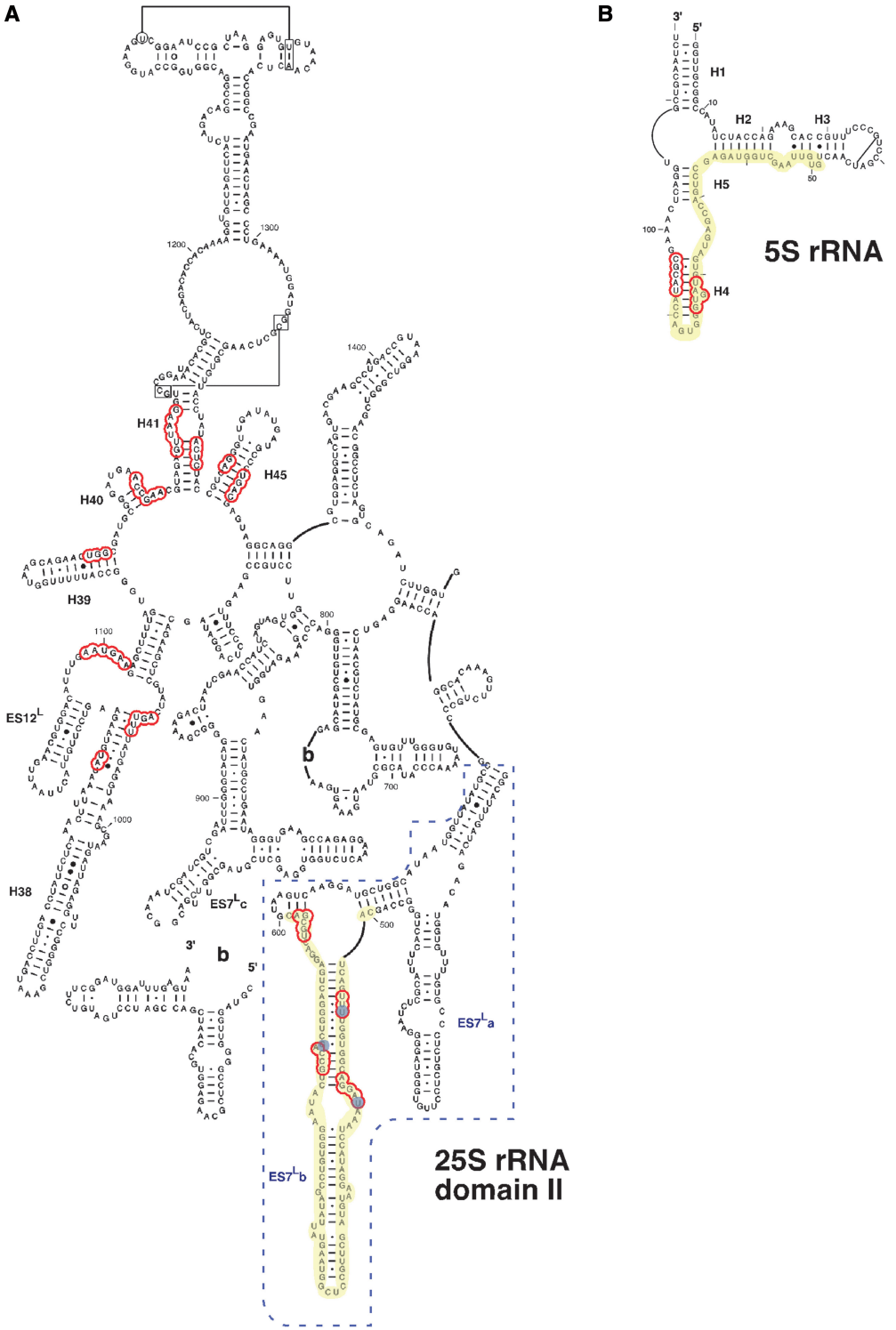


Figure 6. Localization of the CRAC interaction sites of L7 in the secondary structure of domain II of 25S rRNA (**A**) and 5S rRNA (**B**). The structures, residues and helix (H) numbers were taken from the Comparative RNA Web Site (46). The eukaryotic expansion segment ES7^{L} is labelled with a dash box. Red circles indicate rRNA residues situated in close proximity of L7 (closer than 5 Å) in the structure of yeast 60S r-subunit [PDB file 3U5H; (19)]. The CRAC sites are highlighted in yellow; blue circles indicate frequently mutated residues found in the experiments.

particles. Furthermore, genetic and biochemical experiments will be required to unravel the precise function of Rlp7 during ribosome assembly. Recently, it has been found that some archaeal r-proteins have more than one RNA binding site in ribosomes (52). These sites are structurally similar, which explains the promiscuous behaviour of these r-proteins. Whether the hairpin conformation of ITS2 resembles the structure of 25S rRNA domain II involved in the binding of L7 r-protein remains a challenging question for future studies. If this is the case, we will need to address then how proteins whose cores fold into apparently similar 3D structures could be specifically targeted to different locations in pre-ribosomal complexes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [53–58].

ACKNOWLEDGEMENTS

The authors thank J.L. Woolford, Jr. and J. Dembowski for communicating unpublished results, S. Granneman for advices with CRAC, D. Tollervey for providing with models of the ITS2 structure and all colleagues mentioned in the text for their gifts of material used in this study. They are grateful to the proteomic platform of the Pasteur Institute for the availability of the Orbitrap.

FUNDING

Spanish Ministry of Science and Innovation and ERDF [BFU2010-15690 and FR2009-0102]; Andalusian Government [CVI-271 and P08-CVI-03508 to J.d.l.C.]; Agence Nationale de la Recherche [ANR-2011-BSV6-011-785 02]; EGIDE Picasso Programme (to M.F-R.); recipient of an FPI fellowship from the Andalusian Government (to R.B.). Funding for open access charge: Spanish Ministry of Science and Innovation and ERDF [BFU2010-15690 to J.d.l.C.].

Conflict of interest statement. None declared.

REFERENCES

1. Henras,A.K., Soudet,J., Gerus,M., Lebaron,S., Caizergues-Ferrer,M., Mougin,A. and Henry,Y. (2008) The post-transcriptional steps of eukaryotic ribosome biogenesis. *Cell. Mol. Life Sci.*, **65**, 2334–2359.
2. Panse,V.G. and Johnson,A.W. (2010) Maturation of eukaryotic ribosomes: acquisition of functionality. *Trends Biochem. Sci.*, **35**, 260–266.
3. Li,Z., Lee,I., Moradi,E., Hung,N.J., Johnson,A.W. and Marcotte,E.M. (2009) Rational extension of the ribosome biogenesis pathway using network-guided genetics. *PLoS Biol.*, **7**, e1000213.
4. Kressler,D., Hurt,E. and Bassler,J. (2010) Driving ribosome assembly. *Biochim. Biophys. Acta*, **1803**, 673–683.
5. Lebreton,A., Rousselle,J.C., Lenormand,P., Namane,A., Jacquier,A., Fromont-Racine,M. and Saveanu,C. (2008) 60S ribosomal subunit assembly dynamics defined by semi-quantitative mass spectrometry of purified complexes. *Nucleic Acids Res.*, **36**, 4988–4999.
6. Sahasranaman,A., Dembowski,J., Strahler,J., Andrews,P., Maddock,J. and Woolford,J.L. Jr (2011) Assembly of *Saccharomyces cerevisiae* 60S ribosomal subunits: role of factors required for 27S pre–rRNA processing. *EMBO J.*, **30**, 4020–4032.
7. Granneman,S., Petfalski,E. and Tollervey,D. (2011) A cluster of ribosome synthesis factors regulate pre-rRNA folding and 5.8S rRNA maturation by the Rat1 exonuclease. *EMBO J.*, **30**, 4006–4019.
8. Bohnsack,M.T., Martin,R., Granneman,S., Ruprecht,M., Schleiff,E. and Tollervey,D. (2009) Prp43 bound at different sites on the pre-rRNA performs distinct functions in ribosome synthesis. *Mol. Cell*, **36**, 583–592.
9. Granneman,S., Kudla,G., Petfalski,E. and Tollervey,D. (2009) Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc. Natl Acad. Sci. USA*, **106**, 9613–9618.
10. Bradatsch,B., Katahira,J., Kowalinski,E., Bange,G., Yao,W., Sekimoto,T., Baumgartel,V., Boese,G., Bassler,J., Wild,K. *et al.* (2007) Arx1 functions as an unorthodox nuclear export receptor for the 60S preribosomal subunit. *Mol. Cell*, **27**, 767–779.
11. Segerstolpe,A., Granneman,S., Bjork,P., de Lima Alves,F., Rappsilber,J., Andersson,C., Hogbom,M., Tollervey,D. and Wieslander,L. (2013) Multiple RNA interactions position Mrd1 at the site of the small subunit pseudoknot within the 90S pre-ribosome. *Nucleic Acids Res.*, **41**, 1178–1190.
12. Ferreira-Cerca,S., Pöll,G., Gleizes,P.E., Tschochner,H. and Milkereit,P. (2005) Roles of eukaryotic ribosomal proteins in maturation and transport of pre-18S rRNA and ribosome function. *Mol. Cell*, **20**, 263–275.
13. Pöll,G., Braun,T., Jakovljevic,J., Neueder,A., Jakob,S., Woolford,J.L. Jr, Tschochner,H. and Milkereit,P. (2009) rRNA maturation in yeast cells depleted of large ribosomal subunit proteins. *PLoS One*, **4**, e8249.
14. Babiano,R. and de la Cruz,J. (2010) Ribosomal protein L35 is required for 27SB pre-rRNA processing in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **38**, 5177–5192.
15. Jakovljevic,J., Ohmayer,U., Gamalinda,M., Talkish,J., Alexander,L., Linnemann,J., Milkereit,P. and Woolford,J.L. Jr (2012) Ribosomal proteins L7 and L8 function in concert with six A3 assembly factors to propagate assembly of domains I and II of 25S rRNA in yeast 60S ribosomal subunits. *RNA*, **18**, 1805–1822.
16. Ferreira-Cerca,S., Pöll,G., Kuhn,H., Neueder,A., Jakob,S., Tschochner,H. and Milkereit,P. (2007) Analysis of the in vivo assembly pathway of eukaryotic 40S ribosomal proteins. *Mol. Cell*, **28**, 446–457.
17. Fernández-Pevida,A., Rodríguez-Galán,O., Díaz-Quintana,A., Kressler,D. and de la Cruz,J. (2012) Yeast ribosomal protein L40 assembles late into precursor 60S ribosomes and is required for their cytoplasmic maturation. *J. Biol. Chem.*, **287**, 38390–38407.
18. Babiano,R., Gamalinda,M., Woolford,J.L. Jr and de la Cruz,J. (2012) *Saccharomyces cerevisiae* Ribosomal Protein L26 Is Not Essential for Ribosome Assembly and Function. *Mol. Cell. Biol.*, **32**, 3228–3241.
19. Ben-Shem,A., Garreau de Loubresse,N., Melnikov,S., Jenner,L., Yusupova,G. and Yusupov,M. (2011) The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science*, **334**, 1524–1529.
20. Dunbar,D.A., Dragon,F., Lee,S.J. and Baserga,S.J. (2000) A nucleolar protein related to ribosomal protein L7 is required for an early step in large ribosomal subunit biogenesis. *Proc. Natl Acad. Sci. USA*, **97**, 13027–13032.
21. Gadal,O., Strauss,D., Petfalski,E., Gleizes,P.E., Gas,N., Tollervey,D. and Hurt,E. (2002) Rlp7p is associated with 60S preribosomes, restricted to the granular component of the nucleolus, and required for pre-rRNA processing. *J. Cell Biol.*, **157**, 941–951.
22. Saveanu,C., Namane,A., Gleizes,P.E., Lebreton,A., Rousselle,J.C., Noaillac-Depeyre,J., Gas,N., Jacquier,A. and Fromont-Racine,M. (2003) Sequential protein association with nascent 60S ribosomal particles. *Mol. Cell. Biol.*, **23**, 4449–4460.
23. Rodríguez-Mateos,M., Abia,D., García-Gómez,J.J., Morreale,A., de la Cruz,J., Santos,C., Remacha,M. and Ballesta,J.P.G. (2009)

