Institutional Repository - Research Portal
Dépôt Institutionnel - Portail de la Recherche
researchportal.unamur.be
University of Namur

THESIS / THÈSE

MASTER EN BIOCHIMIE ET BIOLOGIE MOLÉCULAIRE ET CELLULAIRE

Implication de la kinase CDK-12 dans le développement postembryonnaire du nématode Caenorhabditis elegans

Martin, Antoine

Award date:
2017

Awarding institution:
Universite de Namur

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 02. May. 2026

Faculté des Sciences

Implication de la kinase CDK-12 dans le développement post-embryonnaire du nématode *Caenorhabditis elegans*

Mémoire présenté pour l'obtention

du grade académique de master 120 en biochimie et biologie moléculaire et cellulaire

Antoine MARTIN

Janvier 2017

Université de Namur
FACULTE DES SCIENCES
Secrétariat du Département de Biologie
Rue de Bruxelles 61 - 5000 NAMUR
Téléphone: + 32(0)81.72.44.18 - Téléfax: + 32(0)81.72.44.20
E-mail: joelle.jonet@unamur.be - http://www.unamur.be

Implication de la kinase CDK-12 dans le développement post-embryonnaire du nématode *Caenorhabditis elegans*

MARTIN Antoine

Résumé

Les kinases qui phosphorylent le domaine C-terminal (CTD) de Rpb1, la plus grosse sous-unité de l'ARN polymérase II, jouent un rôle important pour coordonner la transcription et la maturation des ARN messagers (ARNm). Chez le nématode *C. elegans,* l'inhibition de la kinase de la sérine 2 du CTD, CDK-12, résulte en un arrêt du développement post-embryonnaire réversible au stade larvaire 1 (L1). Cet arrêt, aussi appelé « diapause», est bien décrit et se produit naturellement lorsque les larves éclosent dans un environnement pauvre en nourriture.
Le génome de *C.elegans* contient environ 20% de gènes organisés en opérons, dont l'expression est importante à la sortie de diapause. Le transcrit primaire est co-transcriptionnellement processé en ARNm monocistroniques par un mécanisme de trans-splicing qui différencie les gènes en première position des suivants.
Pour disséquer le rôle de CDK-12, une étude transcriptomique a été réalisée. L'analyse statistique des résultats de RNA-seq a mis en évidence un groupe de gènes downrégulés suite à l'inhibition de CDK-12. Ce groupe est enrichi en gènes en position 2 ou plus des opérons, ce qui suggère un rôle de CDK-12 dans le processus de trans-splicing.
Notre travail a visé à analyser en profondeur les données de séquençage pour analyser l'impact de l'inhibition de CDK-12 sur la maturation des ARN transcrits à partir d'opérons. La conclusion de cette analyse est que l'inhibition de CDK-12 a bien un impact sur le phénomène de trans-splicing des gènes en position 2 ou plus des opérons. Notre travail a aussi révélé les difficultés techniques inhérentes à l'analyse par RNA-seq des extrémités trans-splicées.

Mémoire de master 120 en biochimie et biologie moléculaire et cellulaire

Janvier 2017

Promoteur: D. Hermand

Remerciements

Avant tout, je tiens à remercier Monsieur Hermand pour m'avoir accueilli dans son laboratoire, pour le suivi régulier du travail, pour les discussions à propos de celui-ci.

Je remercie tout particulièrement Carlo Yague-Sanz pour m'avoir suivi et encadré tout au long de ce travail, pour avoir pris le temps de m'expliquer différents logiciels informatiques et pour être resté disponible à toutes questions éventuelles de ma part.

Je remercie aussi Clément Cassart pour m'avoir permis de combiner la pratique et la théorie, notamment grâce à la culture de vers, la technique du chunk ou encore l'observation de résultats d'immunofluorescence.

Un grand merci à Olivier Finet pour m'avoir aidé dans la réalisation de la technique du « primer extension ».

Merci également à Julie Candiracci et Valérie Migeot pour m'avoir intégré dans l'équipe GEMO.

Je remercie aussi Patricia pour sa bonne humeur quotidienne.

Merci à Madame Wauthy, Monsieur De Bolle, Monsieur Silvestre et Monsieur Terwagne pour avoir consacré du temps à la lecture de mon mémoire.

Enfin, je tiens à remercier ma famille et mes amis qui m'ont soutenu et encouragé tout au long de ce travail.

Table des matières

1. Liste des abréviations 5
2. Introduction 6
2.1. Domaine C-terminal de l'ARN polymérase II 6
2.2. *Caenorhabditis elegans* 8
2.2.1. Organisme modèle 8
2.2.2. Cycle de vie 8
2.2.3. Le sexe 8
2.3. La génétique de *C. elegans* 9
2.3.1. Les opérons 9
2.3.2. Le trans-splicing 10
2.4. Études de CDK-12 chez *C. elegans* 12
2.4.1. Les kinases dépendantes de cycline 12
2.4.2. Études précédentes 12
2.4.3. Études actuelles 13
2.5. Programmes bioinformatiques utilisés 15
2.5.1. Tophat2 15
2.5.2. SAMtools 15
2.5.3. blast + ... 15
2.5.4. featureCounts 16
2.5.5. DESeq2 16
2.5.6. Ape 16
2.6. Objectifs 17
3. Matériel et méthodes 18
3.1. Culture de *C. elegans* en laboratoire 18
3.2. Analyse différentielle du transcriptome 19
3.2.1. Version du génome de *C. elegans* utilisée 19
3.2.2. Le modèle utilisé 19
3.2.3. Extraction des gènes d'intérêt 19
3.2.4. Relation entre conditions 19
3.2.5. Représentation des résultats de l'analyse statistique 19
3.2.6. Analyse des gènes disposés en opérons 20

3.3. Quantification des événements de trans-splicing 21
3.3.1. Méthodes testées 21
3.3.2. Détermination de la méthode la plus appropriée 23
3.3.3. Analyse de la quantification des événements de trans-splicing 25
3.4. Primer extension 27
3.4.1. Introduction à la technique du primer extension 27
3.4.2. Gène d'intérêt et primer utilisé 27
3.4.3. Marquage radioactif des primers 28
3.4.4. Marquage radioactif de l'étalon 28
3.4.5. Réaction de primer extension 28
3.4.6. Migration électrophorétique 28
4. Résultats et discussion 29
4.1. Analyse du transcriptome 29
4.1.1. Relation entre conditions 29
4.1.2. Analyse statistique du transcriptome 29
4.1.3. Subsets de gènes affectés suite à l'inhibition de CDK-12 30
4.1.4. Analyse du subset de gènes downrégulés lors de l'inhibition de CDK-12 30
4.1.5. Effet gène spécifique de CDK-12 31
4.2. Quantification des événements de trans-splicing 32
4.2.1. Choix d'une méthode appropriée 32
4.2.2. Analyse des événements de trans-splicing 35
4.3. Détermination expérimentale d'un arrêt de rétro-transcription dans la structure du splice leader par primer extension 39
5. Conclusion et perspectives 41
5.1. Conclusion 41
5.2. Perspectives 42
5.2.1. Quantité d'ARN utilisée lors du primer extension 42
5.2.2. Perfectionnement de la méthode du mapping pour quantifier les événements de trans-splicing 42
5.2.3. Distinction entre gènes trans-splicés SL1 et SL2 en vue de déterminer l'impact de l'inhibition de CDK-12 sur un type de trans-splicing plutôt que l'autre 42
5.2.4. Etude d'un éventuel rôle de CDK-12 dans la méthylation des histones 43
6. Bibliographie 44
7. Annexes 47

Table des figures

Figure 1: Cycle de transcription (figure modifiée à partir de Heideman and Eick, 2012).
Figure 2: Cycle de vie de *C. elegans* à 20°C (figure modifiée à partir de WormAtlas).
Figure 3: Distribution des opérons selon leur longueur.
Figure 4 : Rôle des opérons dans l'induction de la transcription (Zaslaver, 2011).
Figure 5 : Mécanisme du trans-splicing, structure et séquences des splices leader et complexe de trans-splicing.
Figure 6 : Différents types de splicing (figure modifiée à partir de Lasda and Blumenthal, 2011).
Figure 7 : Principe de la méthode « analog-sensitive » (figure modifiée à partir de l'article de Gregan *et al.*).
Figure 8 : Création et caractérisation des différentes conditions testées.
Figure 9 : Principe du RNA-séquençage.
Figure 10 : Principe de TopHat2.
Figure 11 : Overlap complet ou décalé.
Figure 12 : Représentation d'une distribution count comparé à une distribution normale.
Figure 13 : Correction des variances réalisée par DESeq2 (figure modifiée à partir de Love *et al.*, 2014).
Figure 14 : Proportion des différentes natures des reads au sein des résultats bruts du RNA-seq.
Figure 15 : Visualisation de la transformation logarithmique (log2) comparé à la transformation rlog (regularized logarithm transformation ou transformation en logarithme régularisé) (figure modifiée à partir de Love *et al.*, 2014).
Figure 16 : Méthode mise au point pour la quantification des événements de trans-splicing.
Figure 17 : Distribution du qstart, c'est-à-dire le début de l'overlap des reads blastés sur la séquence SL.
Figure 18 : Distribution du qstart, c'est-à-dire le début de l'overlap des reads blastés sur les séquences SL pour différents overlaps.
Figure 19 : Avantage du mapping sur un transcriptome trans-splicé par rapport au blast.
Figure 20 : Principe pour la génération d'un transcriptome virtuel déjà trans-splicé.
Figure 21 : Etapes pour générer un transcriptome trans-splicé.
Figure 22 : Taille des bandes attendues suite au primer extension.
Figure 23 : Heatmap des distances des différentes conditions testées.
Figure 24 : MAplots illustrant le "fold change" des gènes en fonction de leur expression moyenne.
Figure 25 : Profil d'expression génique dans les différentes conditions S, R et RI.
Figure 26 : Proportions des gènes en opérons dans le génome de *C. elegans* et dans le subset de gènes d'intérêt.
Figure 27 : Proportions des positions des gènes situés dans un opéron.
Figure 28 : Gene onthology des gènes dont l'expression est downrégulée suite à l'inhibition de CDK-12 (figure réalisée par Carlo Yague-Sanz).
Figure 29 : Distribution de l'overlap des reads hybrides blastés sur les séquences SL.
Figure 30 : Nuage de points représentant les gènes en fonction de leur nombre d'événements de trans-splicing.
Figure 31 : Courbes de ROC des différentes méthodes.
Figure 32 : Nombre d'événements de trans-splicing détectés avec notre méthode.

Figure 33 : Comparaison de la distribution des ratios $\boldsymbol{nombre\ d'événements\ de\ trans-splicing\ SL2 nombre\ d'événements\ de\ trans-splicing\ SL1}$ de différents subsets de gènes pour les conditions S, R et RI.
Figure 34 : Prédiction de la structure secondaire des splice leader.
Figure 35 : *Genome browser* centré sur l'opéron qui contient *mai-1*, *gpd-2* et *gpd-3*.
Figure 36 : Primer extension réalisé à partir d'ARNm synthétique.
Figure 37 : Réaction potentielle de la rétrotranscriptase face à une structure secondaire.
Figure 38 : Comparaison de la méthode du mapping réalisée au cours de cette étude avec la méthode du mapping effectuée par Blumenthal.

Table des tableaux

Tableau 1 : Code utilisé lors de l'extraction, sa signification et son objectif.
Tableau 2 : Comparaison entre les quatre méthodes. nt signifie nucléotide.

