Université de Montréal

Annotation des ARN non codants du génome de *Candida albicans* par méthode bio-informatique

par Marie Pier Scott-Boyer

Département de biochimie Faculté de Médecine

Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de maîtrise en bio-informatique

février 2009

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

Annotation des ARN non codants du génome de *Candida albicans* par méthode bio-informatique

présentée par :
Marie Pier Scott-Boyer
a été évaluée par un jury composé des personnes suivantes :

François Major, président-rapporteur
Sébastien Lemieux, directeur de recherche
Marcel Turcotte, membre du jury

Résumé

La bio-informatique est un champ pluridisciplinaire qui utilise la biologie, l'informatique, la physique et les mathématiques pour résoudre des problèmes posés par la biologie. L'une des thématiques de la bio-informatique est l'analyse des séquences génomiques et la prédiction de gènes d'ARN non codants. Les ARN non codants sont des molécules d'ARN qui sont transcrites mais pas traduites en protéine et qui ont une fonction dans la cellule. Trouver des gènes d'ARN non codants par des techniques de biochimie et de biologie moléculaire est assez difficile et relativement coûteux. Ainsi, la prédiction des gènes d'ARNnc par des méthodes bio-informatiques est un enjeu important. Cette recherche décrit un travail d'analyse informatique pour chercher des nouveaux ARNnc chez le pathogène *Candida albicans* et d'une validation expérimentale. Nous avons utilisé comme stratégie une analyse informatique combinant plusieurs logiciels d'identification d'ARNnc. Nous avons validé un sous-ensemble des prédictions informatiques avec une expérience de puces à ADN couvrant 1979 régions du génome. Grace à cette expérience nous avons identifié 62 nouveaux transcrits chez *Candida albicans*. Ce travail aussi permit le développement d'une méthode d'analyse pour des puces à ADN de type *tiling array*. Ce travail présente également une tentation d'améliorer de la prédiction d'ARNnc avec une méthode se basant sur la recherche de motifs d'ARN dans les séquences.

Mots-clés : ARNnc, *Candida albicans*, puce à ADN, analyse de *tiling array*

Abstract

Bioinformatics is a multidisciplinary field that uses biology, computer science, physics and mathematics to solve problems in biology. One of the topics of bioinformatics is the analysis of genomic sequences and prediction of genes from non-coding RNA (ncRNA). The non-coding RNAs are RNA molecules that are transcribed but not translated into protein and have a function in the cell. The use of biochemistry and molecular biology techniques in order to find non-coding RNA genes is rather difficult and relatively expensive. Thus, the prediction of genes by bioinformatics methods is an important issue. This research describes a computer analysis to search for new ncRNA in the pathogen *Candida albicans* and an experimental validation. The strategy used was to combine several algorithms and to validate a subset of computer predictions with a microarray experience covering 1979 regions of the genome. We have identified 62 new transcripts in *Candida albicans*. We have also developed an analytical method for tiling array and attempted to improve the prediction of ncRNAs this with a method based on the search of RNA motifs in the sequences.

Keywords : ncRNA, *Candida albicans,* microarray, tilling array analysis

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ADN : acide désoxyribonucléiques

ADNc : ADN complémentaire

ARN : acide ribonucléique

ARNnc : ARN non codant

ARNm : ARN messager

ARNr :ARN ribosomique

snoRNA : ARNnucléolaire

ARNt : ARN de transfert

CGD : Candida Genome Database

EM : algorithme *expectation-maximization*

HMM *: Hidden Markov model* (modèle de Markov caché)

NCM : *nucleotide cyclic motif*

NLS : *Nonlinear Least Sqares*

Nt : nucléotides

ORF : *Open Reading Frame* (cadre ou phase ouverte de lecture)

Pb : Pair de base

SCFG : *stochastic context-free grammars*

SCM : *Structural change model*

SVM : *Support Vector Machine*

UTR : *Untranslated Region* (région non traduite)

VEGF : Vascular endothélial growth factor

Remerciements

Je désire remercier les membres du laboratoire de Bio-informatique Fonctionnelle et Structurelle de l'Institut de Recherche en Immunologie et Cancérologie (IRIC), ainsi que tous les membres de la plateforme de bio-informatique de l'IRIC. Je voudrais remercier tout spécialement mon directeur de recherche Dr Sébastien Lemieux pour son aide et sa disponibilité tout au long de mon projet de maîtrise. Je lui suis également très reconnaissante d'avoir su me communiquer sa passion pour la recherche et la bio-informatique.

Avant-propos

Ce travail sera divisé en 5 chapitres :

1- Le premier chapitre sera une revue de la littérature sur le pathogène *Candida albicans*, les ARN non codants, les méthodes expérimentales et informatiques de détection d'ARN non codants.

2- Le deuxième chapitre présentera la méthodologie de l'analyse informatique pour la recherche d'ARNnc dans le génome de *C. albicans*, la mise en place d'une validation des prédictions d'ARNnc avec une expérience de puces à ADN et le développement d'une analyse statistique des résultats de la validation.

3- Le chapitre 3 sera la présentation des résultats du travail d'annotations ARNnc chez *Candida albicans* présenté sous forme d'un article.

4- Le chapitre 4 sera la présentation du développement d'une nouvelle méthode informatique qui a été utilisé pour détecter les ARNnc directement dans le génome de *Candida albicans*. Cette méthode se base sur la recherche de courts motifs d'ARN.

5- Le dernier chapitre sera une discussion des résultats obtenus et de la bio-informatique des ARNnc, suivi d'une conclusion sur le travail de recherche.

Pour mieux apprécier ce travail certains mots ont été définis dans un glossaire à la fin du document. Ces mots sont suivis du symbole * dans le texte.

Chapitre 1 : Revue de la littérature

Candida albicans

Candida albicans est une levure pathogène. On retrouve *Candida albicans* à l'état naturel dans la bouche et le tube digestif d'environ 80% de la population humaine sans que celui-ci ne cause des maladies et des symptômes. Dans certains cas, ce pathogène cause des infections fongiques (appelé candidoses ou candidases) essentiellement au niveau des muqueuses digestives et gynécologiques. Chez les patients immunodéprimés, comme les patients atteints du sida et les patients cancéreux sous chimiothérapie, les candidoses sont une cause importante de mortalité. *Candida albicans* peut également s'infiltrer dans le flux sanguin, pour causer une infection systémique appelée candidémie. Les candidémies sont des infections plutôt rares, mais elles sont caractérisées par une mortalité de l'ordre de 40% [1]. Quelques médicaments sont efficaces contre les infections fongiques, comme l'Amphotéricine B, les antifongiques de la classe des echinocandins (plus récent) et le plus connue le Fluconazole. Le Fluconazole est une molécule qui inhibe l'activité 14alpha-déméthylase des cytochromes p450. Cette inhibition prévient la conversion de lanosterol en ergostérol, une composante essentielle de la membrane cytoplasmique des levures. L'humain a aussi des cytochromes p450 dans ses cellules, mais l'activité déméthylase mammifère est moins sensible au Fluconazole que l'activité des déméthylases des levures. Ces médicaments ont cependant toutes des limitations et des effets secondaires. Par exemple, une étude de 2007 montre que 13% des souches de Candida recueillies chez des patients hospitalisés étaient résistantes au Fluconazole [2]. Il faut également considérer que les patients immunodéprimés atteints de candidose sont souvent très malades et ne peuvent recevoir une médication trop invasive.

Les particularités génomiques de *Candida albicans*

Candida albicans est un organisme diploïde* qui possède 8 paires de chromosomes, le plus grand contenant les gènes des ribosomes étant appelé R, les suivants étant numérotés de 1 à 7 selon une taille décroissante. Son génome correspond approximativement à 16 Mbp* (haploïde). Le code génétique de *C. albicans* possède une particularité: le codon CUG code pour une sérine et non une leucine.

Pour mieux connaître cet organisme, le génome de la souche SC5314 a été séquencé par Stanford Genome Technology Center [3] et le génome de la souche WO1 a été séquencé par le Broad Institute of MIT et Harvard (*Candida albicans* Sequencing Project. Broad Institute of Harvard and MIT). Le séquençage de la souche SC5314 du génome de *C. albicans* a débuté en octobre 1996. Les efforts successifs de séquençage et des différents assemblages du génome ont marqué les 10 dernières années, pour finalement obtenir l'assemblage 19 en 2004, présentant une version haploïde du génome avec des données sur les différences alléliques observées dans le génome séquencé [3]. L'assemblage 20 (2006) a complété l'assemblage des huit chromosomes de *C. albicans* ; cependant cette annotation comportait des informations de la souche WO1. L'assemblage 20 a été vite remplacé par l'assemblage 21 [4]. Les données de séquençage ont été rendues disponibles à la communauté dans une annotation commune accueillie par la base de données de Candida Genome Database [5]. D'autres génomes d'espèce du genre Candida ont été ou sont présentement séquencés: *C. glabrata*, *C. dubliniensis*, *C.parapsilosis*, *C. guilliermondii, C. lusitaniae* et *C. tropicalis* [6]. Les génomes de ces espèces fourniront des données pour des analyses de génomiques comparatives très informatives. Le génome de *C. albicans* a été annoté et contient environ 6400 gènes codant pour une protéine. Il faut noter que l'annotation a été faite de la façon suivante, un cadre de lectures ouverts de plus de 100 acides aminés est présumé comme un gène codant pour une protéine. Il existe des protéines plus petites que 100 acides

aminés, donc il y a fort probablement d'autres petites protéines à découvrir. Quelques séquences codantes de *Candida albicans* contiennent un intron*. L'annotation des introns est essentielle pour annoter correctement les gènes de ce pathogène et connaître les sites d'épissage alternatifs*. On a identifié jusqu'à présent 415 introns chez *C. albicans* [7]. Il est à noter que très peu d'efforts ont été mis à l'annotation des gènes d'ARN non codants (ARNnc) chez *C. albicans*. Notre projet de recherche consistera à combler ce manque et à identifier les ARNnc de *Candida albicans*.

L'acide ribonucléique

Avant de définir ce que sont les ARN non codants, nous allons débuter par une courte introduction sur la structure et la fonction des ARN en générale. L'acide ribonucléique (ARN) est une macromolécule qui est formée d'une suite de ribonucléotides. Un ribonucléotide comprend une base azotée (soit une Adénine, Cytosine, Guanine ou Uracile), un ribose et un groupement phosphate. La structure primaire de l'ARN rend compte de la séquence nucléique qui est composé des A, G, U, C. Les ARN simple brin se replient sur eux-mêmes pour former une structure stable et compacte, appeler structure secondaire. Cette structure est due à des appariements entre bases complémentaires par exemple : A avec U, G avec C et même parfois à des appariements non canoniques comme G avec U. La structure secondaire peut être complétée par des interactions à longue distance qui définissent une structure tertiaire. Parmi ces interactions, il y a les pseudonoeuds qui est une structure formée par l'interaction d'une boucle* avec une région située à l'extérieur de la tige* qui la délimite. L'existence de structures tertiaires dans les ARN est à la base de ses fonctions.

Les ARN non codants : historique et caractéristiques

La Figure 1 illustre le dogme central de biologie moléculaire décrit par Francis Crick en 1958 [8]. Le dogme central avait été défini comme suit : l'ADN (gènes) est transcrit en ARN, qui est traduit en protéines. Le dogme central met en évidence le premier rôle accordé à l'ARN comme support temporaire de l'information génétique. C'est l'ARN messager qui remplit cette fonction. Celui-ci transmet l'information correspondant à un gène pour synthétiser des protéines. Il existe aussi des ARN qui ne sont pas transcrits en protéines. Ces ARN sont appelés ARN non codants et seront le sujet de ce travail de recherche. Les ARN non codants (ARNnc) sont des molécules ARN (acide ribonucléique) qui sont transcrites à partir de l'ADN génomique, mais non traduites en une protéine et qui ont une fonction dans la cellule. On a longtemps pensé que la plupart des molécules d'ARN étaient seulement des ARN messager qui servaient comme médiateurs entre les gènes et la machinerie de traduction. Le dogme central qui était simple s'est complexifié au cours des dernières décennies, entre autres à cause de la découverte de nouvelles familles d'ARN régulateurs que nous présenterons dans les prochains paragraphes.

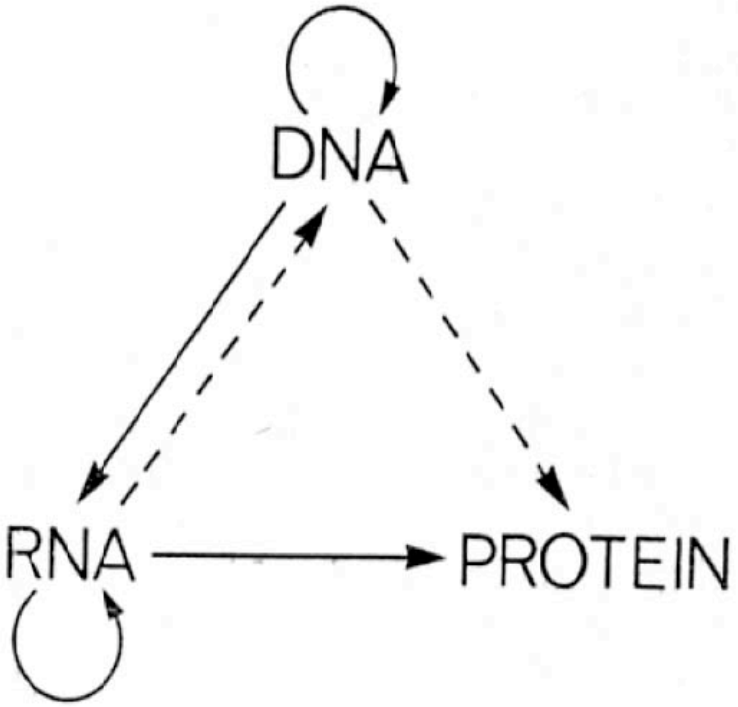

Figure 1 : Le dogme central décrit en 1958 par Crick. À cette époque, on connaissait seulement les processus représentés par les lignes pleines. Les gènes sur l'ADN qui sont transcrits en ARN messagers (ARN) et qui sont traduits en protéines avec l'aide de ribosomes et ARN de transfert. On avait prédit (les lignes pointillées) d'autres interactions possibles mais, elles n'avaient pas encore été observées.

La découverte des ARNnc

Dans les années 1950, on connaissait l'existence de molécules d'ARN messager (ARNm) qui avaient pour rôle d'être l'intermédiaire entre l'ADN et la machinerie pour synthétiser les protéines. Pourtant, les ARNm ne représentent qu'une petite fraction de toute la population d'ARN. A cette époque, on connaissait aussi d'autres ARN présents dans le cytoplasme, comme les ARN ribosomaux (ARNr) qui composent les ribosomes, la machinerie pour synthétiser les protéines [9]. Un troisième type d'ARN fonctionnel a été prévu par l'hypothèse d' « adapteur » [10].

Francis Crick avait prévu l'existence d'une molécule qui négocie entre le code génétique de triplets et les acides aminés. Mahlon Hoagland [11] a biochimiquement observé ces molécules d'ARN « adapteurs ». Ces ARN se sont avérés être les ARN de transfert (ARNt). Les molécules d'ARN avaient maintenant trois rôles, tous impliqués dans la synthèse des protéines. La fraction des ARN qui étaient ni des ARNr et ni des ARNt a été peu étudiée pendant longtemps, car celle-ci était peu abondante et la plupart du temps instable. Il y avait peu de motivation et peu de capacité pour déterminer si cette fraction d'ARN contenait autre chose que des ARNm.

D'autres fonctions pour les ARNnc

Dans les décennies suivantes, plusieurs petits ARN qui n'étaient ni des ARNm, ni des ARNr, ni des ARNt ont été détectés et isolés biochimiquement. Ceuxci incluent la RNase P qui est nécessaire pour la maturation des ARNt [12], les ARN « U » qui participent entres autres à l'épissage des ARNm [13] et les petits ARN nucléolaires (snoRNA) qui guident des modifications chez d'autres ARN. Stark a constaté qu'une fraction pure de RNase P contenait une protéine et aussi un ARN et que cet ARN était nécessaire au fonctionnement de la RNase P [12]. Au milieu des années 80, Sydney Altman et Tom Cech ont démontré que les ARNnc pouvaient catalyser des réactions chimiques par eux-mêmes, un rôle exclusivement réservé aux protéines. Altman a démontré que la composante ARN de la RNase P pouvait cliver l'extrémité 5' des précurseurs d'ARNt. En même temps, le groupe de Tom Cech a démontré que quelques pré-ARNm contenant des introns chez *Tetrahymena* étaient auto-catalytiques. Ils ont ainsi démontré que l'ARN pouvait avoir un rôle catalytique de clivage et ligation sans la présence de cofacteurs de protéine. Cette découverte mène à un prix Nobel de chimie en 1989, car elle a ouvert tout le domaine d'étude des ribozymes et la perspective de thérapies basées sur l'ARN. La capacité des ribozymes à reconnaître et à couper des molécules d'ARN spécifiquement fait d'eux des candidats intéressants pour l'élaboration de thérapies géniques visant à inhiber spécifiquement

l'expression d'un gène. Il a été possible de créer un ribozyme synthétique qui détruit un ARNm qui encode VEGF (vascular endothelial growth factor) [14]. Le VEGF est le facteur de croissance de l'endothélium vasculaire, qui a pour rôle de déclencher la formation de nouveaux vaisseaux sanguins, permettant entres autres la vascularisation des tumeurs.

Les ARNnc et la régulation des gènes

Au début des années 90, le groupe de Victor R. Ambros a identifié le gène lin4 comme étant un régulateur des gènes développementaux lin-14 et lin-28 chez *C. elegans* [15]. Le groupe de Gary Ruvkun a constaté que ce rôle de lin-14 était donné par sa région en 3' qui est non traduite [16]. Deux ans plus tard, le groupe du Dr Ambros a identifié le régulateur du gène lin-4 comme étant un transcrit de 21 nucléotides qui est le complémentaire du 3' UTR* du gène lin-4 [17]. Ils avaient identifié le premier microARN . Sept ans après cette découverte, l'équipe de Gary Ruvkun a identifié un autre gène de microARN, let-7, chez *C. elegans* qui semblait fonctionner d'une façon semblable [18]. Plus tard cette année-là, l'équipe d'Amy E. Pasquinelli a découvert que ces deux microARN n'étaient pas simplement une particularité de *C. elegans* [19]. La séquence du microARN let-7 est conservée et transcrite dans d'autres organismes comme *D. melanogaster* et l'humain. Depuis, des centaines de gènes de microARN ont été trouvés. Beaucoup de ces derniers sont impliqués dans le contrôle des fonctions biologiques de base [20]. Il y a aussi d'autres types d'ARN de régulation qui ont été découverts, comme les *riboswtich.* Les *riboswitch* sont des portions d'un ARNm où se lie un métabolite, qui change la conformation de l'ARNm et par conséquent change l'expression [21].

Classification actuelle des ARN non codant

Actuellement, nous pouvons classifier les fonctions des ARNnc comme appartenant à trois classes, les ARNnc structuraux ou adapteurs, les ARNnc avec un rôle enzymatique (les ribozymes) et les ARNnc régulateurs. Des études récentes sur les transcrits, comme le Projet ENCODE, démontrent que entre 33 et 75% du génome humain serait transcrit [22]. Parmi ces transcrits, on retrouve une importante fraction qui ne code pas pour des protéines. Cette publication suggère que ces transcrits sont non fonctionnels et sont sous sélection neutre peu conservées avec les espèces proches. Des résultats récents démontrent que certains transcrits nécessaires à l'accomplissement de fonctions importantes ne sont pas soumis à une sélection négative lorsqu'on compare à des espèces proches [23] et on pourrait voir apparaître sous peu de nouveaux rôles associés aux ARN. Également, certains transcrits du projet ENCODE pourraient être également être des ARNnc [24].

Les ARNnc connus chez *Candida albicans*

Il y a très peu d'ARNnc connus chez *Candida albicans.* Les ARNnc répertoriés se trouvent en grande partie dans la banque de données Candida Genome Database [5]. On y retrouve 138 ARNnc dont 131 ARN de transfert prédits par outil informatique tRNAscan-SE [25], 1 RNaseP trouvé par homologie avec *S. cerevisia*, 1 snoARN, 1 snRNA U6, 1 signal recognition particle (SRP) trouvé par homologie avec *S. cerevisia* et les 3 ARN ribosomaux. En plus des 138 ARNnc de CGD, les 4 petits ARN nucléaires (snRNA) U1, U2, U4 et U5 membre d'un complexe requis pour épisser les introns, ont été annotés informatiquement [26] . Ces annotations sont résumées à la Annexe A.

Méthode de détection expérimentale

Ces dernières années, la découverte d'ARNnc est au centre d'importantes recherches expérimentales [27] et informatiques [28]. L'appréciation des diverses fonctions des ARNnc est une forte motivation pour découvrir de nouveaux gènes d'ARNnc. Ainsi, plusieurs techniques expérimentales et informatiques ont été développées pour trouver des ARNnc. Historiquement, les ARNnc ont surtout été trouvés un à la fois et souvent de façon accidentelle. Cependant, ces dernières années quelques méthodes à haut débit ont été utilisées pour découvrir et caractériser des ARNnc dans des génomes complets. Trois méthodes expérimentales sont généralement utilisées pour la découverte de nouveaux gènes d'ARNnc: le criblage génétique, le séquençage de banques de ADNc et les puces à ADN.

Criblage génétique

Les criblages génétiques sont des expériences de mutation aléatoire pour identifier des régions du génome (par exemple, un gène) qui possèdent un phénotype d'intérêt que l'on peut sélectionner, par exemple la croissance ou la survie de l'organisme. Les criblages se font généralement en introduisant soit une grosse insertion ou une grosse délétion dans le génome à une position aléatoire. Il est difficile de trouver des gènes d'ARNnc par criblage génétique parce que ceux-ci tendent à être petits. La probabilité d'identifier une mutation aléatoire dans une région de 20 à 100 nucléotides qui est essentielle à la fonction de la cellule est plutôt mince. Les gènes des ARNnc MicF et le DsrA de la bactérie *E. coli* ont été identifiés par criblage génétique. Par exemple, MicF a été découvert en étudiant la régulation génétique des protéines membranaires OmpF et OmpC d'*E. coli* [29]. Plus le génome à étudier est dense, plus d'efforts sont nécessaires pour que toutes les régions du génome soient analysées par des mutations.

Banque de cDNA

Une autre méthode à haut débit pour l'identification des ARNnc est la production et le séquençage d'une banque d'ADN complémentaire (ADNc*). La méthode originale de clonage d'ARNm est basée sur la rétrotranscription des ARNm par une amorce d'oligo(dT) qui se lie à la queue poly-A des messagers et d'une synthèse d'un complément ADN ayant pour résultat une banque d'ADNc. Les ADNc sont séquencés et nous pouvons ainsi connaître les ARNm présents dans une cellule. Cette banque représente idéalement tous les transcrits d'un génome. La différence principale entre les banques conventionnelles d'ADNc et des approches de banque d'ARNnc est la source et le traitement de l'ARN. La fraction des ARNnc est habituellement ignorée dans des banques ADNc car elle ne présente pas toujours des queues poly(A) comme les ARNm. La plupart des ARNm ont une longueur plus 500nt et les ARNnc sont en général plus petits dans un ordre de grandeur ~20 et 500nt. Une stratégie pour séquencer les ARNnc est de faire un enrichissement en filtrant les ARN de plus de 500nt. L'enrichissement d'ARN de petites tailles est réalisé par une séparation de l'ARN total sur un gel dénaturant. On peut également utiliser un anticorps contre une protéine qui lie l'ARN d'intérêt si on veut avoir produit une banque d'ARN qui lient cette protéine. Les ARN ne sont pas choisis par leur taille, mais plutôt quant à leur fonction de lier une protéine commune. Dans les deux cas, un désavantage important de cette méthode est que les ARN seront séquencés en fonction de leur proportion et une majorité des efforts de séquençage sera gaspillée à reséquencer des ARN présents en grande quantité, par exemple les ARNt. Les ARN qui sont en faible nombre de copies risquent de ne pas être séquencés, car l'étape limitant dans cette méthode est le coût et le temps associé au séquençage. Cette technique pourrait devenir plus fréquente avec la venue des nouvelles générations des séquenceurs (pyroséquençage) qui permettent de séquencer à plus haut débit et à moindre coût [30]. Également, des techniques pour filtrer les ARN fréquents ont été mis en place [31].