The amino terminal domain from Mrt4 protein can functionally replace the RNA binding domain of the ribosomal P0 protein. *Nucleic Acids Res.*, **37**, 3514–3521.
24. Lee,S.J. and Baserga,S.J. (1999) Imp3p and Imp4p, two specific components of the U3 small nucleolar ribonucleoprotein that are essential for 18S rRNA processing. *Mol. Cell. Biol.*, **19**, 5441–5452.
25. de la Cruz,J., Kressler,D. and Linder,P. (2004) In: Olson,M.O.J. (ed.), *Nucleolus. Kluwer academic*. Landes Bioscience/eurekah.com, Georgetown, pp. 258–285.
26. Kaiser,C., Michaelis,S. and Mitchell,A. (1994) *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
27. de la Cruz,J., Lacombe,T., Deloche,O., Linder,P. and Kressler,D. (2004) The putative RNA helicase Dbp6p functionally interacts with Rpl3p, Nop8p and the novel trans-acting factor Rsa3p during biogenesis of 60S ribosomal subunits in *Saccharomyces cerevisiae*. *Genetics*, **166**, 1687–1699.
28. Defenouillère,Q., Yao,Y., Mouaikel,J., Namane,A., Galopier,A., Decourty,L., Doyen,A., Malabat,C., Saveanu,C., Jacquier,A. *et al.* (2013) Cdc48-associated complex bound to 60S particles is required for the clearance of aberrant translation products. *Proc. Natl Acad. Sci. USA*, **110**, 5046–5051.
29. Oeffinger,M., Wei,K.E., Rogers,R., DeGrasse,J.A., Chait,B.T., Aitchison,J.D. and Rout,M.P. (2007) Comprehensive analysis of diverse ribonucleoprotein complexes. *Nat. Methods*, **4**, 951–956.
30. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.
31. Cox,J., Matic,I., Hilger,M., Nagaraj,N., Selbach,M., Olsen,J.V. and Mann,M. (2009) A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat. Protoc.*, **4**, 698–705.
32. Cox,J., Neuhauser,N., Michalski,A., Scheltema,R.A., Olsen,J.V. and Mann,M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.*, **10**, 1794–1805.
33. Dez,C., Froment,C., Noaillac-Depeyre,J., Monsarrat,B., Caizergues-Ferrer,M. and Henry,Y. (2004) Npa1p, a component of very early pre-60S ribosomal particles, associates with a subset of small nucleolar RNPs required for peptidyl transferase center modification. *Mol. Cell. Biol.*, **24**, 6324–6337.
34. Rosado,I.V., Dez,C., Lebaron,S., Caizergues-Ferrer,M., Henry,Y. and de la Cruz,J. (2007) Characterization of *Saccharomyces cerevisiae* Npa2p (Urb2p) reveals a low-molecular-mass complex containing Dbp6p, Npa1p (Urb1p), Nop8p, and Rsa3p involved in early steps of 60S ribosomal subunit biogenesis. *Mol. Cell. Biol.*, **27**, 1207–1221.
35. Petitjean,A., Bonneaud,N. and Lacroute,F. (1995) The duplicated *Saccharomyces cerevisiae* gene SSM1 encodes a eucaryotic homolog of the eubacterial and archeabacterial L1 ribsomal protein. *Mol. Cell. Biol.*, **15**, 5071–5081.
36. Frey,S., Pool,M. and Seedorf,M. (2001) Scp160p, an RNA-binding, polysome-associated protein, localizes to the endoplasmic reticulum of *Saccharomyces cerevisiae* in a microtubule-dependent manner. *J. Biol. Chem.*, **276**, 15905–15912.
37. Emery,B., de la Cruz,J., Rocak,S., Deloche,O. and Linder,P. (2004) Has1p, a member of the DEAD-box family, is required for 40S ribosomal subunit biogenesis in *Saccharomyces cerevisiae*. *Mol. Microbiol.*, **52**, 141–158.
38. Thorvaldsdottir,H., Robinson,J.T. and Mesirov,J.P. (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
39. Eswar,N., Eramian,D., Webb,B., Shen,M.Y. and Sali,A. (2008) Protein structure modeling with MODELLER. *Methods Mol. Biol.*, **426**, 145–159.
40. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera-a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
41. Lalo,D., Mariotte,S. and Thuriaux,P. (1993) Two distinct yeast proteins are related to the mammalian ribosomal polypeptide L7. *Yeast*, **9**, 1085–1091.
42. Rodríguez-Mateos,M., García-Gómez,J.J., Francisco-Velilla,R., Remacha,M., de la Cruz,J. and Ballesta,J.P.G. (2009) Role and dynamics of the ribosomal protein P0 and its related trans-acting factor Mrt4 during ribosome assembly in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **37**, 7519–7532.
43. West,M., Hedges,J.B., Chen,A. and Johnson,A.W. (2005) Defining the order in which Nmd3p and Rpl10p load onto nascent 60S ribosomal subunits. *Mol. Cell. Biol.*, **25**, 3802–3813.
44. Ho,J.H.N., Kallstrom,G. and Johnson,A.W. (2000) Nmd3p is a Crm1p-dependent adapter protein for nuclear export of the large ribosomal subunit. *J. Cell. Biol.*, **151**, 1057–1066.
45. Côté,C.A., Greer,C.L. and Peculis,B.A. (2002) Dynamic conformational model for the role of ITS2 in pre-rRNA processing in yeast. *RNA*, **8**, 786–797.
46. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
47. Gamalinda,M., Jakovljevic,J., Babiano,R., Talkish,J., de la Cruz,J. and Woolford,J.L. Jr (2013) Yeast polypeptide exit tunnel ribosomal proteins L17, L35 and L37 are necessary to recruit late-assembling factors required for 27SB pre-rRNA processing. *Nucleic Acids Res.*, **41**, 1965–1983.
48. Côté,C.A. and Peculius,B.A. (2001) Role of the ITS2-proximal stem and evidence for indirect recognition of processing sites in pre-rRNA processing in yeast. *Nucleic Acids Res.*, **29**, 2106–2116.
49. Oeffinger,M., Zenklusen,D., Ferguson,A., Wei,K.E., El Hage,A., Tollervey,D., Chait,B.T., Singer,R.H. and Rout,M.P. (2009) Rrp17p is a eukaryotic exonuclease required for 5' end processing of Pre-60S ribosomal RNA. *Mol. Cell*, **36**, 768–781.
50. Henry,Y., Wood,H., Morrissey,J.P., Petfalski,E., Kearsey,S. and Tollervey,D. (1994) The 5' end of yeast 5.8S rRNA is generated by exonucleases from an upstream cleavage site. *EMBO J.*, **13**, 2452–2463.
51. Mitchell,P., Petfalski,E., Shevchenko,A., Mann,M. and Tollervey,D. (1997) The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-5' exoribonucleases. *Cell*, **91**, 457–466.
52. Armache,J.P., Anger,A.M., Marquez,V., Franckenberg,S., Frohlich,T., Villa,E., Berninghausen,O., Thomm,M., Arnold,G.J., Beckmann,R. *et al.* (2013) Promiscuous behaviour of archaeal ribosomal proteins: implications for eukaryotic ribosome evolution. *Nucleic Acids Res.*, **41**, 1284–1293.
53. Baudin,A., Ozier-Kalogeropoulos,O., Denouel,A., Lacroute,F. and Cullin,C. (1993) A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **21**, 3329–3330.
54. Brachmann,C.B., Davies,A., Cost,G.J., Caputo,E., Li,J., Hieter,P. and Boeke,J.D. (1998) Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast*, **14**, 115–132.
55. Ulbrich,C., Diepholz,M., Bassler,J., Kressler,D., Pertschy,B., Galani,K., Bottcher,B. and Hurt,E. (2009) Mechanochemical removal of ribosome biogenesis factors from nascent 60S ribosomal subunits. *Cell*, **138**, 911–922.
56. Belk,J.P., He,F. and Jacobson,A. (1999) Overexpression of truncated Nmd3p inhibits protein synthesis in yeast. *RNA*, **5**, 1055–1070.
57. Dez,C. and Tollervey,D. (2004) Ribosome synthesis meets the cell cycle. *Curr. Opin. Microbiol.*, **7**, 631–637.
58. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Published online 18 May 2018

Nucleic Acids Research, 2018, Vol. 46, No. 12 ***6009–6025***
doi: 10.1093/nar/gky342

Antisense transcriptional interference mediates condition-specific gene repression in budding yeast

Alicia Nevers[1,2], Antonia Doyen[1], Christophe Malabat[1,3], Bertrand Néron[3], Thomas Kergrohen[1], Alain Jacquier[1,4,*] and Gwenael Badis[1,4,*]

[1]Unité GIM, Institut Pasteur, Paris, France, [2]Sorbonne Université Pierre et Marie Curie, Paris, France, [3]Bioinformatics and Biostatistics Hub, C3BI, Institut Pasteur, USR 3756 IP CNRS, Paris, France and [4]CNRS UMR3525, Paris, France

Received December 12, 2017; Revised April 12, 2018; Editorial Decision April 15, 2018; Accepted April 23, 2018

ABSTRACT

Pervasive transcription generates many unstable non-coding transcripts in budding yeast. The transcription of such noncoding RNAs, in particular antisense RNAs (asRNAs), has been shown in a few examples to repress the expression of the associated mRNAs. Yet, such mechanism is not known to commonly contribute to the regulation of a given class of genes. Using a mutant context that stabilized pervasive transcripts, we observed that the least expressed mRNAs during the exponential phase were associated with high levels of asRNAs. These asRNAs also overlapped their corresponding gene promoters with a much higher frequency than average. Interrupting antisense transcription of a subset of genes corresponding to quiescence-enriched mRNAs restored their expression. The underlying mechanism acts in *cis* and involves several chromatin modifiers. Our results convey that transcription interference represses up to 30% of the 590 least expressed genes, which includes 163 genes with quiescence-enriched mRNAs. We also found that pervasive transcripts constitute a higher fraction of the transcriptome in quiescence relative to the exponential phase, consistent with gene expression itself playing an important role to suppress pervasive transcription. Accordingly, the *HIS1* asRNA, normally only present in quiescence, is expressed in exponential phase upon *HIS1* mRNA transcription interruption.

INTRODUCTION

In steady state, the transcriptome reflects the equilibrium between RNA synthesis and degradation. Eukaryotes have developed sophisticated systems to control the turnover of mRNAs and ncRNAs necessary to the cell, undesired RNA species being rapidly eliminated by quality control mechanisms.

The development of genome-wide techniques such as tiling arrays and cDNA next-generation sequencing to analyse transcriptomes revealed that eukaryotic genomes are pervasively transcribed (1). The genome of budding yeast is particularly compact and it has been hitherto conceded that >70% of it is composed of protein coding ORFs (2). Yet, this is only true if one does not distinguish the two DNA strands. If one takes into account sense and antisense genomic DNA, non protein-coding sequences represent up to 65% of it, leaving room to a large fraction of the genome for the generation of pervasive non-coding transcripts.

In yeast, pervasive transcription has been first reported more than a decade ago. If a fraction of it was uncovered in wild-type cells (3,4), a substantial part of the eukaryotic pervasive transcription is 'hidden' as it generates very short-lived 'cryptic' transcripts. These RNAs are difficult to detect unless they are stabilized by interfering with quality control mechanisms that normally eliminate them (5). Pervasive transcripts detected in wild-type yeast cells have been named 'SUTs' for 'Stable Unannotated Transcripts' (4), and different names have been given to cryptic transcripts depending on which factor was mutated in order to stabilize a particular class of RNAs. For example, CUTs (Cryptic Unstable Transcripts) were characterized upon removal of Rrp6, an exonuclease specific of the nuclear form of the exosome (4,6,7), XUTs were revealed upon removal of the cytoplasmic exonuclease Xrn1 (8) and NUTs correspond to transcripts that accumulate when the nuclear termination factor Nrd1 is depleted (9). Yet, there are in yeast only two main pathways responsible for the efficient elimination of pervasive transcripts: the nuclear Nrd1–Nab3–Sen1 (NNS) pathway, in which the early transcription termination of cryptic transcripts by the NNS complex is coupled to the degradation by the nuclear TRAMP–exosome complex (9–12) and the cytoplasmic non-sense mediated mRNA decay (NMD) pathway (13,14). Many of pervasive transcripts re-

*To whom correspondence should be addressed. Tel: +33 140613331; Fax: +33 140613456; Email: gwenael.badis-breard@pasteur.fr
Correspondence may also be addressed to Alain Jacquier. Tel: +33 140613205; Fax: +33 140613456; Email: jacquier@pasteur.fr

quire both pathways for their efficient and fast elimination (see 13).

Irrespective of which pathway predominates for their degradation, these transcripts all originate from nucleosome free regions (NFRs), which are essentially found 5′ and 3′ of mRNA coding sequences (15). When they originate from 5′ NFRs, they are most often transcribed divergently from mRNAs and result from an intrinsic low polarity of gene promoters (4,7). This divergent transcription has the potential to interfere with the expression of the neighboring upstream gene. Likewise, when a non-coding transcript initiates from the 3′ NFR in an antisense orientation to the upstream gene, its transcription has the potential to interfere with the proper expression of the corresponding mRNA (8,16). Such transcription interference by pervasive transcription is largely prevented genome-wide by the NNS quality control pathway, which ensures the early transcription termination of these transcripts and prevent them to extend into the promoter region of the corresponding antisense genes (9–11,17).

Whether pervasive transcription has a general function is a matter of debate. The fact that highly efficient quality control mechanisms have been selected during evolution to eliminate most of these transcripts argue in favor of the idea that most of them are non functional; however pervasive transcription by itself, more than its product, could play a role. Yet, the existence of the NNS pathway, which, by terminating pervasive transcription early, is key in preserving pervasive transcription from interfering with the expression of many coding genes, also suggests that a large fraction of these events simply result from the low specificity of RNA polymerase II (PolII) transcription initiation.

There are a number of well-documented examples of individual coding gene regulation through the transcription of a non-coding RNA: *SER3* (18), *IME1* and *IME4* (19), *GAL10/GAL1* (20,21), *PHO84* (22), *CDC28* (23) as examples. In the vast majority of cases analysed in budding yeast, the synthesis of a non-coding transcript has only an effect in *cis*. The prevailing model is that repressive chromatin marks are deposited in the promoter regions of genes in the wake of RNA polymerase II (PolII) transcribing the associated non-coding RNAs (24,25). It is thus the act of transcription rather that its product, which is important. Several distinct mechanisms can be at play, but the general theme is that methyltransferases interacting with the the carboxy-terminal domain (CTD) of the PolII large subunit deposit histone methylation marks that recruit repressive chromatin modifiers such as histone deacetylases or nucleosome remodelling complexes. In budding yeast, there are two such CTD associated histone methyl transferases. Set1 methylates histone H3K4 at promoters and gene proximal regions of actively transcribed genes while Set2 methylates H3K36 at more distal gene regions. The role of Set1 is complex. It is responsible for both H3K4 di- and tri-methylation (H3K4me2 and H3K4me3). It has been proposed that H3K4me3 at the beginning of actively transcribed genes could enhance and help maintaining pre-initiation complex assembly and an active acetylated chromatin state, thus playing a positive role on transcription. Conversely, Set1 generates H3K4me2 in the body of gene, which recruits the histone deacetylase complexes SET3 or RPD3L, resulting in transcription initiation repression (see 26 for review). Set2 is responsible for the H3K36 methylation (H3K36me2) in the body of genes, resulting in the recruitment of the Rpd3S deacetylase complex that plays an essential role in preventing improper internal initiation (27,28). Thus both Set1 and Set2 have the potential to mediate transcriptional interference and have been implicated in gene repression by non-coding RNA transcriptional interference (see 24 for review).