1. Liste des abréviations

ADN	Acide désoxyribonucléique
ADNc	ADN complémentaire
ADNr	ADN ribosomique
ARN	Acide ribonucléique
ARNi	Interférence ARN
ARNm	ARN messager
ARN pol	ARN polymérase
ARN pol II	ARN polymérase II
ARNr	ARN ribosomique
ARNt	ARN de transfert
as	Analog-sensitive
ATP	Adénosine triphosphate
BLAST+	Basic Local Alignement Search Tool +
CDK	Cyclin-dependent kinase
cdk-12as	Mutant *cdk-12s*
ChIP	Chromatin immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation-sequencing
CPSF	Cleavage and polyadenylation specifity factor
CstF	Cleavage stimulation factor
CTD	Carboxy-terminal domain
dNTP	Désoxyribonucléotide triphosphate
GAPDH	Glycéraldéhyde-3-phosphate déhydrogénase
FDR	False discovery rate
FPR	False positif rate
NGM	Nematode growth medium
PNK	Polynucléotide kinase
pré-ARNm	Pré-ARN messager
R	Recovery
RI	Recovery plus inhibiteur
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
RT	Rétro-transcription
RTase	Rétrotranscriptase
S	Starvation
SL	Splice leader
S0	Starvation au temps 0
S2	Sérine 2 du CTD
S2P	Phosphorylation de la sérine 2 du CTD
S5	Sérine 5 du CTD
Tm	Température de melting
TPR	True positif rate
WT	Wild-type

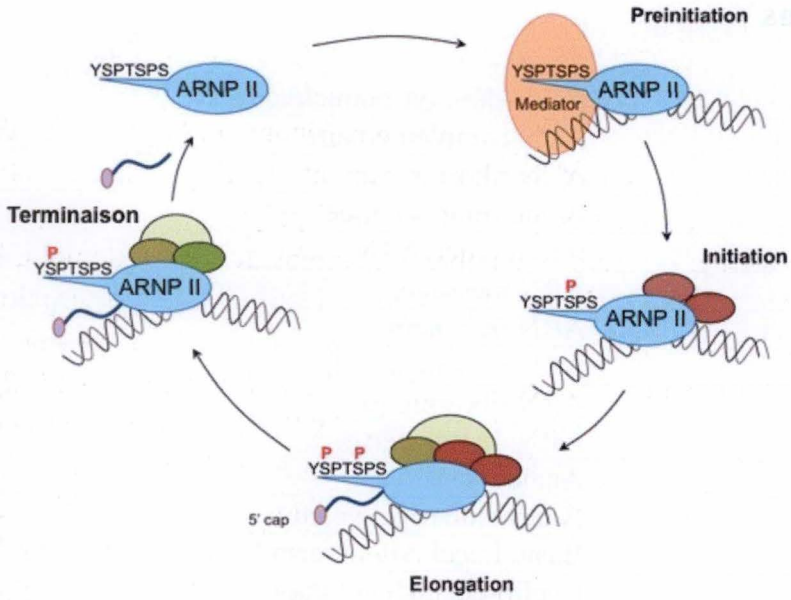


Figure 1 : Cycle de transcription (figure modifiée à partir de Heideman and Eick, 2012). L'ARN polymérase, avec un CTD complètement déphosphorylé, est recrutée par le complexe Mediator. La phosphorylation de la sérine 5 (S5) lors de l'initiation permet le début de la transcription ainsi que le recrutement des enzymes responsables de l'ajout de la coiffe. Ensuite, la sérine 2 (S2) est progressivement phosphorylée pendant l'élongation, ce qui provoque le recrutement des enzymes de splicing. Par contre, la S5 est, elle, progressivement déphosphorylée si bien qu'un pic de sérine 2 phosphorylée (S2P) est observé en 3'. Celui-ci permet le recrutement des enzymes de clivage et de polyadénylation. Enfin, l'ARNm est libéré et exporté hors du noyau tandis que la S2 est déphosphorylée, ce qui permet le recyclage de l'ARN pol II pour permettre un nouveau cycle de transcription.

2. Introduction

2.1. Domaine C-terminal de l'ARN polymérase II

Toute la complexité d'un organisme est inscrite dans son génome et le processus de transcription est la première étape qui permet le développement de l'organisme. Cette transcription est assurée par différents complexes que sont les ARN polymérases (ARN pol). Dans le règne animal, les ARN polymérases sont au nombre d'au moins trois : l'ARN pol I, II et III. L'ARN pol I transcrit l'ADNr en ARNr. L'ARN pol III se restrinct à la transcription des petits ARN, notamment les ARNt tandis que l'ARN pol II est la polymérase qui permet la transcription des gènes codant pour des protéines et un large set d'ARN non codants (*1, 2*). L'ARN pol II est un complexe de 12 sous-unités dont la plus grande, Rpb1, aussi connue sous le nom d'AMA-1 chez *Caenorhabditis elegans* (*C. elegans*), dispose d'un domaine C-terminal (CTD) (*3*). Le CTD se comporte comme une plate-forme qui peut subir des modifications et, en fonction de celles-ci, recruter différentes protéines impliquées dans la maturation des ARN messagers (ARNm) ainsi que dans les modifications de la chromatine (*4*). En effet, le CTD est composé de répétitions de sept acides aminés de consensus $Y_1S_2P_3T_4S_5P_6S_7$, dont la plupart subissent des modifications post-traductionnelles (*5*). Par exemple, les sérines (S) ainsi que la tyrosine (Y) et la thréonine (T) peuvent être phosphorylées alors que les prolines (P) peuvent être isomérisées. Les sérines peuvent également être O-glycosylées (*6, 7*). De plus, on trouve parfois d'autres acides aminés dans la séquence du CTD comme l'arginine (R) et la lysine (K) qui peuvent être méthylées, acétylées ou ubiquitinilées (*8*). Dès lors, un grand nombre de combinaisons est possible, ce qui est communément appelé le code CTD (*6*). En outre, le nombre de répétitions augmente avec la complexité de l'organisme. Il y a, par exemple, 26 répétitions chez la levure alors qu'il y en a 52 chez les mammifères (*9*). Chez *C. elegans*, il y a 42 répétitions (*3*). Cependant, la séquence du CTD reste relativement conservée au fil de l'évolution.

L'état du CTD varie tout au long du cycle de transcription (**figure 1**) (*5*). Tout d'abord, le complexe Mediator participe au recrutement de la polymérase au niveau du promoteur avec un CTD complètement déphosphorylé. Ensuite, l'initiation de la transcription commence avec la phosphorylation de la S5 par la kinase CDK7 associée au facteur de transcription TFIIH. La phosphorylation de la S5 provoque le recrutement des enzymes de capping qui permettent l'addition d'une coiffe 7-méthyl guanosine sur le pré-ARN messager (pré-ARNm). L'ajout d'une coiffe permet de protéger le pré-ARNm de sa dégradation par les exonucléases. Pendant la phase d'élongation, la S2 est phosphorylée alors que la S5 est progressivement déphosphorylée. La phosphorylation de la S2 est réalisée par la kinase CDK-12, tandis que la déphosphorylation se fait par une phosphatase. L'augmentation relative de S2 phosphorylée (S2P) en cours d'élongation aide au recrutement des enzymes de splicing qui permettent l'excision des introns et l'épissage des exons. Le niveau relatif de S2P augmente jusqu'à atteindre un pic en 3' qui, couplé à un motif présent sur l'ARN, provoque le recrutement de complexes impliqués dans la maturation en 3' et la terminaison (*5*). En effet, cela permet d'une part, la polyadénylation, c'est-à-dire l'ajout d'une queue poly(A) à l'extrémité 3' du pré-ARNm, et d'autre part, la libération de la molécule d'ARN par clivage. Puisque le pré-ARNm a été maturé tout au long de ce cycle, la molécule d'ARN finale est appelée ARN

messager (ARNm). Enfin, la S2 est déphosphorylée, ce qui permet de recycler l'ARN pol II pour un nouveau cycle de transcription.

Etant donné que les phosphorylations au niveau du CTD permettent le recrutement de différents facteurs vitaux impliqués dans la maturation des ARN, les kinases du CTD ont été longtemps considérées comme essentielles pour l'expression de tous les gènes. Cependant, des études ont mis en évidence une situation plus complexe. Tout d'abord, chez la levure de fission, *Schizosaccharomyces pombe,* un déletant *lsk1*, le gène codant pour la kinase orthologue de CDK12 chez cette levure, est viable (*7*). Cela suggère donc que la S2P n'est pas essentielle. Un mutant où la phosphorylation ne peut pas avoir lieu sur les sérines 2 du CTD suite à leur remplacement par des alanines (acide aminé qui ne peut pas être phosphorylé), nommé le mutant S2A est également viable malgré un léger retard de croissance de la levure (*10*). Le mutant S2A est stérile suite à son incapacité à induire le programme de différenciation sexuelle. La conclusion de ces travaux est que S2P est requise pour l'expression d'un set de gènes, mais n'est pas indispensable à l'expression génétique générale, contrairement aux étapes de maturation de l'ARNm qui lui sont traditionnellement apparentées.

En ce qui concerne la levure de bourgeonnement, *Saccharomyces cerevesea*, un mutant S2A a aussi été généré. L'étude de ce mutant révèle également un défaut dans la croissance végétative. En parallèle, un mutant *ctk1*, le gène codant pour la kinase orthologue à CDK12 et Lsk1 chez la levure de bourgeonnement, a été créé. Le phénotype de ce mutant présente exactement les mêmes caractéristiques concernant la croissance que le mutant S2A, ce qui soutient la conclusion que Ctk1 est la principale kinase responsable de la S2 chez *Saccharomyces cerevesea* (*11*). Des études ont également été réalisées sur des cellules de drosophile. A nouveau, la conclusion finale est la même, c'est-à-dire que la S2P, par Cdk12 chez la drosophile, est nécessaire à l'expression d'un set de gènes (*12*). Des études ont aussi été réalisées sur des cellules humaines *in vitro* (*13*). Cependant, le mécanisme par lequel la S2P supporte l'expression génique n'est pas bien compris. Enfin, le rôle de la phosphorylation de la sérine 2 n'a jamais été investigué chez un métazoaire *in vivo*. C'est pourquoi notre choix s'est porté au laboratoire sur le nématode *C.elegans,* un modèle établi pour l'analyse génétique.

En résumé, la S2P est nécessaire à l'expression de certains gènes, mais pas indispensable pour la majorité. De plus, les études réalisées jusqu'à présent reposent sur des organismes unicellulaires ou des cellules en culture. La question actuelle est de savoir ce qu'il en est chez un organisme multicellulaire. Pour répondre à cette question, le nématode *C.elegans* est utilisé

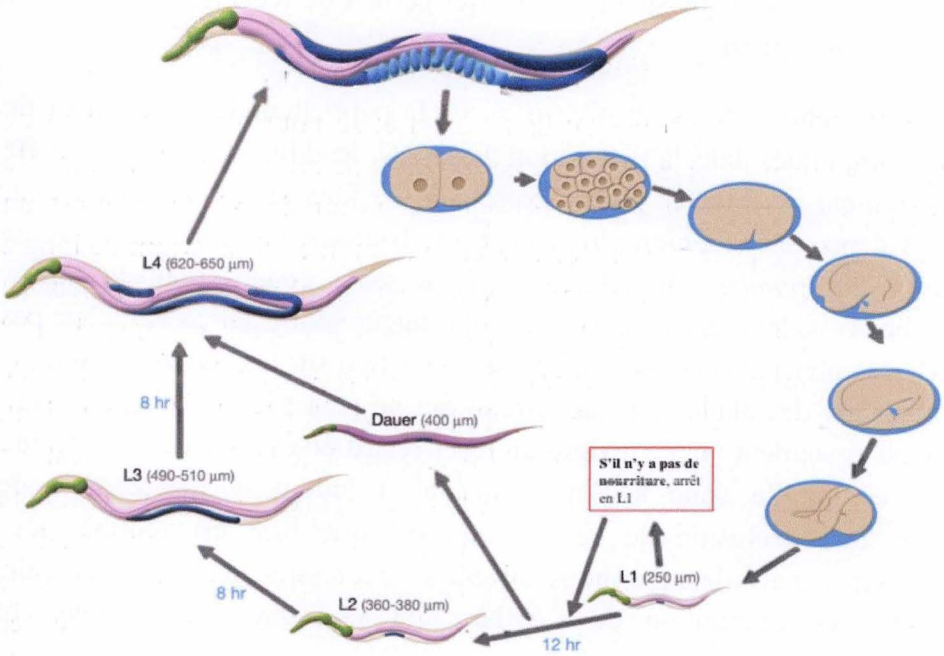


Figure 2 : Cycle de vie de *C. elegans* à 20°C (figure modifiée à partir de WormAtlas). Après 12 heures de développement embryonnaire, une larve de stade L1 éclot de l'œuf. S'il y a de la nourriture disponible pour permettre le développement, alors la larve L1 peut passer successivement par les stades L2, L3, L4 et adultes. Si, par contre, il n'y a pas de nourriture disponible directement après l'éclosion, les larves s'arrêtent précocement en L1 jusqu'à ce que de la nourriture soit de nouveau disponible pour poursuivre leur développement. Cependant, s'il y a de la nourriture après l'éclosion, mais que les conditions s'amenuisent par la suite (surpopulation, stress, …), les larves L1 rentrent dans un stade dauer jusqu'à ce que les conditions s'améliorent et passent alors directement en L4.