Les puces à ADN

Les puces à ADN sont une des méthodes couramment utilisées pour observer les niveaux d'expression de plusieurs gènes en parallèle. Les puces à ADN sont des plaques de verre (ou silicium) sur laquelle des sondes d'ADN* sont déposées à la surface. Des milliers de sondes peuvent être fixées sur une même puce. Pour analyser le niveau des transcrits cellulaires, des échantillons sont marqués et sont hybridés à la plaque. Les échantillons utilisés peuvent être l'ADN génomique, l'ARN converti en ADNc ou de l'ARN. Les échantillons sont généralement marqués avec un marquage fluorescent, tel que Cy3 ou Cy5. Lorsque dans l'échantillon il y a présence d'un transcrit qui est complémentaire à une des sondes, celui-ci s'hybride à la sonde et on peut voir un *spot* fluorescent sur la plaque. La fluorescence des *spots* auxquels l'échantillon a hybridé est lue par un module de balayage (*scanner*). Les résultats sont présentés comme un modèle de couleur, où l'intensité de couleur reflète la quantité de transcriptions qui était présente dans la cellule. Les puces à ADN sont la plupart du temps employées pour quantifier l'expression d'ARNm mais elles peuvent également être des moyens d'étudier l'expression d'ARNnc ou même pour la découverte de ARNnc [32]. Pour cette approche, il faut une puce à ADN avec des sondes couvrant l'ensemble du génome. Des expériences chez *E.coli* ont montré que seulement un sous-ensemble des sondes qui couvrent la région de transcription d'un ARNnc montre un signal [33]. La petite taille de certains ARNnc demande que les sondes recouvrent très précisément la région de transcription. Pour avoir une plus haute probabilité d'observer un signal, les sondes sur la puce doivent être de haute densité, c'est-à-dire qu'il doit y avoir un haut niveau de recouvrement des sondes. Parfois, il est possible d'observer une transcription sur l'autre brin d'un ARNnc ou d'un messager expérimentalement validé, c'est-à-dire un transcrit anti-sens [34]. Il est donc intéressant d'ajouter des sondes dans les deux directions possibles de la transcription. Il faut rester critique face aux transcrits anti-sens car une étude récente a mis en évidence que certains de ces transcrits sont des faux positifs, ou sont erronés et seraient dus à des artéfacts de la rétrotranscription [35]. Ainsi, pour avoir une puce qui permet de voir les petits transcrits, il faut une puce avec des sondes qui couvrent

tout le génome (dans les deux directions) et avec une haute densité de sonde. Ce type de puce est appeler *tiling array*. Pour un génome comme celui de *Candida albicans*, cela consiste à plus d'un million de sondes (pour une couverture à chaque 15nt). Au début de ce travail de recherche, les puces avec un tel nombre de sondes n'étaient pas courantes et très dispendieuses.

Méthode informatique de détection d'ARNnc

Beaucoup d'efforts ont été mis pour développer des outils d'informatiques pour identifier des ARNnc dans des séquences génomiques. L'informatique est un outil très utilisé pour annoter des gènes codants pour des protéines dans des génomes. Nous ne pouvons pas utiliser les mêmes outils pour la recherche des gènes d'ARNnc car ceux-ci n'ont pas de cadre de lecture ouvert qui fournit la majeure partie du signal nécessaire pour l'identification des gènes qui codent pour une protéine. Les méthodes informatiques de détection d'ARNnc peuvent trouver des ARNnc de classe connue avec des outils se basant sur l'homologie ou faire des recherches *de novo* avec des méthodes de thermodynamique, de composition de séquences et de génomique comparative.

Les outils se basant sur l'homologie et descripteur d'ARNnc

Lorsqu'on cherche un gène en particulier dans un nouveau génome, la première étape est de chercher des nouvelles instances de ce qui est déjà connu. Des algorithmes pour la recherche de séquences qui sont statistiquement similaires dans les bases de données de séquences génomiques permettent de trouver des gènes codant pour des protéines connus chez d'autres organismes. Cependant, la recherche d'homologie des ARNnc présente une difficulté particulière, car la conservation de structure secondaire de la molécule semble plus importante que la conservation de la séquence primaire. Les algorithmes doivent prendre en compte la structure secondaire [36]. Pour faire face à ce problème, des programmes de descripteur d'ARN comme RNAmotif [37] et Rnamot [38] ont été créés. Ces outils cherchent dans les bases de données de séquences d'ARN, des

motifs qui décrivent la structure secondaire d'ARN. Pour améliorer les recherches d'ARNnc, des outils probabilistes ont vu le jour. Dans ces outils, les contraintes évolutives pour maintenir la structure secondaire peuvent être modélisées dans un modèle probabiliste SCFC* (*stochastic context-free grammars*) pour un ARNnc particulier, par exemple un ARNt [25]. Le modèle SCFG de l'ARNt a été implanté dans le programme tRNAscan-SE [25] et ses auteurs rapportent qu'il peut trouver les ARNt dans une séquence génomique avec une sensibilité de 99.8% et avec moins de 0.002 % de faux positifs par Mbp de séquences. La plupart des efforts pour identifier des ARNnc dans des séquences génomiques ont été conduits par la recherche de la structure secondaire d'ARN et beaucoup d'outils ont été développés pour rechercher des structures spécifiques dans le génome, telles que ERPIN [39] et INFERNAL [40]. Pour ce type de recherche, il faut bien connaître la structure secondaire de l'ARN recherché. La base de données RFAM [41], [42] contient la structure de plusieurs classes d'ARNnc procaryotes et eucaryotes.

Les outils se basant sur la thermodynamique

Les méthodes de thermodynamique se basent sur le fait que chaque configuration d'une molécule ARN correspond à une quantité d'énergie libre. La configuration la plus stable est celle qui minimise l'énergie libre et elle correspond à la configuration que la molécule d'ARN adopte en se repliant. La recherche de la structure secondaire qui minimise l'énergie libre est en temps N^3 et est réalisable par programmation dynamique. Le programme le plus utilisé qui applique cette approche est Mfold [43]. Les méthodes de recherche d'ARNnc par thermodynamique se basent sur l'hypothèse que l'énergie libre des ARNnc est plus faible que l'énergie libre d'une séquence aléatoire de même composition de nucléotides [44]. Les méthodes de thermodynamique semblent fonctionner pour des structures locales fortes, cependant la majorité des ARNnc (tRNA, rRNA) n'ont pas une énergie libre plus basse que des séquences aléatoires de même composition en nucléotides [45]. Une autre étude plus récente a utilisé des séquences de même composition en dinucléotides* sur 500 ARNt, 581 ARNr et 506 microARN. Leurs résultats révèlent que les

microARN montrent des structures plus stables d'une séquence aléatoire de même composition de dinucléotides [46]. Il y a une controverse à savoir si vraiment la présence de structure secondaire signifie qu'il y a présence d'un ARNnc.

Biais en composition de séquences

Des travaux se sont intéressés à l'existence de biais de séquences dans les ARNnc [47]. Des expériences ont été menées sur des ARNt, des ARNr, des ARN nucléaires, des ARN nucléolaires et des SRP dans trois organismes : la bactérie *Methanococcus jannaschii*, le nématode *Caenorhabditis elegans* et le parasite *Plasmodium falciparum*. Les auteurs se sont focalisés sur le pourcentage en GC et sur la fréquence d'apparition du dinucléotide CG. Les biais de composition ne semblent pas exister dans tous les organismes, mais ils sont observables chez *Methanococcus jannaschi*. L'auteur a cherché à déterminer si une variation locale des compositions pouvait révéler la présence d'un ARNnc pour cet organisme. Dans un génome riche en A+T (ex. bactéries thermophiles), les ARNnc se distinguent nettement en cherchant des régions du génome riche en G+C. Quelques dizaines d'ARNnc ont été ainsi prédits et confirmés expérimentalement chez les espèces *M. jannaschii* et *P. furiosus* avec cette méthode [48]. RNAgenie [49] est un logiciel pour trouver des ARNnc par l'implémentation d'un réseau de neurones qui prend en paramètre d'entrée plusieurs informations. Parmi les entrées du réseau de neurones, on retrouve le pourcentage de A, G, C et T (U), les pourcentages de dinucléotides et les motifs UNCG, GNRA (R=purine), CUYG (Y=pyrimidine), AAR, CUAG et le terme de thermodynamique. La phase d'apprentissage a été effectuée sur l'ensemble des ARNnc d'E. *coli* (principalement des ARNt et des ARNr). Les performances de RNAgenie ont été évaluées sur les génomes de huit organismes. Entre 80% et 90% des ARNnc sont correctement détectés avec une proportion de prédictions positives erronées inférieure, en moyenne, à 15%. Si l'on étudie de plus près ce que le réseau de neurones a appris, on constate que les entrées les plus informatives sont les fréquences des nucléotides, l'information thermodynamique, ainsi que les fréquences d'apparition des dinucléotides

CU, GU et GG. Les motifs structuraux ne participent que faiblement au processus de décision. Il n'y a pas d'autres efforts qui ont été faits pour développer cette méthode dans d'autres organismes et le logiciel est très peu utilisé.

La génomique comparative

Les méthodes de génomique comparative se basent sur le fait que si l'on dispose de plusieurs séquences similaires, alors il y a plus d'informations sur la fonction de ces séquences. Il y a deux raisons pour lesquelles il est intéressant d'utiliser l'information de plusieurs séquences. Premièrement, les ARNnc homologues* ont une fonction induite par une structure commune et deuxièmement parce que les prédictions de structures communes peuvent être confirmées par la présence de mutations compensatoires. Ce type de mutations est engendré par la conservation d'une structure au cours de l'évolution. Donc lorsqu'une base impliquée dans un appariement est mutée, la base complémentaire est mutée pour maintenir l'appariement et la structure. Les approches de génomiques comparatives se basent sur l'idée que la conservation des structures secondaires d'ARNnc fournit l'évidence de fonction biologique. Cette idée a été exploitée dans une étude par Rivas et Eddy [50] dans le développement du logiciel QRNA. Ce logiciel implante un modèle de probabilité de mutations synonymes* pour les gènes codant pour des protéines et un modèle de mutations compensatoires pour les gènes ARNnc dans un alignement de deux séquences. Une limite de cette méthode est que les séquences doivent être entre 65% et 85% similaires car les alignements présentant des niveaux de similarité inférieurs sont souvent incorrects. Les alignements trop similaires ne démontrent pas de mutations compensatoires. Cette idée a été également employée par le programme RNAz [51] combinant la conservation structurale et la stabilité thermodynamique des structures secondaires d'ARN dans des alignements multiples. Dynalign[52] est un autre outil de prédiction *de novo* qui a la particularité de prévoir la plus basse énergie libre d'une structure secondaire commune à deux séquences non alignées et de trouver une structure commune de façon indépendante de la similarité de séquence. Une étude récente [53], qui a

comparé l'efficacité de plusieurs outils de recherche de ARNnc, montre que la méthode employant seulement la stabilité thermodynamique semble donner des résultats comparables à celles qui recherchent également la conservation de structures secondaires ARN pour des séquences de même composition en dinucléotides. Cette étude montre également l'importance d'utiliser un modèle nul de dinucléotides plutôt que de monocluéotides* pour distinguer les vraies prédictions. Ils démontrent que le signal perçu par beaucoup des outils semble venir des fréquences de dinucléotides que le modèle nul de mononucléotides ne considère pas.

Autres méthodes de détection :

Tous les outils présentés utilisent des caractéristiques des séquences d'ARNnc pour trouver un signal permettant de les distinguer. Il y a d'autres caractéristiques possibles qui seraient exploitables pour trouver des ARNnc comme des signaux destinés à la machinerie transcriptionnelle. On peut identifier des gènes d'ARNnc en se basant sur la recherche de motif promoteur de la transcription. Les gènes codant pour des protéines sont généralement transcrits en ARN messager par polymérase II.

Les gènes d'ARNnc sont parfois transcrits par la polymérase II, parfois transcrits par polymérase III (par exemple le ARNt) et par la polymérase I dans le cas des ARNr. Une étude chez *Saccaromyces Cerevisiea* a étudié les promoteurs de la polymérase III pour trouver de nouveaux ARNnc et ont identifié quelques nouveaux ARN [54].

Pour en apprendre davantage sur les motifs promoteurs de la transcription d'ARNnc chez *C. albicans*, il faut d'abord les identifier. Une fois identifiée, des études computationnelles pourraient être faites pour découvrir des motifs promoteurs qui nous informeraient plus sur le mécanisme de transcription des ARNnc chez *Candida albicans*.

L'objectif de l'étude des ARNnc chez *Candida albicans*

Chez les levures, l'organisme modèle *S. cerevisiae* a été le plus étudié. *C. albicans* et *S. cerevisiae* ont divergé il y a approximativement 800 millions années et

donc montrent probablement une divergence évolutionnaire importante [55]. En utilisant l'algorithme Blast [56] à une *e-value** $< 1\times10^{-20}$, nous pouvons conclure qu'approximativement 40% de gènes de *C. albicans* n'ont aucun homologue chez *S. cerevisiae* [57]. Avec ces informations, nous sommes tentés de croire que nous trouverons des ARNnc chez *C. albicans* qui n'auront pas d'homologue chez *S. cerevisiae.* Pour cette recherche, nous ne voulons pas nous limiter à seulement trouver des homologues d'ARNnc de *S. cerevisias*. Les efforts pour la recherche d'ARNnc seront faits directement avec le génome de *C. albicans*.

Nous voulons étudier les ARNnc chez *Candida albicans* pour deux objectifs : obtenir plus de connaissances à leur sujet et trouver des cibles thérapeutiques pour la mise en place de thérapies contre *Candida albicans.* L'intérêt d'étudier les ARNnc est que ceux-ci ont des rôles fondamentaux dans la cellule et que nous ne connaissons pas encore tous les rôles associés à ces molécules. Nous avons espoir de pouvoir trouver des microARN dans *Candida albicans* à cause de la présence d'un homologue de la protéine Argonaute (reliée au mécanisme de régulation des microARN) dans son génome malgré qu'aucun phénomène relié à la régulation des gènes par de petits ARN n'a pas été observé. Il est à noter que cette protéine n'a pas d'homologue chez *S. cevevisiae*.

Un intérêt d'étudier les ARNnc particulièrement chez *C. albicans* est que ceux-ci peuvent être utilisés comme cibles thérapeutiques. Le développement de nouveaux antifongiques est restreint par le nombre de cibles potentielles. Chez les procaryotes, l'ARNr (suffisamment différent de celui des eucaryotes) a été utilisé comme cible antibiotique, par exemple dans les classes d'antibiotiques Aminoglycosides, Tetracyclines, Macrolides, etc. Chez *Candida albicans*, la molécule Hoechst 33258 inhibe les introns du groupe 1* en affectant leurs repliements [58]. Il faut considérer que d'un point de vue pharmacologique, il est plus difficile de trouver des cibles pour des médicaments contre des levures que pour les bactéries. Comme *Candida albicans* est un eucaryote, elle a beaucoup de similarité avec les cellules humaines. Ainsi, nous voulons une cible qui affecte la survie

de *C. albicans* mais qui n'affecte pas les cellules humaines et nous avons espoir que les ARNnc sont de bons candidats.

Cette recherche décrit un travail d'analyse informatique pour chercher de nouveaux ARNnc chez *C. albicans*. Notre but est d'identifier un sous-ensemble de candidats d'ARNnc avec une analyse informatique et de les valider expérimentalement.

Chapitre 2 : Méthodologie de rechercher et validation d'ARNnc chez Candida albicans

Introduction

Ce chapitre vous présentera, dans un premier lieu, les détails de l'analyse informatique effectuée pour trouver des prédictions d'ARN non codants dans le génome de *Candida albicans* et deuxièmement la validation expérimentale d'un sous-ensemble de prédictions basées sur la conception d'une puce à ADN. Nous discuterons également de l'analyse et du développement d'une méthode pour trouver les transcrits à partir des données obtenues des puces à ADN. Il y a un chevauchement entre le chapitre 2 et la chapitre 3, afin de présenter les détails la méthodologie qui sera présenté de façon superficiel dans le l'article.

Analyse informatique

Une analyse informatique pour identifier des ARNnc a été faite avec les logiciels suivants: INFERNAL/RFAM [40],[41, 42], QRNA [50], RNAz [51] et Dynalign [52]. Ces outils ont été choisis parmi les outils disponibles afin de représenter des approches différentes et possiblement augmenter la diversité des prévisions obtenues. Les paramètres des divers outils ont été optimisés à partir d'espèces procaryotes ou de vertébrés et aucun consensus n'existe quant à quel outil est le meilleur pour identifier des ARNnc dans les levures. Sauf indication contraire, toutes les méthodes ont été appliquées sur l'assemblage 19 du génome de *C. albicans* [3] avec les régions de basse complexité masquées avec DUST [59] car ces régions semblent interférer avec les outils de prédictions ARNnc. Les régions de basse complexité représentent 4.3 % (649080 nt sur 15098438 nt) du génome de *C. albicans*. Pour réduire le temps de calcul, nous avons également enlevé les séquences génomiques correspondant aux copies alléliques car les variantes étaient mineures. La taille finale des séquences génomiques est de 15.1 Mbp.

RFAM/INFERNAL

Le premier outil utilisé est RFAM/INFERNAL [40,41,42]. RFAM 8.0 est une collection de 574 modèles de covariance d'ARNnc des eucaryotes et procaryotes. Le logiciel Infernal-0.81 recherche des régions dans des séquences génomiques qui se conforment aux propriétés des modèles. Nous avons indépendamment recherché chaque modèle de covariance dans le génome de *C. albicans*. Nous avons gardé les prédictions au-dessus des 3 niveaux de seuils (trusted, gathering et noise) prescrits par les auteurs pour chaque modèle.

QRNA

Le deuxième outil utilisé est QRNA-2.0.3. Ce logiciel identifie les structures secondaires conservées entre deux espèces. Le génome de *C. albicans* a été découpé en séquences de 1500 nucléotides qui se chevauchent par 150 nucléotides. Ces séquences ont été alignées avec les génomes des espèces étroitement liées suivantes : *C. Dubliniensis*, *C. tropicalis*, *C. parasilosis*, *C. guilliermondii*, *C. lusitaniae*, *C. glabrata*, *Debaryomyces hansenii* et *S. cerevisiae* avec WU-blast-2.0 [60] en utilisant les paramètres par défaut. Nous avons gardé le sous-ensemble des alignements avec une similarité entre 75 et 90 % et qui ont une longueur de 100 nucléotides et plus. Les alignements ont été encore divisés en séquences de 150 nucléotides se chevauchant de 50 nucléotides et le logiciel QRNA a été exécuté sur ces alignements. Le logiciel QRNA évalue si un alignement appartient au modèle RNA (mutations compensatoires) pour les ARN non codants, au modèle COD pour les protéines et au modèle neutre OTH qui représente les mutations aléatoires. Pour chaque alignement, QRNA retourne un score de probabilité pour chacun de ces modèles et le modèle dont la probabilité est la plus élevée est retenu. Nous avons gardé tous les alignements ayant une probabilité d'appartenir au modèle d'ARN. Les distributions des scores obtenus ont été calculées pour chacune des espèces et employées pour déterminer un seuil spécifique à l'espèce.

RNAz

Le troisième outil utilisé était RNAz-1.0 [51]. Ce programme trouve des structures secondaires d'ARN conservées et thermodynamiquement stables dans des alignements multiples. L'utilisation des alignements multiples a comme conséquence de fournir plus d'informations disponibles à l'algorithme pour identifier la conservation de structure. Nous avons construit des alignements multiples avec l'outil Multiz [61]. Les alignements ont été construits en cherchant des séquences homologues de *C. albicans* de 500 nt (se recouvrant 150 nt) dans les génomes *C. dubliniensis*, *C. tropicalis* et *C. parasilosis*. Nous avons décidé de nous limiter à ces espèces, car quand des espèces plus éloignées ont été employées nous avons obtenu des alignements multiples de pauvre qualité. Nous avons expérimenté avec les espèces *C. guilliermondii*, des *C. lusitaniae*, *C. glabrata*, *Debaryomyces hansenii*, et *S. cerevisiae*. Le logiciel RNAz a été exécuté sur les alignements plus longs que 50 nucléotides. Le logiciel retourne un score Z associé à l'énergie libre et un score d'indice de conservation de structure. RNAz utilise un SVM (Support Vector Machines) pour classifier si l'alignement est un ARN en se basant sur les deux scores. Tous les alignements qui ont été prédits comme un ARN avec une probabilité plus grande que 0.995% ont été gardés.

Dynalign

Le logiciel a été conçu pour identifier de l'ARN structuré en combinant la minimisation d'énergie libre et l'analyse comparative de séquences pour trouver une structure à basse énergie libre commune à deux séquences sans exiger une similitude de séquence. L'équipe du Dr. David Mathews (Université de Rochester) a effectué cette analyse avec un logiciel basé sur Dynalign [52], qui n'est pas encore publiée. Les alignements du génome de *C. albicans* contre les génomes de C. *dubliniensis*, *C. glabrata,* et *C. parapsilosis* ont été faits avec le logiciel MUMMER [62]. L'assemblage 20 de *C. albicans* a été utilisé pour construire les alignements. Les résultats de Dynalign ont été transposés sur l'assemblage 19 avec BLAST. Un petit sous-ensemble des prédictions, 1.89% (6996 sur 369930), n'ont pas pu être retrouvées sur l'assemblage 19 et n'ont pas pu

être employées pour les stades plus avancés de l'analyse. Dynalign retourne des scores Z pour chaque alignement. La distribution de score Z a été alors calculée pour chaque espèce comparée et employée pour déterminer un seuil spécifique à l'espèce.

Choix des régions de validation

Nous avons intentionnellement gardé les seuils très permissifs pour les divers algorithmes pour obtenir un grand nombre de prédictions. Les prédictions ont été priorisées pour la validation selon le procédé suivant:

1- Pour faire une comparaison quantitative de l'exactitude des méthodes, nous avons validé les 200 prévisions avec les plus hauts scores pour chacune des 4 méthodes, incluant celles trouvées dans les ORF. Ces 800 prévisions représentent 32.5% des sondes de la puce.

2- Les régions restantes de validation ont été choisies en donnant la priorité à des régions contenant des prédictions qui se chevauchent, et qui ne sont pas dans un ORF ou dans une basse région de complexité. Ceci représente 58% des sondes.

3- Comme contrôle négatif, nous avons également ajouté 200 régions de 200 nucléotides dans les régions intergéniques (à plus de 150 nucléotides du début et de la fin d'ORF) qui ne chevauchent aucune des prédictions, incluant celles qui n'ont pas été retenues en 1 et 2. Ces régions représentent 9.7% des sondes.

4- Les ARNnc annotés dans la banque de données Candida Genome Database [5] ont été ajoutés comme contrôle positif. Les contrôles positifs représentent 2.4% des sondes.

Nous avons fusionné les régions de validation qui se chevauchent. Au total, il y a 1979 régions de validation avec une longueur de moyenne de 374 nucléotides, y incluant les contrôles.