Does pervasive transcription, and in particular antisense transcription, play a larger role in gene regulation? If so, the act of transcription by itself may constitute a critical step in that pathway. If not, apart from a few exceptions, pervasive transcription may only represent transcriptional noise.

Several large-scale studies attempted to answer this question. Genes with large expression variability (such as stress response and environment specific genes) often have antisense expression suggesting a general regulatory effect of antisense on gene expression (29). Others correlated antisense expression with chromatin marks, either in a wild-type context or with a *rrp6* mutant (17,25,30,31) but no global anti-correlated trend was found between asRNA and mRNA expression.

Very recently, NETseq experiments in the fission yeast pointed out the widespread existence of antisense diversity, and the observation of a global anti-correlation between sense mRNA and antisense level of transcription (32). Antisense expression is higher for poorly expressed genes, which also show a specific histone modification pattern.

To which extent asRNA transcription could act on gene regulation was examined lately by measuring, under various conditions, the effect of specific antisense SUTs transcription interruption on the expression of the corresponding proteins fused to GFP (33). This study showed that, for 12–25% of genes associated with an antisense SUT, a detectable but weak antisense-dependent gene regulation could be observed under at least one condition. Although no specific biological pathway seemed enriched in the tested asRNA responsive genes, the analysis showed that repression by asRNA transcription interference helps reducing somehow mRNA expression basal levels, especially for genes expressed at a low level, reinforcing complete gene shut off. However, the analysis was restricted to SUTs, i.e. non-coding RNAs readily detectable in wild-type cells, which are limited compared to the reality of antisense transcription in the cell as we know that SUTs represent only a minority of the pervasive transcripts, most of which are too unstable to be detected in wild-type cells (4,7,8,34).

The nuclear NNS quality control pathway prematurely terminates the transcription of many of the pervasive RNAs to prevent them from interfering with mRNA expression (9,35). However, many pervasive transcripts escape, at least in part, this first surveillance pathway and are extended up to cryptic cleavage and polyadenylation sites (polyA sites), potentially over the transcription start site (TSS) of their associated genes. This can lead to the export of long non-coding RNAs into the cytoplasm, where they are rapidly degraded by the NMD pathway (13,14).

In order to measure a relevant ‘antisense transcriptome’, we analysed genome-wide the amount of asRNAs associated to each mRNA in a NMD mutant context (*upf1Δ*). In

this mutant, hidden pervasive transcripts that escaped the nuclear NNS surveillance accumulate in the cytoplasm and can thus be quantified. An important fraction of the less expressed genes are associated with asRNAs, especially if the asRNAs overlapped the associated sense gene promoter. In addition, many of these genes with promoter-overlapping asRNAs were enriched for genes up-regulated in chromatin remodelling mutants such as *set2Δ* or *set1Δ*. Interestingly, the majority of mRNAs enriched during the stationary phase (G0) fall in the category of genes poorly expressed during the exponential phase and 30% of them are associated with antisense RNAs overlapping their promoter, a much higher proportion than overall average (9.5%). These observations strongly suggest that this particular class of genes is frequently subjected to asRNA transcription interference for full repression during exponential growth, a prediction we validated experimentally for a subset of genes.

Our study showed that antisense-mediated transcriptional interference is, in budding yeast, a mechanism more frequently used than anticipated when mRNA expression needs to be tightly repressed under specific conditions.

MATERIALS AND METHODS

Yeast strains and cultures

All strains are listed in Supplementary Table S1, are derivative of BY4741 or BY4742, and were obtained from the Euroscarf deletion collection (http://www.euroscarf.de/). A 37 nucleotides sequence constituting the NNS terminators (GTAATGAATTAAGTCTTGATATATAACA ATTAGCTTG construct 78-wt in (36)) was inserted into BY4741 or BY4742 strains using the seamless cloning-free PCR-based allele replacement methods as described in (37).

Briefly, gene-specific PCR products containing adaptamer A or adaptamer B and NNS terminator were reconstituted with two successive PCR using A-GENE primer and GENE_NNS_S/AS_rev (PCR1) and GENE_B and NNS_AS_GENE_fwd (PCR2), followed by A_GENE and GENE_B (PCR3). GENE stands for ARO10, PET10 NNS_S, SHH3 MOH1 and CLD1 (see Supplementary Table S2). In parallel Fragment L and R were obtained using primer CS1199/CS1200 and CS1201/CS1202 on a URA3 K. lactis DNA template from plasmid pBS1539, (38). All PCR were done with a high-fidelity Phusion® High-Fidelity (NEBiolabs), following the manufacturer's instructions.

For PET10_NNS antisense, a PCR product GB988/GB989 obtained using pFL38 (from http://seq.yeastgenome.org/vectordb) as a DNA template was used to transform BY4741 plated on SC-URA medium. [URA3+] clones were transformed with 100 pmol of annealed GB990/GB991 primers, plated on YPD at 30°C overnight, and replicated on 5FOA medium in order to select *URA3* popped-out constructs. All the constructs were sequence-verified.

Supplementary Figure S1 lists the position of NNS terminator insertion, in both sense and antisense orientation. Strains and oligonucleotides are listed in Supplementary Table S1 and Table S3 respectively.

Cells were grown to mid-exponential phase in YPD-rich medium at 30°C, and homogeneous populations were purified as 'Quiescent cells' (or G0), obtained from a stationary phase culture after 10 days of growth at 30°C in YPD and purification of the dense fraction on percoll gradient according (39).

RNA extraction

Total RNA from logarithmic and G0 cells were extracted with guanidium thiocyanate phenol-chloroform following (40) with the addition of 500 μl of glass beads prior to solution D addition, and vortex in a MagNA lyser (Roche) 90 s at 4800 rpm after solution D addition.

Libraries preparation

3′ Long SAGE libraries were constructed as described in (41), except than total RNA were extracted from BY4741 logarithmic and G0 cells using the guanidium thiocyanate phenol-chloroform procedure described in (40).

TruSeq stranded mRNA LT sample prep kits (Illumina) were used to prepare RNAseq libraries, on RiboZero gold (Illumina) treated RNA according the manufacturer's instruction. Single read 50 (SR50) sequencing were performed on an Illumina Hiseq 2500 (Pasteur Transcriptomic Platform PF2).

Northern blot

Northern blots were carried out on 4 μg Total RNA as described in (7) using strand specific 32P-labeled riboprobes (see Supplementary Table S3) except for *SCR1* for which a 32P-labeled oligonucleotide was used (GB987).

Strand-specific RT-qPCR

Turbo DNase-treated RNA (Ambion) from exponentially growing yeast cells was used after acid Phenol Chloroform purification as an input for reverse transcription using 2 pmol of each gene-specific primers and 1 μg RNA using 0.5 μl of SuperScript III reverse transcriptase (Invitrogen) according to the manufacturer's instructions but supplemented with 20 mg/ml actinomycin D (Thermo Fisher) to ensure strand specificity of the reverse transcription. For *PET10* and *SHH3* strand specific reverse transcription, a mix of 2 μM FF3033, AC407, AC429, AC500 and GB1038 primers was used for sense-specific, and FF3033, AC430, AC431 and AC63 primers for antisense-specific measurement (see Supplementary Table S3 for gene correspondence). For qPCR, cDNA samples and -RT controls were diluted 10 times, and 2 μl were amplified using the qPCR Mix 2X Lo-Rox (Eurobiogreen). *CPS1* mRNA was used as the reference gene as its level does not change between exponential and G0 phases.

Data analysis

Illumina reads treatments. For RNAseq libraries, duplicated reads were first filtered out using fqduplicate (ftp://ftp.pasteur.fr/pub/gensoft/projects/fqtools/fqtools-1.1.tar.gz). Then sequencing error were corrected using Musket ((42); version 1.1). Reads of

bad quality were removed using fastq_qual_trimmer (https://github.com/ivars-silamikelis/fastq_qual_trimmer, version 1.0) with a threshold of 20. Illumina adpaters were finally removed using Flexbar ((43); version 2.7). After removal of the random sequence tag, resulting reads were mapped using bowtie ((44); version 2.2.3 with the following parameters: –N 1 –p 1 |-no-unal –D 15 –R 2 –L 22 –I S,1,1.15) and a compilation of *Saccharomyces cerevisiae* genome (S288C reference sequence, Release 64 obtained from the Saccharomyces Genome Database (SGD) [http://www.yeastgenome.org/]) and *Schizosaccharomyces pombe* genome (ASM294 reference sequence, v2.19 obtained from PomBase [http://www.pombase.org/]) as reference genomes.

For 3′Long SAGE libraries, duplicated reads were first filtered out using fqduplicate. Illumina adpaters were then removed using AlienTrimmer (45). Reads corresponding the 3′ end of transcripts were identified by detection of a polyA sequence at the end of the reads with a minimal size of 6 nucleotides. After Poly A removal, the resulting reads were mapped using bowtie (same version and parameters that above) and the *S. cerevisiae* genome (S288C reference sequence, Release 64 obtained from the Saccharomyces Genome Database (SGD) [http://www.yeastgenome.org/]). False positive reads (i.e. reads identified by a ≥6nt encoded polyA sequence but not a true 3′end) were filtered out by matching with encoded PolyA sequence in the genome.

Mapped reads processing. For 3′ Long SAGE libraries the 3′-end positions of the resulting mapped reads were used as TTS positions and extracted to wig files. For RNAseq libraries, reads corresponding to the whole transcripts and full read coverage were extracted to wig files.

Normalization and differential expression. Transcript differential expressions were calculated using DESeq2 (4) within the SARTools pipeline ((46); version 1.4.1).

Sample corresponding to cells in exponential phase were first treated together as a separated group, as well as samples corresponding to cells in G0. During the process, SARTools performed a normalization step. Normalization factors were extracted and used to produce normalized wig files.

G0 samples were normalized in a second time against exponential phase sample using the spike-in of *S. pombe* transcripts. *S. pombe* transcripts median reads counts were determined for each sample after the first normalization step. Then a global mean for *S. pombe* transcripts reads counts was calculated for quiescent and exponential phase samples. A ratio Exponential/Quiescent was calculated and applied to all G0 samples (wig files and transcripts reads counts).

Heatmap counting and visualization. Antisense / mRNA coverage was counted and visualized in a −50-+200 nucleotides windows using the Counter RNAseq window (CRAW) package version 0.9.0 (https://pypi.python.org/pypi/craw/0.9.0).

RESULTS

Characterization of antisense transcription in *upf1Δ* cells

In order to reveal antisense transcription that escaped the nuclear NNS surveillance pathway, we quantified the asRNAs levels in the proximal region of the protein coding genes using a +1 (mRNA TSS) to +200 nucleotides window with a strain impaired for the cytoplasmic NMD surveillance pathway (*upf1Δ* mutant). Figure 1A shows these values (y axis) plotted against the average mRNA levels per gene (number of reads per nucleotide; x axis). Similar to previously reported data (25), a linear regression analysis did not reveal any correlation between the levels of antisense transcription and that of the corresponding mRNAs (Pearson correlation coefficient $R^2 = 0.07$). Yet, the less expressed mRNAs appeared to be generally associated with high levels of asRNAs. In order to quantify this observation, we partitioned the genes according to their mRNAs levels in ten bins with an equal numbers of genes. The less expressed genes (bin 1) had significantly higher levels of asRNAs than the genes within higher mRNA expression categories (bins 2 to 10; see Figure 1B, Dataset 1 and Supplementary Table S3).

At least two non-exclusive phenomena could explain this observation. First, gene transcription itself could have a repressive effect on asRNA transcription initiation from their corresponding gene-3′ NFRs (29). Hence, asRNAs initiating within NFRs situated downstream of non-expressed genes should be less subjected to such repression by coding gene transcription. Conversely, antisense-transcription from 3′ NFRs could be a common mean to contribute to a tight gene repression. If the former explanation is correct, asRNAs associated with non-expressed genes should not show different termination characteristics than other asRNAs. In contrast, it was shown that repression by asRNAs correlates with mRNA TSS overlap (33). If asRNAs associated with the less expressed genes contribute to their tight repression, these asRNAs should overlap the mRNA TSSs more often than other asRNAs. We thus categorized genes depending on the occurrence of their associated asRNAs across TSSs by analysing a window between −50 nucleotides to +200 nucleotides relative to the mRNA TSS. We defined three types of genes. Genes without substantial asRNAs over the +1 to +200 nucleotide region (arbitrarily set below three reads per base over this window) defined class N (No antisense). Genes with asRNAs but with an average read number below three in the −50 to −1 nucleotide region, thus terminating before the mRNA TSS, defined class M (mRNA antisense). Conversely, genes with asRNAs with an average read number above three in the −50 to −1 nucleotide region defined genes with TSS overlapping asRNAs (class O—overlapping antisense) (Figure 1C; Dataset 1). Figure 1D shows a heat map of the sense and antisense transcripts over a −200 to +200 nucleotides window around the mRNA TSSs, classified according to the three classes. Among the 5892 protein coding genes analysed, 5076 belong to class N, 259 to class M and 557 have an overlapping asRNA (class O—Figure 1E). The higher proportion of asRNAs in bin 1 mostly resulted from the over-representation of class O asRNAs, which represent 30% of bin1 (181 class