2.2. *Caenorhabditis elegans*

2.2.1. Organisme modèle

C. elegans est un nématode (ver rond) qui vit dans le sol et se nourrit naturellement de bactéries. Ce nématode possède plusieurs caractéristiques qui font de lui un organisme modèle très intéressant. Premièrement, comme il se nourrit exclusivement de bactéries, c'est un organisme facilement cultivable en laboratoire. Deuxièmement, il fait environ 1 mm de long à l'âge adulte. Troisièmement, il se reproduit assez vite, ce qui permet d'obtenir rapidement un grand nombre d'individus. Quatrièmement, il dispose également d'un petit génome de 100 Mb contenant 20000 gènes codant pour des protéines. De plus, comparé aux bactéries, l'étude de *C. elegans* apporte des informations plus proches de l'homme phylogénétiquement parlant. En effet, ce nématode est un eucaryote multicellulaire, comparé à la levure qui est un eucaryote unicellulaire. Cinquièmement, *C. elegans* a été le premier organisme multicellulaire séquencé (*14*). Son génome est donc connu et largement annoté, ce qui rend l'analyse génétique plus aisée. Sixièmement, ce ver dispose d'un corps transparent, ce qui permet de voir un signal luminescent à travers celui-ci et est donc fortement utilisé en immunofluorescence (*15*). *C. elegans* est d'ailleurs très utilisé pour les études concernant la neurobiologie, la sénescence ou encore le développement. Septièmement, l'anatomie de *C. elegans* a été décrite. Ainsi, on sait, par exemple, qu'il y a 959 cellules somatiques et 302 neurones (*14*). Huitièmement, beaucoup de protéines de *C. elegans* possèdent un orthologue chez l'être humain. Corsi (2015) les estime entre 60% et 80%.

2.2.2. Cycle de vie

Le cycle de vie de *C. elegans* est composé de différents stades (**figure 2**). Dans des conditions optimales, le développement embryonnaire s'étale sur 12 heures (3h *in utero* et 9h *ex utero*). Après éclosion des œufs, l'embryon passe par différents stades larvaires séparés par des mues : L1, L2, L3 et L4. Cependant, s'il n'y a pas assez de nourriture au moment de l'éclosion, les larves restent bloquées en L1 (*16*). Notons que l'ajout de nourriture entraîne une reprise rapide du développement post-embryonnaire. Par contre, si les conditions sont favorables au moment de l'éclosion, mais qu'elles s'amenuisent progressivement (surpopulation, stress, ...) alors les larves L1 rentrent dans un état de dormance relative nommé « dauer » et, quand les conditions s'améliorent, les larves dauer passsent directement au stade L4 (*16, 17*).

2.2.3. Le sexe

Deux types d'organismes sexuellement différents existent chez *C. elegans* : les hermaphrodites (XX) qui portent deux X sur leur paire de chromosomes sexuels et les mâles (X0) qui ne possèdent qu'un seul X au niveau du chromosome sexuel (*15*). En fait, les mâles proviennent d'une non-disjonction lors de la méiose (Lints, R. and Hall, D.H. 2009. Male introduction. In *WormAtlas*.). Les hermaphrodites sont capables de produire les gamètes mâles et femelles. Les hermaphrodites possèdent un ovotestis qui va produire, dans un premier temps, des spermatozoïdes qu'ils stockeront ensuite dans la spermathèque au cours du stade L4, et dans un second temps, des ovocytes. Le passage graduel des ovocytes à travers la spermathèque assure la reproduction sexuée. Puisque les hermaphrodites sont

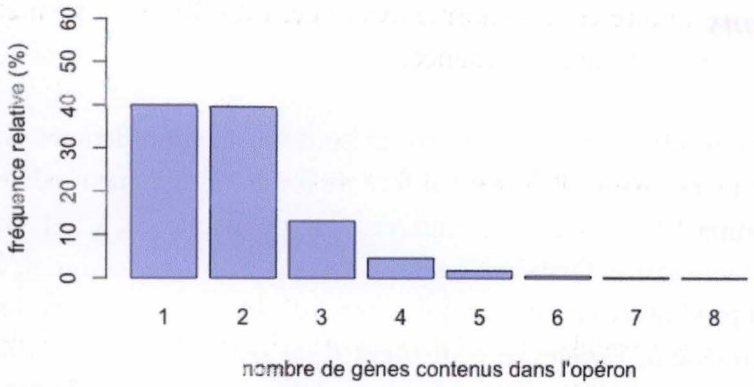


Figure 3: Distribution des opérons selon leur longueur. Tous les opérons contiennent au moins deux gènes et certains en contiennent jusque 8 bien que ceux-ci sont peu fréquents.

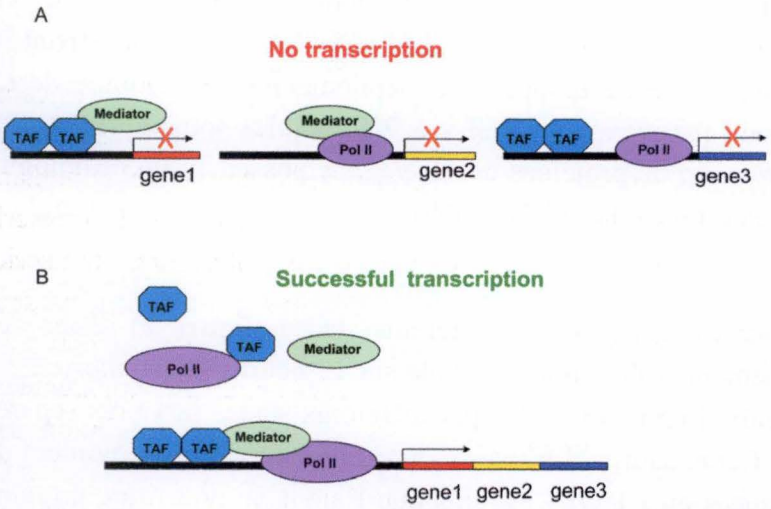


Figure 4 : Rôle des opérons dans l'induction de la transcription (Zaslaver, 2011). (**A**) Le fait d'avoir des gènes et des promoteurs isolés engendre une compétition pour les différents composants de la machinerie de transcription. Ainsi, quand une transcription de plusieurs gènes est nécessaire, la transcription ne sait pas se faire. (**B**) Par contre, quand les gènes sont rassemblés en opérons, c'est-à-dire que les gènes sont rassemblés sous le contrôle d'un seul promoteur, la compétition diminue si bien qu'une transcription massive est possible. TAF : facteur activateur de la transcription. Pol II signifie l'ARN polymérase II.

autofécondables, leur utilisation en laboratoire est un énorme avantage. En effet, cela permet la conservation des souches avec un minimum de maintenance.

2.3. La génétique de *C. elegans*

Le génome de *C. elegans* comporte deux particularités : il contient des opérons et la plupart des ARNm sont trans-splicés (*18*).

2.3.1. Les opérons

Le génome de tout organisme contient des gènes et en amont de ceux-ci leur promoteur, région qui permet d'initier la transcription. Ces gènes peuvent être soit isolés soit être disposés en opérons, c'est-à-dire des séries de gènes en cluster sous le contrôle d'un seul promoteur (*19*). Bien que cette notion soit plus commune chez les procaryotes, il existe aussi des opérons chez de nombreux eucaryotes. Les nématodes dont *C. elegans* en disposent. En effet, chez ce dernier, 15 à 20% des gènes codant pour des protéines sont en opérons (*20*). Pour exemple, les gènes *gpd-2* et *gpd-3*, gènes qui codent pour l'isoenzyme GADH-2 fait partie d'un opéron (*21*). Il est à noter que, chez *C. elegans*, les opérons comportent maximum huit gènes (**figure 3).**

Il y a deux grandes différences entre les opérons rencontrés chez les bactéries et ceux rencontrés chez *C. elegans*. Tout d'abord, le transcrit des opérons des bactéries est un transcrit polycistronique, c'est-à-dire que la transcription génère une seule molécule d'ARN contenant les phases ouvertes de lecture des différents gènes. Par contre, chez *C. elegans*, le transcrit primaire des opérons est co-transcriptionnellement clivé en plusieurs transcrits monocistroniques, c'est-à-dire différentes molécules contenant chacune le transcrit d'un seul gène et ce, grâce au trans-splicing (*22*). Ceci sera expliqué par la suite. Ensuite, les gènes au sein d'un opéron bactérien encodent souvent des protéines fonctionnellement apparentées. Autrement dit, cette structure en opérons permet chez les bactéries une co-régulation des gènes qui interviennent dans un même pathway. Ceci n'est pas nécessairement le cas chez *C. elegans* (*23*) (*24*). Le modèle proposé par Zaslaver (2011) pour expliquer l'organisation en opérons chez *C. elegans*, propose un avantage transcriptionnel (**figure 4**). En effet, dans le but de permettre une transcription optimale des gènes, plusieurs composants de la machinerie de transcription doivent s'assembler au niveau du promoteur. Il y a donc une compétition naturelle entre les différents promoteurs pour la machinerie et cette compétition est d'autant plus grande qu'il y a de promoteurs, rendant la polymérase limitante. Cela est davantage marqué dans le cas où le besoin de transcription est important. C'est notamment le cas lors de la sortie de starvation où il y a un besoin massif de transcription, typiquement lors de la sortie d'un arrêt en L1. Dans le cas de gènes isolés, la compétition est tellement importante que la transcription ne se fait pas de manière optimale. Supportant cette hypothèse, les gènes généralement impliqués dans le développement et la croissance sont souvent organisés en opérons (*23, 24*).