Constructions de la puce à ADN

La validation expérimentale est basée sur la mise au point, le design et la conception d’une puce à ADN. Pour chaque région du génome à valider, nous avons étendu les régions

de 60 nucléotides à chacune de leurs extrémités et nous avons conçu des sondes d'environs 60 nt à chaque 20 nt dans les deux directions. Le design résultant a été synthétisé par Roche NimbleGen Systems Inc., validant 1979 prévisions en utilisant 72000 sondes uniques. Nous considérerons comme une validation, la démonstration que la région est exprimée.

Protocole expérimental de la puce à ADN

Pour valider la présence d'une sélection de prédictions d'ARNnc dans *C. albicans* nous avons regardé le profil d'expression au niveau des régions de validation. Nous avons fait une expérience avec 8 réplicats biologiques avec comme condition expérimentale la croissance de *C. albicans* à 30 °C. Sur la puce, nous avons hybridé de l'ARN total marqué avec un fluorochrome et un enrichissement pour les petits ARN marqués avec un fluorochrome différent. Les détails du protocole expérimental se trouvent au chapitre 3.

Analyse de la puce à ADN

Analyse des résultats bruts

Nous récupérons des expériences de puces à ADN un fichier texte contenant pour chaque sonde présente sur la puce, les intensités de fluorescence des deux fluorochromes correspondant à l'extraction d'ARN total et de petits ARN. Nous avons premièrement effectué une transformation logarithmique sur les données brutes. La transformation log permet de diminuer l'influence des valeurs très élevées sur la moyenne des valeurs parce que la distribution des valeurs est biaisée vers les valeurs hautes. Cette transformation est couramment appliquée à des résultats obtenus d'un tube photomultiplicateur (le détecteur qui se trouve dans les *scanners* de puces) car le signal produit par un tube photomultiplicateur est exponentiel par rapport au signal amplifié.

On corrige les biais expérimentaux par une normalisation des données. Nous avons appliqué une normalisation quantile entre les données des huit puces (pour les petits ARN et les ARN totaux) pour que nous puissions les comparer (Voir Figure 2). Une

normalisation quantile force la similitude des deux distributions [63]. Il s'agit d'une normalisation grandement utilisée en analyse de puces et qui, contrairement à une normalisation Loess (la normalisation la plus utilisée), est directement applicable sur un jeu de données contenant plus de 2 puces.

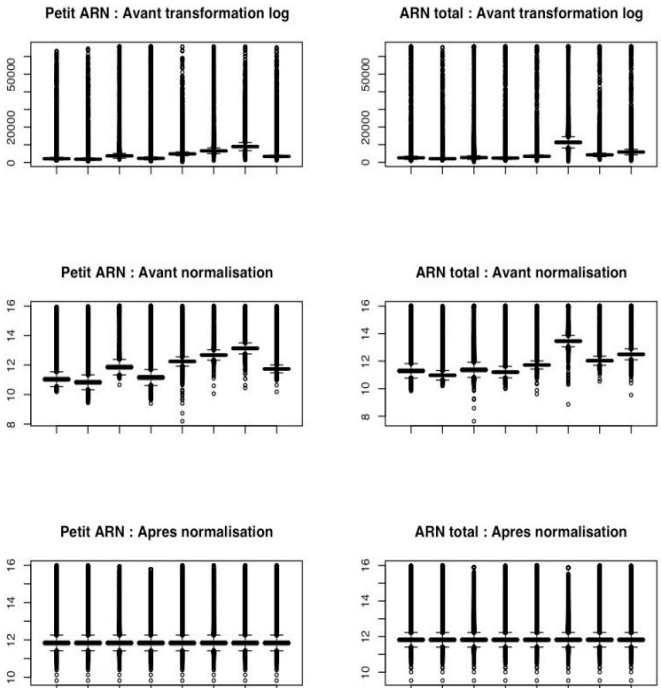

Figure 2 : Les *box-plots* avant et après la transformation des données brutes. Le *box-plot* est un diagramme qui permet d'observer la valeur médiane, maximale et minimale des intensités des sondes présentes sur les différentes puces. En haut, nous pouvons observer les *box-plots* des huit puces avant la transformation log des données des petits ARN et les ARN totaux. Au milieu, nous pouvons observer les *box-plots* des huit puces avant normalisation quantile et après la transformation log. En bas, nous pouvons observer les *box-plots* des puces après la normalisation quantile. Après la normalisation, nous pouvons observer que les intensités médianes des puces sont les mêmes.

Identification des régions transcrites

A partir des données brutes de la puce, nous nous sommes intéressés à trouver les régions qui sont transcrites dans la cellule. Une méthode consiste à trouver des transcrits en se basant sur un seuil d'intensité, par exemple la valeur qui représente le 97ième quantile des intensités des sondes (environ 3% d'expression). On définit une région exprimée lorsqu'un certain nombre de sondes consécutives se trouve au dessus de ce seuil [64]. Cette méthode n'est pas très robuste car nous ne savons pas *a priori* quel pourcentage du génome est transcrit. Il existe aussi des algorithmes pour identifier les transcrits qui se basent sur l'algorithme SCM (*Structural Change Model*) [65]. Ces algorithmes consistent à trouver où placer des bornes qui délimitent les débuts et les fins de transcrits, dans le but de minimiser la somme des résiduels par rapport à la moyenne d'intensité des sondes à l'intérieur et à l'extérieur des transcrits.

Pour identifier des transcrits dans les régions validées par la puce à ADN, nous avons développé un outil spécifiquement adapté à la structure de notre puce. Premièrement, nous avons supposé qu'il y a deux états dans le génome, un état transcrit (comportant un niveau de fluorescence élevé) et un état non transcrit (sans fluorescence) que nous appelons *background.* On présume que pour chaque région de validation, il y a zéro ou un transcrit avec une série de sondes adjacentes avec un niveau de fluorescence élevé. Deuxièmement, nous avons considéré que nous chercherons à placer deux bornes (une borne pour le début du transcrit et une deuxième borne à la fin du transcrit) pour délimiter la région qui est transcrite pour chaque zone de validation. Contrairement au modèle SCM et afin de nous permettre d'exploiter le chevauchement de nos sondes, nous modéliserons le début et la fin du transcrit au niveau des nucléotides. Une sonde chevauchant, c'est-à-dire ne s'hybridant que partiellement à un transcrit, présentera un niveau de fluorescence intermédiaire entre le background et le niveau d'une sonde s'hybridant entièrement au transcrit. Ainsi, une sonde peut contribuer à l'état transcrit et à l'état *background.* Nous avons ajouté la contrainte qu'il doit y avoir au moins 60 nt entre les deux bornes, pour s'assurer d'avoir des transcrits avec une longueur minimum de 60 nt.

L'équation (1) modélise l'intensité d'une sonde par rapport à :

$$X^{ij} = f(w^{ij})\,\mu t + (1-f(w^{ij}))\,\mu b \quad (1)$$

où X^{ij} est l'intensité de la sonde i dans la région j, w^{ij} est le pourcentage de chevauchement de la sonde dans la région transcrite, $f(w^{ij})$ est la fonction qui détermine le poids de la valeur w (entre 0 et 1) qui représente la transition entre le niveau de background et le niveau d'intensité du transcrit par rapport au pourcentage de chevauchement, µb la moyenne de l'intensité du background et µt la moyenne de l'intensité du transcrit. Suivant ce modèle, nous plaçons les bornes de transcription (B1 et B2) dans le but de minimiser la somme des résiduels.

L'équation des résiduels (2) :

$$r^i = X^i - [(f(w^i)\mu_t) + ((1 - f(w^i))\mu_b)]^2 \quad (2)$$

L'équation de la somme des résiduels (3) :

$$\sum(r^i) = \sum X^i - [(f(w^i)\mu_t) + ((1 - f(w^i))\mu_b)]^2 \quad (3)$$

L'algorithme recherche pour les N nucléotides chaque région de validation j, toutes les combinaisons possibles pour placer les bornes de transcription (B1 et B2). L'algorithme est une recherche exhaustive $O(N^2)$ puisque les régions sont

relativement courtes (moyenne de 374 nt). Voici comment nous avons implanté le modèle pour trouver les bornes optimales:

Pour B1 dans [1, N - 60]:
Pour B2 dans [B1+60, N] :

$$\mu_t = \frac{\sum f(w^i)X^i}{\sum f(w^i)}$$

$$\mu_b = \frac{\sum (1 - f(w^i))X^i}{\sum (1 - f(w^i))}$$

$$\sum r^i = \sum X^i - [(f(w^i)\mu_t) + ((1 - f(w^i))\mu_b)]^2$$

On retourne le minimum ($\sum r^i$) pour les paires B1 et B2 testées

Fonction des poids

Nous avons premièrement considéré que la distribution des résiduels r^{ij} par rapport au chevauchement *w* était linéaire, c'est-à-dire que si une sonde chevauche le transcrit de 10% alors elle contribue à 10% du signal du transcrit. En observant la distribution des résiduels r^{ij} par rapport au chevauchement *w* (Voir Figure 3, panneau du haut), nous avons observé un biais suggérant que la véritable relation correspondait à une transition plus rapide que ce qui est modélisé par la relation linéaire. Avec les résultats de la fonction linéaire, nous avons tenté d'optimiser la fonction f0(w^i), en calculant le poids optimal (w^o) pour un chevauchement donné, selon cette formule (4):

$$X^i = w\mu t + (1-w^o)\mu b$$

$$X^i = w^o\mu t + \mu b - w^o\mu b$$

$$X^i = w^o(\mu t - \mu b) + \mu b$$

$$w^o = (X^i - \mu t)/(\mu b - \mu t) \quad (4)$$

Nous observons que le chevauchement des sondes sur le transcrit, n'est pas une fonction linéaire mais les données suivent une tendance sigmoïde (voir Figure 3).

Nous utilisons une fonction sigmoïde car elle permet de bien modéliser nos données. Avec le package Nonlinear Least Squares (NLS) de R [66], nous avons estimé les paramètres de la sigmoïde à partir de nos données sauf les asymptotes horizontales qui ont été fixé à A=0 et le B=1, correspondant à attribuer une intensité de μt dans le transcrit et μb en dehors, pour que la fonction nous donne un chiffre entre 0 et 1. L'algorithme Gauss-Newton a été utilisé pour déterminer les estimations des moindres carrés des paramètres du modèle non-linéaire. Nous présumons que notre premier estimé de f^1 (w^j) ne devant pas être très loin des résultats optimaux, nous utilisons donc ces bornes pour optimiser la fonction $f^2(w^j)$ comme un algorithme expectation-maximization (EM) [67].Nous avons recalculé les résiduels avec cette fonction jusqu'à convergence, c'est-à-dire lorsque l'erreur résiduel moyen retourné par NLS ne changeait plus. Deux itérations ont été nécessaires à partir de la fonction linéaire $f^0(w^j)$, soit $f^1(w^j)$, $f^2(w^j)$. Ceci nous permet finalement d'avoir une fonction sigmoïde qui prend un pourcentage de chevauchement avec le transcrit et nous retourne un poids modélisant mieux nos observations. Notre équation finale (6) :

$$f^2(w^i) = \frac{1}{1 + e^{\frac{(0.5155) - w^i}{0.04359}}} \tag{6}$$

Nous classifions le niveau d'expression en calculant la différence en du transcrit est μt-μb, car nous présumons que le μb est très semblable à chaque région j.

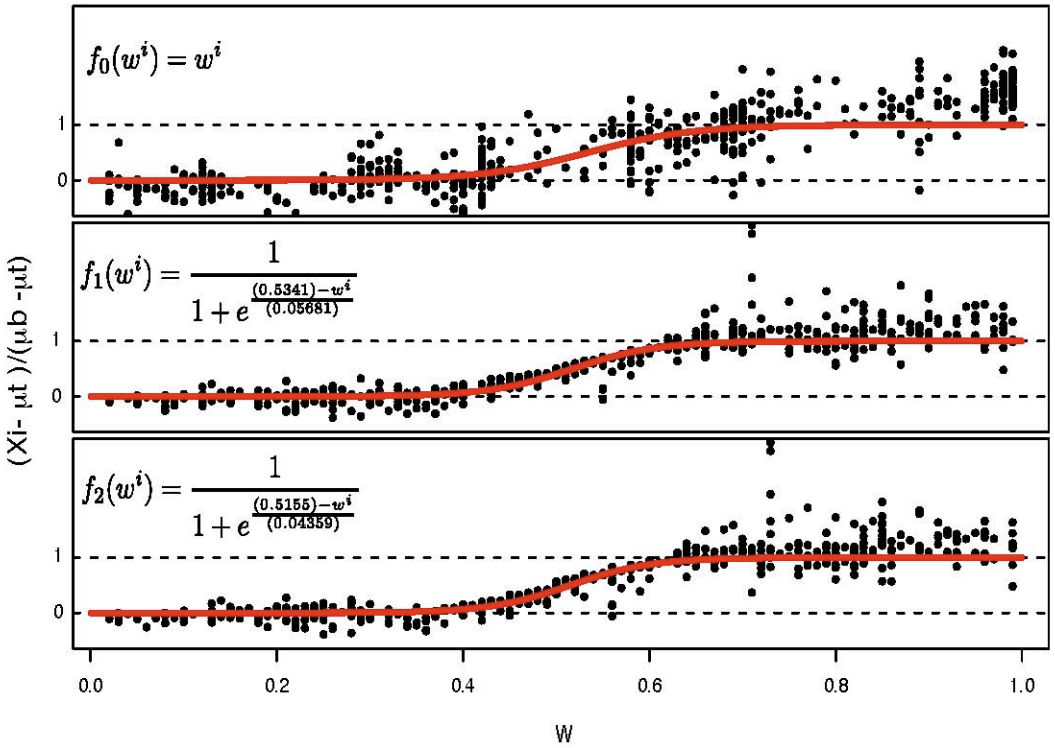

Figure 3 : L'optimisation de la fonction *f* (w). Sur l'axe des X, on retrouve les valeurs du chevauchement et sur l'axe des Y le poids optimal. En rouge, la fonction sigmoïde qui modélise des données. Avec la fonction linéaire f0(w^i), l'erreur résiduelle moyenne est de 0.3184; avec une fonction sigmoïde après une première optimisation f1(w^i) l'erreur résiduelle moyenne est de 0.2627 et une fonction sigmoïde après une deuxième optimisation f2(w^i) avec une erreur résiduelle moyenne de 0.2674.

Chapitre 3 : Résultats de recherche et validation d'ARNnc

Avant–propos

Ce chapitre présente un article en préparation pour la revue Plos Computational Biology. Ma contribution spécifique à cet article est d'avoir réalisé l'analyse computationnelle des différents outils de recherche d'ARNnc (sauf DYNALIGN), la priorisassions des régions à valider, le développement de la méthode d'analyse des puces à ADN et l'analyse des résultats des puces à ADN. Ma contribution à la rédaction de l'article est d'environ 50%.

Genome-wide Annotation of Non-coding RNAs in *Candida albicans*: from *in Silico* Prediction to Validation

Marie Pier Scott-Boyer, Guillaume Bouvet ,Martine Raymond and Sébastien Lemieux
Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, P.O. Box 6128, Station Centre-Ville, Montréal, QC H3C 3J7, Canada.

Abstract

The non-coding RNAs are functional molecules that are transcribed but not translated into proteins. This research describes a computational analysis followed by experimental validation to identify novel non-coding RNAs in the pathogenic yeast *Candida albicans*.

We have applied and combined the results of four computational tools to identify ncRNAs and have experimentally validated a subset of the predictions using a tiling array covering 1,979 short regions representing approximately 5% of the genome. We have developed an algorithm to analyze tiling arrays data that makes efficient use of overlapping probes typical in high-density tiling arrays. From this analysis, we have identified 62 new transcripts in *Candida albicans*.

We have compared the performance of the four computational tools for predicting non-coding RNA based on the expressed transcript they have found. We have observed that RFAM/INFERNAL shows the highest performance for finding known transcripts and Dynalign shows the highest performance to find novel transcripts *de novo*. Furthermore, we have observed that each method identified unique candidates, making a multi-algorithms approach a better solution at the present time. Finally, this research demonstrated the combination of computational approaches and microarray-based validations can be an effective ways to identify new transcripts in an important human pathogen.

Author summary

In the past years, non-coding RNAs have become an emerging focus of molecular biology. Non-coding RNAs are transcripts that, without being translated into proteins play important roles in the cell. In this work, we used a combination of four computational approaches followed by a large-scale experimental validation to identify 62 non-annotated and expressed transcripts from the *Candida albicans* genome, an important human pathogen. Results obtained from our experimental validation were used to compare the performance of the different computational methods.

Introduction

Among the fungal pathogens responsible for human systemic diseases, *Candida albicans* is one of the most important and causes a wide variety of infections, ranging from mucosal infections in healthy individuals to life-threatening systemic infections in individuals with an impaired immune system. The publication, three years ago, of the *C. albicans* genome [3] has considerably increased the possibilities for genomic and transcriptomic studies, including computational screen for discovering novel non-coding RNAs (ncRNAs) and design of tiling microarrays. Non-coding RNAs are functional RNA molecules that are transcribed but not translated into a protein. In eukaryotic cells, they have been shown to be key players in several essential processes such as translation [11], splicing [68]and regulation [69]. In the Candida Genome Database (CGD) [5], a set of 135 ncRNAs are annotated for assembly 19 including 131 putative tRNAs predicted by the tRNAscan software [25]. The other 4 are predicted by sequence homology with *S. cerevisiae* and correspond to the RNAseP, a snRNA U6, a signal recognition particle (SRP) and a snoRNA. The rRNA has been annotated as part of the assembly 21 of the *Candida albicans* genome [4]. And, not yet in CGD, the snRNAs U1 to U6 have been annotated by computational approaches [26].

Expression profiling using tiling microarrays, and sequencing of cDNA libraries are often used to discover new ncRNA genes. To be comprehensive on a genome-wide scale, these methods need to be repeated for a number of growth conditions and multiple replicates need to be performed, resulting in a high overall cost. The development of computational methods to identification of ncRNAs from genomic sequences is attractive since it could be use to focus experimental validation on a fraction of the genome, bringing the overall cost within more practical limits. Several algorithms have been developed, but their low specificity and the lack of a gold standard to assess the quality of their predictions has lead to a lack of consensus on the choice of algorithm [28]. In most ncRNA families, the conservation of the primary sequence is weak and, following the work of Chen *et al.* [44], prediction tools tend to rely mostly on secondary structure searches. A first class of tools searches for a defined structural motif in the genome, examples of this class are ERPIN [39] and INFERNAL [40]. These tools are flexible and fairly efficient but require a precise knowledge of the structural features of the ncRNA family searched for. The computational time is not only proportional to the size of the genome but also to the number of families searched for. To indentify new ncRNA families, thermodynamic profiling experiments [70] have been done based on the assumption that computational predictions of RNA secondary structures will return lower free energies when applied to ncRNA sequences compared to surrounding genomic sequences. This fundamental assumption has been questioned [45] and recent implementations (see zMFold in [53]) tend to remove free energy biases induced by the local dinucleotides composition. Comparative genomics approaches have been found to be sensitive for predicting ncRNAs genes since evolutionary conservation of secondary structures provides further evidence for biological function of the predicted structure. This idea has been exploited in a study by Rivas *et al.* [50] while developing the software QRNA and is also used by the program RNAz [51], combining structural conservation and thermodynamic stability of RNA secondary structures in

multiple sequence alignments. Dynalign [52] also attempts *de novo* prediction of ncRNA, but is unique in computing, for a pair of sequences, the lowest free energy secondary structure simultaneously to the optimal alignment. A recent study [53], showed the importance of background modeling as a neutral model to discriminate real predictions from noise. They also showed that the local dinucleotide frequencies are highly biased between different tools, suggesting that the lack of proper background modeling severely hampers the accuracies of the tools used.

Based the application of four computational tools, we report on the computational screen of the 16 Mbp haploid genome of *C. albicans*. Following this screen, more than 1,700 predictions were experimentally validated using a custom-designed focused tiling array covering 5% of the interrogated genome, resulting in the discovery of 62 putative ncRNAs in *C. albicans*. These results allowed us to compare the accuracies of the tools used, to provide guidance on how to best use these tools and confirm the existence of unannotated ncRNAs in an important human pathogen.

Results

Computational screen

We have performed a computation screen on the *Candida albicans* genome with the following ncRNA prediction methods: INFERNAL / RFAM, QRNA, RNAZ and Dynalign. As expected by the low stringency of the thresholds recommended by the algorithms used, a large number of predictions were obtained from the computational screen. Among the programs applied, some are very computationally intensive, in the most extreme case the application of RFAM / INFERNAL took 1.6 years of CPU time (AMD Opteron 2.2GHz). Thus, the systematic approach chosen would hardly be applicable to larger genomes without access to tremendous computational resources or introducing heuristic filters. To limit the number of regions analyzed, we applied

adjusted cutoffs for each methods based on the distribution of the scores observed. The distributions of the scores were compared and thresholds were determined to separate promising predictions from the method's background noise (see Figure 4). For the four methods tested, we applied the following cutoffs: i) For the INFERNAL/RFAM screen we applied the different levels of cutoff (trusted, gathering and noise) specified for each model by the RFAM team as part of the release 8.0 [42]. This resulted in a total of 132 predictions above the trusted cutoff, 175 above the gathering cutoff and 2731 above the noise cutoff. ii) For the QRNA screen, we applied a cutoff of 5 bits. This threshold has been used in a similar screen on the *S. cerevisiae* genome [71], and resulted in our case in 3603 predictions. iii) All alignments predicted as RNA with a probability above or equal to 0.995 were kept for the RNAz screen and resulted in 1883 predictions. iv) Finally, Dynalign thresholds were specified on a per species basis to obtain a comparable number of predictions between the species compared. We kept all predictions with a score under -7.32 for *C. albicans* vs. *C. dubliniensis*, -1.87 for *C. albicans* vs. *C. glabrata* and -3.99 for *C. albicans* vs. *C. parapsilosis*. This resulted in a total of 12,180 predictions based on Dynalign's results.