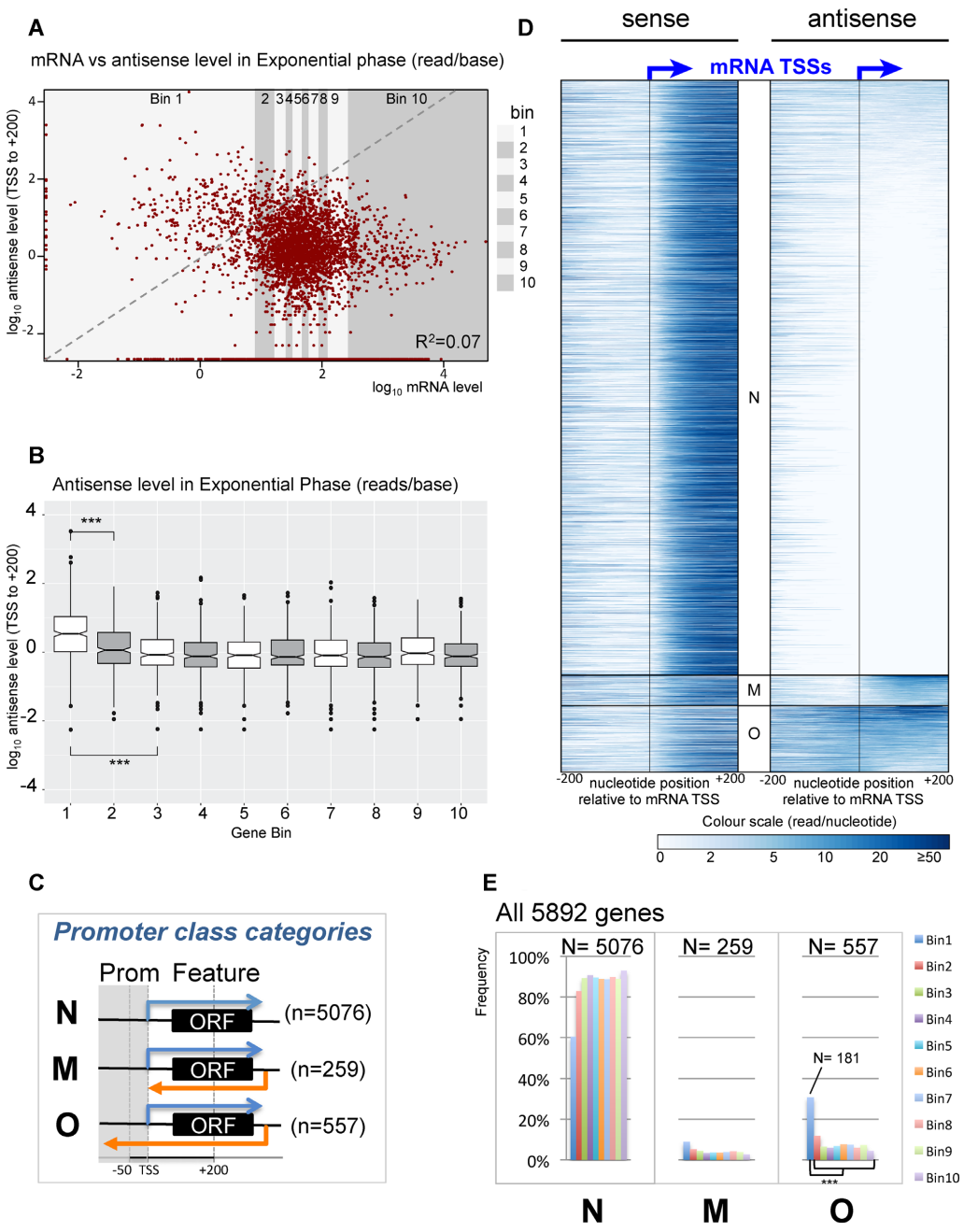


Figure 1. Antisense ncRNAs are over-represented in lowly expressed genes. (**A**) Scatter plot representing the antisense level (ordinate) function of the corresponding mRNA level (abscissa) in $\log_{10}$ read/base in an *upf1Δ* background. Average read counts per nucleotides were determined for each 5892 genes, and divided into ten bins (grey strips) of equal length ($N = 589$ genes per bin for bin1 to bin9; $N = 591$ genes for bin10). The Pearson correlation coefficient $R^2 = 0.07$. (**B**) Comparison of the average antisense level distribution between bins. Boxplots show the distribution of the average antisense levels within each bin. Brackets indicate the results of an Anova test on pairs of distributions, with $^{***}P < 0.001$. (**C**) Schematic of the gene-associated promoter class categories depending of the presence and the characteristics of asRNA: N = No asRNA, M = asRNA within the mRNA, O = TSS-overlapping asRNA. An arbitrary threshold of at least three RNA sequencing reads per nucleotide, in a +1 to +200 nucleotide window relative to the mRNA TSS position, was used to define the presence of an asRNA. (**D**) Heatmap distribution of mRNA (left) and antisense (right) around the mRNA TSS of all genes (from position −200 to +200 relative to the mRNA TSS), sorted by antisense and promoter class categories. Depending of the class of promoter defined in C, a category N, M or O was assigned to each gene. (**E**) Promoter class categories count per bin. The total number of genes that belong to each class of promoter is indicated (class N: $N = 5076$; class M: $N = 259$; class O: $N = 557$). Bar charts represent the percentage of each class within the 10 bins defined in A (see also Dataset 1). Brackets indicate the results of a statistical inference test on pairs of distributions between bin1 and each other bins, with $^{***}P < 0.001$.

O genes among the 590 genes in bin1). Class O asRNAs represented 78% of all asRNAs of bin 1 (181/233), while this proportion was only of 64.5% (376/583) within all the other bins (Figure 1E, Supplementary Figure S2A and Dataset 1). Consistent with our results, the heat map obtained from NETseq experiment (47) in a wild type strain looks similar to Figure 1D (Supplementary Figure S2B). A global analysis by computing the means of the NETseq reads in each strand and for each nucleotide between positions −200 and +200 relative to mRNA TSSs in each promoter classes (N, M and O) confirmed that, overall, the mRNAs (sense NETseq data) that belong to classes M and O are less expressed that those belonging to class N. Moreover, it also validated the overall promoter classification, with antisense RNAs associated with class O being substantially more transcribed that those associated with class N and the antisense RNAs associated with class M being mostly expressed within the ORF region, the signal becoming very weak upstream of the mRNA TSS (Supplementary Figure S2C).

It strongly suggested that the higher number of asRNAs associated genes in bin 1 relative to the others bins reflected a potential regulatory role associated with a number of these asRNAs.

Genes up-regulated in the absence of chromatin regulators are enriched in the class of poorly expressed genes with TSS-overlapping asRNAs

If antisense transcription can affect sense transcription, one should expect that genes associated with asRNAs be more up-regulated in chromatin modifier mutants implicated in transcriptional interference, in particular *set1* and *set2* mutants. Given that antisense transcriptional interference involves the extension of asRNA up to the promoter regions of the genes, Set2, which promotes H3K36me2 at late stages of PolII elongation, seemed a good candidate to mediate gene repression by asRNA transcription. *SET2* mutants are intrinsically difficult to analyse by RNAseq or tilling arrays since a major role of Set2 is to suppress both sense and antisense internal initiation within gene transcribed regions (13,27,28,48). The cryptic initiation events observed in *set2Δ* mutants in the sense orientation can thus lead to misleading quantitation due to the overall increase of sense RNAseq counts (13). We thus took advantage of the analysis of individual TSSs in the Malabat *et al.* study, which allows the quantitative analysis of the specific mRNA TSSs, irrespective of internal transcriptional initiation. We considered a gene as up-regulated upon *SET2* deletion when its strongest mRNA-linked TSS cluster was induced at least two fold with a *P*-value ≤0.05 (Supplementary file 3 in (13)). Ninety-five of 5228 genes analysed in this dataset were up-regulated in a *set2Δ strain* (see Dataset 1) Strikingly, genes with TSS-overlapping asRNAs (class O) showed the highest percentage of up regulation in a strain lacking *SET2* (9.1% compared to 1.8% for all genes; Figure 2A). Combining gene promoter classes with the mRNA expression level categories drastically increased this bias since class O of bin 1 showed the highest proportion of genes up-regulated in *set2Δ* cells (Figure 2B, right panel).

Direct measurement of transcription levels by NETseq have been analysed in a *set2Δ* mutant (47). Although a higher number of genes were found to be up-regulated in absence of Set2 in this dataset, possibly due to internal initiation events not being filtered out, the same trend was observed (Supplementary Figure S3A). This prompted us to analyse the data for the *set1Δ*, as well as *rco1Δ* and *eaf3Δ* (two components of the Rpd3S deacetylase complex) mutants from the same dataset, as these factors have also been found to be involved in transcriptional interference. Genes up-regulated upon deletion of these genes were also clearly over represented in class O (Supplementary Figure S3B–D). Altogether these results suggest that repression by antisense transcriptional interference is frequent for poorly expressed genes, a process mediated by several chromatin-modifying factors linked to elongating PolII. Supplementary Figure S3E reports the large number of up-regulated genes overlapping in the different mutant strains, which highlights the redundancy of these processes (47). Interestingly, the number of asRNA up-regulated in a *set2Δ* strain present an opposite trend than corresponding senses, and is significantly lower when associated to genes with TSS-overlapping asRNA which were shown particularly up-regulated in *set2Δ* (Supplementary Figure S3A). This suggests that asRNAs tend to be down regulated in a *set2Δ* strain when the associated mRNA is up-regulated.

Quiescence enriched genes are associated with high levels of asRNAs

We next determined if poorly expressed genes (bin 1) belong to a particular category of regulated genes. An expected category of genes strongly repressed during exponential growth are those found enriched in stationary phase and/or in quiescence (G0). We thus analysed a dataset reporting the time course of mRNA expression of a wild-type strain over a complete 10-days growth. Figure 3A shows that the stationary phase-enriched genes (SP-enriched in (49)) are the most abundant in bin 1. However in stationary phase, the cell population might not be homogeneous since it is composed of dead, senescent and quiescent cells (39,50). To circumvent this problem, we performed a genome-wide RNAseq analysis using a homogenous population of quiescent cells derived from wild-type or *upf1Δ* strains in order to analyse both gene and pervasive transcription (see Materials and Methods). To normalize the overall level of transcripts per genome, we spiked in the budding yeast cultures before RNA extraction with identical reference aliquots of a *Schizosaccharomyces pombe* exponential culture (see Materials and Methods for the normalization procedure). We defined quiescence-enriched (Q-enriched) mRNAs as being, following normalization, five times more abundant in the G0 population than the exponential growing phase (total of 261 genes, Supplementary Table S4). As anticipated, Q-enriched mRNAs were found in majority within bin 1 (163 genes in bin1 among the 261 Q-enriched genes; Figure 3B and C). Accordingly, Figure 3D shows that, as for genes within bin 1, Q-enriched mRNAs were associated with higher asRNA levels than random ($***P = 1.5\ 10^{-4}$) and their distribution in the different asRNA associated genes classes (classes N, M and O) was similar to that of bin 1 (Supplementary Figure S4A). While, upon deletion of *SET2*, 1.8% of all genes are up reg-

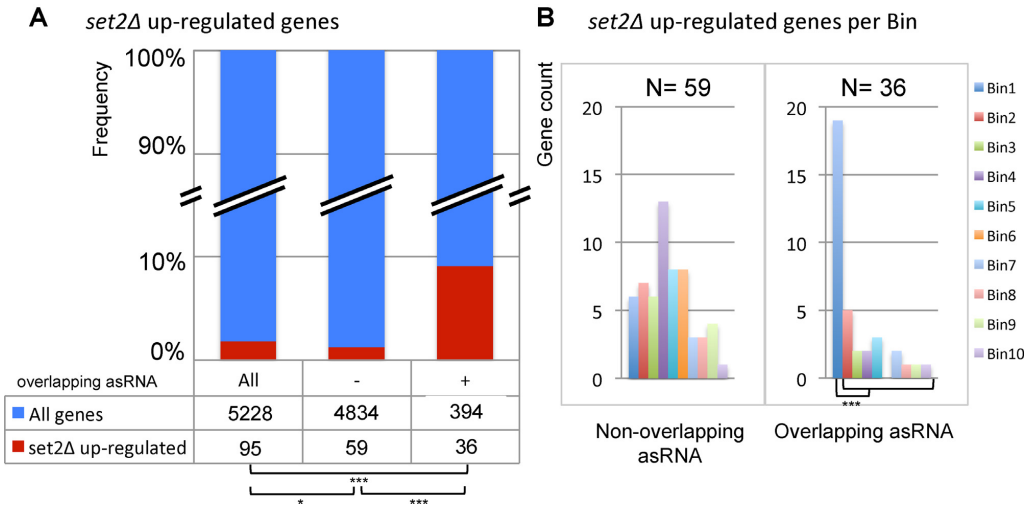


Figure 2. Promoter overlapping antisenses are overrepresented in *set2Δ* targets. (**A**) Gene distribution across promoter categories in *set2Δ* up-regulated genes (dataset from 13). Stacked histograms represent the proportion of *set2Δ* up-regulated genes across all genes (All), or depending on the presence of a TSS-overlapping asRNA ('+' = class O) or not ('–' = classes N + M). Brackets indicate the results of a statistical inference test on pairs of distributions, with $*P < 0.05$, and $***P < 0.001$. (**B**) *Set2Δ* up-regulated genes count depending the presence of a TSS-overlapping asRNA or not and per bin. Brackets indicate the results of a statistical inference test on pairs of distributions between bin1 and each other bins, with $***P < 0.001$

ulated, this fraction rises to 9.4% of all genes with TSS-overlapping asRNAs (Figure 2A) and to more than 35% when only considering Q-enriched genes (Supplementary Figure S4B). Breaking down these figures by bins and promoter classes showed that this strikingly high proportion was primarily contributed by class O genes, representing 13 out of 22 (59%) of the *set2Δ* up-regulated Q-enriched genes (Figure 3E). These observations strongly suggested that asRNA transcriptional interference could be a frequent mechanism of tight repression for this specific class of genes. In order to directly test this hypothesis, we chose for further analysis five representative examples of Q-enriched genes associated with an asRNA spanning the mRNA TSS: *PET10*, *SHH3*, *MOH1*, *CLD1* and *ARO10*. Among these genes, only *ARO10* was previously tested for asRNA mediated transcription interference (33).