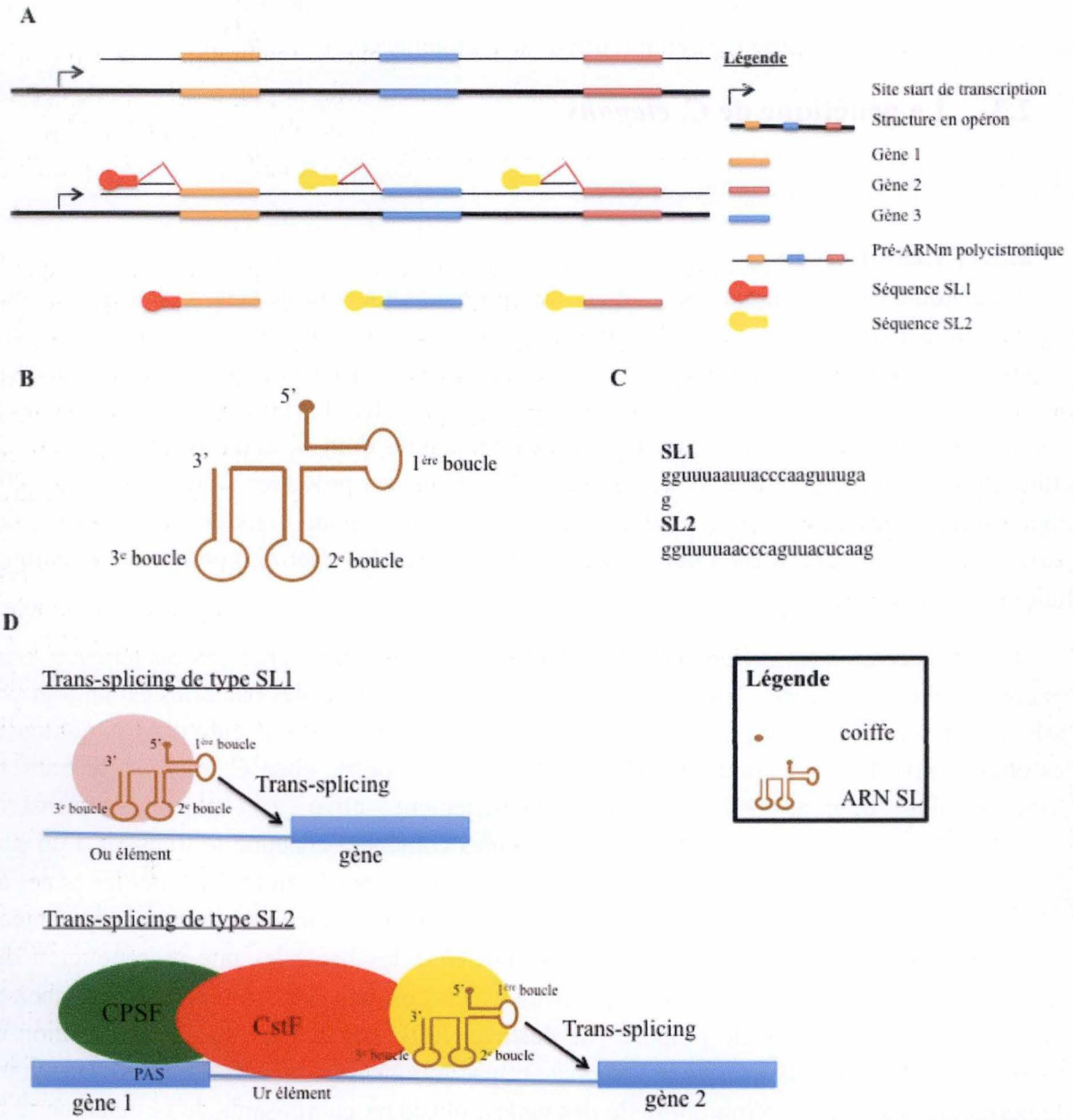


Figure 5 : Mécanisme du trans-splicing, structure et séquences des splice leader et complexe de trans-splicing. (**A**) Mécanisme de trans-splicing au niveau des opérons. La transcription des opérons mène à la formation d'un transcrit polycistronique. Le trans-splicing, de type SL1 ou SL2 en fonction de la position du gène dans l'opéron, permet de juxtaposer la séquence SL en 5' du pré-ARNm. De cette façon, on obtient plusieurs transcrits monocistroniques. Cependant, le trans-splicing se fait de manière co-transcriptionnelle. La formation d'un transcrit polycistronique est donc virtuelle *in vivo*. (**B**) Structure de l'ARN SL. L'ARN SL, qu'il soit de type SL1 ou SL2, comporte trois boucles et une coiffe en 5'. (**C**) Séquences de SL1 et SL2 (séquences s'étalant de l'extrémité 5' jusque la base de la première boucle). Ainsi, les séquences des deux SL ont la même longueur et sont fortement similaires bien qu'il existe quelques différences. (**D**) Machineries de trans-splicing différentes en fonction du SL. En ce qui concerne le trans-splicing de type SL1, c'est directement la protéine associée au ARN SL qui va interagir avec l'élément Ou, ce qui va permettre le transfert de la séquence SL en 5' du gène. En ce qui concerne le trans-splicing SL2, celui-ci est couplé à la polyadénylation du gène en amont. Tout d'abord, le complexe de facteurs stimulant le clivage (CstF) va interagir avec l'élément Ur. Ensuite, il y a d'une part recrutement des facteurs stimulateurs du clivage et de polyadénylation (CSTF) qui va interagir à la fois avec CstF et le site de polyadénylation (PAS). D'autre part, le CstF interagit avec la protéine associée au SL2 ARN, ce qui va permettre le transfert de la séquence SL2 en 5' du gène. Autrement dit, le trans-splicing de type SL2 est couplé à la polyadénylation du gène en amont.

2.3.2. Le trans-splicing

L'épissage ou le splicing est un phénomène requis pour la maturation d'une majorité d'ARN eucaryotes qui comprend également l'ajout d'une coiffe, la polyadénylation, c'est-à-dire l'ajout d'une queue riche en adénosine (queue poly A) et plus rarement l'édition. Le trans-splicing est une variante d'épissage. Environ 70% des gènes codant pour des protéines sont trans-splicés chez *C. elegans* (*25*).

Le trans-splicing est similaire à l'épissage classique, le cis-splicing, à la différence que le trans-splicing se déroule entre deux molécules d'ARN indépendantes. Chez *C. elegans*, une des deux molécules sera toujours un ARN particulier appelé "splice leader" (SL), l'autre est un ARNm prématuré (pré-ARNm). Ce type de trans-splicing est plus communément appelé le SL trans-splicing. Plus précisément, le SL ARN, d'une longueur d'environ 100 nucléotides (nt), sert de donneur, à hauteur de 22 nt, pour l'extension de l'extrémité 5' d'un pré-ARNm (**figure 5A**). Ce phénomène est assuré d'une part, grâce de petites protéines ribonucléaires (snRNP) associées au SL ARN et d'autre part, grâce à deux réactions de trans-estérification (*26*). Chez *C. elegans*, les SL présentent des structures secondaires (**figure 5B**) (*26*) et possèdent une coiffe à l'extrémité 5' de la molécule. Lors du trans-splicing, seule la boucle I est ajoutée au pré-ARNm. Chez *C. elegans*, il y a deux types de SL : SL1 et SL2 (*27*). La **figure 5C** illustre leur séquence. Il est à souligner que, bien qu'il s'agisse de deux molécules différentes, leurs séquences sont fortement similaires. Cependant, alors que le trans-splicing de type SL1 s'opère pour les gènes à proximité d'un promoteur, c'est-à-dire les gènes isolés et les premiers gènes des opérons, le trans-splicing de type SL2 est spécifique aux gènes en position deux ou plus des opérons (*19*). Cette spécificité du type de transsplicing est assurée par des séquences consensus qui permettent le recrutement de complexes différents (**figure 5D**) (*22, 28*). En effet, en ce qui concerne le trans-plicing de type SL1, des séquences consensus disposées au niveau de l'outron, région similaire à un intron qui se situe en amont de l'exon à trans-splicer, permet le recrutement spécifique du SL1 ARN. Par contre, en ce qui concerne le trans-splicing de type SL2, cela est un peu plus complexe. Les régions intergéniques des opérons, régions entre les gènes de l'opéron, contiennent des séquences consensus riches en uridine. C'est pourquoi cela porte le nom d'élément Ur. Le facteur stimulant le clivage (CstF) reconnait cet élément. Ce complexe va interagir d'un côté, avec la snRNP du SL2 et d'un autre côté, avec le facteur de stimulation de la polyadénylation et du clivage (CPSF) du gène en amont. Ainsi, au sein d'un opéron, la polyadénylation d'un gène donné est couplée au trans-splicing du gène localisé directement en aval.

Le trans-splicing a plusieurs rôles. Tout d'abord, par l'addition d'une séquence SL, il permet de transférer une coiffe hyperméthylée au pré-ARNm et ainsi, de le protéger de l'action des RNAses (*29*). Il permet également de générer des transcrits monocistroniques à partir des opérons (**figure 5A**) (*22*) (*29*).

Dans le but de faciliter la compréhension du mécanisme, deux complexités du trans-splicing ont été précédemment volontairement omises. Premièrement, *C. elegans* dispose de différentes isoformes des SL. En effet, il existe plusieurs gènes qui codent pour différents variants de la séquence de SL2 (*19*). Par contre, malgré le fait que plusieurs gènes codent pour SL1, ceux-ci proviennent de duplication de la même séquence. Donc, une seule forme de SL1

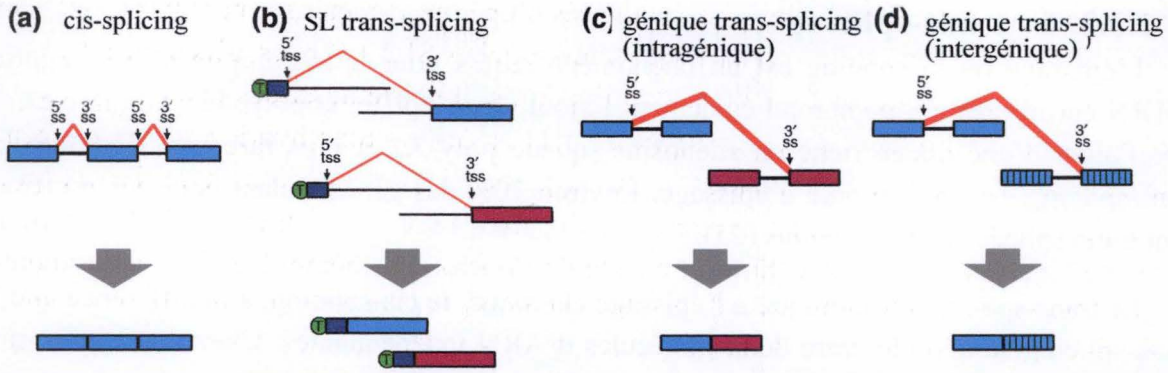


Figure 6 : Différents types de splicing (figure modifiée à partir de Lasda and Blumenthal, 2011). (a) le cis-splicing est un épissage des exons au sein d'une même molécule d'ARN. **(b)** Le SL trans-splicing fait intervenir deux molécules d'ARN : le SL ARN et le pré-ARNm. **(c-d)** le génique trans-splicing fait intervenir deux molécules d'ARN qui ne sont ni l'une ni l'autre un SL ARN et qui proviennent soit du même gène soit de gènes différents. « ss » signifie « site de splicing », « tss » signifie « site de trans-splicing ».

existe. Deuxièmement, jusqu'à présent, seul le cas d'opérons caractérisés par un trans-splicing SL1 pour premier gène de l'opéron et un trans-splicing SL2 pour les gènes en aval a été abordé. Cependant, il existe d'autres sortes d'opérons. Il existe aussi des opérons SL1 (dont tous les gènes de l'opéron sont trans-splicés SL1) et des opérons hybrides (dont les gènes peuvent être trans-splicés SL1 ou SL2) (*18*). La particularité des opérons hybrides est qu'ils disposent de promoteurs alternatifs internes à l'opéron. Dès lors, la transcription de certains gènes peut avoir lieu de manière indépendante de l'opéron lui-même. De plus, ce promoteur alternatif permet d'avoir un trans-splicing SL1 du gène (*30*). En effet, lorsque la transcription se fait sur l'opéron entier, les gènes en position 2 ou plus de l'opéron sont trans-splicés SL2. Par contre, lorsque c'est le promoteur alternatif qui initie la transcription, le gène est trans-splicé SL1. Cela est, par exemple, le cas du gène splt-1 qui est trans-splicé à 71% SL1 et 29% SL2. Ce type d'opéron représente un quart des opérons et 5% du génome de *C. elegans* (*19*).