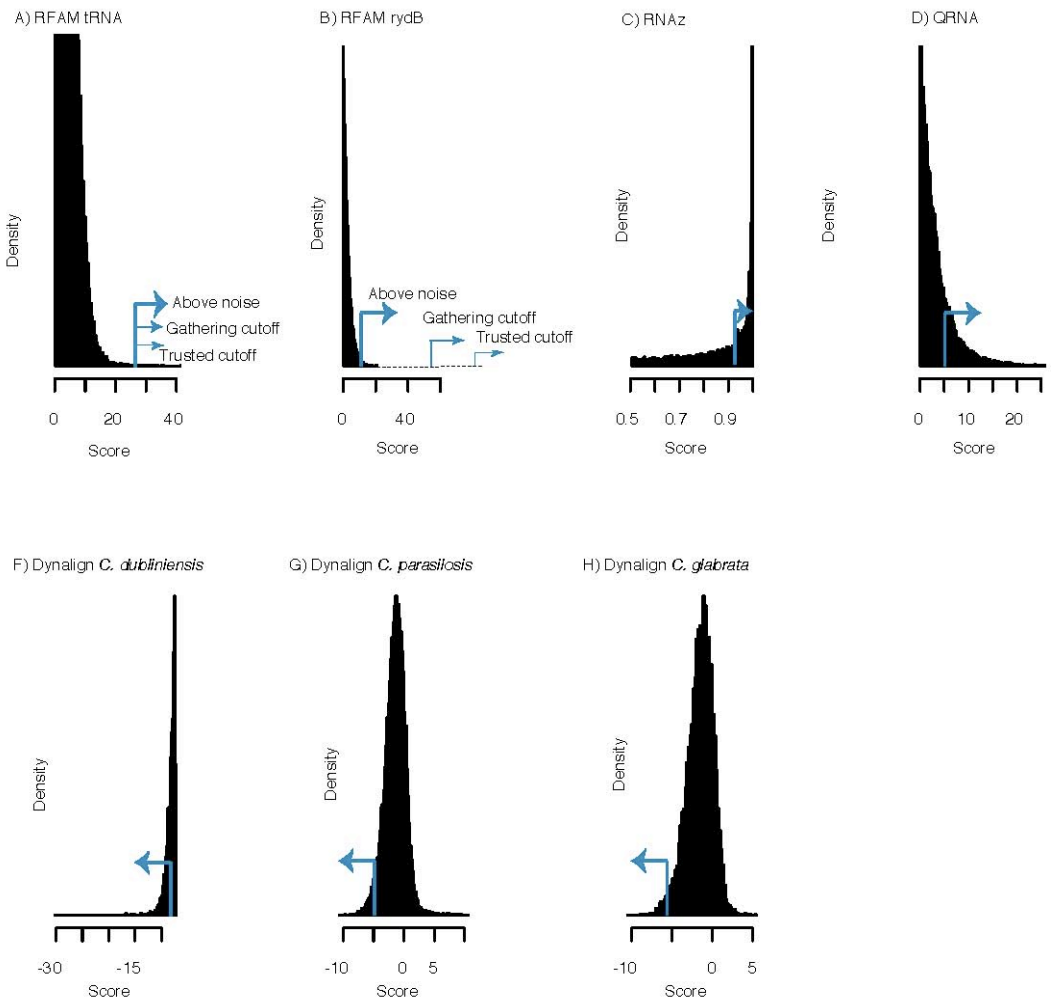

Figure 4: The distribution of scores for each computational method. A) Distribution of scores for the tRNA model of RFAM. The trusted cutoff for this model is 25 (25 for the gathering and 24.99 for the noise cutoff) B) Distribution of scores for the rydB model of RFAM. The trusted cutoff for this model is 75 (50 for the gathering and 12 for the noise cutoff) C) Distribution of scores for RNAz where the selected cutoff for inclusion in the later stages is shown.D) Distribution of scores for QRNA. The distribution of score for QRNA for all the species. To have species specifics distribution see supplementary material. E,F,G) Distribution of score for DYNALIGN. The distribution are species specifics since different cutoff of have been chosen for each species.

As a first attempt to computationally validate our results, we looked at the number of predictions from each method that overlapped with the set of 135 ncRNAs annotated in CGD. This set of annotated ncRNAs represents a very limited number of families but they also correspond to the best-known examples. Among the 135 annotated ncRNAs, 56% (65 / 135) were identified by INFERNAL/RFAM, 30% (40 / 135) by QRNA, 16% (21 / 135) by RNAz and 42% (57 / 135) by Dynalign. Without surprise, RFAM identified the largest fraction from this subset. This can be explained by the presence, in RFAM, of a specific model for tRNAs, the most abundant ncRNA from the CGD set (131 out of 135). The low performance from RNAz is likely to be due to a lower number of predictions that were kept above the chosen threshold. To account for these variations between algorithms, we counted the number of annotated ncRNAs per prediction. We found that, on average, RFAM/INFERNAL yielded 0.021 known ncRNA per prediction, 0.011 for QRNA, 0.016 for RNAz and 0.0047 for Dynalign. Finally, 23% (31 / 135) of the CGD set were not found by any method based on the cutoffs we choose, and a low 2.2% (3 / 135) were found by all four methods. These results indicate that different methods tend to identify a different subset of the known ncRNAs and that consensus between methods can't be used in practice.

To better understand the complementarities between methods, we looked at the number of nucleotides that overlap in two or more sets of predictions. We expected to isolate some of the regions were all methods would agree to detect a structural signal. The observed overlaps are reported in Figure 5 A), the number of nucleotides overlapped by the four methods is slightly below 300 making this criterion far too stringent for any practical purpose. We computed the expected overlap between methods when assuming no correlation between them and found that they still had a fair tendency (average enrichment of 6 fold) to identify similar regions, confirming that they do recognize a common signal. To exclude the possibility that this signal is as trivial as the nucleotide or dinucleotide compositions, we computed these distributions on the sets of predictions found by each method. Figure

5 B) presents these distributions relative to the genome, demonstrating the strong preferences exhibited by the different methods. Similar results were obtained by Babak *et al.* [53] using alignments of known ncRNAs in various species. We observed that the distributions are drastically different for each method but the Dynalign predictions seem to be closest to the set of annotated ncRNAs. This last observation should be taken with the caveat that most of the annotated ncRNAs are tRNAs (131 out of 135), which induces a strong bias on the distributions obtained from this set. Whether this bias should be part of the prediction algorithm (as in [47]) or should be removed (as in [53]) is still up for debate, but it definitely has an impact on the specific sequences identified. Given the fact that different methods yield a distinct set of predictions and that there is no consensus on how to reconcile or explain these differences, this favors the use of a combination of methods to take advantage of their complementarities.

A)

INFERNAL	QRNA	RNAz	Dynalign	% of nt	% expected
				91.8	
x				0.71	
	x			1.17	
		x		2.51	
			x	3.29	
x	x			0.03	0.0084
x		x		0.027	0.0179
x			x	0.364	0.0236
	x	x		0.060	0.0294
	x		x	0.122	0.0387
		x	x	0.063	0.0825
x	x	x		0.003	0.0002
x		x	x	0.003	0.0006
x		x	x	0.004	0.0003
	x	x	x	0.010	0.0010
x	x	x	x	0.001	0.00001

B)

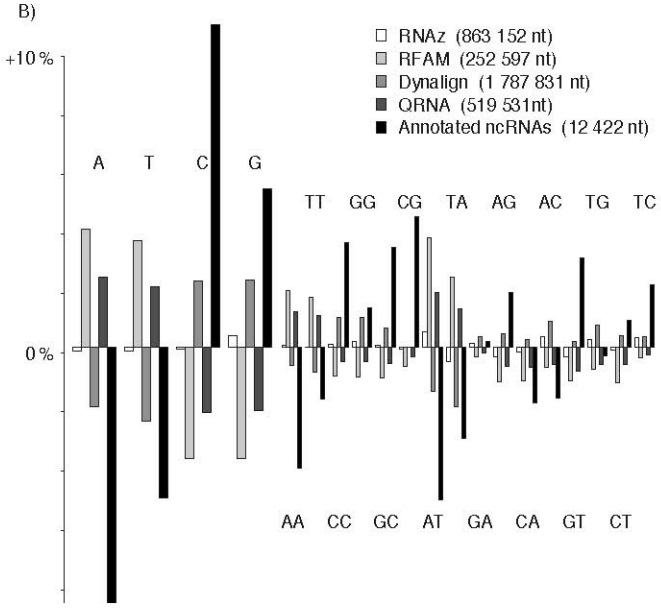

Figure 5: Biases from computational predictions. A) Distribution of the nucleotides between all possible combinations of prediction methods. The expected frequencies indicate the amount of overlap expected in absence of biases. The unusually high observed frequencies for combinations of methods indicate a common bias underlying the four methods tested. B) The distribution of nucleotide and dinucleotide frequencies for sequences predicted by the different methods and annotated ncRNA from CGD. To highlight the variations, the frequencies are reported as a difference compared to the whole genome.

Experimental validations by custom tiling arrays

We designed a 71,968-probes DNA microarray (Roche NimbleGen, Inc.), tiling a set of validation regions using 60-nt probes starting every 20 nucleotides in both directions. The selection of regions to validate was made by keeping the 200 top-scoring predictions for each method, a set of 200 random regions of 200 nucleotides selected to not overlap ORFs, known ncRNAs or predictions, and the set of 135 annotated ncRNAs from CGD (among which 90% are tRNAs). The remaining 1,148 regions were selected by prioritizing regions where several methods overlapped (Fig. 6 shows the details of the tiling strategy). We isolated total RNA from the reference strain SC5314 after standard growth at 30°C. We used direct labeling of the RNAs (see Material and methods) to obtained strand-specific results. For each of the eight biological replicates, we hybridized both a small RNA extraction (see Material and methods) and a total RNA extraction to avoid any possible length-induced labeling bias.

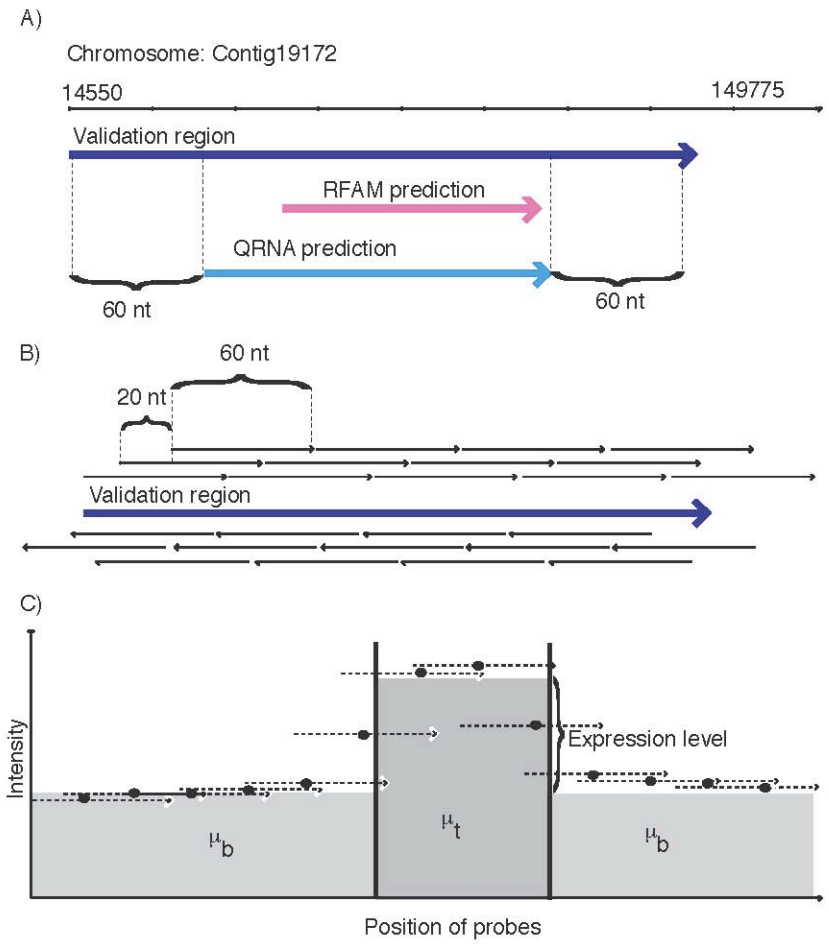

Figure 6: Selection of the validation regions, microarray design, boundaries identification. A) An example of a validation region chosen from an overlap between RFAM and RNAz predictions. The predictions have been extended by 60 nucleotides in both directions. B) The diagram shows the strategy used to tile the validation regions, resulting in the design of 71968 probes averaging 58 nucleotides at every 20 nucleotides on both the forward and reverse strand, validating 1979 regions. C) Typical experimental results obtained for a validation region. 5' and 3' ends of the observed transcript were determined by fitting a model in which partial hybridization between probe and transcript is accounted for. The expression level is the difference between μt and μb, respectively quantifying the transcript and background intensities.

Microarrays intensities were log-transformed and quantile-normalized, keeping the preprocessing step to a minimum. Transcripts were identified using a novel algorithm (see Material and methods) that takes advantage of partial hybridizations between probes and transcripts to yield nucleotide-level transcript boundaries (see Materials and methods). The level of expression is taken relative to local background signal computed for each validation region. Figure 6 C) gives a schematic overview of the approach implemented. Typical results are shown in Figure 7 for six validation regions, highlighting the near absence of noise relative to the expression signal and perfect strand specificity.

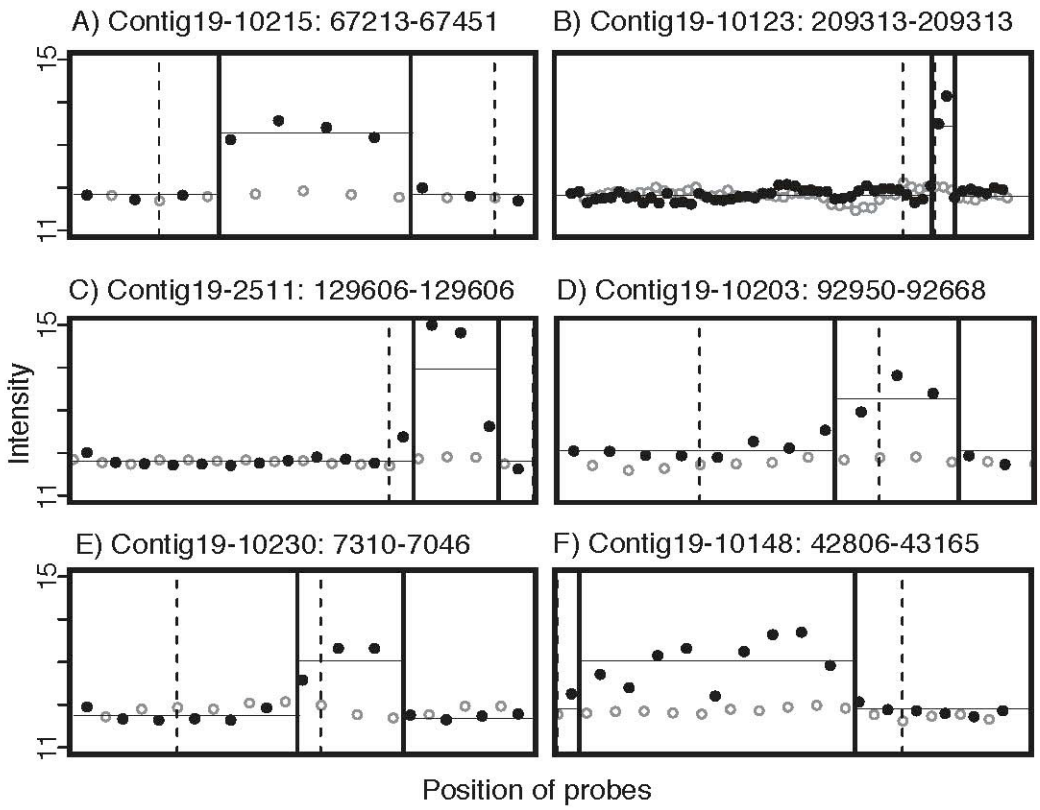

Figure 7: Detailed view of results obtained for six validation regions, showing both the probe-level data and the boundaries obtained from their analysis. Dots are positioned at the center of the probes and indicate the average intensities obtained from eight biological replicates. Black and gray dots are respectively for sense and antisense probes. The vertical lines indicate transcript boundaries obtained by the initial probe-based algorithm (dashed) and by the refined nucleotide-based algorithm (solid). The solid horizontal lines are the μt and μb obtained from the nucleotide-based algorithm. A) A 80-nt intergenic trancript predictied by RFAM (model: Small_nucleolar_RNA_SNORD14). B) A 60-nt intergenic transcript predicted by RNAz. C) A 60-nt transcript antisense to the N-terminal of ORF19.7342 (orthologous to *S. cerevisiae* Axl1). D) A 69-nt transcript antisense to the 3' UTR of ORF19.4392 (orthologous to *S. cerevisiae* Dem1), predicted by Dynalign. E) A 60-nt transcript antisense to the 5' UTR of ORF19.5516 (orthologous to S. cerevisiae Srp72), predicted by QRNA. F) A 192-nt transcript corresponding to the U1 snRNA [26].

Validated transcripts

To identify regions where the level of expression is sufficient to confirm transcription, we used the 99th percentile of the expression level among the negative controls as a threshold. A transcript with an expression level above 0.8 was thus considered expressed. Using this threshold, we identified the expression of 62 novel small RNAs, and confirmed the expression of 105 (out of 135) annotated ncRNAs from CGD. The 62 novel transcripts, tentatively assigned as new ncRNAs, are described in Table 1 and classified in four different groups, depending on their localization with respect to an adjacent coding region: i) 17 transcripts are located in intergenic regions. This classification is based on a median UTR length in *S. cerevisae* of 68 and 91, respectively for 5' and 3' UTR [65]. To include a safety margin, all transcripts located farther away than 150 nucleotides from any ORF were defined as intergenic. ii) 8 transcripts are located in UTRs, one sense to a 3' UTR, 6 antisense to 5' UTRs, and 1 antisense to a 3' UTR. iii) 2 transcripts are located in annotated introns. iv) 35 transcripts are located inside the coding regions (ORFs) themselves, 12 sense and 23 antisense to their ORF.

	μ1-μ2	Contig location		Related ORF	Ortholog *S.cerevisiae*	ORF function
ORFs antisense	2,43	Contig19-2511:123669..123728(-)	x	orf19.7342 (AXL1)	AXL1	Haploid specific endoprotease
	2,15	Contig19-2511:129847..129906(+)		orf19.7345	YGR205	ATP-binding protein of unknown function
	2,02	Contig19-10126:1529..1588(-)		orf19.1788	XKS1	Xylulokinase, converts D-xylulose and ATP to xylulose 5-phosphate and ADP
	1,73	Contig19-2335:6839..6902(-)		orf19.5504	YMR317	Hypothetical protein unknown function
	1,70	Contig19-10183:85824..85883(-)	x	orf19.3572		Hypothetical protein, pas d'orthologue chez Sacc.
	1,61	Contig19-10242:669..739(-)		orf19.6282	NSR1	Nucleolar protein that binds nuclear localization sequences
	1,58	Contig19-2485:51788..51847(-)		orf19.6919		Hypothetical protein unknown function
	1,25	Contig19-10171:21168..21227(+)		orf19.3216		Orf uncharacterized
	1,23	Contig19-10241:26232..26291(+)	x	orf19.6205		Orf uncharacterized
	1,22	Contig19-10064:18903..18962(+)		orf19.661 (KRR1)	KRR1	Essential nucleolar protein required for the synthesis of 18S rRNA
	1,17	Contig19-1748:2390..2449(+)		orf19.1291 (ABZ1)	ABZ1	Para-aminobenzoate (PABA) synthase
	1,16	Contig19-10172:19913..19972(+)		orf19.3228	TMN3	Predicted ORF in Assemblies 19, 20 and 21
	1,08	Contig19-10256:16418..16477(+)	x	orf19.6846 (PHO85)	PHO85	Cyclin-dependent kinase, with ten cyclin partners
	1,06	Contig19-10246:8368..8427(+)	x	orf19.6345 (RPG1A)	RPG1	Subunit of the core complex of translation initiation factor 3 (eIF3)
	1,05	Contig19-10186:46006..46065(+)		orf19.3679	YNL200	Putative protein of unknown function
	1,03	Contig19-10162:51420..51492(+)		orf19.2887 (MET13)	YMR295	Protein of unknown function that associates with ribosomes
	0,99	Contig19-10248:136316..136375(+)		orf19.6525	INP1	Peripheral membrane protein of peroxisomes
	0,98	Contig19-10104:6123..6182(+)		orf19.1229	CSE1	Nuclear envelope protein that mediates the nuclear export of importin alpha (Srp1p)
	0,93	Contig19-10234:63606..63674(+)		orf19.5813	YOR052	Nuclear protein of unknown function
	0,91	Contig19-10212:213934..214006(-)		orf19.4673 (BMT9)		Putative beta-mannosyltransferase,
	0,90	Contig19-10184:95871..95930(-)	x	orf19.3638 (PGA46)		Putative GPI-anchored protein of unknown function
	0,84	Contig19-10161:71914..71973(+)		orf19.2842 (GZF3)	GZF3	GATA zinc finger protein
	0,83	Contig19-10216:90494..90587(+)		orf19.4949		Hypothetical protein unknown function
	0,81	Contig19-10183:85192..85251(+)	x	orf19.3572		Hypothetical protein unknown function
ORFs sense	1,91	Contig19-10208:990..1049(-)		orf19.4488	SWI3	Subunit of the SWI/SNF chromatin remodeling complex
	1,78	Contig19-10174:9600..9659(+)	x	orf19.3366.1		Hypothetical protein unknown function
	1,59	*Contig19-10192:126713..126772(+)*	*x*	*orf19.3838 (EFB1)*	*EF1B*	*Elongation factor EF1 beta*
	1,16	Contig19-10057:341..400(-)	x	orf19.567 (TFB3)	TFB3	Subunit of TFIIH and nucleotide excision repair factor 3 complexes
	1,10	*Contig19-10150:95361..95420(-)*		*orf19.2478*		*Ribosomal L30 unit*
	1,09	Contig19-10215:133274..133333(-)		orf19.4777 (DAK2)	DAK2	Dihydroxyacetone kinase, required for detoxification of dihydroxyacetone
	1,03	Contig19-10123:20829..20888(-)	x	orf19.1614 (MEP1)	MEP3	Ammonium permease of high capacity and low affinity
	1,03	Contig19-10171:20648..20707(-)		orf19.3216		Orf uncharacterized
	0,90	Contig19-1748:2418..2477(-)		orf19.1291 (ABZ1)	ABZ1	Para-aminobenzoate (PABA) synthase
	0,89	Contig19-10129:10767..10826(-)		orf19.1814	STT4	Phosphatidylinositol-4-kinase
	0,87	Contig19-10259:12581..12641(-)		orf19.6926 (CSC25)	CDC25	Membrane bound guanine nucleotide exchange factor (GEF or GDP-release factor)
	0,87	Contig19-10246:6418..6485(-)		orf19.6344 (RBK1)	RBK1	Putative ribokinase
	0,81	Contig19-10256:16443..16502(-)	x	orf19.6846 (PHO85)	PHO85	Cyclin-dependent kinase

Table A: List of expressed transcripts observed using the *C. albicans* focused tiling array. We classified the predicted transcripts in four different groups, depending on their localization with respect to the closest ORF: i) the intergenic group where 17 transcripts are located more than 150 nt away from an ORF which is considered to be the average UTR size in *S. cerevisae*, ii) the UTR group where transcripts are located between the beginning or end of an ORF and 150 nt upstream or downstream, in which one transcript is in sense of 3'-UTR whereas are 6 antisense of 5'-UTR and 1 antisense of 3'-UTR , iii) the intronic groups, including 2 trancripts (in italic), and iv) the ORF group, where 12 transcripts were found in sense of ORFs and 23 found in antisense of ORFs. For each expressed transcript, the exact location (column 3), the adjacent ORF in *C. albicans* (column 4) and its homologue in *S. cerevisiae* (column 5) are given. When it was possible, the ORF function was succinctly described (column 6).