Time course of quiescence-enriched mRNAs and corresponding asRNAs show an inverse expression pattern

In order to examine the relative behavior of these mRNAs in relation to their associated asRNAs, we performed Northern-blots time course experiments starting (t0') by the addition of rich medium to quiescence purified cells and using strand-specific RNA probes. The five selected Q-enriched mRNAs were not only accumulating during quiescence but were in fact strongly induced after ∼48 h of culture (Figure 4), which coincides with the post diauxic shift transition (49). The asRNAs started to accumulate between 5 and 30 min upon rich medium addition to reach a peak of expression at ∼24 h, after which they rapidly disappeared. The mRNAs followed the inverse trend with the exception of *ARO10* that was not as substantially repressed during the exponential phase. These observations are compatible with the asRNA transcription contributing to mRNA repression. Conversely, they are also compatible with induction of the mRNA repressing the associated asRNAs.

Interruption of antisense transcription results in de-repression of quiescence-enriched genes

One of the main effects of the NNS pathway is to prevent the expression of most pervasive transcription from interfering with the normal expression of genes genome-wide (9). This mechanism is thus intrinsically optimized to result in an early termination and in a strand specific manner. We choose to use it in order to specifically terminate asRNA transcription close to their transcription start by introducing in the TSS proximal region of the asRNAs a short ((37) nucleotides) optimal NNS termination signal (NNS-ter; (36)). In order to perturb as little as possible the corresponding mRNAs, we introduced this NNS-ter sequence seamlessly using a cloning-free method allowing chromosomal modifications without leaving selection markers (37). This sequence was introduced in the proximal region of the asRNAs, corresponding to the terminal region of the mRNAs (see Supplementary Figure S1 and Supplementary Table S2). The introduction of the NNS-ter signal resulted in the proper elimination of all asRNAs and in a strong up-regulation of the corresponding mRNAs, except for *ARO10* (Figure 5A). We note that *ARO10* is also, out of the five genes examined, the one that showed the weakest mRNA repression during the exponential phase (see Figure 4 and Discussion). We then verified that for NNS constructs inserted upstream the stop codon, the observed effect was not due to a NMD effect -a consequence of the ORF disruption that could insert a premature stop codon that could be recognized like a NMD substrate- but to the effect of the antisense interruption (Figure 5B lanes 'if' for 'in frame'). We also verified that the use of a scrambled, inactive version of the NNS terminator, which could not interrupt antisense

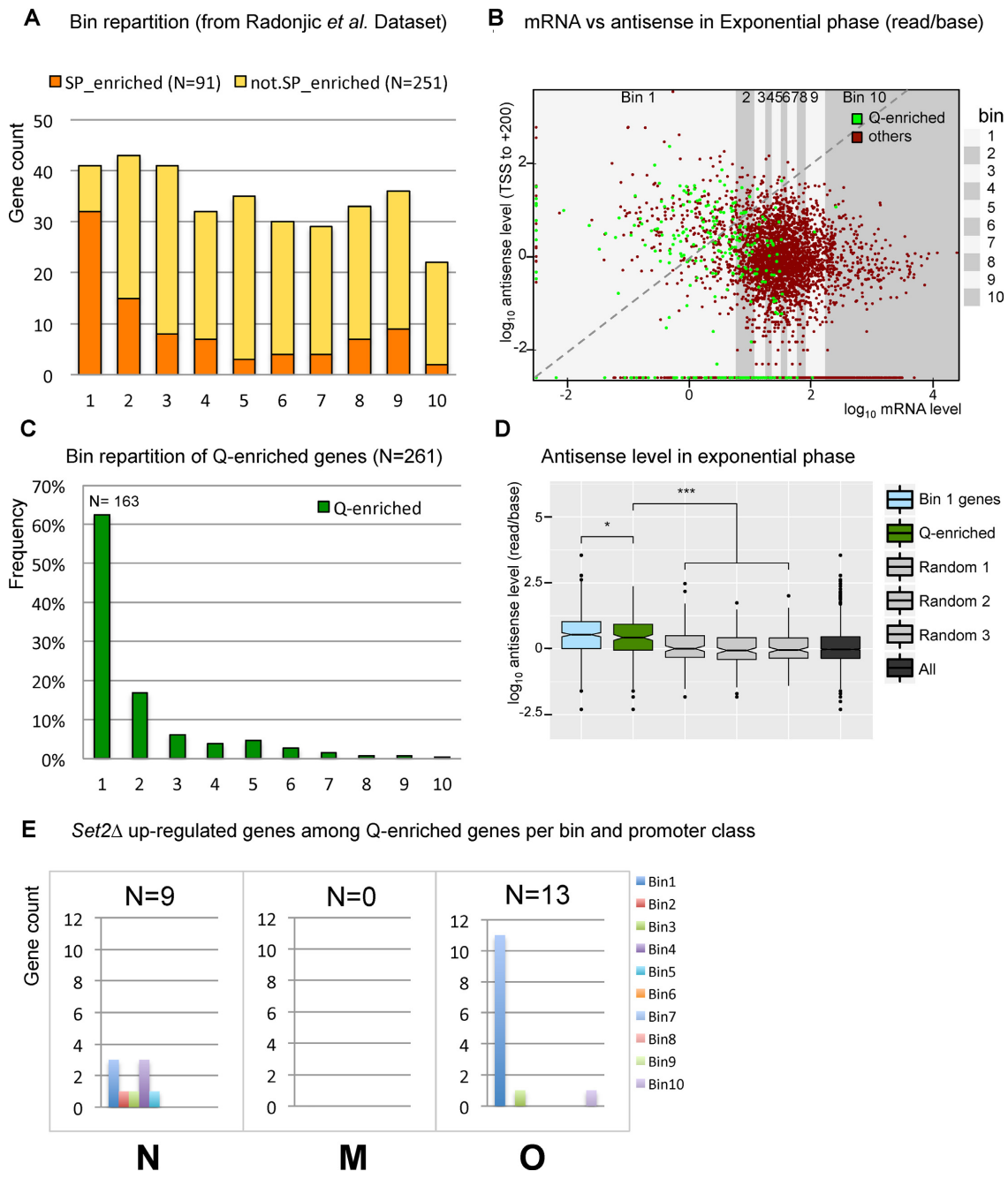


Figure 3. Quiescent-enriched genes are associated with high antisense level. (**A**) Bar plot of stationary phase-enriched genes count (SP-enriched) versus other genes count (not SP-enriched) within each bin (dataset from 49). (**B**) Distribution of quiescence-enriched genes among the 5892 yeast genes. Scatter plot of the antisense level as a function of the corresponding mRNA level. 261 genes were found enriched at least 5 times between exponential and quiescence, defining the quiescence-enriched genes (Q-enriched, green dots, see also Dataset 1 and Materials and Methods). (**C**) Bar chart of the 261 Q-enriched genes within the 10 bins. (**D**) Distributions of antisense level for different gene categories in exponential phase. Boxplots show the mean antisense level of 261 corresponding Q-enriched genes (green) or 'Random' (gray) genes. Random-1, -2 and -3 were defined by random sampling of 261 genes among all the 5892 genes. 'Bin_1' (blue) or 'All' categories (black) are the measures of all 589 genes from bin1 or all 5892 genes respectively. Brackets indicate the results of an Anova test on pairs, with $*P < 0.05$, and $***P < 0.001$. (**E**) *set2Δ* up-regulated genes count among Q-enriched genes per promoter class and bin. The bar charts represent the count of *set2Δ* up-regulated genes within each category of promoter and each bin (see also Dataset 1). The total number of genes that belong to each promoter class is indicated (N).

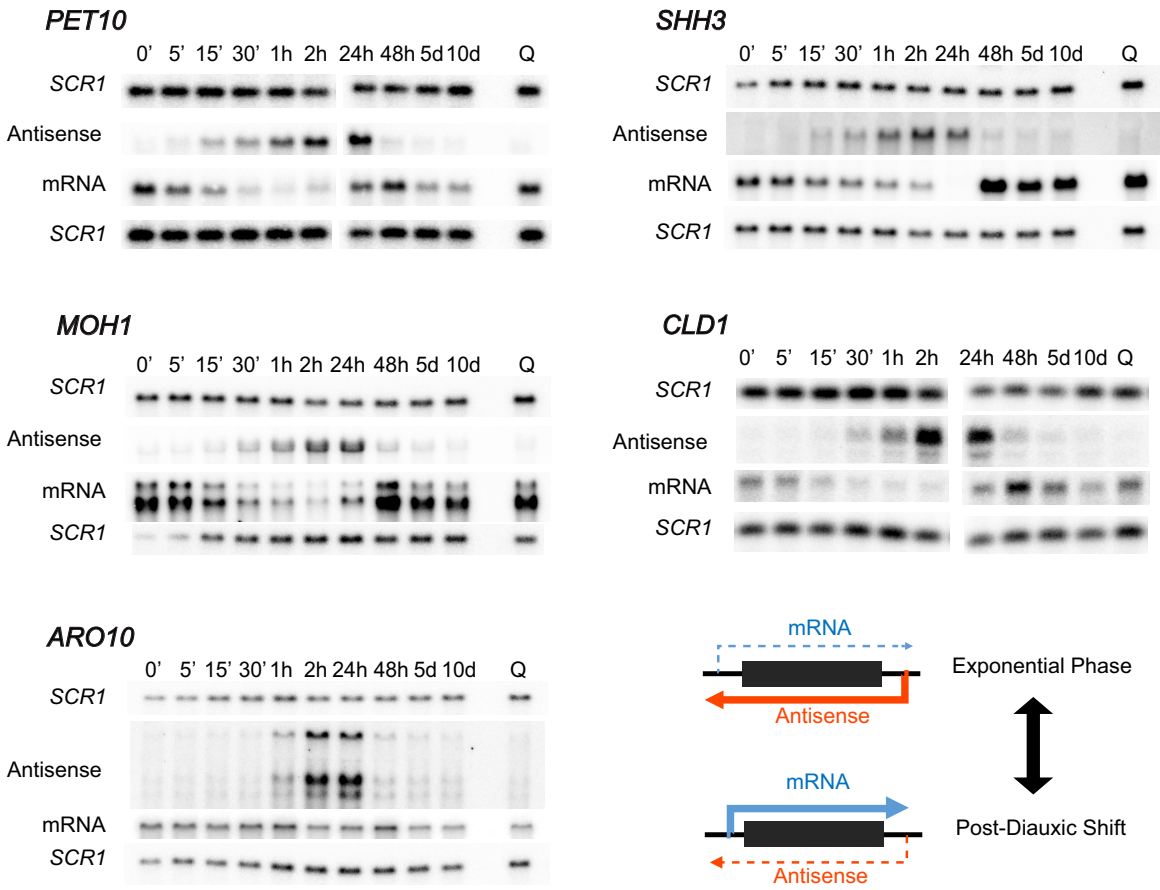


Figure 4. Quiescence-enriched genes mRNA and corresponding asRNAs are anti-regulated. Northern-Blot probing for time course mRNA and antisense transcripts in a Δ*upf1* strain for five examples of Q-enriched genes: *PET10*, *SHH3*, *MOH1*, *CLD1* and *ARO10*. Time point 0' is the time at which quiescent-arrested cells are restarted in rich YPD medium. *SCR1* is used as a loading control. RNA probes are described in Supplementary Figure S1 and Supplementary Table S2.

transcription anymore, had no effect on mRNA expression (Figure 5B lanes 'sc' for scrambled).

The asRNA associated gene repression acts in *cis*

Although the majority of non-coding RNA associated gene regulation has been shown to act only in *cis*, a *trans* effect of the asRNA itself has been invoked in a few cases (see 24 for a review). We directly addressed this question on *PET10* by comparing the mRNA and the asRNA expression in *cis* and in *trans*. To this end, we built diploid strains where *PET10* sense and antisense transcripts were disrupted on one or two of the homologous chromosomes, allowing the expression of the asRNA either from the same chromosome as the mRNA (in *cis*), from the opposite chromosome (in *trans*) or without asRNA expression (Figure 6A). RT-qPCR measurement showed that *PET10* mRNA is repressed only when its asRNA is expressed in *cis* (blue) but not in *trans* (green). In this case the mRNA level reached the same level as observed in the control strain without antisense (red). This is fully consistent with the hypothesis that the antisense transcription and not the asRNA itself, acts to repress mRNA by a transcriptional interference mechanism.