Il est dès lors intéressant de remettre ce type de splicing dans son contexte évolutif. Tout d'abord, on distingue différents types de splicing. Alors que le cis-splicing consiste en un épissage des exons au sein d'une même molécule d'ARN, le trans-splicing se passe entre deux molécules d'ARN. Cependant, on distingue encore le SL trans-splicing (une des deux molécules est le splice leader) du génique trans-splicing qui est lui-même subdivisé en intragénique (les deux molécules proviennent du même gène) et intergénique (les deux molécules d'ARN proviennent de deux gènes différents). En résumé, il y a donc quatre sortes de splicing (**figure 6**) (*29*).

Blumenthal (2005) a établi que le SL trans-splicing est associé avec l'évolution des opérons. En effet, le SL trans-splicing se passe au sein d'organismes disposant de structures en opérons alors que ces dernières sont plus rares chez les vertébrés. Ainsi, une analyse phylogénétique des événements de trans-splicing a révélé que ceux-ci sont relativement présents chez les protozoaires, certains radiaires ainsi que certains protostomiens tels les nématodes et que leur nombre diminue chez les eucaryotes supérieurs (*31*).

2.4. Études de CDK-12 chez *C. elegans*

2.4.1. Les kinases dépendantes de cycline

Les kinases dépendantes de cyclines (CDK) sont notamment connues pour leur rôle au niveau du cycle cellulaire. Ces kinases s'associent à une cycline et régulent les transitions du cycle en phosphorylant un large set de substrats (*32*). Plus tard, d'autres CDK ont été découvertes et impliquées dans le contrôle de la transcription (*32*). Parmi celles-ci, on compte notamment CDK7, CDK9 et CDK12.

2.4.2. Études précédentes

C'est en 2010 que Greenleaf a démontré le rôle de CDK-12 dans la phosphorylation de la sérine 2 du CTD (*33*). Cependant, un mutant homozygote knockout du gène chez *C. elegans* ne permet pas d'étudier le rôle physiologique de CDK-12. En effet, ce mutant présente un développement embryonnaire normal, mais arrête son développement en L1 (thèse de C. Cassart). Dès lors, une perte totale et permanente de l'activité de CDK-12 bloque le développement des vers en L1 directement après éclosion. C'est pourquoi les mutants de délétion de *cdk-12* ne peuvent pas être maintenus à l'état homozygote.

Afin de contourner ce problème, une approche basée sur l'interférence ARN (ARNi) peut être utilisée, dans le but de diminuer l'expression de *cdk-12* dans des vers sauvages. Notons que, contrairement à l'approche génétique, l'ARNi a rarement un effet complet, le silençage dépend de la présence d'ARN exogènes. Le principe de l'ARNi est qu'un ARN exogène va empêcher la traduction d'ARNm endogène et, par conséquent, diminuer le taux d'expression de la protéine (*34*). Pour réaliser cela, de l'ARN double brin est injecté à des vers en stade L4 et ensuite, leur descendance est étudiée. Dans le cas de CDK-12, un arrêt des vers au stade L1 est observé ((*3*) ; thèse de C. Cassart). Les phénotypes observés après ARNi ou délétion du gène sont donc identiques.

Un arrêt en L1 est également observé dans des conditions physiologiques lorsque les œufs éclosent en absence de nourriture, ce qui suggère que les mutants *cdk-12* se comportent comme des vers starvés malgré la présence de nourriture. La question biologique qui se pose alors est de savoir comment l'absence de *cdk-12* résulte en un phénotype similaire à la starvation.

L'inconvénient de l'approche ARNi dans le contexte de l'étude d'une protéine est qu'elle provoque le plus souvent un silençage incomplet et pas toujours reproductible (*35*). Une approche génétique résoudrait ce problème, mais comme dit plus haut, le mutant est létal.

Le laboratoire s'est récemment attelé à générer un mutant conditionnel pour pouvoir analyser en détail le phénotype généré par l'inactivation de CDK-12.

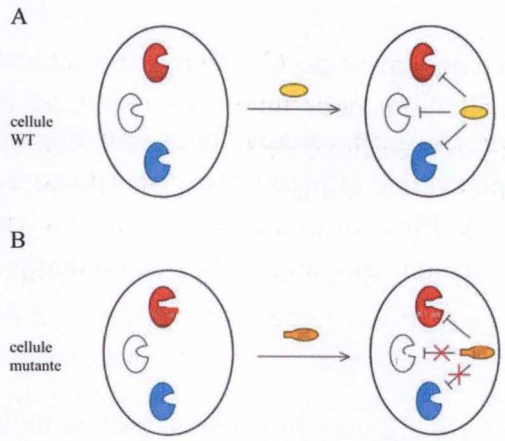
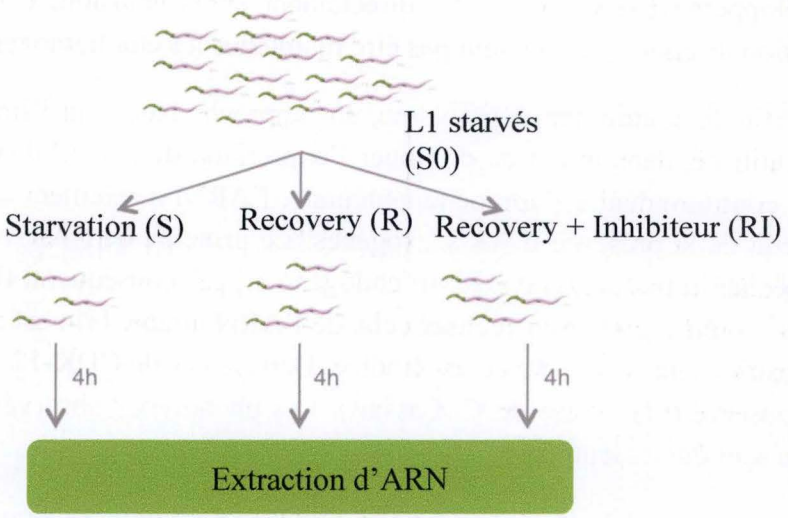


Figure 7 : Principe de la méthode « analog-sensitive » (figure modifiée à partir de l'article de Gregan *et al.*). (**A**) Utiliser un inhibiteur contre le site catalytique d'une kinase mène à l'inhibition de toutes les kinases présentes dans la cellule. Autrement dit, cela n'est pas spécifique. (**B**) La méthode « analog-sensitive » vise à introduire une mutation dans le site catalytique de la kinase cible en vue de pouvoir l'inhiber spécifiquement grâce à l'utilisation d'un inhibiteur adapté.

A



B

	SO	S	R	RI
DMSO	Non	Oui	Oui	Oui
Nourriture	Non	Non	Oui	Oui
Inhibiteur	Non	Non	Non	Oui

Figure 8 : Création et caractérisation des différentes conditions testées. (**A**) Génération des différentes conditions. On voit sur cette figure qu'une population initiale de vers starvés en L1 est divisée pour créer les trois conditions : S, R et RI (schéma modifié à partir du power point de C. Cassart). (**B**) Caractéristiques des différentes conditions.

2.4.3. Études actuelles

Afin d'approfondir le rôle de CDK-12 dans le développement de *C. elegans*, un mutant conditionel « *analog-sensitive* » (*as*) a été créé. La méthode a pour intérêt de combiner la génétique et la chimie de manière à permettre l'inhibition spécifique de CDK-12 (**figure 7**). Les kinases ont un site catalytique fortement conservé si bien qu'un inhibiteur le ciblant va inhiber toutes les kinases. Le principe de la méthode *as* est d'introduire une mutation dans la séquence du site catalytique de CDK-12 de sorte à obtenir une kinase avec un site catalytique un petit peu plus grand. Dans le cas présent, il s'agit de remplacer une phénylalanine (acide aminé avec une grande chaîne latérale) par une glycine (acide aminé avec une courte chaîne latérale). Une fois la souche mutante créée, un analogue plus volumineux de l'ATP sert d'inhibiteur spécifique de la kinase d'intérêt (*36*) (*37*) .

Lorsque des vers adultes homozygotes pour *cdk-12as* sont mis en présence d'inhibiteur, un arrêt en L1 est observé, ce qui confirme le phénotype observé suite à la délétion ou par ARNi. De manière remarquable, lorsque l'inhibiteur est éliminé par lavage, les vers poursuivent un développement normal. En outre, en absence d'inhibiteur, aucun défaut de croissance des vers n'est remarqué. Cela indique que l'inhibition de CDK-12 est spécifique et réversible.

Pour investiguer l'implication de CDK-12 dans le développement, une étude transcriptomique du mutant *cdk-12as* a été réalisée. Le changement de niveau d'expression de chaque transcrit suite à l'inhibition de CDK-12 peut dès lors être analysé. Pour permettre cette étude, des vers starvés en L1 sont utilisés. Ceux-ci sont ensuite divisés en trois conditions:

- des L1 qui sont maintenus en starvation (condition de starvation ou S),
- des L1 qui sont mis en présence de nourriture de manière à permettre leur sortie de starvation (condition de recovery ou R)
- des L1 en recovery en présence d'inhibiteur (condition recovery + inhibiteur ou RI) (**figure 8**). Un RNA-seq est réalisé dans les différentes conditions.

Le RNA-seq permet de séquencer le transcriptome. Il comprend 4 étapes : la génération d'une librairie, le séquençage en lui-même, le mapping des reads séquencés et la "summarization", c'est-à-dire la détermination du nombre de reads mappant sur un gène (**figure 9A, verso de la page suivante**). Premièrement, l'ARN total est extrait. A partir de cette fraction, une ribodéplétion est réalisée en vue d'enrichir la fraction avec d'autres ARN que les ARNr, qui sont initialement présents en grand nombre et qui représentent, dans le cas d'études liées à l'ARN polymérase II, une contamination (**figure 9B, verso de la page suivante**). Ensuite, ces ARN sont fragmentés en plus petites séquences (100 – 300 bp) et une rétro-transcription est réalisée afin d'obtenir l'ADN complémentaire des ARN (ADNc). L'ajout d'une seconde séquence heptanucléotidique permet de générer le second brin d'ADNc. De plus, une adénosine est ajoutée à l'extrémité 3' par l'adénylate ligase, ce qui permet d'obtenir des extrémités « chunk » et de lier les primers de séquençage. Deuxièmement, le séquençage est fait de manière « paired-end », c'est-à-dire que les deux extrémités des fragments sont séquencés. On obtient donc des reads pairés. Troisièmement, les reads pairés sont mappés sur le génome. Notons que, si un fragment provient de la partie 5' d'un ARN trans-splicé, un read (50bp) porte la séquence du splice leader (~22bp) et par