UTRs	1,09	Contig19-10148:42836..43028(+)	orf19.2404,135	POP1	Subunit of both RNase MRP
			orf19.2406,246	GTR2	Putative GTP binding protein
	1,25	Contig19-10203:92826..92895(-) x	orf19.4392 (DEM1),57	DEM1	Defects in cell morphologyprotein of unknown function
	1,28	Contig19-10230:7160..7219(-)	orf19.5515,168	CBP3	Mitochondrial protein required for assembly of ubiquinol cytochrome-c reductase complex
			orf19.5516,65	SRP72	Core component of the signal recognition particle (SRP) ribonucleoprotein (RNP) complex
	1,09	Contig19-10247:27833..27892(+)	orf19.6385,59	ACO1	Aconitase
	0,95	Contig19-10225:93357..93492(+)	orf19.5333 (GCN1),133	GCN1	Positive regulator of the Gcn2p kinase activity
			orf19.5334,153	TSI11	mRNA-binding protein expressed during iron starvation
	0,91	Contig19-10254:192850..192969(+)	orf19.6814 (TDH3),95	TDH3	Glyceraldehyde-3-phosphate dehydrogenase
	0,87	Contig19-10147:27317..27376(+)	orf19.2361,1	SPT10	Putative histone acetylase
	0,80	Contig19-10187:24278..24337(+)	orf19.3701,89		nothing
Intergenics	2,11	Contig19-2479:4630..4763(-) x	orf19.6834,1469		nothing
	1,62	Contig19-10123:209100..209159(-)	orf19.1700 (RPS7A),392	RPS7A	Protein component of the small (40S) ribosomal subunit
			orf19.1701,147	RKI1	Ribose-5-phosphate ketol-isomerase
	1,52	Contig19-10217:9048..9107(-)	orf19.5039,303	RRP42	Protein involved in rRNA processing
			orf19.5040,205	ASM4	Nuclear pore complex subunit
	1,46	Contig19-10215:67292..67372(+)	orf19.4753 (PFK26),1330	PFK26	6-phosphofructo-2-kinase
			orf19.4754 (ZWF1),281	ZWF1	Glucose-6-phosphate dehydrogenase (G6PD)
	1,42	Contig19-2500:25221..25280(+)	orf19.6973,1511	PIM1	ATP-dependent Lon protease
			orf19.6975 (YST1),819	RPS0A	Protein component of the small (40S) ribosomal subunit
	1,40	Contig19-2479:12242..12506(+)	orf19.6834.1 (TAR1),699	TAR1	Mitochondrial protein involved in regulation of respiratory metabolism
	1,20	Contig19-2500:55197..55256(+)	orf19.6983,4872	REG1	Regulatory subunit of type 1 protein phosphatase Glc7p
			orf19.6984,4477	AVT3	Vacuolar transporter
	1,17	Contig19-10176:76470..76543(+) x	orf19.3430,1084		nothing
			orf19.3431,851	MIP1	Catalytic subunit of the mitochondrial DNA polymerase
	1,14	Contig19-10123:33549..33608(+)	orf19.1618 (GFA1),875	GFA1	Glutamine-fructose-6-phosphate amidotransferase
			orf19.1619,1574	CTK1	Catalytic (alpha) subunit of C-terminal domain kinase I (CTDK-I)
	1,07	Contig19-2485:35595..35654(+)	orf19.6911,806		nothing
			orf19.6912,181	CKI1	Choline kinase
	1,00	Contig19-10135:63421..63493(+) x	orf19.1902 (NOC2),340	NOC4	Nucleolar protein
			orf19.1903 (TOR1),460	TOR2	PIK-related protein kinase and rapamycin target
	0,97	Contig19-10163:17464..17523(+)	orf19.2929 (GSC1)3099	GSC2	Catalytic subunit of 1,3-beta-glucan synthase
			orf19.2930,1775	YGR054	Eukaryotic initiation factor (eIF) 2A
	0,97	Contig19-10184:27609..27668(-)	orf19.3606,405	SNA4	Protein of unknown function
			orf19.3607,356	ECM18	EProtein of unknown function
	0,95	Contig19-10215:294902..294961(+	x orf19.4864,710	YJU3	Serine hydrolase with sequence similarity to monoglyceride lipase
			orf19.4865,398	SAC1	Phosphatidylinositol phosphate (PtdInsP) phosphatase
	0,86	Contig19-10236:10299..10619(+)	orf19.5858 (EGD2),306	EGD2	Alpha subunit of the heteromeric nascent polypeptide-associated complex
			orf19.5859 (DAL8),1110	DAL5	Allantoin permease
	0,83	Contig19-2479:8508..8648(+)	orf19.6835,2129		nothing
	0,80	Contig19-10238:48221..48280(-)	orf19.6143,1381	CLG1	Cyclin-like protein that interacts with Pho85p
			orf19.6146 (CLG1),185		nothing

Comparison of method performance

We computed the number of validated transcripts that were predicted as the best 200 predictions for each method. For the 105 transcripts overlapping annotated ncRNAs from CGD, INFERNAL/RFAM predicted 58, QRNA 9, RNAz 7 and Dynalign 0. These results confirms the capacity of INFERNAL/RFAM to recognize known families and seems to favor QRNA and RNAz over the more computationally intensive Dynalign. For the 62 previously unannotated transcripts, the picture is quite different: only 5 were among the 200 best predictions of RFAM, 4 for QRNA, 4 for RNAz and 10 for Dynalign. Our set of annotated ncRNAs being mostly tRNAs, RFAM gains most of its hits by correctly predicting tRNAs within its 200 best predictions, which is of limited interest for the discovery of novel classes. By limiting ourselves to novel transcripts (essentially excluding tRNAs), we showed that Dynalign is able to identify twice more novel transcripts than any other method. Unfortunately, the small sampling observed doesn't allow a Fisher exact test to statistically establish the superiority of Dynalign *vs.* the other three methods.

To better take advantage of the 200 predictions obtained from each method, we designed a second comparison based on a graphical representation similar to a ROC curve (shown in Figure 8). The curves are obtained by first ordering the validation regions according to their expression level (x-axis) and identifying the locations in this ordering of transcripts that were predicted by each method (y-axis). Using this representation a poor prediction method, completely uncorrelated with the observed expression, would follow the dashed diagonal. Deviation from the diagonal toward the upper left corner indicates that the method has a non-random tendency to predict expressed regions. This can be observed, at varying degrees, for the four methods used. Using this representation, the INFERNAL/RFAM method clearly appears as more successful at identifying expressed transcripts. Again, this success is mostly due to the efficacy of the RFAM tRNA model to identify its target with high

confidence (shown with black dots on Figure 8). Among the *de novo* prediction methods (Dynalign, QRNA and RNAz), Dynalign rapidly identifies a larger number of expressed transcripts demonstrating a superior capacity at identifying novel transcripts. The poor performance of RNAz is likely attributed to our lack of manual curation on the multiple alignments used (see Material and methods). The tendency of the negative control curve toward the lower right corner indicates an important bias for identifying regions that are not expressed. This can have two explanations: i) the nonrandom nature of our negative controls, they were specifically selected in intergenic regions and avoided all computational predictions (including the numerous predictions that were not selected for validation). ii) The 1,779 non-random regions were all selected by at least one of the algorithms, thus our dashed diagonal is more representative of the average accuracy of this overall selection method. This observation on the negative control can be interpreted as an indication that there is no pervasive transcription as suggested for higher eukaryotes [22].

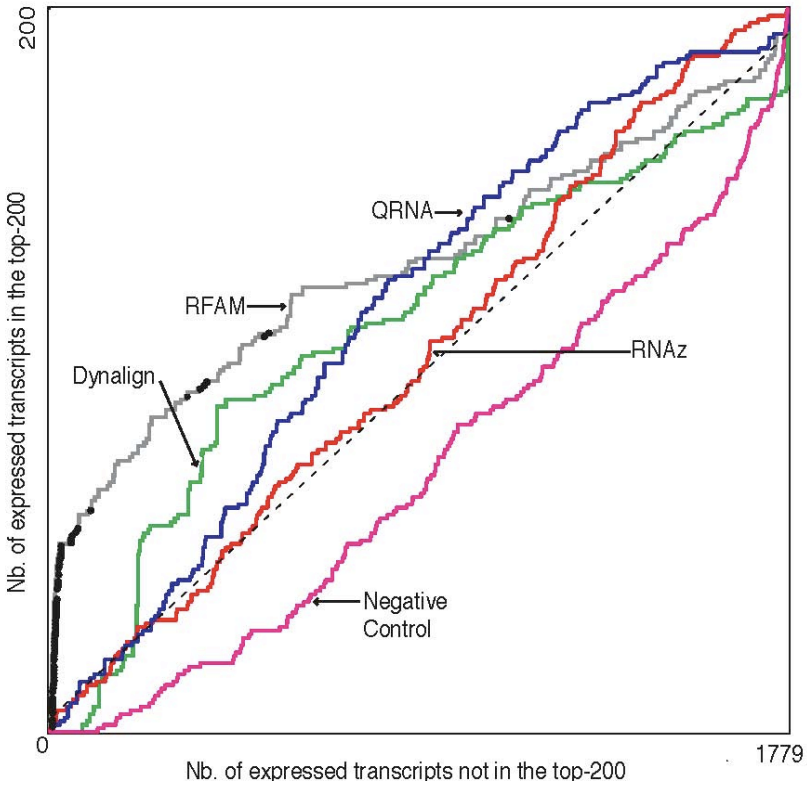

Figure 8: Comparison of accuracies for the four methods. To avoid the complexity of identifying equivalent thresholds for each method, we used a graphical representation similar to the ROC curve. The validation regions were ordered according to their expression level (µt-µb). The curves are obtained by identifying along this ordering the number of expressed transcripts that were predicted by a given method on the y-axis. The accuracy of a method is reflected by how far its curve stretches toward the upper left corner. The results from RFAM are strongly dominated by its facility to quickly identify tRNAs (shown with black dots) with high confidence. Among the *de novo* methods (Dynalign, QRNA, RNAz), Dynalign shows a significantly better capacity for identifying expressed transcripts. The dashed line indicates the expected curve in the absence of correlation between predictions and expression. Our criteria for selecting negative controls (more than 150 nt away from any ORF and not overlapping any prediction) introduced a strong bias against expression.

Despite the overall poor performance of some of the software tested, it is important to notice that if we consider predictions that were unique to each method, 6 out of the 62 new transcripts were only predicted by RFAM, 3 by QRNA, 6 by RNAz and 3 by Dynalign. These results show the complementarities of the methods tested and argues in favor of using a combination of algorithms for computational screens.

Importance of the enrichment of small RNA

To look at the necessity of small RNA enrichment, we compared the expression level difference between the small RNA and the total RNA extractions (see Figure 9). The boundaries of the transcripts were optimized on the small RNA extractions. An important portion of the expressed small RNAs is not expressed in the total RNA, shown in Figure 9 by the vertical spread of data points. This fraction represents most of the annotated transcripts from CGD, confirming the pertinence of using a small RNA extraction. A number of transcripts are observed both in the small and total RNA extractions (shown by the diagonal spread of data points), significantly breaking the trends observed for annotated ncRNAs. This could be explained if these transcripts were sufficiently long to incorporate more fluors (making them brighter than short RNAs in the total RNA extraction), but still sufficiently short (or having cleaved sub-species) to be present in the small RNA extraction. A few known transcripts were identified in this group: the 123-nt snR52 (putative C/D snoRNA), the 264-nt scr1 (putative SRP) and a few regions overlapping the ribosomal subunits 18S (1787-nt) and 25S (3361-nt). The transcripts for which we couldn't identify both boundaries were tagged as "incomplete validation". This reflects the possibility that the actual transcript is longer than observed, but we found no link between this property and whether the transcript were observed in the small or total RNA extractions. Besides raising interesting technical questions regarding the differences between these types of transcripts, these results argue in favor of not using total RNA extraction in future experiments.

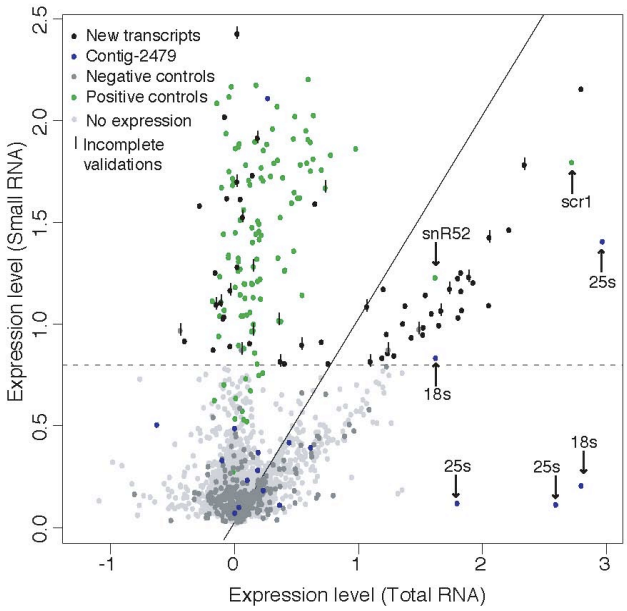

Figure 9: Expression levels obtained from small and total RNA fractions. The horizontal dashed line represent an expression threshold of 0.8, obtained by taking the 99 percentile of the negative controls. In order to obtain comparable values, the expression levels of the total RNA fraction were computed using the same boundaries identified on the small RNA fraction. The solid line represent equal expression levels between small and total RNA fractions. Left of this line are transcripts only identified in the small RNA fraction, very likely to be ncRNAs. Among this group we found most of the annotated small ncRNAs (green dots). Slightly below this line are a number of regions observed both in the small and total RNA fraction. Among these, are transcripts overlapping the 123-nt snR52 (putative C/D snoRNA), the 264-nt scr1 (putative SRP) and ribosomal subunits 18S (1787-nt) and 25S (3361-nt). Likely to be hybridizing with large transcripts, the predictions from Contig19-2479 (covering ribosomal RNAs) have been highlighted in blue. Three regions overlapping the 25S and 18S were appropriately observed only in the total RNA fraction, other predictions from this contig appeared in all regions of the graph. Some transcripts were identified on the edge of the validation regions (vertical dash), resulting in the possibility that the actual transcript is longer than the one identified by our algorithm.

Discussion

Computational screen

Our computational screen of the *C. albicans* genome was performed with the following tools: INFERNAL\RFAM, QRNA, RNAZ and Dynalign, using protocols proposed by the authors or used in the literature for the application of these methods. However, the determination of appropriate score thresholds remained a difficult and arbitrary task. The distribution of scores returned by all methods showed no sign of being separable between background signal and true positives (see Figure 4). In an ideal setting, scores corresponding to actual ncRNAs would appear as clear outliers to the background distribution. In the absence of a reliable noise model for these distributions, we guided our choice of thresholds to have enough predictions to fill our validation microarray. This lack of separation between true positives and the background distributions can be attributed to the weakness of the structure or conservation signals present in ncRNAs. The quality of the structure signal is strongly affected by our limited capacity at accurately predicting RNA secondary structure. This could be refined by introducing advance folding techniques, such as the use of nucleotide cyclic motifs [72] [73]. Unfortunately, this type of approach is, and by a large margin, too computationally intensive to be applied on a genome-wide basis and significant efforts are needed to reach that capacity. The conservation signal depends on the quality of the alignments used, but also on the choice of species included in the comparison. This choice is dictated by a tradeoff between having sufficient variations to identify co-variations (compensatory mutations that preserve a secondary structure) and sufficient similarities to allow for an accurate alignment. Since these features are bound to vary across a genome, an ideal approach would use multiple genomes and for each location on the target genome keep sequences that are likely to yield good predictions. Finally, besides early interesting results using a simple approach [54], no recent efforts were made to take advantage and integrate other

types of signal such as the presence of promoters, transcription factor binding sites or splicing signals. It can be argued that these signals by themselves will be much less discriminating compared to folding and conservation, but would undoubtedly add confidence when used in conjunction with these.

Results of the computational screen demonstrated a surprisingly low overlap between the different methods. We have observed a preference of nucleotide and dinucleotide composition for the predictions of each method, similar to the observations made by Babak et al. [53] in the context of multiple eukaryotes species. Two explanations can be made for the presence of these biases: i) They can arise from the training set used to parameterize the algorithms. In the case of RFAM, these are explicit by the enumeration of the sequences aligned to derive each model. For the other programs, it is more complex to trace back or adjust the training set. ii) They also appear because of the lack of correction with respect to local composition. As an example, Dynalign does make this normalization by using a predicted distribution of scores for dinucleotide shuffles and converts the raw score into a z-score. RNAz also makes this correction but only using the nucleotide composition. Recent work from Gesell and Washietl [74] now allows for this correction in RNAz. Our results suggest that this correction makes predictions from Dynalign closer to known ncRNA in dinucleotide composition while RNAz, QRNA and RFAM are all showing biases very different from the set of annotated ncRNAs. This argues in favor of systematically normalizing scores by local dinucleotide composition, which is also the conclusion obtained by Babak *et al.* from a very different dataset with their zMFold algorithm. In opposition to these results, an interesting study made in prokaryotes [47] has shown that ncRNAs can be predicted by local variation in nucleotide and dinucleotide composition. The approach was very successfully applied in the A-T rich *Methanococcus jannaschii*, where the local shift in CpG frequency was shown to be sufficient to identify ncRNAs with a sensitivity above 97%. In this case, the exact

signal that is used for prediction would be removed by the normalization used in Dynalign, zMFold [53], and proposed for RNAz [74]. Further attention must be devoted to reconcile these observations since it indicates that by normalizing the scores by dinucleotide compositions we are removing a significant signal that could increase the accuracy of the predictions.

Non-coding RNAs in *Candida albicans*

Although the *C. albicans* genome was published with all annotated ORFs, it is likely that non-annotated protein-coding genes still remain that encode for proteins smaller than 100 amino acids. Even if the regions on the array were selected based on their potential to form and conserve secondary structure, we can't rule out the possibility that some of the transcripts observed on our microarrays could be small ORFs. Unfortunately, the simple presence of a short ORF in an expressed transcript represent limited evidence for claiming that the protein product is expressed.

We have identified 62 novel transcripts containing computationally predicted RNA secondary structures, in some cases conserved across multiple species. These transcripts occurred not only in intergenic regions but also overlapping coding regions and untranslated regions of protein-coding genes**,** both in sense and antisense. First, around 77% (101/131) of the annotated tRNAs were found to be expressed in cells grown at 30°C. The 23% remainder could reflect false positives in algorithm predictions or an absence of expression of these regions in standard condition, further tests in multiple conditions will be required to confirm this. The other positive controls, corresponding to the RNAse P, ribosomal subunits and predicted snRNAs, were all expressed at high level, confirming the sensitivity of our approach.

Our custom tiling array represents 4.9% (740,908 out of 15.1 Mbp) of the genome. Extrapolating our results would predict that approximately 1,265 short novel transcripts could still be identified. However, the 4.9% that we have covered in our validation is not representative of the whole genome since it has been selected in a

biased way based on the result of the computational screen. The more accurate our algorithms are, the more optimistic this extrapolation will be. Still, this can be considered as an approximate upper bound to the number of novel transcripts yet to identify. To approximate a lower bound on this number, one could be tempted to use the 200 random regions that we have included as negative controls (0.26% of the genome), with 4 regions being expressed. There are several reasons why these results can't be extrapolated to the rest of the genome: i) Since these regions were used to determine our expression threshold (99^{th} percentile of the observed expression levels, μt-μb), it is by design that we obtain 4 that are expressed (1% of 200 regions analyzed in 2 directions). ii) Using only 0.26% of the genome (200 regions of 200 bp) for this extrapolation will result in an inacceptable approximation. Probes dedicated to these negative controls account for 5.4% of our array (3,942 probes). Increasing this proportion would reduce our focus on regions with potential for expressing ncRNAs. iii) Finally, our criteria to be considered as a negative control included the absence of overlap with any computational predictions. We have shown (see Figure 8) that these selection criteria are enough to significantly avoid expressed transcripts, making these regions inappropriate to represent the whole genome.

Our analyses revealed that one particular transcript was found to be sense to the 5' UTR of orf19.5515, a homologue to *CBP3*, a mitochondrial protein required for assembly of the ubiquinol cytochrome-c reductase complex of the mitochondrial respiratory chain in *S. cerevisiae*. This observation is reminiscent of the regulation mechanism of SER3 in *S. cerevisiae* that is down-regulated by the *SRG1* non-coding RNA [75].The author described the presence of TATA-boxes and upstream activating sequence (UAS) defining the transcription starts of both *SRG1* and *SER3*. Sequence analysis of the region upstream to orf19.5515 revealed TATA-boxes and putative UAS compatible with two transcription start sites, suggesting a similar mechanism regulating the expression of orf19.5515. Two short transcripts were observed and correspond to annotated introns in *C.*

albicans. Both are homologous snR18 and snR39, two C/D box small nucleolus RNAs (snoRNAs) annotated in *S. cerevisiae*. These snoRNAs guide 2'-O-methylation of large subunit (LSU) rRNA at positions A649 and C650 and position A807, respectively. The first snoRNA is found in the orf19.3838, the translation elongation factor EF-1 beta (*EFB1*) and the second one in the orf19.2478.1, the homologous protein to *RPL7A*, a protein component of the large (60S) ribosomal subunit, nearly identical to *Rpl7Bp* and has similarity to *E. coli L30* and rat *L7* ribosomal proteins.

Experimental design

We have found that our experimental design provides a sharp contrast between expression levels derived between sense and antisense (see Figure 7), allowing us to identify with high confidence the strand specificity of all transcripts validated. Two technical decisions distinguish our work from typical tiling arrays application: i) We have used platinum-based chemical labeling of the RNA which do not require reverse transcription or amplification of the RNA, thus completely eliminating the noise, biases and reverse transcription artifacts [76] introduced by these steps. ii) We have selected a platform that offers long-oligonucleotide (60-mer on average) probes which, by increasing the specificity of the hybridization, gives us a high signal-to-noise ratio. The algorithm we developed to identify transcript boundaries at the nucleotide-level alleviates the loss of resolution sometimes attributed to the use of long probes in tiling microarrays.

Comparison of computational methods using experimental validation

Different alignment algorithms were used by the three *de novo* prediction tools: QRNA used WU-blast [60], RNAz used Multiz [61], and Dynalign used MUMmer [77]. It is thus possible that the observed variations in performance are due to the alignment tools rather than to the ncRNA prediction algorithm. For simplicity, we will use the name of the structure prediction tools to identify the systems made by combining the structure prediction and alignment tools.

The performance of Dynalign may be explained by one feature that sets it apart from the other three tools: sequence alignments and RNA secondary structure predictions are computed in a single step. Other *de novo* tools require to first perform a sequence alignment that doesn't take into account structural conservation, then they look for a conserved structure. It is surprising to observe that the transcripts with the highest expression levels (mostly tRNA) are not efficiently identified by Dynalign. The counter performance of RNAz may be due to the quality of the multiple alignments, since the alignments used were not curated by hand. RNAz also requires that at least three species share a given window before it has a chance to evaluate this sequence. With the set of species that was available to us, algorithms relying on pairwise alignments (QRNA and Dynalign) were more accurate and a large fraction of their predictions were obtained from comparison with a single species, *C. dubliniensis*. The analysis presented in Figure 8, besides providing a threshold-free comparison of the computational methods tested, indicates that these methods clearly tend to identify regions that are expressed when compared with the results obtained from random intergenic regions. Since all four methods are guided by two signals: the presence of RNA secondary structure and sequence conservation, these expressed transcripts are biased toward structured and evolutionarily conserved RNAs. This observation is made even more robust by considering (see Figure 5) that the predictions made by the four algorithms are poorly correlated. This evidence for the presence of conserved secondary structure in expressed transcripts provides a reasonable argument indicating that these transcripts are under selective pressure and thus functional.

Conclusion

With high-throughput sequencing of transcriptome becoming more accessible, future applications of ncRNA prediction algorithms will shift from the detection in genomic sequences toward recognition and classification from transcriptome sequencing projects. To this end, computationally more intensive approaches [73] will

become usable, and efforts should be devoted to develop unsupervised classification of novel transcripts. For genomes of small sizes, whole genome tiling array are also becoming popular tools and the need is already urgent for refined analysis of raw intensities and downstream computational analysis of expressed transcripts. Examples of this are the 17 transcripts found to be expressed in intergenic regions: while they represent more than a quarter of the novel transcripts we have observed, we have currently no starting point as to their possible function. There is currently an important gap between what can be rapidly inferred from protein sequences compared to what is possible with RNA sequences. Finally, we should emphasize that we have demonstrated that an accessible combination of computational approaches and microarray-based validations can be used to specifically identify novel transcripts in an important human pathogen.