Several PolII elongation-associated chromatin modification factors cooperate to mediate antisense transcriptional interference

As described above, TSS-overlapping asRNA associated genes were more prone to be up-regulated upon deletion of chromatin modifiers such as *SET2* (Figure 2 and Supplementary Figure S3) or *SET1*, *RCO1* and *EAF3* (Supplementary Figure S3A–D) than the other categories of genes. Although the effects of *set2Δ*, *rco1Δ* and *eaf3Δ* are expected to be largely redundant as these factors act in the same chromatin modification pathway (27,28,47), we also observed that more than half (80 out of the 155) of the genes that we computed in the Churchmann dataset (47) as the most up-regulated in *set2Δ* were also up-regulated in *set1Δ* (Supplementary Figure S3E). This suggested that these different chromatin modifiers might cooperate to mediate asRNA transcriptional gene repression. The interpretation of such data are complicated by the fact that chromatin mod-

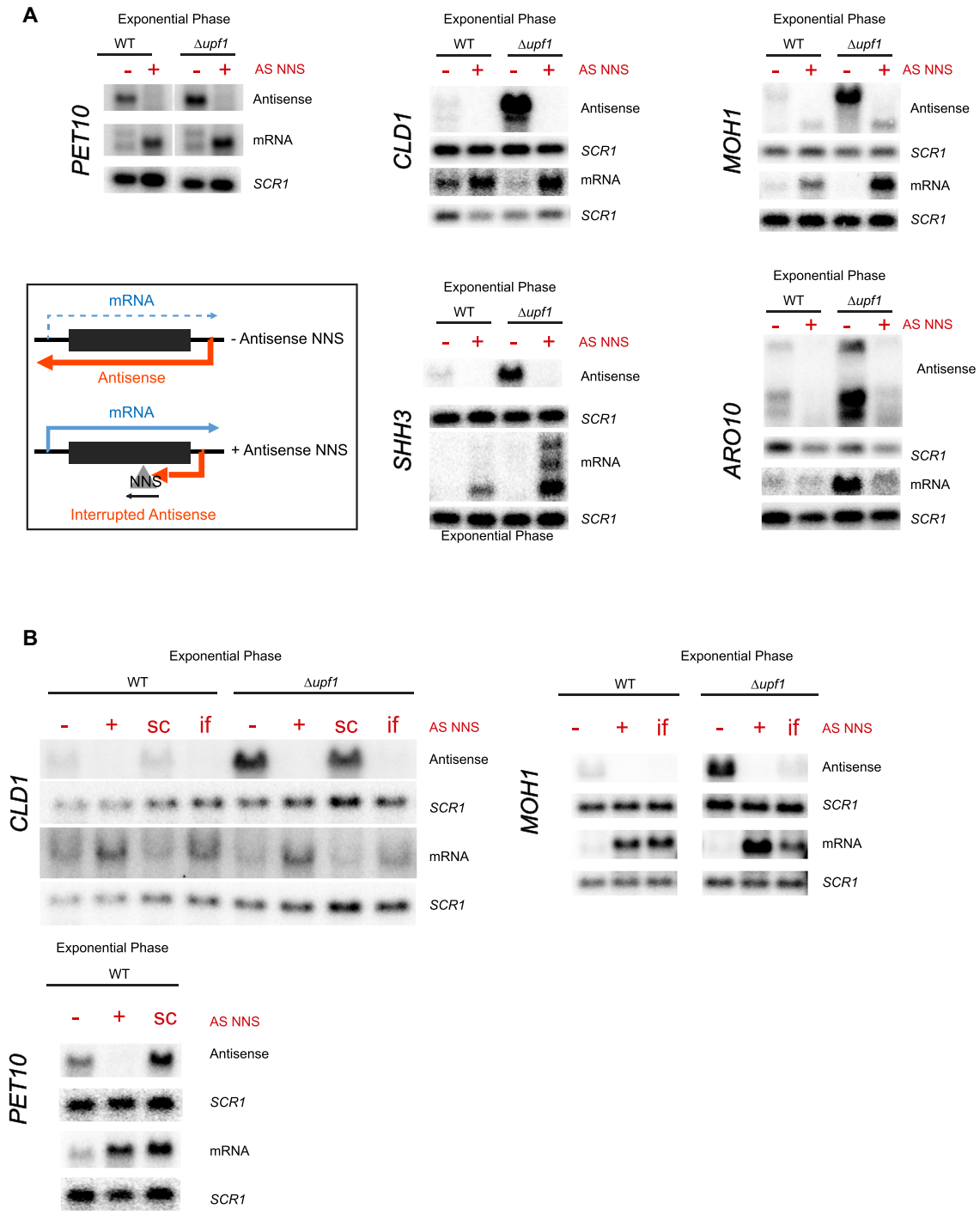


Figure 5. Antisense transcription interruption during the exponential growth relieves repression of quiescence-enriched genes. (**A**) Northern blot probing for the *PET10, CLD1, MOH1, SHH3* and *ARO10* mRNA and antisense transcripts in the WT and *Δupf1* strains with (+) or without (–) the insertion of an antisense Nrd1-Nab3-Sen1 terminator (AS NNS). *SCR1* is used as a loading control. RNA probes and NNS insertion are described in Supplementary Figure S1 and Supplementary Table S2 (see also Materials and Methods for strain construction and AS NNS-corresponding strains in Supplementary Table S1). (**B**) Northern blot probing for the *PET10, CLD1*, and *MOH1* mRNA and antisense transcripts with scrambled ('sc') NNS controls (corresponding to a scrambled NNS sequence resulting in a non functional NNS terminator for *PET10* and *CLD1*) and/or in frame ('if') NNS insertion in order to maintain the frame in the open reading frame on the opposite strand of the NNS terminator for *CLD1* and *MOH1*) (see also Materials and Methods).

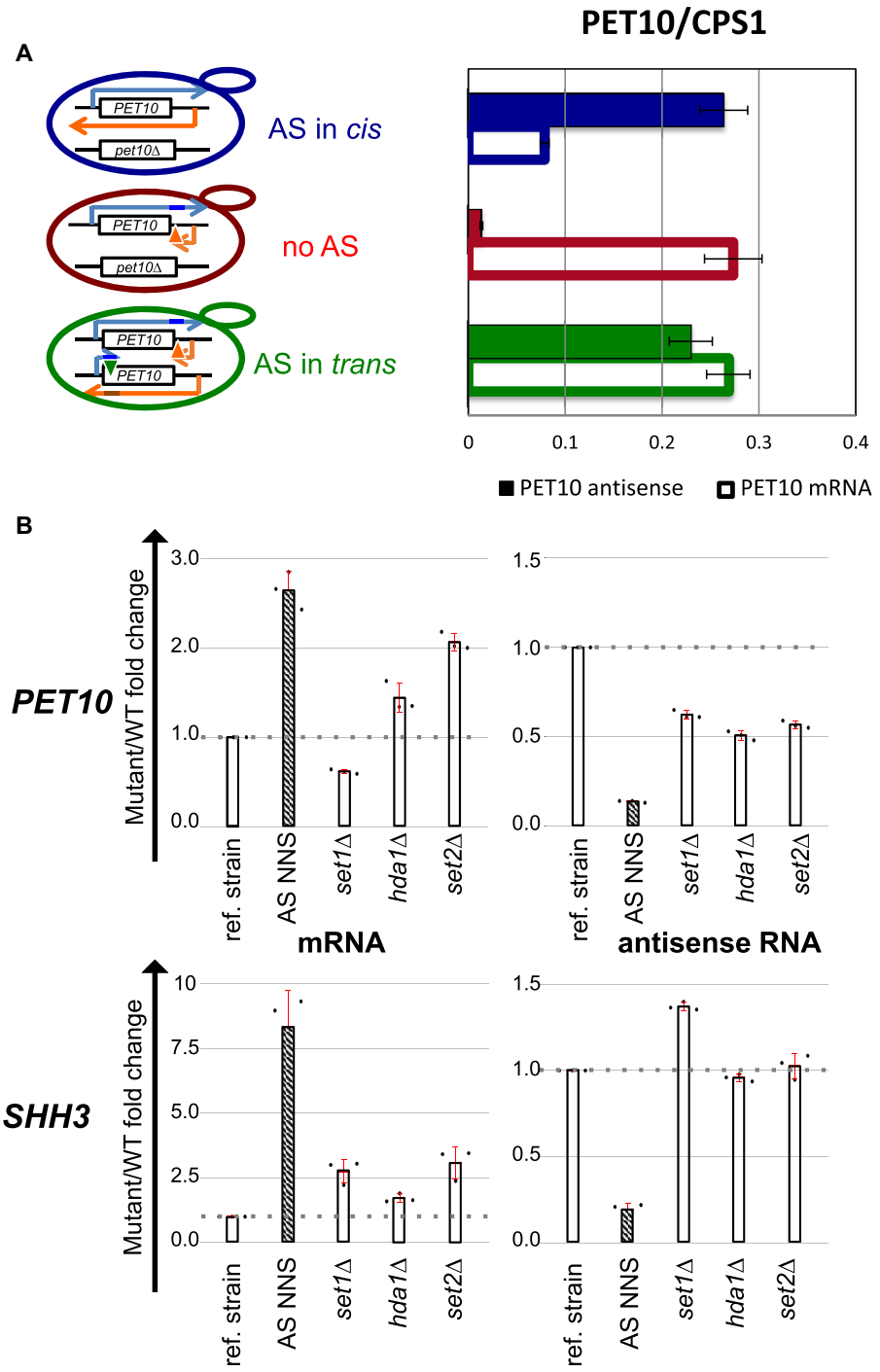


Figure 6. Antisense repression is mediated by transcriptional interference mechanisms. (**A**) Strand-specific RT-qPCR analysis of *PET10* mRNA and antisense RNA abundance in diploid strains. PET10 antisense is transcribed *in cis* (blue), in trans (green) or not produced (red). The triangles symbolise the insertion of the NNS signal, in orange for the NNS signal specific to the asRNA, in green for the mRNA. (**B**) Strand-specific RT-qPCR analysis analysis of *PET10* (upper panel) and *SHH3* (lower panel) mRNAs and antisense abundances in a mutant strain where the deletion of *UPF1* (ref. strain) is either combined to an antisense-specific NNS terminator insertion (AS NNS, dashed; positive control), or to the deletion of a chromatin modification factor (*set1Δ, hda1Δ* and *set2Δ*).

ifiers can affect both the mRNAs and their associated antisense (24,25). To address this question, we directly measured the effects of *set1Δ*, *set2Δ* and *hda1Δ* on both the asRNA and the mRNA by strand-specific RT-qPCR (see Materials and Methods). The analysis of the *PET10* locus shows a complex picture. Not only the mRNAs were positively affected in different mutants, but the levels of asRNA were also impaired in all these mutants, making the evaluation of the antisense transcription interference on the mRNA difficult. In contrast, the *SHH3* asRNA was not repressed in these deletion strains while the mRNA was significantly de-repressed in all the mutants, although not at the level of the control strain in which the asRNA transcription elongation is restricted by the NNS-terminator (Figure 6B). This suggests that several chromatin modification pathways cooperate to mediate an efficient transcriptional interference to repress gene expression.

Sense and antisense transcription can mutually repress each other

As discussed above, the fact that the category of genes associated with asRNAs was enriched in the least expressed genes (bin 1 or quiescence-enriched genes) could result from two non-mutually exclusive phenomena: the asRNA transcription represses the mRNA or the absence of mRNA expression spares pervasive asRNA from transcription interference by the sense transcription. We showed that, in four out of five genes tested, asRNA transcription interruption led to an increase of sense mRNA levels, indicating a strong repressive effect of asRNA transcription on mRNA levels. As observed in Figure 4, the expression time courses of the mRNAs and asRNAs present inversed expression patterns, which is compatible with the mutual repression of sense and antisense transcription.

Analysis of chromatin modification mutants in the NET-seq dataset (47) showed that, in the absence of chromatin modifiers, the asRNAs associated with class M or class O genes behaved markedly differently. While the majority of the asRNAs did not vary in the mutants relative to wild-type, class M asRNA were more often up regulated in the mutants than class O asRNAs (Supplementary Figure S5A). One possible explanation for this observation could be that class O asRNAs are more enriched in the category of the less expressed mRNAs (bin 1; Figure 1E), thus less susceptible to be subjected to transcriptional repression by their cognate mRNAs and thus less susceptible to be de-repressed in the absence of chromatin modifiers. If this hypothesis is correct, the tendencies of asRNAs in bins of low mRNA expression versus high expression should show opposite trends, irrespective of whether they belong to class M or class O. Supplementary Figure S5B shows that, indeed, in the *set2Δ*, *rco1Δ* and *eaf3Δ* mutants, asRNAs associated with the high expressed genes are more often up regulated in the mutants that the ones associated with the less expressed genes. Conversely, asRNAs associated with the less expressed gene are more often down regulated, which could reflect the up regulation of their associated mRNA. The fact that these trends are not observed in the *set1Δ* mutant could possibly reflect the fact that this factor mainly acts early during transcription elongation, making it less susceptible to affect the promoter of their associated antisense transcripts (see 24,26 for review).

These observations strongly suggested that, as anticipated, not only asRNAs transcription is able to repress mRNA expression but, conversely, mRNA expression has the potential to repress asRNA transcription. We wanted to directly assess this prediction by using, on the same subset of genes, the same experimental strategy as in Figure 5, but now specifically restricting transcription of the mRNAs by introducing an NNS terminator at the beginning of *PET10*, *SHH3*, *ARO10* and *MOH1*. Figure 7A shows that in all cases but for *MOH1*, restricting mRNA transcription elongation led to an up-regulation of the corresponding asRNA at 48 h (post-diauxic shift). The absence of effect observed for *MOH1* can be explained by the fact that its polyA site is the only one (out of the four genes analysed) located upstream of its associated asRNA TSS (see Supplementary Figure S1).

The mutual repression of sense and antisense transcription can thus be observed but could be dependent on *locus*-specific architecture.

As shown above, the mRNAs and their associated asRNAs exhibited an inverse pattern of expression between the exponential phase and quiescence, as expected if the expression of sense and antisense were mutually exclusive. Q-enriched genes were associated with more antisense transcription than average during the exponential phase, i.e. when these genes are repressed. Conversely, one could expect that they would be less associated with asRNA than average in quiescence since they are the genes whose mRNAs are most abundant under this condition. This turned out not to be the case. Indeed, the quiescence-enriched genes remained associated with slightly higher asRNA levels than average even during quiescence (Figure 7B). This is consistent with the observation that there is no obligatory repression of asRNA transcription when sense transcription is induced and with the observation that the quiescence-enriched genes are, overall, more associated with asRNAs than other genes.

The analysis of sense and antisense expression in quiescence also revealed that if the mRNA levels are strongly decreased in quiescence, as expected, the global level of asRNAs did not change markedly (Supplementary Figure S6). A likely explanation is that mRNA transcription interferes with pervasive asRNA transcription during exponential phase. The global repression of transcription in quiescence (51) could then be compensated for the asRNAs reduced interference from mRNA transcription. This was verified at the *HIS1* locus where the strong asRNA observed only in quiescence (Figure 8A) could be revealed during the exponential phase by interrupting the *HIS1* gene transcription by a strand specific NNS terminator insertion (Figure 8B).