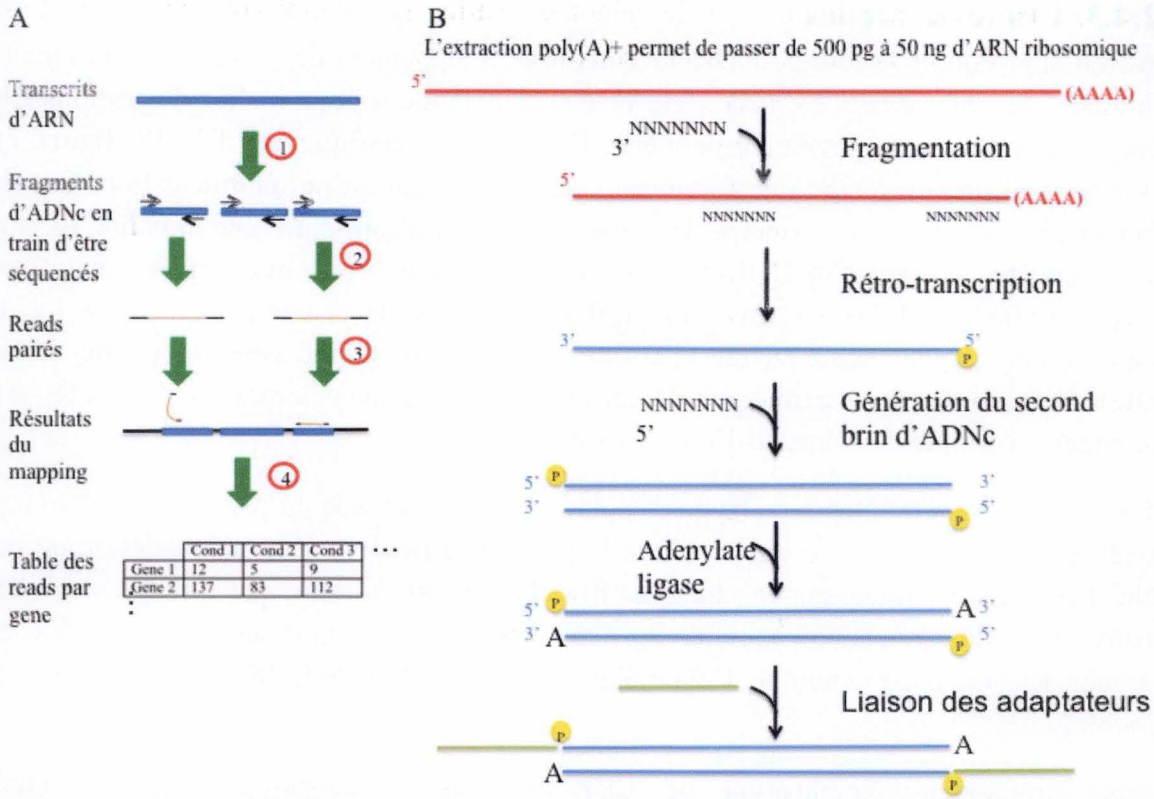


Figure 9 : Principes du RNA-séquençage. (**A**) L'ARN total est extrait, (1) fragmenté avant de subir une rétro-transcription en vue d'obtenir l'ADN complémentaire (ADNc) (2) Ensuite, ces fragments d'ADNc sont séquencés de manière « paired-end », c'est-à-dire que les deux extrémités du fragment sont séquencées. Ainsi, des reads pairés sont obtenus, où chaque read fait 50 nucléotides (nt). (3) Suite au séquençage, ces reads pairés sont mappés sur le génome. Il faut souligner le fait que chez *C. elegans*, il y a un phénomène de trans-splicing. Le read provenant de l'extrémité 5' d'un transcrit trans-splicé porte donc les 22 nt de la séquence SL et 23 nt du pré-ARNm, c'est pourquoi ce read dit hybride ne mappe pas sur le génome. Finalement, une summarization permet de générer la table count. (**B**) Description de la première étape. L'extraction poly(A) permet d'éliminer entre 500 pg et 50 ng d'ARN ribosomique. Ensuite, l'ARN est fragmenté et un heptanucléotide est ajouté. Ce dernier permet d'obtenir un brin d'ADNc suite à une étape de rétro-transcription tandis que l'ARN est dégradé. L'ajout d'un second heptanucléotide permet d'obtenir le deuxième brin d'ADNc et l'addition d'adénylate ligase a pour intérêt d'ajouter une adénosine (A) à l'extrémité 3'. Cet ajout d'un A en 3' permet la liaison spécifique des adaptateurs et le phosphate en 5' permet de lier ce primer à l'ADNc.

conséquent, ne peut être aligné sur le génome. Enfin, la "summarization" permet de déterminer l'expression relative de chaque gène (*38*).

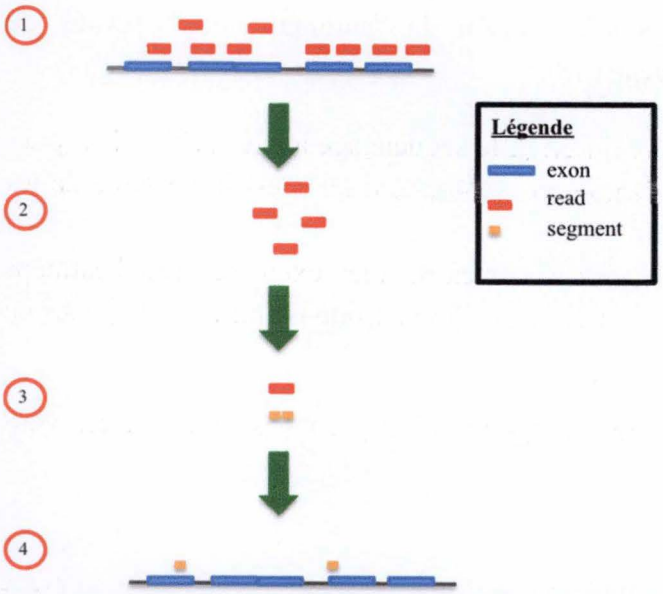


Figure 10 : Principe de TopHat2. Le programme TopHat2 va permettre de faire un mapping en tenant compte du splicing. Pour cela, il fonctionne en quatre étapes. (1) Les reads sont mappés sur le génome. (2) Les reads chevauchants sont récupérés. (3) Ces reads sont subdivisés en fragments. (4) Les fragments sont mappés sur le génome.

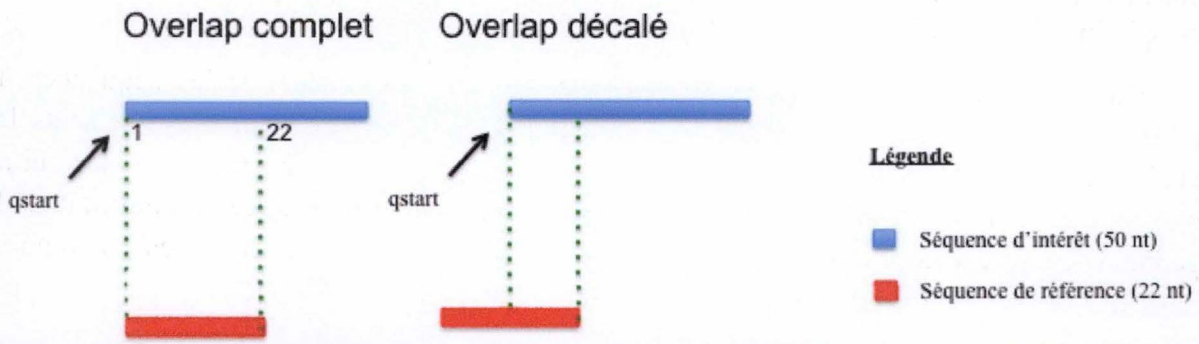


Figure 11 : Overlap complet ou décalé. Lors d'un blast, un overlap est établi entre la séquence d'intérêt et la séquence de référence. La position du début de l'overlap sur la séquence d'intérêt porte le nom de qstart.

2.5. Programmes bioinformatiques utilisés

Les méthodes bioinformatiques permettent

- de traiter une grande quantité de données. En effet, le séquençage à haut débit (HTS pour high througout sequencing) génère un tel nombre de données qu'il est impossible de les analyser autrement.
- de concrétiser et représenter des données complexes. Par exemple, les heatmaps d'expression permettent de visualiser les différences d'expression de milliers de gènes en un seul graphique.

Au cours de ce mémoire, différents logiciels, programmes et packages sont utilisés. Nous avons fait le choix de les introduire ici.

2.5.1. Tophat2

TopHat2 est un programme qui permet de mapper des données brutes de RNA-seq sur une séquence de référence, dans notre cas, le génome de *C.elegans*, en tenant compte du cis-splicing. En effet, les reads provenant du RNA-seq ne contiennent que des exons tandis que les séquences génomiques de référence contiennent une alternance d'exons et d'introns. De plus, l'ordre des exons au niveau du génome et celui du transcrit n'est pas le même, en raison du splicing. TopHat2 doit donc tenir compte de ces deux subtilités.

Concrètement, TopHat2 réalise un premier mapping des reads sur les séquences de référence (**figure 10**) (*39*). Le génome étant une alternance d'introns et d'exons, seuls les reads dont la séquence ne provient que d'un seul exon mappent directement. Les reads qui ne mappent pas sont eux découpés en plus petites séquences et sont mappés à nouveau sur le génome de référence. Ce second mapping permet de trouver la provenance de reads composés de séquences provenant d'exons différents.

Finalement, un fichier d'alignement est généré à la fin du mappping, fichier qui contient les reads, leur position sur le génome de référence, ainsi que différents "flags" (ex : *read paired, first read, read reverse strand, ...)* donnant des informations supplémentaires quant au mapping.

2.5.2. SAMtools

SAMtools est un programme qui permet de manipuler les fichiers SAM ou BAM et d'en extraire spécifiquement certains reads qui correspondent à certains critères basés sur les "flags" du fichier d'alignement (*40*).

2.5.3. blast +

Le principe de blast + (Basic Local Allignement Search Tool +) est de diviser la séquence d'intérêt, la query, en plus petites séquences, de mapper ces dernières sur une séquence de référence et d'élargir ce mapping aussi loin que possible en déterminant un score (*41*). Ce score va prendre en compte les matchs, les insertions, les délétions, ... Ainsi, un overlap est défini (zone de recouvrement de la séquence de référence par la query). Cet overlap est caractérisé par un qstart, c'est-à-dire la position de la query à laquelle l'overlap commence (**figure 11**).

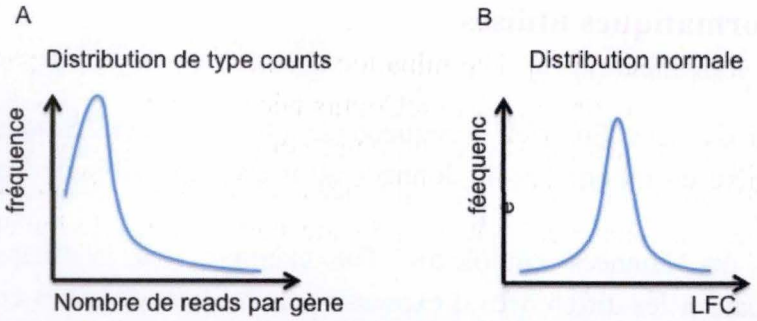


Figure 12 : Représentation d'une distribution count comparé à une distribution normale. LFC est l'acronyme du logarithme du fold change, c'est-à-dire $\log(\frac{\textit{expression du gène dans la condition A}}{\textit{expression du gène dans la condition B}})$. (**A**) La distribution de type count se caractérise par une fréquence importante de gènes avec un faible nombre de reads. (**B**) Distribution normale obtenue en urtilisant le LFC sur l'axe des abscisses.

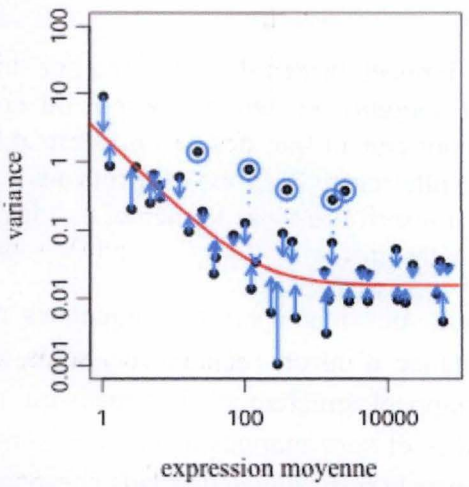


Figure 13 : Correction des variances réalisée par DESeq2 (figure modifiée à partir de Love *et al.*, 2014). Cette correction se base sur la relation entre la variance de l'expression des gènes et leur expression moyenne. Les points noirs représentent les estimations de variance avant correction et les flèches bleues permettent de se rendre compte de la correction qu'opère DESeq2. Notons que pour les gènes avec une estimation de variance élevée (points noirs entourés par un cercle bleu), la correction n'est pas réalisée. Cette correction a pour but de limiter les faux positifs.