Materials and Methods

Computational screen

A computational screen for ncRNAs was performed with the following software: INFERNAL/RFAM [42], QRNA [50], RNAz [51]and DYNALIGN [52]. These tools were chosen among the tools available to represent a variety of approaches and thus increase the diversity of predictions obtained. We also knew that the parameterization of the various tools was made from variety of species (most being trained with either prokaryotes or vertebrate sequences) and it was unclear which would best perform in a unicellular fungi. Except when specified, all methods were ran on assembly 19 of the *Candida albicans* genome [3] with repeated regions masked with DUST (DustMasker program [59]). Since allelic variants were minor and to reduce computational time, we removed from the genomic sequences contigs corresponding to allelic copies. The final size of genomic sequence pool was 15,1 Mbp, masking with DUST removed 4.3 % (649080 out of 15098438) of the genomic sequence.

The first tool used was RFAM/INFERNAL. RFAM is a collection of ncRNA families covariance models from eukaryotes and prokaryotes, we used version 8.0, containing 574 models [42]. Infernal-0.81 is the software that searches genomic sequences for occurrences of these models using stochastic context-free grammars (SCFGs) [40].We independently searched each model vs. the genomic sequence, and to maximize sensitivity we did not applied any filter based on homology.

The second tool used was QRNA-2.0.3 [50]. This algorithm identifies conserved secondary structures between two species. The genome was split in sequences of 1500 nucleotides overlapping by 150 nucleotides. These sequences were aligned with the genomes of the following closely related species: *C. dubliniensis, C. tropicalis, C. parasilosis, C. guilliermondii, C. lusitaniae, Debaryomyces hansenii, C. glabrata*, and *S. cerevisiae* with WU-blast-2.0 using default parameters. QRNA was run on the subset of those alignments longer than 100 nucleotides and showing between 75-90 % identity. These alignments were further divided into sequences of 150 nucleotides overlapping by 50 nucleotides and then submitted to QRNA. From the output of QRNA, we kept all alignments predicted as structured RNA. The log-odds posterior scores distributions were then computed for each compared species and used to determine a species-specific threshold.

The third tool used was RNAz-1.0 [51]. This program predicted structurally conserved and thermodynamically stable RNA secondary structures in multiple sequence alignments. Using multiple alignments results in more information being available to the algorithm to identify structure conservation. Using sequences of *C. albicans* of 500 nt overlapping by 150 nt, we did multiple alignments with Multiz [61] with species *C. dubliniensis, C. tropicalis* and *C. parasilosis*. We decided to limit ourselves to those species, because when more distant species were used (we experimented with: *C. guilliermondii, C. lusitaniae, Debaryomyces hansenii, C. glabrata*, and *S. cerevisiae*), we obtained poor multiple alignments due to lower conservation. RNAz was run on

alignments longer then 50 nucleotides including gaps. For each alignment, RNAz computes a Z-score using a support vector machine regression, all alignments that were predicted as RNA were kept.

The software DYNALIGN[52] was designed to predicts RNA structure by combining in a single step free energy minimization and comparative sequence analysis to find a low free energy structure common to two sequences. Results from the C. ablicans screen were provided by R. Tyagi and D. Mathews (Rochester University). The genome alignments of *C. albicans* against *C. dubliniensis, C. glabrata* and *C. parapsilosis* were made with Mummer. DYNALIGN [52] was ran on the assembly 20 of *Candida albicans* and predictions were transposed on assembly 19 using blast (exact match). 1.89% (6996 out of 369930) of the predictions were not mapped to assembly 19 and couldn't be used for the later stages of analysis. The z-score distributions were then computed for each compared species and used to determine a species-specific threshold.

We intentionally kept very permissive thresholds on the various algorithms, resulting in a large number of predictions. Predictions were then prioritized for validation according to the following procedure. The validations of the 200 predictions with highest scores for each method were first included to allow for a quantitative comparison of the accuracy of these methods. For this comparison the predictions found in ORFs were kept. These regions represented 32.5 % of the probes present on the validation microarray. The remaining validation regions were chosen by prioritizing regions containing predictions that overlap between methods, and that are not contained in an ORF or in a low complexity region. These represent the bulk of the probes present on the array, 58 %. As a set of negative controls, we also added in our design 200 regions of 200 nucleotides in intergenic regions (more then 150 nucleotides from the beginning and end of ORFs) that are not contained in the predictions sets (represents 9.7 % of the probes). The tRNA and rRNA already annotated in CGD were also added has positive control (2.4 % of the probes). Overlapping validation regions were merged to obtain a total of 1979 regions, including

controls, with an average of 374 nucleotides covered per region. The number of predictions selected for experimental validation induced by the results of RFAM is 565, QRNA is 1505, RNAZ 1312, and Dynalign 2725.

Microarray design

A DNA microarray based on a focused tilling design was built and used to validate the computational predictions. For each region to validate, we added 60 nucleotides at the beginning and the end of the region. For each expanded validation region we designed probes of 60-nt at every 20-nt, in both the forward and reverse direction. The resulting array was synthesized by Nimblegen (Roche NimbleGen, Madison, WI) to validate 1979 predictions regions using 72000 unique probes.

Strains, media, culture conditions

The reference SC5314 *C. albicans* strain was used to validate expression of predicted ncRNAs since it corresponds to the strained used to derive the genomic sequence used for our computational analyses and array design. The cells were grown at in yeast extract / peptone / dextrose (YPD) with glucose to mid-log phase (30°C, 240 rpm). The fresh cell pellets were then crushed with liquid nitrogen using mortar and pestle and instantaneously frozen and stored at -80°C prior to RNA extraction to maintain RNA integrity.

Nucleic acid extractions and labeling

Total RNA was extracted using the TRIZOL® protocol (Invitrogen, Burlington, ON). The small RNA enriched fraction was obtained using the MirVana isolation kit (Ambion Biosystems Canada, Streetville, ON) following standard conditions described by the manufacturer except for the second RNA precipitation, in which one volume of isopropanol was used. The purity and concentration of RNA samples were determined from A260/A280 readings and RNA integrity was determined by capillary electrophoresis using an RNA 6000 Nano Laboratory-on-a-Chip kit and Bioanalyzer

2100 (Agilent Technologies, Santa Clara, CA) per the manufacturer's instructions. 7µg of total RNAs and small RNAs were direct-labeled using 7µl of Alexa fluor 532 and Alexa fluor 647 with the Ulysis nucleic acid labeling kits (Invitrogen, Burlington, ON) following the standard protocol described by the manufacturer excepted for the labeling reaction where, for RNA, the incubation was performed at 90°C for 15 minutes. The labeled RNAs were separated from unincorporated dyes using a BIOSPIN P30 column (Bio-Rad, Hercules, CA) and ethanol precipitated at –20°C overnight. Arrays were hybridized following a dye swap design to eliminate any bias related to the dye / nucleic acid labeling. Hybridizations were performed on the MAUI Hybridization System (BioMicro System Inc, Salt Lake City, UT) which allows for mixing of the labeled probes on the array surface, at 42°C overnight. Three different stringency washes were performed using Nimblegen Hybridization and Wash buffers (Roche Nimblegen, Madison, WI) following standard conditions for 4-plex arrays. Scanning was performed on a GenePix® 4000B scanner after adjusting the dye ratio at approximately 1. Image acquisitions and fluorescence intensities were obtained using GenePix Pro 6.0 and NimbleScan 4.0, respectively (Roche Nimblegen, Madison, WI). This protocol was applied to 8 biologically independent replicate to derive statistically robust results.

Data processing and analysis

We first performed a logarithmic transformation on the raw data. To remove array-specific biases, a quantile normalization was then applied on the two sets of eight arrays chips, keeping small and total RNAs apart.

To identify transcripts in our validation regions, we developed an algorithm specifically designed to identify a single transcript within a validation region. Our model relies on the assumption that within a region all probe intensities are obtained from a linear combination of two intensities: a background level observed for probes hybridizing to no specific transcript and a "transcript" level observed when a probe

hybridizes a transcript over its whole length. We observed that probes located in-between background-level and transcript-level probes displayed an intermediary intensity value and postulated that it results from a partial complementarity between the probe and transcript. We modeled the probe intensity from probe *i* of the region analyzed by: $Xi = f(wi)\,\mu t + (1 - f(wi))\,\mu b$ where *wi* is the fraction of that probe that hybridizes to the transcript, $f(\cdot)$ is a function that converts this fraction into a linear weight, μb and μt are respectively the background and transcript level intensities in the region. Since our validation regions are relatively small, we evaluated the complete space of transcript boundaries at the nucleotide level in each region, returning the set of boundaries that best fit our model over our set of eight biological replicates. The best fit was based on minimizing the sum of squared residuals, SSR, between the predicted intensities from the model, *Xi*, and observed probe intensities, *Oi*. To avoid signal obtained from a single outlier, we added the constraint that boundaries should be at least 60 nt apart. For each set of boundaries evaluated, the SSR is computed as follows:

$$\begin{aligned}
\mu_t &= \frac{\sum_i f(w_i)O_i}{\sum_i f(w_i)} \\
\mu_b &= \frac{\sum_i (1 - f(w_i))O_i}{\sum_i (1 - f(w_i))} \\
\mathrm{SSR} &= \sum_i (O_i - X_i)^2 \\
&= \sum_i (O_i - [f(w_i)\mu_t + (1 - f(w_i))\mu_b])^2
\end{aligned}$$

By first assuming that $f(wi) = wi$, we observed that intensities observed on the array follows a sigmoid relation. Determination of the best fitting parameters for the sigmoid function was obtained by implementing an EM algorithm [67].

Acknowledgments

We would like to acknowledge the assistance of Patrick Gendron and Geneviève

Boucher for providing database development and support that facilitated the work presented here. We are grateful to Rahul Tyagi and David H. Mathews (University of Rochester) for providing us with their *C. albicans* screen results using Dynalign. This work was supported by the Canadian Institutes of Health Research (CIHR) Team Grant on Fungal Pathogenesis (CTP-79843). IRIC is supported in part by the Canadian center of excellence in commercialization and research (CECR), the Canada Foundation for Innovation (CFI) and by the Fonds de Recherche en Santé du Québec (FRSQ).

Chapitre 4 : Nouvelle méthode de détection d'ARNnc

Introduction

Mise en contexte

Au premier chapitre, nous avons présenté plusieurs approches informatiques d'indentification d'ARNnc. Toutefois, ces méthodes permettent seulement d'identifier des ARNnc d'une classe connue (c'est-à-dire pour laquelle la structure est connue) ou qui sont bien conservés d'espèce en espèce. Un des objectifs de notre recherche étant de trouver des ARNnc constituant de bonnes cibles thérapeutiques chez *Candida albicans,* le gène correspondant devrait préférablement ne pas être conservé chez l'humain.

Nous aimerions donc trouver de nouveaux ARNnc propres à *Candida albicans*, car si un ARNnc est fortement conservé chez les levures, les chances sont bonnes pour qu'il le soit également pour l'ensemble des eucaryotes. Nous voulons développer une méthode *de novo* d'identification des ARNnc, sans utiliser la comparaison de séquences entre espèces ni de modèle pré-établi pour trouver des ARNnc non conservés.

La méthode proposée se base sur l'idée que les ARNnc sont possiblement plus structurés que des régions sans fonction. La prédiction d'ARNnc à l'aide de la thermodynamique étant controversée [44], nous proposons donc une méthode alternative pour trouver de l'ARN structuré. Notre approche observe d'abord la distribution des motifs structuraux d'ARN qu'il est possible de former à partir d'une séquence génomique et la distribution est alors employée pour discriminer si la séquence génomique contient un signal d'ARN structuré. La contrainte aux classes de motifs considérées est le *nucleotide cyclic motif* (NCM), tel que présenté par le groupe de François Major [60,72] (voir Figure 10).

Hypothèse

Nous posons l'hypothèse que la distribution des NCM diffère si une structure secondaire se cache dans une séquence analysée. Nous espérons ainsi pouvoir trouver un biais dans la distribution des NCM des séquences d'ARNnc. Si l'hypothèse est

bonne, nous tenterons de déterminer s'il y a une variation locale des compositions de NCM dans le génome de *C. albicans* (dans des fenêtres de 100 nt) peut révéler la présence d'un ARNnc [72].

Méthodologie

Les nucleotide cyclic motif (NCM)

Voici la façon de définir les NCM : Une structure secondaire d'ARN peut être représentée par un graphe où les sommets correspondent aux nucléotides, et les arêtes aux relations d'adjacences ou aux paires de bases entre les nucléotides. Ce graphe peut être décomposé en cycles, c'est-à-dire un chemin d'arêtes consécutives dont les deux sommets aux extrémités sont le même. Les NCM sont les cycles minimaux d'un graphe d'ARN, c'est-à-dire que ces cycles ne sont pas décomposables en plus petits cycles. Nous avons choisi d'utiliser ces motifs car ils représentent les blocs de construction d'un ARN structuré et modélisent ainsi les interactions secondaires (canonique et non canonique) que l'on peut observer dans l'ARN.

La méthode observe les NCM à quatre ou cinq nucléotides et une ou deux appariements produites à partir d'une séquence d'ARN (voir Figure 10 pour des exemples de NCM). Nous avons choisi ces NCM car ceux-ci sont parmi les plus observés dans l'ARN structuré.

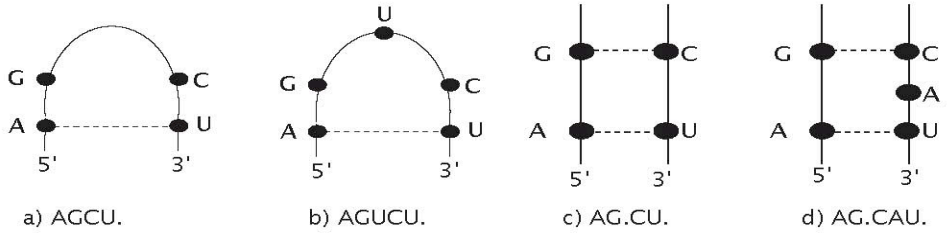

Figure 10: Représentation graphique des quatre types de NCM étudiés. Les lignes pleines représente une adjacence et une ligne pointillé un appariement. a) Un NCM de 4 nucléotides avec un appariement. b) Un NCM de 5 nucléotides avec un appariement. c) Un NCM de 4 nucléotides avec deux appariements. d) Un NCM de 5 nucléotides avec deux appariements. Sous chaque représentation graphique se trouve la correspondance textuelle utilisée pour représenter un NCM. Dans cette représentation, les nucléotides adjacents représentent une adjacence dans la structure alors qu'un point représente un appariement. Un point en fin de séquence représente un appariement avec le premier nucléotide.

Génération des distributions de NCM

À partir d'une séquence génomique donnée, nous avons énuméré tous les NCM de 4 et 5 nucléotides et une ou deux paires que la structure secondaire de cette séquence peut théoriquement contenir. Par exemple pour la courte séquence AUAGCA on retrouve les NCM suivant:«AUAG.»,«UAGC.», «AGCA.», «AU.CA», «.AU.GC»et « UA.CA » .

Les NCM sont générés de la façon suivante (voir Figure 11): pour une séquence génomique donnée, nous avons énuméré toutes les sous-séquences de deux et trois nucléotides, puis effectué toutes les combinaisons possibles entre ces sous-séquences (à l'exception des combinaisons entre les sous-séquences de 3 nt et 3 nt). Ceci représente les NCM de 4-5 nucléotides avec deux appariements. Pour les NCM à un seul appariement, nous avons énuméré toutes les sous-séquences de 4 et 5 nucléotides. Un même NCM peut avoir plusieurs représentations textuelles à cause de l'arrangement spatiale des nucléotides. Par exemple le NCM suivant : « UA.CA » peut être représenté par les façons textuelles suivantes « A.CA.U », « CA.UA », « A.UA.C ».

Pour cette raison nous avons canonisé chaque NCM, c'est à dire que pour chaque NCM, une liste de toutes ses représentations linéaires possibles est donc générée et triée en ordre alphabétique, pour finalement retourner le premier de la liste. De cette façon, c'est toujours la même représentation linéaire d'un même NCM qui est retournée. Ces NCM sont stockés dans une structure de données (une *map*) avec le nombre d'occurrences dans la séquence. Le résultat final est une distribution de NCM observée. L'algorithme pour générer la distribution de NCM est dans l'ordre de complexité de N^2, où N est le nombre de nucléotide sde la séquence, le temps pour générer une distribution de NCM pour une séquence de 100 nt est d'environ 0.169 secondes.

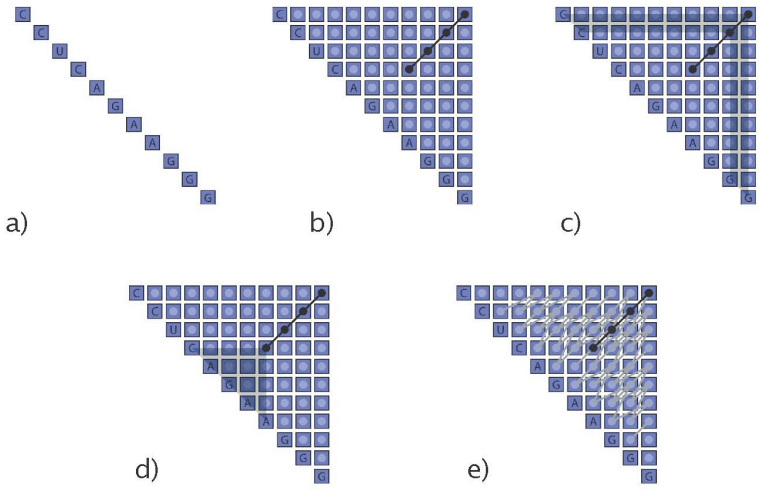

Figure 11 : Génération d'une distribution de NCM. a) Pour une séquence génomique, b) dans une représentation de matrice de points, une structure secondaire d'ARN est un ensemble de liens entre les paires (montrée en noir). c) Chaque lien interne est un NCM à deux appariements (en mauve foncé) d) et les liens terminaux sont des NCM d'un appariement (boucles). e) Tous les NCM possibles de un et deux appariements peuvent être obtenus par l'énumération de tous les liens de la matrice.

Application au génome de *C. albicans*

Nous présentons ici le protocole développé pour la recherche des ARNnc dans le génome de *C. albicans*. Premièrement, nous avons établi un modèle nul pour la distribution des NCM à partir des régions intergéniques* du génome de *C. albicans*, découpées en 17 000 séquences de 100 nucléotides. La taille de 100 nucléotides est, quant à elle, représentative de la taille moyenne des ARNnc tout en offrant des temps de calculs raisonnables. Pour construire ce modèle nul, nous avons supposé que toutes les régions intergéniques non répétées ne sont pas structurées, et nous avons retiré les séquences présentant des régions de basse complexité avec l'outil « DUST » [56]. Ensuite, nous avons calculé les scores de χ^2 (chi-carré) entre notre modèle de distribution de NCM de séquence intergénique et chacune des distributions de NCM obtenues à partir de séquences qui ont servi à construire le modèle. Nous utilisons ici le score de χ^2 de Pearson afin de déterminer si la distribution observée des fréquences de cycles pour la séquence peut avoir été engendrée par le modèle. Le score χ^2 (équation 1) prend en entrée une distribution de probabilité de cycles d'un modèle, *E*, et une table de fréquence de cycles d'une séquence, *O* :

$$X^2 = \sum \frac{(E_i - O_i)^2}{E_i} \qquad (1)$$

où Ei = probabilité du NCM i (pris du modèle) x Nb de NCM totaux de la séquence et Oi = Nb de NCM i de la séquence.

Pour mieux analyser les scores obtenus nous avons calculé une *p-value* * empirique

afin de quantifier le niveau de pertinence de ce score. Pour ce faire, nous avons besoin d'un grand nombre de scores, soit 1×10^6 séquences. Le génome de *C. albicans* n'est pas assez grand pour avoir un million de séquences intergénique de 100 nt différente (le génome compte 15.1×10^6 pb), nous avons généré 1×10^6 séquences à partir d'une chaîne de Markov d'ordre 0. La chaîne de Markov a été entraînée sur les régions intergéniques de *C. albicans*. Des séquences générées par une chaine de Markov d'ordre 0 nous permet d'avoir la même composition en nucléotides que les séquences du génome de *C. albicans*. Nous avons calculé les 1×10^6 scores de χ^2 entre le modèle neutre et les séquences générées à partir de la chaîne de Markov d'ordre 0. Ensuite, nous avons déterminé la p-value en calculant le nombre de fois que des scores plus hauts sont observés parmi les 1×10^6 scores de Markov. La Figure 12 résume la méthodologie que nous avons utilisée. Comme contrôle positif nous avons également calculé les *p-value* des scores du modèle neutre contre des séquences de d'ARNt (131 séquences) et comme contrôle négatif des séquences aléatoires généré par une chaine de Markov d'ordre 0.

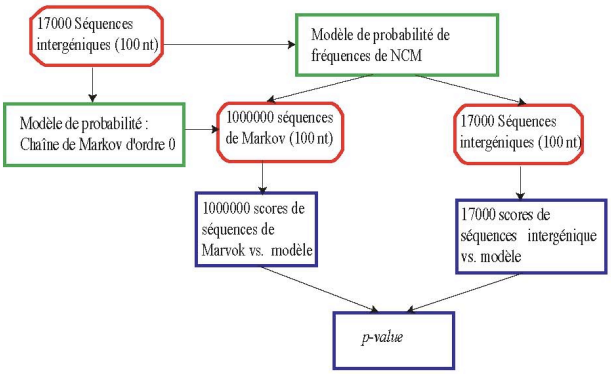

Figure 12: Résumé de la méthode de comparaison des fréquences observées de NCM au modèle de probabilité de NCM pour obtenir une *p-value*. Les boîtes rouges représentent les ensembles de séquences, les boîtes bleues représentent les distributions de score et les boîtes vertes sont les modèles de probabilités.

Application au génome de *S. cerevisiae*

L'application a été répétée avec des séquences *S. cerevisiae* pour tirer profit d'une meilleure annotation des ARNnc (418 ARNnc). Cette fois-ci le modèle nul a été construit avec des séquences intergéniques de *S. cerevisiea* dont l'ordre des dinucléodites a été changé de façon aléatoire (*dinucleotide shuffle*). Nous avons comparé les scores χ^2 du modèle nul des distributions de NCM avec des séquences intergéniques, des séquences ARNnc et des séquences aléatoires.

Résultats

Résultats de l'application chez *C. albicans*

Nous avons regardé la distribution des *p-values* pour les scores de χ^2 obtenus en comparant des distributions de NCM pour des séquences du génome de *C. albicans*, des séquences d'ARNt et des séquences aléatoires contre la distribution modèle établie (voir la Figure 13). Les séquences avec des *p-values* plus basses sont les séquences les plus différentes du modèle de probabilité. Nous avons observé que les ARNt annotés ont des *p-values* basses et nous avons aussi remarqué qu'un certain nombre de séquences intergéniques montrent une distribution semblable. Ceci nous indique que notre modèle permettre de distingué les ARNt et également qu'un certains nombres de séquences intergénique montre des différences du modèle nul. Comme prévu, le contrôle négatif (séquences aléatoires) produit une distribution uniforme de *p-value*.

Donc, nous semblons voir un biais dans les distributions de NCM pour les séquences de ARNt vers des valeurs basses de *p-value* et donc on pourrais croire qu'il ont une tendance à diverger d'un modèle nul de NCM. Un certain nombre de séquences intergéniques semble avoir le même biais, donc on pourrait penser que ces séquences contient de la structure comme les séquences d'ARNt.

Figure 13: Les résultats de distribution de NCM dans *C. albicans*. La distribution des *p-*

values pour les scores de χ^2 des séquences intergéniques sont en noir, les séquences neutres sont en orange et des séquences des ARNt sont en bleu. Nous pouvons observer que les séquences ARNt annotés des *p-values* basses.