DISCUSSION

Targeted studies have previously described a ten or so of specific examples, in which the transcription of a non-coding RNA was mediating gene regulation ((24), see for reviews (52)). To what extent antisense-mediated transcription interference affects gene expression genome-wide re-

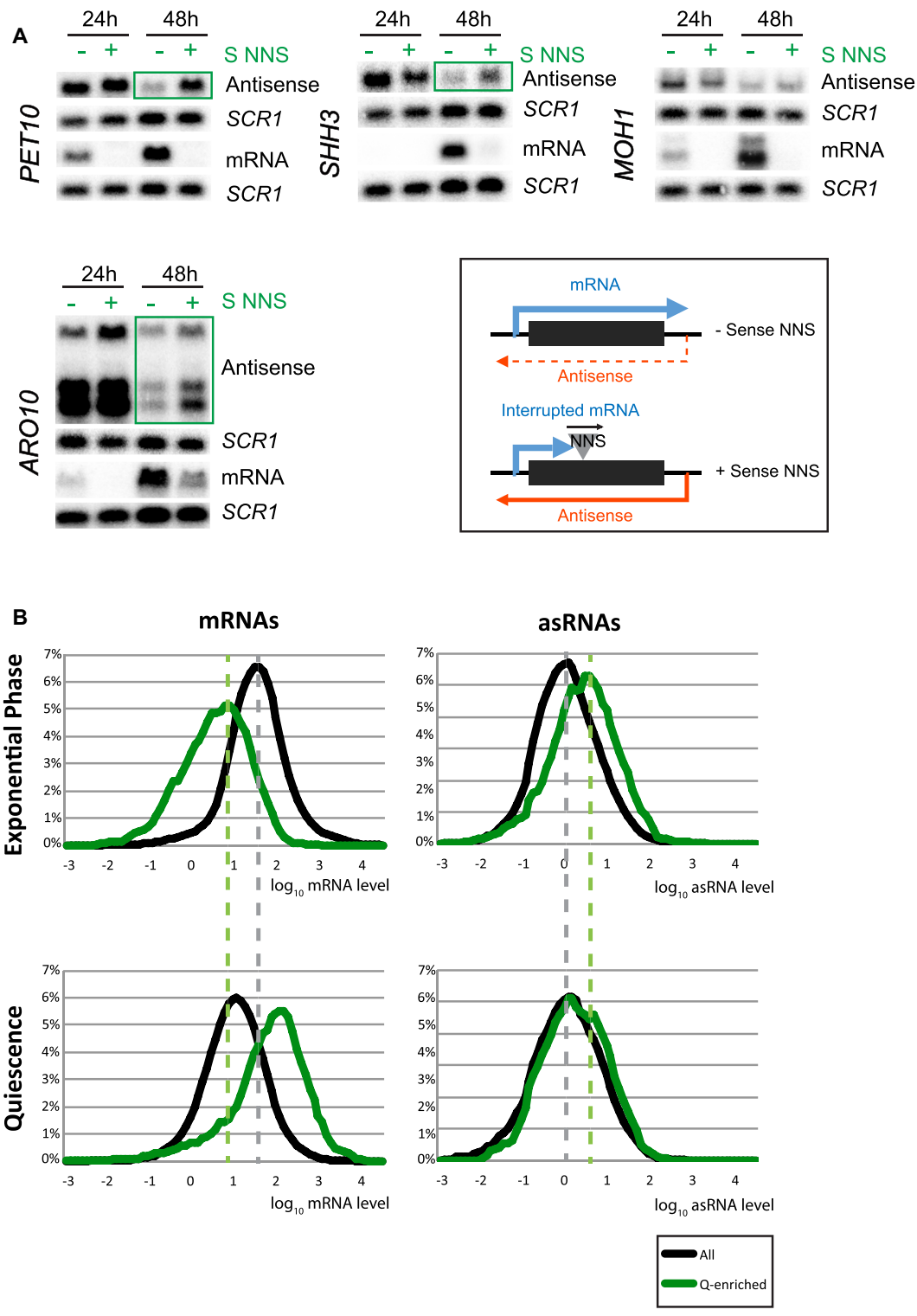


Figure 7. Gene expression is repressive for antisense non-coding transcription. (**A**) Northern blot analysis of *PET10*, *MOH1*, *SHH3* and *ARO10* mRNA and antisense RNAs in *Δupf1* strain, after 24 h or 48 h of growth in YPD and with (+) or without (–) the insertion of a sense Nrd1-Nab3-Sen1 terminator (NNS S). RNA probes and NNS insertion are described in Supplementary Figure S4 and Supplementary Table S2 (see also Material and Methods for strain construction) and NNS S -corresponding strains in Supplementary Table S1. *SCR1* is used as a loading control. (**B**) Comparison of density plots between all (black lines) and Q-enriched genes (green lines) for mRNAs (left panels) or associated asRNA (right panels) from cultures harvested in exponential (upper panels) or G0 (lower panels) phases. Log_{10} RNA levels are plotted (abscissa) function of the frequency (ordinate).

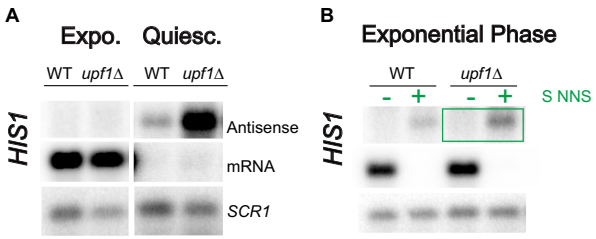


Figure 8. *HIS1* associated asRNA is induced in quiescence or when HIS1 mRNA transcription is interrupted. (A, B) Northern blot analysis of *HIS1* mRNA and antisense RNAs in WT and *Δupf1* strains in exponential phase or quiescence (**A**) or in exponential phase with (+) or without (–) the insertion of a sense Nrd1-Nab3-Sen1 terminator (NNS S; **B**). RNA probes and NNS insertion are described in Supplementary Figure S1 and Supplementary Table S2 (see also Materials and Methods for strain construction) and NNS S-corresponding strains in Supplementary Table S1. *SCR1* is used as a loading control.

mains poorly defined. A recent large-scale approach (33), which was used to address this question, only focused on genes associated with asRNAs sufficiently stable to be readily detected in wild type cells (SUTs; (4)). It showed that antisense transcription weakly affected the expression of only 12–25% of the SUTs associated genes and no particular class of genes was found to be specifically affected. Here, we addressed the question from a different angle by searching classes of genes frequently presenting characteristics associated with asRNA transcription interference.

In order to identify genes potentially repressed by asRNA transcription interference, we analysed the transcriptome of NMD deficient cells (*upf1Δ*) since abrogating NMD reveals non-coding RNAs normally efficiently degraded by this quality control pathway (13). Using a relatively stringent threshold for antisense detection (see Figure 1C), we defined 816 asRNAs. The less expressed genes were more often associated with asRNA than average and, most interestingly, this bias essentially resulted from a higher number of TSS overlapping asRNAs, reaching 30% of the genes in the bin corresponding to the least expressed genes (bin 1; Figure 1E, Supplementary Figure S2A and Dataset 1). This strongly suggested that antisense mediated transcription interference could contribute to the repression of up to 30% of these least expressed genes (bin 1). Remarkably, genes whose mRNAs were enriched in quiescence relative to exponential growth are mostly found in bin 1 and behaved similarly (Figure 3). It thus defined a family of genes potentially associated with frequent asRNA transcription mediated repression. This hypothesis was strengthened by the observation that genes associated with TSS overlapping asRNAs in bin 1 were also subjected to a regulation by chromatin modification factors much more often than average (Figure 2 and Supplementary Figure S3) and this was particularly true for quiescence-enriched genes (Figure 3E and Supplementary Figure S4B). This was especially noteworthy since asRNA transcription mediated regulation was previously found to affect single genes that belong to diverse genes families.

To test the hypothesis that the full repression of quiescence-enriched genes during the exponential phase often relies on interference by antisense transcription, we directly analysed five of these Q-enriched genes associated with TSS-overlapping asRNA (*PET10, CLD1, MOH1, SHH3* and *ARO10*). For four out of these five genes, specifically interrupting asRNA transcription resulted in a strong induction of the corresponding mRNAs during the exponential phase (Figure 5). Interestingly, the only gene that did not respond was *ARO10* but it was also the least repressed gene in our conditions during the exponential phase (Figure 4). Interestingly, this was the only gene we analysed that was also analysed in the Huber study (33). Consistently with our observations, although they could not find a repressive effect of its associated asRNA in rich medium, they found it to be regulated by antisense transcription when the cells were grown in synthetic complete medium. It thus turns out that the TSS-overlapping asRNAs associated to all five Q-enriched genes we tested can have a repressive role on gene transcription. Some of the mRNAs, such as *SHH3* and *MOH1* revealed upon restricting asRNA transcription, exhibit a higher signal in the *upf1Δ* background. In the absence of substantial expression of these genes during the exponential phase, their TSSs are not robustly defined but they could have multiple TSSs (13), some of which being upstream of potential uORFs, which could explain mRNA stabilization in absence of NMD. The *ARO10* mRNA appears stabilised in the upf1Δ mutant. This mRNA exhibited a slight stabilization in the absence of NMD in previous studies (see for example 13). This effect is less pronounced in the presence of the AS NNS. This might reflect a destabilization of this specific mRNA upon insertion of the NNS terminator sequence within its ORF.

If, in a few instance, the asRNA itself was suggested to play a direct role in gene repression, in the majority of cases examined thus far this repressive effect was shown to be mediated in *cis* by antisense transcription interference, the asRNA being only a by-product of this process (see for review (52)). Using strand specific NNS terminators in diploid strains, we directly showed, on the *PET10* locus, that the effects we observed act only in *cis*, which confirmed that transcriptional interference is likely the mechanism at play in these examples (Figure 6A).

Chromatin modifiers are though to be key players of transcriptional interference (24). Considering the high redundancy of chromatin modifiers, we can extrapolate that the number of genes submitted to a regulation by them is underestimated (Supplementary Figure S3E). We could effectively measure this redundancy in two examples (*PET10* and *SHH3*). The single deletion of each factor we tested couldn't reach the complete de-repression that was observed when the asRNA was interrupted (Figure 6B). Taken together, these results strongly suggest that the observed asRNA-transcription mediated repression involves several redundant chromatin modification/remodelling pathways. This is reminiscent of previous observations showing that gene silencing is mediated by redundant mechanisms involving multiple histone modifiers (53).

Interestingly, we found the asRNA repression upon induction of the mRNA to be frequent, although not obligatory and depending on the fact that the induced mRNA transcription overlaps the asRNA TSS (Figure 7 and Supplementary Figure S1). Overall, the asRNA levels remained high in quiescence, even slightly higher than average (Figure 7B, right panels). It suggests a model by which, in con-

trast to previously studied examples (see for examples 18, 19), the asRNA expression is not regulated by specific transcription regulators. Rather these RNAs would be constitutively expressed, unless repressed by sense transcription when mRNAs are induced and overlap their TSSs. Their transcription would thus act 'passively' as an amplifier of gene regulation, turning an non-induction into repression, as previously suggested for the SUR7 gene (29). Consistent with this model, blocking asRNA transcription elongation during the exponential phase resulted in a 2.6 and 8.3 fold increase of *PET10* and *SHH3* mRNAs respectively (Figure 6B), which is markedly lower than the induction estimated by comparing the increase of their relative expression levels measured from the quiescence versus exponential phase transcriptome datasets (5.9 and 264.6 fold increase respectively in the *upf1Δ* background; Dataset 1).

In our study, we demonstrated that TSS-overlaping antisense-mediated transcriptional interference is a frequent mechanism used for full gene repression. This mechanism is often hidden since these antisense transcripts are rapidly degraded by the NMD pathway and therefore not detected in wild type conditions.

Making more complex the overall picture, we and others reported the existence of conditional asRNAs, such as for example the *HIS1* antisense RNA specifically expressed during G0 (Figure 8A), or Meiotic Unannotated transcripts (MUTs; (54)). In addition, asRNAs were shown to mediate protein expression regulation depending on various growth conditions (33). Widespread antisense transcription has thus the potential to repress the synthesis of sense RNA and participate to differential gene expression and adaptation to various environmental and growth conditions.

Interestingly, the presence of a PET10 asRNA with an inversed expression profile compared to the mRNA was shown to be conserved in all five analysed *Saccharomyces* species, supporting its functional role (55). More generally, a phylogenetic conservation study of lncRNAs in budding yeasts has shown that, since the divergence with *N. castellii*, which has retained a functional RNAi machinery, the level of asRNAs and their extent has globally increased. Accordingly, this suggested that the lack of RNAi favored the development of asRNA transcription mediated gene regulation (56). Remarkably, a recent study in *S. pombe* showed the existence of numerous antisense transcription genome wide, despite the presence of the RNAi machinery in *S. pombe* (32). Nevertheless, in this study, small transcriptome analysis couldn't detect dsRNA issued from antisense RNAs, suggesting that asRNAs and RNAi coexistence was possible in this organism without deleterious effect. A conceivable hypothesis to explain this coexistence is that these asRNAs are mainly cryptic and efficiently cleaned up by nuclear and cytoplasmic surveillance pathways in the cell before they have the opportunity to accumulate and form detectable dsRNA.

Revealing the importance of antisense pervasive transcription and its interplay with gene expression, our study highlighted for the first time the importance of antisense-mediated transcriptional interference and the mutual repression on gene versus antisense transcription depending on growth conditions. We could estimate that this mechanism concerned up to 30% of the least expressed genes and resulted in a strong and efficient gene repression.