2.5.4. featureCounts

FeatureCounts est un programme qui permet la détermination du nombre de reads par gène à partir des résultats du mapping (*42*). Pour cela, featureCounts nécessite deux types de fichier. un fichier qui contient les reads alignés (.BAM) et un autre comportant l'annotation simplifiée du génome, c'est-à-dire un fichier en cinq colonnes : le nom du gène, le chromosome sur lequel il est situé, le brin concerné, la position du site start de transcription et celle de la fin de transcription.

2.5.5. DESeq2

Pour réaliser une analyse différentielle des données de RNA-seq, c'est-à-dire une analyse de la significativité des différences entre les conditions, le package DESeq2 disponible sur le logiciel R est utilisé. En effet, deux problèmes majeurs sont rencontrés lors de l'analyse de telles données. Premièrement, la distribution de données RNA-seq (nombre de reads mappés par gène) est de type count alors que la plupart des tests statistiques se basent sur une distribution normale (**figure 12**). Deuxièmement, il y a peu de réplicats utilisés lors du RNA-seq. Ainsi, la confiance envers les estimations de la moyenne et de la variance est faible et le fold change mesuré, c'est-à-dire $\frac{l'expression\ génique\ dans\ la\ condition\ A}{l'expressiongénique\ dans\ la\ condition\ B}$, est peu fiable. DESeq2 réalise une série d'améliorations par rapport à un test classique (une simple ANOVA par exemple) afin de résoudre au mieux ces problèmes (*43*).

Pour commencer, DESeq2 corrige l'estimation de la variance en prenant en compte la relation entre la variance et l'expression moyenne des gènes (**figure 13**).

Ensuite, DESeq 2 décompose l'expression génique selon les différents facteurs considérés dans l'étude. La contribution de chaque facteur est ensuite estimée en logarithme de fold change (LFC) et sa significativité est évaluée (Wald test pvalue).

Finalement, DESeq2 fait une correction du logarithme du fold change (LFC). Ainsi, DESeq2 permet d'être plus stringent et d'éliminer d'éventuels faux positifs dus au faible nombre de réplicats (*43*) (*44*).

2.5.6. Ape

Le programme Ape permet de pouvoir manipuler une séquence. Ape peut, par exemple, être utilisé pour visualiser les produits de restriction d'une enzyme et ainsi pouvoir prédéfinir les enzymes de restriction à employer. Ici, il sera utilisé dans le but de déterminer des primers dans le cadre du primer extension.

2.6. Objectifs

Le premier objectif de mon mémoire est de comprendre comment CDK-12 influence le processus de transcription et le développement de *C. elegans*, en analysant l'effet de l'inhibition de la kinase sur le transcriptome. Comme le phénotype résultant de l'inhibition de CDK-12 est similaire à celui d'une diapause, il est intéressant d'analyser le transcriptome dans une condition de starvation/recovery, avec et sans inhibition de CDK-12.

Le deuxième objectif est d'étudier le mécanisme moléculaire par lequel CDK-12 agit sur la transcription. Pour cela, différentes techniques sont mises au point en vue d'essayer d'utiliser les données de RNA-seq pour investiguer le rôle de CDK-12 dans le trans-splicing

3. Matériel et méthodes

3.1. Culture de *C. elegans* en laboratoire

Deux types de cultures du nématode peuvent être faites : la culture sur boîte de pétri et celle en milieu liquide (*45*).

Concernant la culture sur boîte de pétri, on commence par préparer un milieu Nematode Growth Milieu (NGM) qui contient de l'agar, du NaCl, du bactopecton et de l'eau. Ensuite, on va mettre en culture une souche bactérienne qui servira de nourriture pour les vers. La souche bactérienne utilisée est *E. coli* OP50. En effet, la souche d'*E. coli sp.* prolifère plus rapidement que *C. elegans* si bien que l'observation des nématodes au sein de la boîte de pétri est rendue difficile. La souche OP50 est une souche d'*E. coli* manipulée génétiquement afin d'optimaliser la croissance bactérienne par rapport à celle de *C. elegans* et ainsi, de permettre une observation plus aisée. Les vers sont ajoutés et conservés à 20°C pour permettre un bon développement. Dans le but d'éviter la starvation des nématodes, un transfert est réalisé tous les 3-4 jours. Deux grandes méthodes sont utilisées au cours de ce mémoire : le chunk et le transfert ver par ver. Le premier consiste à utiliser un scalpel stérile, à découper un morceau du milieu de culture (classiquement un carré de plus ou moins un demi centimètre) et à le transférer sur une nouvelle boîte de pétri (contenant du milieu NGM et une population d'*E. coli* OP50). Ainsi, les vers prélevés de la première boîte vont se répandre sur la nouvelle boîte. Notons que cette technique peut être utilisée dans notre cas, car il s'agit de boîtes homogènes où tous les individus ont le même phénotype, soit 100 % de wild-type (WT) ou 100 % de mutant *cdk-12as*. Si, par contre, les boîtes sont hétérogènes, cette technique ne peut être utilisée, car le transfert aurait favorisé le développement d'un des phénotypes, c'est-à-dire qu'en terme de proportion, on aurait peut-être prélevé un carré avec 70% d'individus WT et 30% de *cdk-12as*, ce qui aurait favorisé par la suite le développement des WT. Une seconde technique utilisée est le transfert individuel des vers. Pour cela, on utilise un pick terminé par un fil de platine. Concrètement, on effleure légèrement le milieu afin de récupérer un agrégat de bactéries et attraper le ver qui va alors coller à l'extrémité du fil. Ensuite, on pose doucement le nématode sur la nouvelle boîte de pétri. L'avantage de la platine est qu'elle peut être passée sous flamme entre chaque transfert pour éviter toute contamination et refroidir rapidement. Cette technique est plus précise que la précédente et est utilisée pour plusieurs manipulations dont notamment l'extraction d'ADN du nématode qui nécessite la lyse individuelle de vers.

Le deuxième moyen de faire une culture de nématodes au laboratoire est de la réaliser en milieu liquide. Celle-ci a pour avantage de réaliser une culture d'un plus grand nombre de vers comparé à la culture sur boîte. Pour cela, un milieu appelé S Medium et une solution concentrée d'*E. coli* OP50 sont utilisés. Il est cependant conseillé de ne pas laisser se développer plus d'une génération de nématodes, car on observe alors la formation du stade dauer du ver malgré la présence de nourriture.

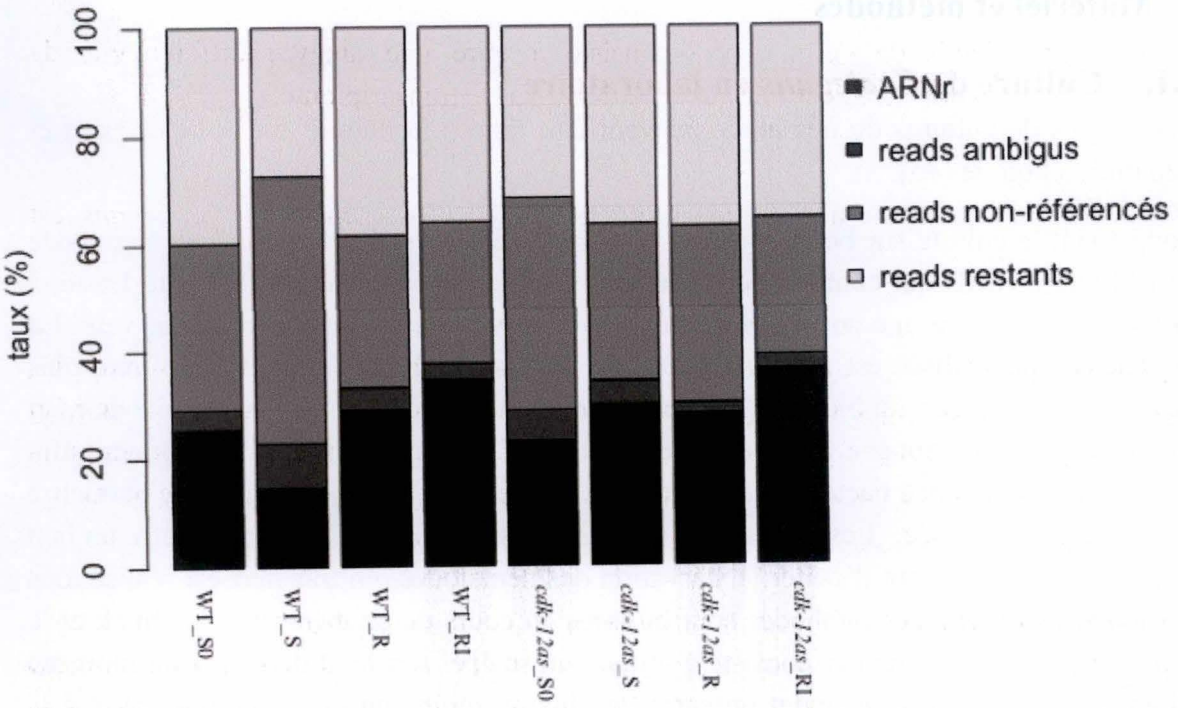


Figure 14 : Proportion des différentes natures des reads au sein des résultats bruts du RNA-seq. L'axe des abscisses permet d'identifier la condition (S0, S, R ou RI) ainsi que la souche (WT ou *cdk-12as*) concernée. L'axe des ordonnées est le taux exprimé en pourcent. Les différentes nuances de gris permettent d'identifier la nature des reads : reads correspondant à de l'ARN ribosomique (ARNr), reads correspondant à une séquence ambiguë (par exemple, reads chevauchants), reads correspondant à une séquence non-référencée et les reads restants (correspondant donc à des ARN transcrits).

3.2. Analyse différentielle du transcriptome

Le premier objectif de ce mémoire consiste à faire une analyse différentielle de l'expression génique afin d'évaluer le rôle de CDK-12 sur le transcriptome.

3.2.1. Version du génome de *C. elegans* utilisée

Pour réaliser cette analyse différentielle, la version WS235 du génome de *C. elegans*, est utilisée (www.wormbase.org).

3.2.2. Le modèle utilisé

L'analyse statistique de données RNA-seq nécessite de corriger l'estimation des variances. Pour réaliser une analyse différentielle de l'expression, le package DESeq2 disponible sur R est utilisé. En effet, l'intérêt de DESeq2 est qu'il permet, à partir d'un modèle, de considérer l'effet sur l'expression de chaque facteur séparément. Dans le cas présent le modèle employé est

```
Design=~strain+DMSO+food+drug+batch+strain:food+strain:drug
```

Ce qui signifie que la souche, le DMSO, la nourriture, l'inhibiteur et le réplicat peuvent avoir un effet sur l'expression génique, auxquels viennent s'ajouter les interactions (un terme peut influencer l'autre) entre la souche et la nourriture ainsi que celles de la souche et l'inhibiteur.

3.2.3. Extraction des gènes d'intérêt

L'analyse des résultats de featureCounts permet d'identifier différentes séquences sur lesquelles mappent les reads parmi lesquels les ARN ribosomiques (ARNr) et ARN messagers (**figure 14**). Afin d'éliminer les reads correspondant aux ARNr, les résultats de featuresCounts sont filtrés.