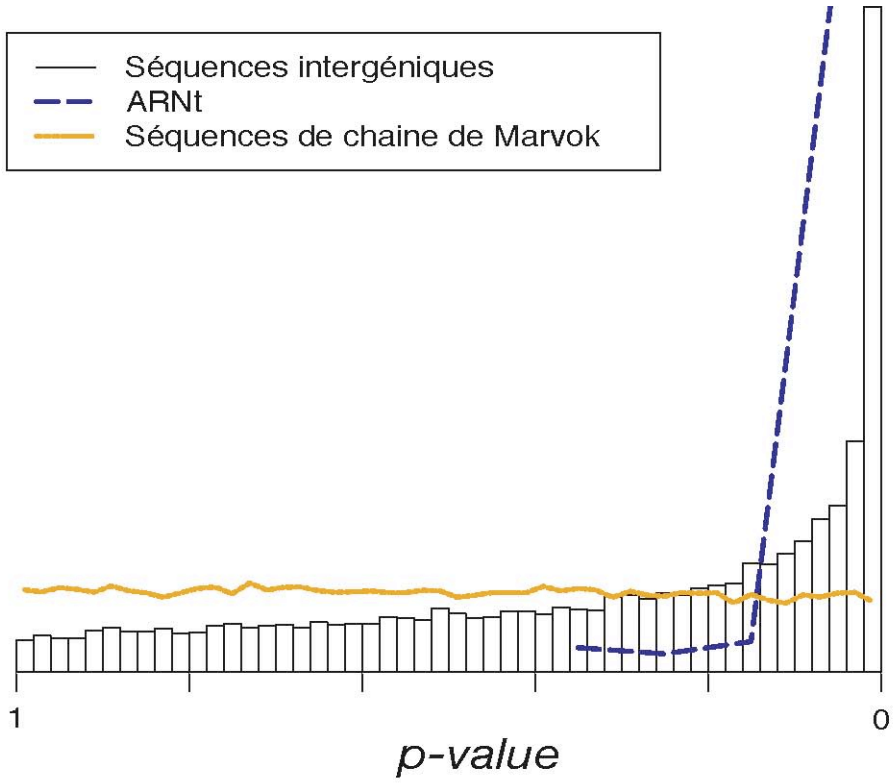

Figure 13: Les résultats de distribution de NCM dans *C. albicans*. La distribution des *p-values* pour les scores de χ^2 des séquences intergéniques sont en noir, les séquences neutres sont en orange et des séquences des ARNt sont en bleu. Nous pouvons observer que les séquences ARNt annotés des *p-values* basses.

Résultats de l'application dans *S. cerevisiea*

Nous avons calculé la distribution des scores de χ^2 obtenue en comparant des distributions de NCM pour des séquences intergéniques, des séquences d'ARNnc et

des séquences aléatoires du génome de *S. cerevisiae* contre la distribution modèle nul établie contruit de séquences intergéniques dont l'ordre des dinucléodites a été changé de façon aléatoire (voir Figure 14). Cette fois-ci, nous n'observons pas de biais dans les distributions de NCM d'ARNnc et des séquences intergéniques. Pour savoir si ces observations étaient dûes à l'organisme ou à la différence de modèle neutre, nous avons répété l'application avec des séquences de *C. albicans*, c'est-à-dire avec un modèle nul construit avec de séquences intergéniques de *C. albicans* dont l'ordre des dinucléodites a été changé de façon aléatoire (voir Figure 15).

Les résultats pour l'application *C. albicans* avec comme modèle nul un *dinucléotide shuffle* démontre qu'on ne distingue pas de différence entre les scores obtenus des ARNnc et des séquences intergéniques. Ceci suggère les distributions de dinucléotide joue un rôle important dans la détection des ARNnc.

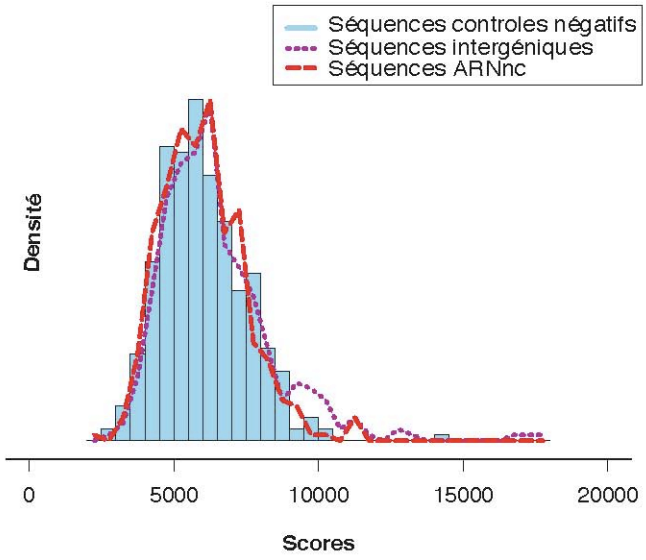

Figure 14: Les résultats des expériences de distribution de NCM dans *S. cerevisiae*. La distribution des scores de χ^2 de séquences intergéniques est en mauve, des séquences d'ARNnc est en rouge et des séquences aléatoires est en bleu. Dans cette expérience, le modèle nul utilisé est construit à partir de séquences où la composition en dinucléotides a été conservée.

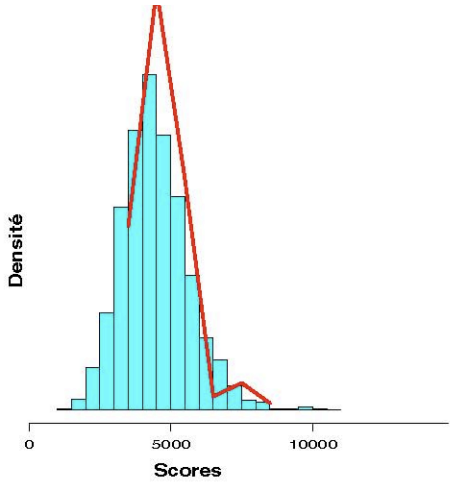

Figure 15 : Les résultats de l'application de distribution de NCM dans *C. albicans* avec un modèle nul construit à partir de séquences intergénique où l'ordre de composition en dinucléotides à été mélangé aléatoirement. La distribution des scores de χ^2 de séquences intergéniques est en bleu et celle des séquences d'ARNnc est en rouge.

Discussions et conclusion

Dans l'application chez *C. albicans* avec un modèle nul construit à partir de séquences intergéniques, nous avons observer une différence dans la composition en NCM pour les séquences de ARNt. Quand le travail a été refait dans le génome *S. cerevisiae* et de *C. albicans* avec comme modèle nul des séquences dont l'ordre des dinucléotides a été changé, on n'observe pas de biais dans les compositions en NCM. Ceci suggère que la distribution de dinucléotides porte la majeure partie du signal que nous avons observé dans les séquences de ARNt de *C. albicans*. Le biais de composition en dinucléotides dans les ARN peut être expliqué par l'adjacence des appariements de nucléotides qui est importante pour l'empilement* d'appariements qui contribue à la stabilité des structures secondaire et tertiaire d'ARN.

Cette méthode nous appris l'importance des distributions de dinucléotide et d'avoir un bon modèle nul dans la recherche des ARNcn, mais nous ne poursuivrons à explorer l'approche de distribution de NCM. La méthodologie est plutôt compliquée pour seulement en retirer des statistiques sur les distributions de dinucléodites et trinucléotides entre des séquences intergéniques et des séquences d'ARNnc. Malheureusement, le signal des distributions de dinucléodites et trinucléotides n'est pas un puissants prédicteurs d'ARNnc.

La problématique de développer une méthode *de novo* pour l'identification d'ARNnc sans utiliser la comparaison de séquences entre espèces reste ouverte. Les distributions de dinucléotides doivent probablement être pris en considération pour trouver une solution à cette problématique. Nous en discuterons d'avantage au chapitre 5 de l'importance des dinucléotides dans la recherche d'ARNnc et d'autres voix à explorer pour trouver une solution à cette problématique.

Ce travaille à été fais en collaboration avec Rémi Planel dans le cadre d'un stage d'été dans le laboratoire de Bioinformatique Structurelle et Fonctionnelle de L'IRIC. Rémi Planel a écris le code du dinucleodite shuffle.

Chapitre 5 : Discussion et conclusion

Les transcrits identifiés

Nous avons atteint notre objectif de trouver des nouveaux transcrits chez *Candida albicans*. Sur les 1979 régions du génome que nous avons validé sur la puce à ADN, nous avons identifié 172 transcrits exprimés chez *C. albicans*, dont 105 transcrits qui sont des ARNnc répertorié dans la base de données de CDG [5]. Malgré l'utilisation de méthodes informatiques à basse sensibilité et l'application d'un seuil très rigoureux au niveau de notre analyse d'expression, nous avons identifié 62 nouveaux transcrits. Nous croyons que ces transcrits sont potentiellement des ARNnc de par leurs longueurs (qui sont en moyenne 100nt) et leurs locations dans le génome qui sont en majorité dans des régions intergéniques. De plus, la présence de structure secondaire conservée fournit un argument raisonnable indiquant que ces transcrits sont sous pression sélective et donc fonctionnel. Il est a noté qu'une seule condition de croissance a été analysé, donc si ces ARN sont des régulateurs on peut s'attendre à ce que leurs expressions fluctuent d'une condition à l'autre et que beaucoup de nos résultats négatifs s'avèreent simplement exprimés dans d'autres conditions de croissance. Des expériences de *Rapid Amplification of cDNA Ends* **(**RACE*) seront fait pour déterminer la longueurs des transcrits et s'ils sont spécifiques à une seule région du génome. Le gros du travail pour découvrir nouvelles cibles thérapeutiques a été fait, il ne reste qu'à vérifier si ces ARNnc identifiés sont essentiels à la survie de *Candida albicans* selon le protocole GRACE qui a déjà été appliqué aux région codante pour des protéines [57].

Discussion sur l'analyse informatique d'ARNnc

L'analyse informatique avec les outils RFAM/Infernal, QRNA, RNAz et Dynalign a donné plusieurs résultats. Parmi les outils appliqués, certains sont très longs à

exécuter, le cas le plus extrême étant 1.6 année de temps CPU (AMD Opteron 2.2GHz) pour faire l'analyse avec INFERNAL/RFAM sur le génome *Candida albicans*. Ainsi, l'approche choisie peut être difficilement applicable à de plus grands génomes sans accès à d'énormes ressources de calcul.

Les résultats de l'analyse informatique a démontré un faible chevauchement des prédictions entre les différentes méthodes, mais le chevauchement observé est plus grand qu'attendu en l'absence de biais. Pour exclure la possibilité que le signal observé était aussi trivial que la composition de nucléotides ou dinucleotide, nous avons comparé les distributions de prédictions trouvées par chaque méthode. Nous avons observé que la composition en dinucléotide des séquences prédites par les différentes méthodes est très variable. Des observations semblables ont été fais dans une études récente [53] qui a comparé l'efficacité de plusieurs outils de recherche d'ARNnc et montre que le signal perçu par beaucoup des outils semble venir des fréquences de dinucléotides. Notre méthodologie avec les NCM (chapitre 4) nous a également donné des résultats semblables. Cependant d'autres travaux qui se sont intéressés à l'existence de biais de séquences dans les ARNnc [47-49] ont conclu que dans certains organismes la fréquence d'apparition du dinucléotide CG peut être utilisé pour la détection des ARNnc. Les fréquences de dinucléotides ont une importance dans le détection d'ARNnc, mais il n'est pas clair s'il faut retirer leurs signaux comme le suggère les travaux de Baback [53] ou les utiliser comme le suggère les travaux de Schaffner[47].

Il est possible que les fréquences de dinucléotides sont un signal approprié comme certain type de ARNnc, comme les tRNA, mais pas pour d'autres types ARNnc. Ceci expliquerait les mauvais résultats pour la détection des ARNt de l'outil Dynalign qui utilise une normalisation de dinucléotide dans son modèle nulle. Anisi, il serait intéressant de regarder les distributions de dinucléotide par classe d'ARNnc. Pour l'instant ceci serait difficile car certaine classe d'ARNnc n'ont pas beaucoup de réprésentants les banques de données. Egalement, la composition spécifique à chaque espèce risque d'affecter le résultat.

Une autre importante observation de cette recherche est que nos résultats indiquent clairement que les méthodes testées ont tendance à identifier les régions qui sont exprimées par rapport aux résultats obtenus à partir de régions intergéniques aléatoires(Voir figure 8). De plus, nous avons observé que chaque méthode que nous avons testé a trouvé des transcrits exprimés n'ayant été prédit que par elle, confirmant la complémentarité des méthodes informatiques. Donc, dans l'état actuel des outils informatiques nous avons avantage à utiliser plusieurs méthodes de détection. Il est a noté qu'il est possible que cette observation soit due à la basse sensibilité des méthodes, car nous avons seulement validé les meilleures prédictions de chaque méthode. Ceci nous démontre que la sequence contient bien une information permettant d'indentifier les ARNnc.

Proposition d'une nouvelle technique de recherche d'ARNnc

Les observations sur les analyses informatiques et de la méthode des NCM (Chapitre 4) nous permettent de conclure que l'intégration de plusieurs signaux est probablement nécessaire à la prédiction d'ARNnc.

Je suggère pour améliorer la recherche d'ARNnc qu'il faut une méthode qui combine différentes sources d'information, soit les signaux de la séquence et les signaux de la machinerie transcriptionnelle. Les signaux de la machineries transcriptionnelle, comme les promoteurs de la transcription, ajouterais une information sur la possibilité que la cellule recrute la machinerie de la transcription (nécessaire pour transcrits un ARN à cet endroit dans le génome). L'approche serait de trouver les régions du génome qui décrivent significativement une région promotrice de la transcription et de regarder si la séquence en aval à le potentiel d'être un ARNnc.

Les signaux de la machinerie transcriptionnelle pour les ARNnc n'ont pas été très étudié et caractérisé, donc dans un premier temps il faut définir les motifs

promoteurs de la polymérase II (Pol II) et de la polymérase III (Pol III). Avec les données disponibles sur RFAM [41] on pourrait trouver les promoteurs en cherchant un motif commun dans les séquences en amont des gènes connus ARNnc. On pourrait construire un modèle de Markov caché (*HMM*)* pour les promoteurs de Pol II ou Pol III pour les eucaryote et les procaryotes. Le HMM pour les promoteurs de Pol II, pourraient être entrainé sur les sequences en amont des gènes codant pour des protéines qui sont également transcript par Pol II.

On chercherait dans le génome des régions qui décrivent significativement nos modèles de promoteur. Pour ces régions, on chercherait dans la séquence en aval la probabilité d'avoir un ARNnc. Pour calculer la probabilité d'être un ARNnc, nous pourrions utiliséer un réseau de neurones qui qui tient compte de la fréquence de nucléotides, de dinucléotides et des paramètres de la thermodynamique. On pourrais entraîner 3 réseaux de neurone, sois un avec des données de ARNnc, un avec des données de protéine et un avec des séquences neutre.

Un problème avec ces approches est la nécessitée d'avoir un ensemble de données d'entraînement, mais avons dans la base de données RFAM des centaines d'ARNnc. Donc dans un premier lieu, il faudrait regarder s'il y il effectivement un biais de fréquence de nucléotide, de dinucléotide et de thermodynamique entre les séquences de ARNnc, de protéine et intergénique et si ces biais sont les mêmes entre les différentes espèces.

L'ajout de la recherche des motifs promoteur de la transcription serait un ajout important à la recherche *de novo* d'ARN non codant.

La bio-informatique des ARNnc

Depuis le début de ce projet de recherche, des nouvelles techniques expérimentales ont été développées, comme le pyroséquençage. Cette technique de séquençage permet d'obtenir à haut débit avec des coûts moindres de grande quantité de séquences. Ce travail de recherche aurait pu être effectué en séquençant les petits ARN dans différentes

conditions. Également, la quantité de sonde que l'on peut mettre sur des puces à ADN s'est beaucoup amélioré et change le paysage de ce travail. Il est important de noter que ce qui est observé par ces deux techniques apporteront des informations sur se qui est exprimé ou non dans la cellule, mais n'apporteront pas d'informations sur la fonction de ces ARN.

Donc les prédicateurs informatiques d'ARNnc seront complémentaires aux nouvelles technologies pour nous informer sur la fonction potentielle de ces transcrits. La nouvelle génération d'outils de prédicateurs d'ARNnc en plus de déterminer si le transcrit est exprimé devras déterminer si ce transcrit non-codant est fonctionnel. Également la bioinformatique joue un rôle important dans le développement et l'implantation de modèles d'analyses, par exemple le développement que nous avons effectué pour l'analyse de notre puce à ADN. Ces nouvelles technologies, demanderons également autant d'effort pour le développement d'analyse des données à haut débit.

La bio-informatique a beaucoup plus apporté au domaine des ARN autre que la découverte d'ARNnc dans les séquences génomiques. Dans des recherches postérieurs, j'aimerais pouvoir utiliser les connaissances acquises sur les ARNnc, pour les intégrer au système d'interactions des protéines, afin d'obtenir un système d'interaction plus globale qui modélise mieux le fonctionnement de la cellule. Pour l'instant les ARNnc sont des inconnus dans ce type de modèles de système.

Conclusion

Ce travail de recherche a d'identifier et de confirmer la présence de 62 nouveaux transcrits chez la levure pathogène *Candida albicans*.

Pour trouver de nouveau ARNnc chez *C. albicans*, nous avons utilisé comme stratégie une analyse informatique combinant différentes méthodes existantes suivit d'une validation par puce à ADN. Les résultats de cette méthodologie nous ont appris la complémentaire des méthodes utilisées et que les méthodes testées ont tendance à identifier

les régions qui sont exprimées par rapport aux résultats obtenus à partir de régions intergéniques aléatoires

Ce travail a aussi contribué au développement d'une méthode pour identifier les transcrits exprimer à partir de données de puces à ADN (type *tiling array*). Cette méthode à la particularité de tenir compte de l'hybridation partielle des sondes à un transcrit et de donner les limites du transcrits au niveau des nucléotides. Cette méthode pourrais être appliquer avec d'autre plateforme d'analyse, comme un *tilling array* d'un génome complet.

Il est cependant important de noter que cette méthodologie ne tient pas compte des cas où la région de validation est transcrite. Pour résoudre ce problème il faudrait il faudrait considérer le niveau d'intensité du background de toute la puce plutôt que pour chaque région de validation.

Nous avons tenté de développer une méthode *de novo* d'identification des ARNnc avec une méthodologie se basant sur l'observation de motifs structuraux (NCM) dans les séquences d'ARNnc. Cette méthodologie s'est relevée inefficace pour identifier des ARNnc dans des séquences des génomes mais nous a démontré l'importance des distributions de dinucléotide dans la recherche des ARNnc.

Glossaire

ARN homologues :

Des ARN homologues sont des gènes qui ont une origine commune.

Boucle : Topologie particulière au sein d'une molécule ARN composée de régions non-appariées.

cDNA : Un simple brin artificiellement synthétisé à partir des ARNm. Il est obtenu après une réaction de transcription inverse d'un ARNm mature et représente ainsi la copie de l'ARNm en ADN.

Dinucléotide : Une molécule composée principalement de deux unités de nucléotides. Dans une séquence la fréquence des dinucléodites est le nombre de fois que se dinucleotide est retrouvé.

Diploïde: Une cellule est diploïde si elle possède 2*n* chromosomes organisés en *n* pairs.

Empilement *(staking*): Les bases azotées de l'ARN ont tendance à s'empiler sous l'action de plusieurs forces dont la principale est l'hydrophobicité.

Épissage alternatif: L'épissage est un processus qui consiste en l'excision des introns et en la ligature des exons. L'épissage alternatif permet à un gène de coder plusieurs ARNm mature ou protéines différentes.

E-value : Le e-value indique combien de fois vous pouvez vous attendre à une certaine séquence de se produire au hasard.

HMM: Modèle statistique dans lequel le système modélisé est supposé être un processus stochastique de paramètres inconnus.

Intergénique :

ADN entre les gènes donc techniquement non transcrit.

Intron:

Un intron est un fragment non codant d'un gène.

Introns du groupe 1 : Un intron qui est un grand ribozyme qui peut s'auto-épissé. C'est un ribozyme particulier puisqu'il a deux activités catalytiques: clivage et ligation.

Mbp : 10^6 pairs de base.

Mononucléotide :

Une molécule composée principalement d'un nucléotide.

Mutation synonyme :

Substitution d'un codon par un autre codon qui code le correspond au même acide aminé.

P-value **:** La probabilité d'obtenir au hasard un résultat extrême que la valeur réellement observée.

RACE (**Rapid Amplification of cDNA Ends):** Une technique utilisée en biologie moléculaire pour obtenir la longueur d'une séquence d'ARN transcription trouvée dans une cellule. Le résultat est la production d'une copie d'ADNc de l'ARN d'intérêt, produit par la transcription inverse, suivie d'amplification par PCR. Les copies d'ADNc amplifiés sont ensuite séquencées pour obtenir une longueur de séquence de l'ARN d'origine.

SCFG (*stochastic context-free grammars*): Modèle statistique dans laquelle chaque unité de production est augmentée avec une probabilité**.**

Sondes d'ADN : Séquence d'ADN ou d'ARN que l'on utilise pour détecter des séquences homologues (complémentaires) par hybridation.

Tige : Une série de paire de bases canonique (hélice).

UTR : (UnTranslated Region) sont les parties de l'ARNm qui ne sont pas traduites en protéines. La région 5'UTR contient la coiffe et le 3'UTR la queue polyA.

Bibliographie

1. Ryan KJ RG: **Sherris Medical Microbiology**, 4th edn; 2004.

2. A. Kalkanci EB, B. Aykan , K. Caglar , K. Hizel ,, D. Arman SK: **Epidemiology and antifungal susceptibility of Candida species isolated from hospitalized patients**. *Journal de Mycologie Medicale* 2007, **17**:16—20.

3. Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT *et al*: **The diploid genome sequence of Candida albicans**. *Proc Natl Acad Sci U S A* 2004, **101**(19):7329-7334.

4. van het Hoog M, Rast TJ, Martchenko M, Grindle S, Dignard D, Hogues H, Cuomo C, Berriman M, Scherer S, Magee BB *et al*: **Assembly of the Candida albicans genome into sixteen supercontigs aligned on the eight chromosomes**. *Genome Biol* 2007, **8**(4):R52.

5. Arnaud MB, Costanzo MC, Skrzypek MS, Binkley G, Lane C, Miyasato SR, Sherlock G: **The Candida Genome Database (CGD), a community resource for Candida albicans gene and protein information**. *Nucleic Acids Res* 2005, **33**(Database issue):D358-363.

6. Rossignol T, Lechat P, Cuomo C, Zeng Q, Moszer I, d'Enfert C: **CandidaDB: a multi-genome database for Candida species and related Saccharomycotina**. *Nucleic Acids Res* 2008, **36**(Database issue):D557-561.

7. Mitrovich QM, Tuch BB, Guthrie C, Johnson AD: **Computational and experimental approaches double the number of known introns in the pathogenic yeast Candida albicans**. *Genome Res* 2007, **17**(4):492-502.

8. Crick F: **[Central dogma of mollecular biology]**. *Tsitologiia* 1971, **13**(7):906-910.

9. Brenner S: **RNA, ribosomes, and protein synthesis**. *Cold Spring Harb Symp Quant*

Biol 1961, **26**:101-110.

10. Crick FH: **On protein synthesis**. *Symp Soc Exp Biol* 1958, **12**:138-163.

11. Hoagland MB, Stephenson ML, Scott JF, Hecht LI, Zamecnik PC: **A soluble ribonucleic acid intermediate in protein synthesis**. *J Biol Chem* 1958, **231**(1):241-257.