DATA AVAILABILITY

The data reported here have been deposited in NCBI GEO under the accession number GSE101368. Genomic data reported here have been deposited in NCBI GEO under the accession number GSE101368 and are accessible at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE101368. It will remain private until publication yet accessible using the following secure token that can be transmitted to the reviewers: svgvmsoorvcjrqz

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Frank Feuerbach for providing strains LMA2811 and LMA2819. We thank Bernard Turcotte, Cosmin Saveanu and Micheline Fromont-Racine for discussions and critical reading of the manuscript. We acknowledge Jean Yves Coppee and Caroline Proux for the facilities and expertise of the Transcriptomic Platform (PF2) for cDNA libraries sequencing.

FUNDING

Pasteur Institute, the Centre National de la Recherche Scientifique and the Agence Nationale pour la Recherche [ANR-14-CE-10-0014-01, 2014]; A.N. received Fellowships from the French Ministry of Research and from the Fondation pour la Recherche Médicale [FDT20160435375, 2016]. Funding for open access charge: Agence Nationale pour la Recherche [ANR-14-CE-10-0014-01, 2014].
Conflict of interest statement. None declared.

REFERENCES

1. Carninci,P. and Hayashizaki,Y. (2007) Noncoding RNA transcription beyond annotated genes. *Curr. Opin. Genet. Dev.*, **17**, 139–144.
2. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.
3. David,L., Huber,W., Granovskaia,M., Toedling,J., Palm,C.J., Bofkin,L., Jones,T., Davis,R.W. and Steinmetz,L.M. (2006) A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 5320–5325.
4. Xu,Z., Wei,W., Gagneur,J., Perocchi,F., Clauder-Münster,S., Camblong,J., Guffanti,E., Stutz,F., Huber,W. and Steinmetz,L.M. (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature*, **457**, 1033–1037.
5. Jensen,T.H., Jacquier,A. and Libri,D. (2013) Dealing with pervasive transcription. *Mol. Cell*, **52**, 473–484.
6. Wyers,F., Rougemaille,M., Badis,G., Rousselle,J.-C., Dufour,M.-E., Boulay,J., Régnault,B., Devaux,F., Namane,A., Séraphin,B. *et al.* (2005) Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell*, **121**, 725–737.
7. Neil,H., Malabat,C., d' Aubenton-Carafa,Y., Xu,Z., Steinmetz,L.M. and Jacquier,A. (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, **457**, 1038–1042.

8. Van Dijk,E.L., Chen,C.L., d' Aubenton-Carafa,Y., Gourvennec,S., Kwapisz,M., Roche,V., Bertrand,C., Silvain,M., Legoix-Né,P., Loeillet,S. *et al.* (2011) XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature*, **475**, 114–117.
9. Schulz,D., Schwalb,B., Kiesel,A., Baejen,C., Torkler,P., Gagneur,J., Soeding,J. and Cramer,P. (2013) Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell*, **155**, 1075–1087.
10. Arigo,J.T., Eyler,D.E., Carroll,K.L. and Corden,J.L. (2006) Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. *Mol. Cell*, **23**, 841–851.
11. Thiebaut,M., Kisseleva-Romanova,E., Rougemaille,M., Boulay,J. and Libri,D. (2006) Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. *Mol. Cell*, **23**, 853–864.
12. Tudek,A., Porrua,O., Kabzinski,T., Lidschreiber,M., Kubicek,K., Fortova,A., Lacroute,F., Vanacova,S., Cramer,P., Stefl,R. *et al.* (2014) Molecular basis for coordinating transcription termination with noncoding RNA degradation. *Mol. Cell*, **55**, 467–481.
13. Malabat,C., Feuerbach,F., Ma,L., Saveanu,C. and Jacquier,A. (2015) Quality control of transcription start site selection by nonsense-mediated-mRNA decay. *eLife*, **4**, e06722.
14. Wery,M., Descrimes,M., Vogt,N., Dallongeville,A.-S., Gautheret,D. and Morillon,A. (2016) Nonsense-Mediated decay restricts LncRNA levels in yeast unless blocked by Double-Stranded RNA structure. *Mol. Cell*, **61**, 379–392.
15. Jiang,C. and Pugh,B.F. (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.*, **10**, 161–172.
16. Murray,S.C., Serra Barros,A., Brown,D.A., Dudek,P., Ayling,J. and Mellor,J. (2012) A pre-initiation complex at the 3′-end of genes drives antisense transcription independent of divergent sense transcription. *Nucleic Acids Res.*, **40**, 2432–2444.
17. Castelnuovo,M., Zaugg,J.B., Guffanti,E., Maffioletti,A., Camblong,J., Xu,Z., Clauder-Münster,S., Steinmetz,L.M., Luscombe,N.M. and Stutz,F. (2014) Role of histone modifications and early termination in pervasive transcription and antisense-mediated gene silencing in yeast. *Nucleic Acids Res.*, **42**, 4348–4362.
18. Martens,J.A., Laprade,L. and Winston,F. (2004) Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature*, **429**, 571–574.
19. Van Werven,F.J., Neuert,G., Hendrick,N., Lardenois,A., Buratowski,S., van Oudenaarden,A., Primig,M. and Amon,A. (2012) Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast. *Cell*, **150**, 1170–1181.
20. Houseley,J., Rubbi,L., Grunstein,M., Tollervey,D. and Vogelauer,M. (2008) A ncRNA modulates histone modification and mRNA induction in the yeast GAL gene cluster. *Mol. Cell*, **32**, 685–695.
21. Pinskaya,M., Gourvennec,S. and Morillon,A. (2009) H3 lysine 4 di- and tri-methylation deposited by cryptic transcription attenuates promoter activation. *EMBO J.*, **28**, 1697–1707.
22. Castelnuovo,M., Rahman,S., Guffanti,E., Infantino,V., Stutz,F. and Zenklusen,D. (2013) Bimodal expression of PHO84 is modulated by early termination of antisense transcription. *Nat. Struct. Mol. Biol.*, **20**, 851–858.
23. Nadal-Ribelles,M., Solé,C., Xu,Z., Steinmetz,L.M., de Nadal,E. and Posas,F. (2014) Control of Cdc28 CDK1 by a stress-induced lncRNA. *Mol. Cell*, **53**, 549–561.
24. Castelnuovo,M. and Stutz,F. (2015) Role of chromatin, environmental changes and single cell heterogeneity in non-coding transcription and gene regulation. *Curr. Opin. Cell Biol.*, **34**, 16–22.
25. Murray,S.C., Haenni,S., Howe,F.S., Fischl,H., Chocian,K., Nair,A. and Mellor,J. (2015) Sense and antisense transcription are associated with distinct chromatin architectures across genes. *Nucleic Acids Res.*, **43**, 7823–7837.
26. Venters,B.J. and Pugh,B.F. (2009) How eukaryotic genes are transcribed. *Crit. Rev. Biochem. Mol. Biol.*, **44**, 117–141.
27. Carrozza,M.J., Li,B., Florens,L., Suganuma,T., Swanson,S.K., Lee,K.K., Shia,W.-J., Anderson,S., Yates,J., Washburn,M.P. *et al.* (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, **123**, 581–592.
28. Keogh,M.-C., Kurdistani,S.K., Morris,S.A., Ahn,S.H., Podolny,V., Collins,S.R., Schuldiner,M., Chin,K., Punna,T., Thompson,N.J. *et al.* (2005) Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell*, **123**, 593–605.
29. Xu,Z., Wei,W., Gagneur,J., Clauder-Münster,S., Smolik,M., Huber,W. and Steinmetz,L.M. (2011) Antisense expression increases gene expression variability and locus interdependency. *Mol. Syst. Biol.*, **7**, 468.
30. Gu,M., Naiyachit,Y., Wood,T.J. and Millar,C.B. (2015) H2A.Z marks antisense promoters and has positive effects on antisense transcript levels in budding yeast. *BMC Genomics*, **16**, 99.
31. Kim,J.H., Lee,B.B., Oh,Y.M., Zhu,C., Steinmetz,L.M., Lee,Y., Kim,W.K., Lee,S.B., Buratowski,S. and Kim,T. (2016) Modulation of mRNA and lncRNA expression dynamics by the Set2-Rpd3S pathway. *Nat. Commun.*, **7**, 13534.
32. Wery,M., Gautier,C., Descrimes,M., Yoda,M., Vennin-Rendos,H., Migeot,V., Gautheret,D., Hermand,D. and Morillon,A. (2018) Native elongating transcript sequencing reveals global anti-correlation between sense and antisense nascent transcription in fission yeast. *RNA N. Y.*, **24**, 196–208.
33. Huber,F., Bunina,D., Gupta,I., Khmelinskii,A., Meurer,M., Theer,P., Steinmetz,L.M. and Knop,M. (2016) Protein abundance control by Non-coding antisense transcription. *Cell Rep.*, **15**, 2625–2636.
34. Geisler,S., Lojek,L., Khalil,A.M., Baker,K.E. and Coller,J. (2012) Decapping of long noncoding RNAs regulates inducible genes. *Mol. Cell*, **45**, 279–291.
35. Vasiljeva,L., Kim,M., Mutschler,H., Buratowski,S. and Meinhart,A. (2008) The Nrd1–Nab3–Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nat. Struct. Mol. Biol.*, **15**, 795–804.
36. Porrua,O., Hobor,F., Boulay,J., Kubicek,K., D'Aubenton-Carafa,Y., Gudipati,R.K., Stefl,R. and Libri,D. (2012) In vivo SELEX reveals novel sequence and structural determinants of Nrd1-Nab3-Sen1-dependent transcription termination. *EMBO J.*, **31**, 3935–3948.
37. Erdeniz,N., Mortensen,U.H. and Rothstein,R. (1997) Cloning-free PCR-based allele replacement methods. *Genome Res.*, **7**, 1174–1183.
38. Rigaut,G., Shevchenko,A., Rutz,B., Wilm,M., Mann,M. and Séraphin,B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.*, **17**, 1030–1032.
39. Allen,C., Büttner,S., Aragon,A.D., Thomas,J.A., Meirelles,O., Jaetao,J.E., Benn,D., Ruby,S.W., Veenhuis,M., Madeo,F. *et al.* (2006) Isolation of quiescent and nonquiescent cells from yeast stationary-phase cultures. *J. Cell Biol.*, **174**, 89–100.
40. Chomczynski,P. and Sacchi,N. (2006) The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. *Nat. Protoc.*, **1**, 581–585.
41. Neil,H. and Jacquier,A. (2011) Enrichment of unstable non-coding RNAs and their genome-wide identification. *Methods Mol. Biol. Clifton NJ*, **759**, 87–106.
42. Liu,Y., Schröder,J. and Schmidt,B. (2013) Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinforma. Oxf. Engl.*, **29**, 308–315.
43. Dodt,M., Roehr,J.T., Ahmed,R. and Dieterich,C. (2012) FLEXBAR-Flexible barcode and adapter processing for Next-Generation sequencing platforms. *Biology*, **1**, 895–905.
44. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
45. Criscuolo,A. and Brisse,S. (2013) AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics*, **102**, 500–506.
46. Varet,H., Brillet-Guéguen,L., Coppée,J.-Y. and Dillies,M.-A. (2016) SARTools: A DESeq2- and EdgeR-Based R pipeline for comprehensive differential analysis of RNA-Seq data. *PLoS One*, **11**, e0157022.
47. Churchman,L.S. and Weissman,J.S. (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, **469**, 368–373.
48. Venkatesh,S., Smolle,M., Li,H., Gogol,M.M., Saint,M., Kumar,S., Natarajan,K. and Workman,J.L. (2012) Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes. *Nature*, **489**, 452–455.

49. Radonjic,M., Andrau,J.-C., Lijnzaad,P., Kemmeren,P., Kockelkorn,T.T.J.P., van Leenen,D., van Berkum,N.L. and Holstege,F.C.P. (2005) Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon S. cerevisiae stationary phase exit. *Mol. Cell*, **18**, 171–183.
50. Aragon,A.D., Rodriguez,A.L., Meirelles,O., Roy,S., Davidson,G.S., Tapia,P.H., Allen,C., Joe,R., Benn,D. and Werner-Washburne,M. (2008) Characterization of differentiated quiescent and nonquiescent cells in yeast Stationary-Phase cultures. *Mol. Biol. Cell*, **19**, 1271–1280.
51. McKnight,J.N., Boerma,J.W., Breeden,L.L. and Tsukiyama,T. (2015) Global promoter targeting of a conserved lysine deacetylase for transcriptional shutoff during quiescence entry. *Mol. Cell*, **59**, 732–743.
52. Donaldson,M.E. and Saville,B.J. (2012) Natural antisense transcripts in fungi. *Mol. Microbiol.*, **85**, 405–417.
53. Verzijlbergen,K.F., Faber,A.W., Stulemeijer,I.J. and van Leeuwen,F. (2009) Multiple histone modifications in euchromatin promote heterochromatin formation by redundant mechanisms in Saccharomyces cerevisiae. *BMC Mol. Biol.*, **10**, 76.
54. Lardenois,A., Liu,Y., Walther,T., Chalmel,F., Evrard,B., Granovskaia,M., Chu,A., Davis,R.W., Steinmetz,L.M. and Primig,M. (2011) Execution of the meiotic noncoding RNA expression program and the onset of gametogenesis in yeast require the conserved exosome subunit Rrp6. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 1058–1063.
55. Yassour,M., Pfiffner,J., Levin,J.Z., Adiconis,X., Gnirke,A., Nusbaum,C., Thompson,D.-A., Friedman,N. and Regev,A. (2010) Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. *Genome Biol.*, **11**, R87.
56. Alcid,E.A. and Tsukiyama,T. (2016) Expansion of antisense lncRNA transcriptomes in budding yeast species since the loss of RNAi. *Nat. Struct. Mol. Biol.*, **23**, 450–455.