12. Stark BC, Kole R, Bowman EJ, Altman S: **Ribonuclease P: an enzyme with an essential RNA component**. *Proc Natl Acad Sci U S A* 1978, **75**(8):3717-3721.

13. Zieve G, Penman S: **Subnuclear particles containing a small nuclear RNA and heterogeneous nuclear RNA**. *J Mol Biol* 1981, **145**(3):501-523.

14. Li LH, Guo ZJ, Yan LL, Yang JC, Xie YF, Sheng WH, Huang ZH, Wang XH: **Antitumor and antiangiogenic activities of anti-vascular endothelial growth factor hairpin ribozyme in human hepatocellular carcinoma cell cultures and xenografts**. *World J Gastroenterol* 2007, **13**(47):6425-6432.

15. Ruvkun G, Ambros V, Coulson A, Waterston R, Sulston J, Horvitz HR: **Molecular genetics of the Caenorhabditis elegans heterochronic gene lin-14**. *Genetics* 1989, **121**(3):501-516.

16. Wightman B, Burglin TR, Gatto J, Arasu P, Ruvkun G: **Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development**. *Genes Dev* 1991, **5**(10):1813-1824.

17. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14**. *Cell* 1993, **75**(5):843-854.

18. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans**. *Nature* 2000, **403**(6772):901-906.

19. Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Muller P *et al*: **Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA**. *Nature* 2000, **408**(6808):86-89.

20. He L, Hannon GJ: **MicroRNAs: small RNAs with a big role in gene regulation**. *Nat Rev Genet* 2004, **5**(7):522-531.

21. Epshtein V, Mironov AS, Nudler E: **The riboswitch-mediated control of sulfur metabolism in bacteria**. *Proc Natl Acad Sci U S A* 2003, **100**(9):5052-5056.

22. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.

23. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A *et al*: **An RNA gene expressed during cortical development evolved rapidly in humans**. *Nature* 2006, **443**(7108):167172.

24. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermuller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A *et al*: **Structured RNAs in the ENCODE selected regions of the human genome**. *Genome Res* 2007, **17**(6):852-864.

25. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. *Nucleic Acids Res* 1997, **25**(5):955964.

26. Mitrovich QM, Guthrie C: **Evolution of small nuclear RNAs in S. cerevisiae, C. albicans, and other hemiascomycetous yeasts**. *RNA* 2007, **13**(12):2066-2080.

27. Huttenhofer A, Vogel J: **Experimental approaches to identify non-coding RNAs**. *Nucleic Acids Res* 2006, **34**(2):635-646.

28. Backofen R, Bernhart SH, Flamm C, Fried C, Fritzsch G, Hackermuller J, Hertel J, Hofacker IL, Missal K, Mosig A *et al*: **RNAs everywhere: genome-wide annotation of structured RNAs**. *J Exp Zoolog B Mol Dev Evol* 2007, **308**(1):1-25.

29. Jin T, Inouye M: **Identification of the genes in multicopy plasmids affecting ompC and ompF expression in Escherichia coli**. *FEMS Microbiol Lett* 1995, **133**(3):225-231.

30. Novaes E, Drost DR, Farmerie WG, Pappas GJ, Jr., Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome**. *BMC Genomics* 2008, **9**:312.

31. Jochl C, Rederstorff M, Hertel J, Stadler PF, Hofacker IL, Schrettl M, Haas H, Huttenhofer A: **Small ncRNA transcriptome analysis from Aspergillus fumigatus suggests a novel mechanism for regulation of protein synthesis**. *Nucleic Acids Res* 2008, **36**(8):2677-2689.

32. Stricklin SL, Griffiths-Jones S, Eddy SR: **C. elegans noncoding RNA genes**. *WormBook* 2005:1-7.

33. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S: **Identification of novel small RNAs using comparative genomics and microarrays**. *Genes Dev* 2001, **15**(13):1637-1651.

34. Tjaden B: **An approach for clustering gene expression data with error information**. *BMC Bioinformatics* 2006, **7**:17.

35. Emanuelsson O, Nagalakshmi U, Zheng D, Rozowsky JS, Urban AE, Du J, Lian Z, Stolc V, Weissman S, Snyder M *et al*: **Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome**. *Genome Res* 2007, **17**(6):886-897.

36. Eddy SR, Durbin R: **RNA sequence analysis using covariance models**. *Nucleic Acids Res* 1994, **22**(11):2079-2088.

37. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R: **RNAMotif, an RNA secondary structure definition and search algorithm**. *Nucleic Acids Res* 2001, **29**(22):4724-4735.

38. Gautheret D, Major F, Cedergren R: **Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA**. *Comput Appl Biosci* 1990, **6**(4):325-331.

39. Gautheret D, Lambert A: **Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles**. *J Mol Biol* 2001, **313**(5):1003-1011.

40. Eddy SR: **A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure**. *BMC Bioinformatics* 2002, **3**:18.

41. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database**. *Nucleic Acids Res* 2003, **31**(1):439-441.

42. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes**. *Nucleic Acids Res* 2005, **33**(Database issue):D121-124.

43. Zuker M: **On finding all suboptimal foldings of an RNA molecule**. *Science* 1989, **244**(4900):48-52.

44. Chen JH, Le SY, Shapiro B, Currey KM, Maizel JV: **A computational procedure for assessing the significance of RNA secondary structure**. *Comput Appl Biosci* 1990, **6**(1):7-18.

45. Rivas E, Eddy SR: **Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs**. *Bioinformatics* 2000, **16**(7):583605.

46. Bonnet E, Wuyts J, Rouze P, Van de Peer Y: **Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences**. *Bioinformatics* 2004, **20**(17):2911-2917.

47. Schattner P: **Searching for RNA genes using base-composition statistics**. *Nucleic*

Acids Res 2002, **30**(9):2076-2082.

48. Klein RJ, Misulovin Z, Eddy SR: **Noncoding RNA genes identified in AT-rich hyperthermophiles**. *Proc Natl Acad Sci U S A* 2002, **99**(11):7542-7547.

49. Carter RJ, Dubchak I, Holbrook SR: **A computational approach to identify genes for functional RNAs in genomic sequences**. *Nucleic Acids Res* 2001, **29**(19):39283938.

50. Rivas E, Eddy SR: **Noncoding RNA gene detection using comparative sequence analysis**. *BMC Bioinformatics* 2001, **2**:8.

51. Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs**. *Proc Natl Acad Sci U S A* 2005, **102**(7):2454-2459.

52. Mathews DH, Turner DH: **Dynalign: an algorithm for finding the secondary structure common to two RNA sequences**. *J Mol Biol* 2002, **317**(2):191-203.

53. Babak T, Blencowe BJ, Hughes TR: **Considerations in the identification of functional RNA structural elements in genomic alignments**. *BMC Bioinformatics* 2007, **8**:33.

54. Olivas WM, Muhlrad D, Parker R: **Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs**. *Nucleic Acids Res* 1997, **25**(22):4619-4625.

55. Heckman DS, Geiser DM, Eidell BR, Stauffer RL, Kardos NL, Hedges SB: **Molecular evidence for the early colonization of land by fungi and plants**. *Science* 2001, **293**(5532):1129-1133.

56. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

57. Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C *et al*: **Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery**. *Mol Microbiol* 2003, **50**(1):167-181.

58. Disney MD, Childs JL, Turner DH: **Hoechst 33258 selectively inhibits group I intron self-splicing by affecting RNA folding**. *Chembiochem* 2004, **5**(12):16471652.

59. Morgulis A, Gertz EM, Schaffer AA, Agarwala R: **A fast and symmetric DUST implementation to mask low-complexity DNA sequences**. *J Comput Biol* 2006, **13**(5):1028-1040.

60. Lopez R, Silventoinen V, Robinson S, Kibria A, Gish W: **WU-Blast2 server at the European Bioinformatics Institute**. *Nucleic Acids Res* 2003, **31**(13):3795-3798.

61. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED *et al*: **Aligning multiple genomic sequences with the threaded blockset aligner**. *Genome Res* 2004, **14**(4):708-715.

62. Delcher AL, Phillippy A, Carlton J, Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison**. *Nucleic Acids Res* 2002, **30**(11):2478-2483.

63. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.

64. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S *et al*: **Global identification of human transcribed sequences with genome tiling arrays**. *Science* 2004, **306**(5705):22422246.

65. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM: **A high-resolution map of transcription in the yeast genome**. *Proc Natl Acad Sci U S A* 2006, **103**(14):5320-5325.

66. Bates DMaW, D.G. : **Nonlinear Regression Analysis and Its Applications**. New York; 1989.

67. Dempster AP, Laird NM, Rubin DB: **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *Journal of the Royal Statistical Society Series B (Methodological)*

1977, **39**(1):1-38.

68. Valadkhan S: **snRNAs as the catalysts of pre-mRNA splicing**. *Curr Opin Chem Biol* 2005, **9**(6):603-608.

69. Singh SK, Pal Bhadra M, Girschick HJ, Bhadra U: **MicroRNAs--micro in size but macro in function**. *FEBS J* 2008, **275**(20):4929-4944.

70. Le SV, Chen JH, Currey KM, Maizel JV, Jr.: **A program for predicting significant RNA secondary structures**. *Comput Appl Biosci* 1988, **4**(1):153-159.

71. McCutcheon JP, Eddy SR: **Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics**. *Nucleic Acids Res* 2003, **31**(14):4119-4128.

72. Lemieux S, Major F: **Automated extraction and classification of RNA tertiary structure cyclic motifs**. *Nucleic Acids Res* 2006, **34**(8):2340-2346.

73. Parisien M, Major F: **The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data**. *Nature* 2008, **452**(7183):51-55.

74. Gesell T, Washietl S: **Dinucleotide controlled null models for comparative RNA gene prediction**. *BMC Bioinformatics* 2008, **9**:248.

75. Martens JA, Laprade L, Winston F: **Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene**. *Nature* 2004, **429**(6991):571574.

76. Perocchi F, Xu Z, Clauder-Munster S, Steinmetz LM: **Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D**. *Nucleic Acids Res* 2007, **35**(19):e128.

77. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes**. *Genome Biol* 2004, **5**(2):R12.

Annexe A : Les ARNnc connus chez *C. albicans* avant cette recherche

Nom	Position dans l'assemblage 19				Type d' ARNnc	Méthode de prédiction
tI(AAU)1	Contig19-10014	20476	20562	+	ARNt	tRNAscan-SE
tT(CGU)1	Contig19-1279	120	207	+	ARNt	tRNAscan-SE
tE(UUC)1	Contig19-10046	176800	176893	+	ARNt	tRNAscan-SE
tF(GAA)1	Contig19-10046	65635	65722	-	ARNt	tRNAscan-SE
tK(CUU)1	Contig19-10046	154886	154975	-	ARNt	tRNAscan-SE
tM(CAU)1	Contig19-10046	24716	24787	+	ARNt	tRNAscan-SE
tR(UCU)1	Contig19-10046	65731	65803	-	ARNt	tRNAscan-SE
tD(GUC)1	Contig19-10051	47765	47836	+	ARNt	tRNAscan-SE
tN(GUU)1	Contig19-10052	1121	1223	-	ARNt	tRNAscan-SE
tN(GUU)2	Contig19-10052	1310	1415	-	ARNt	tRNAscan-SE
tP(UGG)1	Contig19-10052	53264	53356	-	ARNt	tRNAscan-SE
tS(AGA)1	Contig19-10052	119648	119729	+	ARNt	tRNAscan-SE
tT(AGU)1	Contig19-10052	497	569	-	ARNt	tRNAscan-SE
SCR1	Contig19-10053	92936	93200	+	SRP	Orthologue de *S. cerevisiae*
tG(GCC)1	Contig19-10057	92358	92428	+	ARNt	tRNAscan-SE
tH(GUG)1	Contig19-10063	31610	31707	-	ARNt	tRNAscan-SE
tE(CUC)1	Contig19-10076	135856	135927	-	ARNt	tRNAscan-SE
tF(GAA)2	Contig19-10076	117311	117398	-	ARNt	tRNAscan-SE
tF(GAA)3	Contig19-10076	97161	97248	-	ARNt	tRNAscan-SE
tR(UCU)2	Contig19-10076	117405	117477	-	ARNt	tRNAscan-SE
tR(UCU)3	Contig19-10076	97255	97327	-	ARNt	tRNAscan-SE
tS(CAG)1	Contig19-10080	91580	91661	+	ARNt	tRNAscan-SE
tY(GUA)1	Contig19-10097	348	456	-	ARNt	tRNAscan-SE
tD(GUC)2	Contig19-10109	26294	26365	-	ARNt	tRNAscan-SE
tD(GUC)3	Contig19-10109	26449	26520	-	ARNt	tRNAscan-SE
tD(GUC)4	Contig19-10109	27976	28047	+	ARNt	tRNAscan-SE
tG(GCC)2	Contig19-10109	28719	28789	+	ARNt	tRNAscan-SE
tQ(UUG)1	Contig19-10109	36510	36596	+	ARNt	tRNAscan-SE
tL(CAA)1	Contig19-10119	86430	86548	-	ARNt	tRNAscan-SE
tV(AAC)1	Contig19-10119	259226	259299	+	ARNt	tRNAscan-SE
tE(UUC)2	Contig19-10123	78432	78525	+	ARNt	tRNAscan-SE

tK(UUU)1	Contig19-10123	113928	114002	+	ARNt	tRNAscan-SE
tR(CCU)1	Contig19-10123	264160	264248	+	ARNt	tRNAscan-SE
tV(AAC)2	Contig19-10123	118026	118099	+	ARNt	tRNAscan-SE
tY(GUA)2	Contig19-10123	102259	102373	+	ARNt	tRNAscan-SE
tS(GCU)1	Contig19-10124	5627	5741	+	ARNt	tRNAscan-SE
tS(GCU)2	Contig19-10124	5835	5950	+	ARNt	tRNAscan-SE
tS(AGA)2	Contig19-10133	11914	11995	+	ARNt	tRNAscan-SE
tV(CAC)1	Contig19-10136	12167	12238	+	ARNt	tRNAscan-SE
tE(UUC)3	Contig19-10137	96852	96945	-	ARNt	tRNAscan-SE
tE(UUC)4	Contig19-10137	96959	97052	-	ARNt	tRNAscan-SE
tQ(CUG)1	Contig19-10137	94231	94302	-	ARNt	tRNAscan-SE
tL(CAA)2	Contig19-10139	60588	60706	+	ARNt	tRNAscan-SE
tL(UAA)1	Contig19-10139	19583	19698	+	ARNt	tRNAscan-SE
tL(CAA)3	Contig19-10140	17658	17772	+	ARNt	tRNAscan-SE
RPR1	Contig19-10141	31955	32291	-	Rnase P	Orthologue de *S. cerevisiae*
tA(AGC)1	Contig19-10141	48467	48539	-	ARNt	tRNAscan-SE
tI(AAU)2	Contig19-10141	49790	49876	-	ARNt	tRNAscan-SE
tM(CAU)2	Contig19-10141	69680	69751	+	ARNt	tRNAscan-SE
tQ(UUG)2	Contig19-10141	121648	121734	+	ARNt	tRNAscan-SE
tI(UAU)1	Contig19-10146	50600	50707	+	ARNt	tRNAscan-SE
tG(UCC)1	Contig19-10150	130442	130512	+	ARNt	tRNAscan-SE
tT(AGU)2	Contig19-10150	122417	122489	-	ARNt	tRNAscan-SE
tL(AAG)1	Contig19-10151	134829	134949	+	ARNt	tRNAscan-SE
tL(UAA)2	Contig19-10151	17771	17886	-	ARNt	tRNAscan-SE
tE(UUC)5	Contig19-10155	9485	9578	-	ARNt	tRNAscan-SE
tT(UGU)1	Contig19-10158	66081	66152	-	ARNt	tRNAscan-SE
tA(AGC)2	Contig19-10160	5209	5281	-	ARNt	tRNAscan-SE
tE(UUC)6	Contig19-10160	15146	15242	+	ARNt	tRNAscan-SE
tL(UAA)3	Contig19-10163	250962	251077	-	ARNt	tRNAscan-SE
tN(GUU)3	Contig19-10163	221938	222042	+	ARNt	tRNAscan-SE
tR(UCU)4	Contig19-10163	39251	39323	-	ARNt	tRNAscan-SE
tT(AGU)3	Contig19-10172	149634	149706	+	ARNt	tRNAscan-SE
tA(AGC)3	Contig19-10173	154524	154596	+	ARNt	tRNAscan-SE
tA(AGC)4	Contig19-10173	5292	5364	+	ARNt	tRNAscan-SE
tS(UGA)1	Contig19-10173	136439	136520	-	ARNt	tRNAscan-SE
tA(UGC)1	Contig19-10176	16333	16405	-	ARNt	tRNAscan-SE
tY(GUA)4	Contig19-2020	2997	3090	-	ARNt	tRNAscan-SE
tY(GUA)5	Contig19-2020	3468	3576	-	ARNt	tRNAscan-SE

tD(GUC)5	Contig19-10183	782	853	-	ARNt	tRNAscan-SE
tG(GCC)3	Contig19-10183	3933	4003	+	ARNt	tRNAscan-SE
tI(AAU)3	Contig19-10183	88993	89079	-	ARNt	tRNAscan-SE
tQ(UUG)3	Contig19-10183	54170	54255	-	ARNt	tRNAscan-SE
tT(AGU)4	Contig19-10183	19217	19289	+	ARNt	tRNAscan-SE
tR(CCG)1	Contig19-10186	70826	70930	-	ARNt	tRNAscan-SE
tV(AAC)3	Contig19-10190	34783	34856	+	ARNt	tRNAscan-SE
tW(CCA)1	Contig19-10191	23165	23261	+	ARNt	tRNAscan-SE
tI(AAU)4	Contig19-10192	74602	74688	-	ARNt	tRNAscan-SE
tQ(UUG)4	Contig19-10192	83268	83350	+	ARNt	tRNAscan-SE
tQ(UUG)5	Contig19-10192	83361	83446	+	ARNt	tRNAscan-SE
tA(AGC)5	Contig19-10194	117715	117787	+	ARNt	tRNAscan-SE
tM(CAU)3	Contig19-10196	76991	77064	+	ARNt	tRNAscan-SE
tP(UGG)2	Contig19-10196	34939	35030	+	ARNt	tRNAscan-SE
tG(CCC)1	Contig19-10200	52516	52586	+	ARNt	tRNAscan-SE
tS(UGA)2	Contig19-10204	24345	24426	+	ARNt	tRNAscan-SE
tT(UGU)2	Contig19-10204	17769	17840	-	ARNt	tRNAscan-SE
tL(AAG)2	Contig19-10205	44733	44853	+	ARNt	tRNAscan-SE
tA(AGC)6	Contig19-10209	63368	63440	-	ARNt	tRNAscan-SE
tD(GUC)6	Contig19-10212	222675	222746	-	ARNt	tRNAscan-SE
tG(GCC)4	Contig19-10212	282240	282310	-	ARNt	tRNAscan-SE
tG(GCC)5	Contig19-10212	286368	286438	+	ARNt	tRNAscan-SE
tG(UCC)2	Contig19-10212	293397	293467	-	ARNt	tRNAscan-SE
tS(UGA)3	Contig19-10212	204758	204839	+	ARNt	tRNAscan-SE
tS(CGA)1	Contig19-10215	278569	278663	-	ARNt	tRNAscan-SE
tA(AGC)7	Contig19-10216	268416	268488	-	ARNt	tRNAscan-SE
tR(ACG)1	Contig19-2285	1816	1909	+	ARNt	tRNAscan-SE
tR(ACG)2	Contig19-2285	1940	2035	+	ARNt	tRNAscan-SE
tF(GAA)4	Contig19-10225	82343	82430	+	ARNt	tRNAscan-SE
tV(UAC)1	Contig19-10225	81610	81705	-	ARNt	tRNAscan-SE
tA(UGC)2	Contig19-10233	124039	124111	-	ARNt	tRNAscan-SE
tC(GCA)1	Contig19-10235	5082	5179	+	ARNt	tRNAscan-SE
tC(GCA)2	Contig19-10235	5257	5351	+	ARNt	tRNAscan-SE
tK(CUU)2	Contig19-10236	183417	183501	-	ARNt	tRNAscan-SE
tP(UGG)3	Contig19-10237	72995	73086	-	ARNt	tRNAscan-SE
tT(AGU)6	Contig19-10237	157434	157506	+	ARNt	tRNAscan-SE
tK(UUU)2	Contig19-10238	18142	18216	+	ARNt	tRNAscan-SE
tL(UAA)4	Contig19-10241	30271	30386	+	ARNt	tRNAscan-SE
tP(AGG)1	Contig19-10241	35909	36019	+	ARNt	tRNAscan-SE

tL(CAA)5	Contig19-2413	13186	13296	-	ARNt	tRNAscan-SE
tE(UUC)7	Contig19-10247	122943	123036	+	ARNt	tRNAscan-SE
tG(GCC)6	Contig19-10247	124607	124677	-	ARNt	tRNAscan-SE
tI(AAU)5	Contig19-10247	55173	55258	+	ARNt	tRNAscan-SE
tK(UUU)3	Contig19-10247	28919	29004	-	ARNt	tRNAscan-SE
tK(UUU)4	Contig19-10247	29677	29751	-	ARNt	tRNAscan-SE
tK(UUU)5	Contig19-10247	30682	30756	+	ARNt	tRNAscan-SE
tL(CAA)4	Contig19-10247	83801	83915	-	ARNt	tRNAscan-SE
tN(GUU)4	Contig19-2449	20493	20586	-	ARNt	tRNAscan-SE
tP(UGG)4	Contig19-10251	11575	11665	+	ARNt	tRNAscan-SE
tP(UGG)5	Contig19-10251	12162	12252	+	ARNt	tRNAscan-SE
tS(AGA)3	Contig19-10251	7042	7123	-	ARNt	tRNAscan-SE
tV(AAC)4	Contig19-10251	8436	8509	-	ARNt	tRNAscan-SE
snR52	Contig19-10254	52700	52823	+	Petit ARN nucléolaire	Orthologue de *S. cerevisiae*
tF(GAA)5	Contig19-10254	51803	51890	-	ARNt	tRNAscan-SE
tR(UCU)5	Contig19-10254	51899	51971	-	ARNt	tRNAscan-SE
tW(CCA)2	Contig19-10254	206677	206772	+	ARNt	tRNAscan-SE
tY(GUA)3	Contig19-10254	172427	172542	+	ARNt	tRNAscan-SE
tS(AGA)4	Contig19-10257	16681	16762	-	ARNt	tRNAscan-SE
snR6	Contig19-2500	66590	66690	+	Petit ARN nucléaire U6	Orthologue de *S. cerevisiae*
tV(AAC)5	Contig19-2500	65564	65637	-	ARNt	tRNAscan-SE
tV(AAC)6	Contig19-2500	65653	65726	-	ARNt	tRNAscan-SE
tL(UAA)5	Contig19-2506	83596	83711	+	ARNt	tRNAscan-SE
tD(GUC)7	Contig19-2511	111301	111372	-	ARNt	tRNAscan-SE
tM(CAU)4	Contig19-2511	21962	22035	+	ARNt	tRNAscan-SE
tH(GUG)3	Contig19-2514	64593	64690	-	ARNt	tRNAscan-SE
tL(CAA)6	Contig19-2516	157011	157121	-	ARNt	tRNAscan-SE
U1_snRNA	Contig19-10148	42822	43065	+	snRNA	
U2_snRNA	Contig19-10076	49694	49908	+	snRNA	
U4_snRNA	Contig19-10162	105435	105574	+	snRNA	
U5_snRNA	Contig19-2516	46589	46723	+	snRNA	
RDN5	Contig19-2479	4625	4745	+	5S ARNr	
RDN18	Contig19-2479	6927	8713	+	18S ARNr	
RDN25	Contig19-2479	9161	12521		25S RNAr	

I