3.2.4. Relation entre conditions

Dans le but d'analyser la relation entre les réplicats et entre les conditions, la distance euclidienne entre ceux-ci est calculée et les résultats sont représentés à l'aide d'une heatmap des distances. En effet, il existe une relation inversément proportionnelle entre cette distance et la corrélation. La heatmap des distances illustrée dans ce travail a été réalisée grâce au package heatmap disponible sur R et elle a été colorée grâce au package RcolorBrewer également disponible sur R.

3.2.5. Représentation des résultats de l'analyse statistique

Pour mettre en évidence les différences au niveau de l'expression génique, un Wald test est effectué et les résultats sont représentés sous forme de MAplots. Les Maplots sont des graphiques du log de fold change des gènes en fonction de leur expression. De plus, les gènes y sont représentés de différentes couleurs en fonction de leur significativité (valeur par défaut : p-value ajustée <0,1). La p-value ajustée est une correction de la p-value par rapport au nombre de tests statistiques effectués. En effet, la p-value n'est pas dépendante du nombre de tests tandis que la p-value ajustée l'est. Pour concrétiser cela, prenons par exemple 1000 valeurs dont 250 sont positives sur base de tests statistiques. Avoir une p-value de 5% signifie que 50 valeurs seront, par hasard, positives alors qu'avoir une p-value ajustée de 5% signifie que, parmi les 250 positifs, 10 sont dus au hasard. Autrement dit, la p-value ajustée est plus

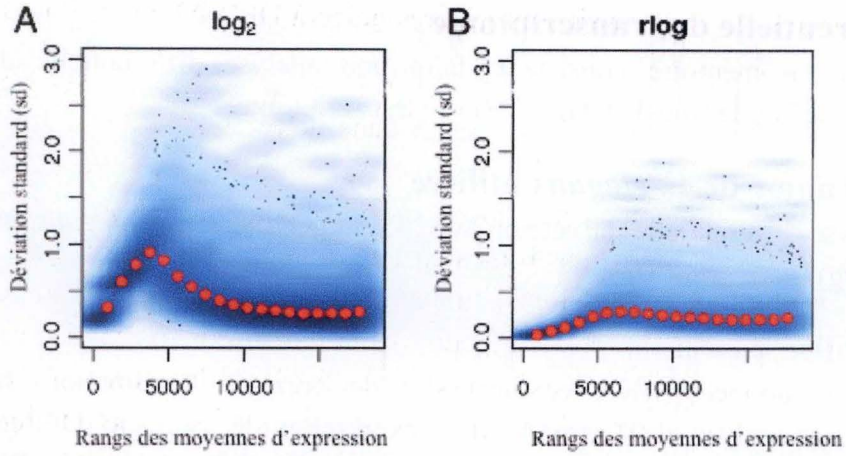


Figure 15 : Visualisation de la transformation logarithmique (log2) comparé à la transformation rlog (regularized logarithm transformation ou transformation en logarithme régularisé) (figure modifiée à partir de Love *et al.*, 2014). La courbe rouge représente la relation entre la déviation standard (sd) et le rang des moyennes d'expression. (**A**) La transformation logarithmique présente une hétérogénéité entre les sd des valeurs caractérisées par une faible expression et celles des valeurs caractérisées par une expression plus importante. (**B**) La transformation rlog mène à une meilleure homogénéité.

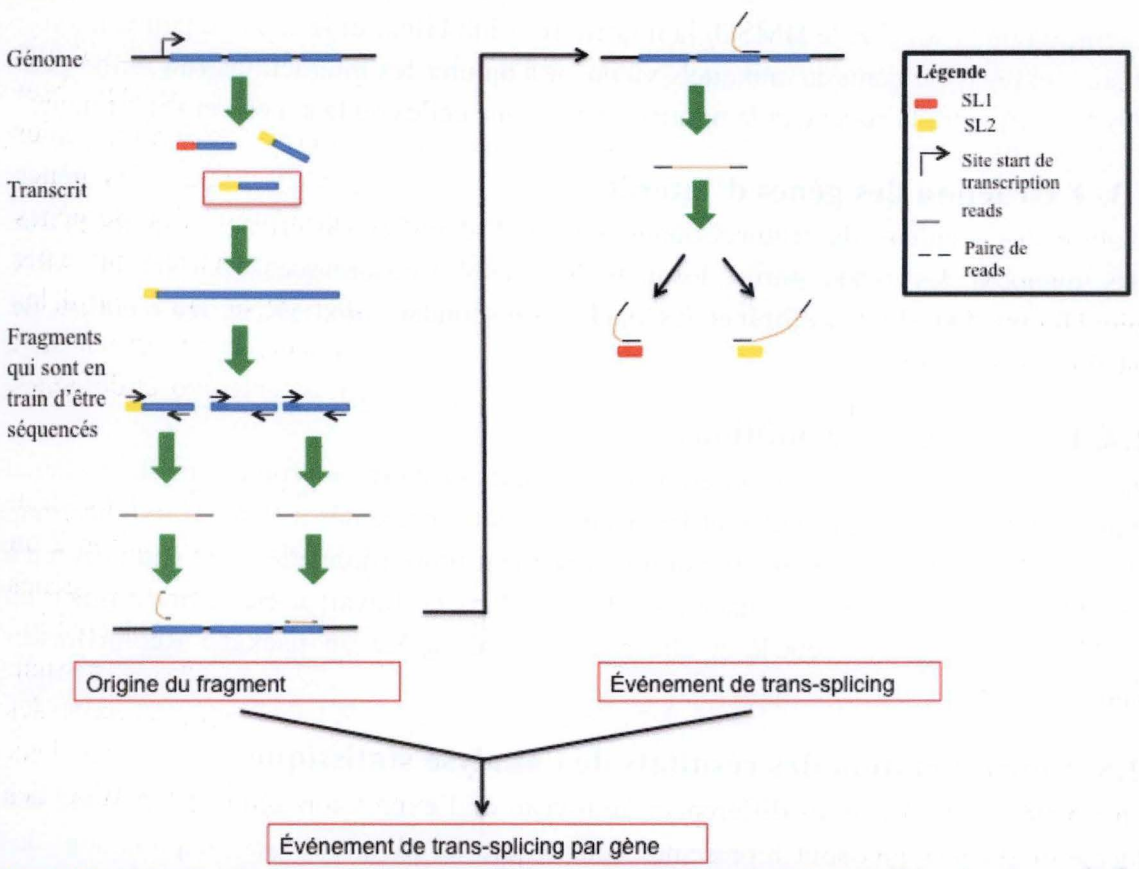


Figure 16 : Méthode mise au point pour la quantification des événements de trans-splicing. Le génome de *C. elegans* est naturellement transcrit et trans-splicé. Ensuite, l'ARN total est extrait et les ARN messagers sont fragmentés et rétro-transcrits. Par la suite, ces fragments sont séquencés de manière « paired-end », c'est-à-dire que les deux extrémités du fragment sont séquencées sur 50 nucléotides (nt). Après ce séquençage, les reads pairés ainsi obtenus sont mappés sur le génome. Cependant, si un read provient de l'extrémité 5' d'un ARN trans-splicé, le read porte la séquence SL (22 nt) et une partie de la séquence du transcrit (23 nt). Par conséquent, ce read ne mappe pas sur le génome. Néanmoins, le second read permet d'obtenir une information sur l'origine du fragment. La quantification des événements de trans-splicing repose sur la récupération de ces reads pairés dont un read mappe et l'autre non, de blaster ce dernier contre les séquences SL en vue de déterminer s'il provient d'un événement de trans-splicing SL1 ou SL2 et enfin, d'utiliser le programme featureCounts pour combiner les deux informations disponibles (l'origine du fragment et le type d'événement de trans-splicing) en vue de quantifier le nombre d'événements de trans-splicing par gène.

stringente. Les Maplots présentés dans ce travail sont générés grâce à DESeq2 en utilisant le même modèle que ci-dessus (point 2.2.2.).

Pour avoir une vue globale des variations d'expression dans notre étude, une heatmap d'expression à travers les conditions a été réalisée. Pour effectuer cela, le nombre de reads des gènes différentiellement exprimés sont extraits spécifiquement grâce aux critères padj < 0,05 et LFC < log2(1,5). Ensuite, ces reads sont transformés en logarithme régularisé (rlog). La transformation rlog est une transformation similaire à la transformation logarithmique pour les valeurs avec une expression élevée (**figure 15**). Cependant, cette transformation réalise une correction pour les valeurs caractérisées par une expression assez faible. Cela permet d'avoir une certaine homogénéité des valeurs contrairement à la transformation logarithmique traditionnelle. Pour convertir les reads en rlog, la fonction rlog() disponible dans le package DESeq2 est utilisée. Par la suite, le Z score de ces dernières valeurs est calculé, c'est-à-dire que les données rlog sont centrées et réduites. Enfin, les résultats sont représentés sous forme de haetmap grâce à la fonction heatmap.2 du package gplots.

3.2.6. Analyse des gènes disposés en opérons

Lors de l'analyse de la heatmap d'expression, un subset de gènes downrégulés a été mis en évidence comme étant affecté suite à l'inhibition de CDK-12. Il s'est avéré que ce subset de gènes présente un enrichissement en gènes disposés en opérons. Pour le vérifier, un fichier d'annotation du génome de *C. elegans* (WS235) comportant une annotation des gènes disposés en opérons, a été téléchargé à partir de "wormbase". Ainsi, la proportion de gènes disposés en opérons au niveau du génome de *C. elegans* peut être évaluée. Par la suite, cette même annotation est utilisée pour déterminer la proportion de gènes en opérons parmi les gènes downrégulés suite à l'ajout de l'inhibiteur de CDK-12. Enfin, ces deux proportions sont comparées en vue de mettre en évidence un éventuel enrichissement dans le subset de gènes downrégulés.

Dans le but de mettre en évidence un effet plus important de l'inhibiteur selon la position du gène dans l'opéron, la proportion de gènes en position 1 des opérons ou en position 2 ou plus est analysée au niveau du génome et au niveau du subset. Pour cela, les gènes appartenant à un même opéron sont isolés. Ensuite, ces gènes sont classés selon leur position dans le génome. Ainsi, l'ordre établi est soit croissant, s'il s'agit du brin +, soit décroissant, s'il s'agit du brin -. Enfin, les gènes sont classés en deux catégories : ceux en position 1 des opérons et ceux en position 2 ou plus. Ainsi, en représentant le taux de gènes dans ces deux catégories sous forme de "stacked barplot", on peut mettre en évidence la proportion des gènes en position 1 des opérons par rapport à ceux en position 2 ou plus.

Une troisième analyse de ce subset consiste à déterminer la fonction des gènes concernés par la downrégulation. Ainsi, un "gene ontology enrichment analysis" a été réalisé par Carlo Yague-Sanz. Le principe est de regarder si une fonction biologique est enrichie dans un groupe de gènes par rapport au génome complet grâce à un test de Fisher.

Tableau 1 : Code utilisé lors de l'extraction, sa signification et son objectif.

code	*samtools view -f 133 -F 8 input > output*	*samtools view -f 73 -F 260 input > output*
signification	**Ce que nous souhaitons** : que les reads soient pairés, que le read ne mappe pas et qu'il s'agisse d'un read 2 **Ce que nous ne souhaitons pas** : que le read qui lui est pairé (son mate) ne mappe pas	**Ce que nous souhaitons** : que les reads soient pairés, que le read qui lui est pairé (son mate) ne mappe pas et qu'il s'agisse d'un read 1 **Ce que nous ne souhaitons pas** : que le mate mappe et qu'il s'agisse d'un alignement secondaire
objectif	Récupérer les reads qui portent potentiellement la séquence SL	Récupérer les mates des reads qui portent potentiellement la séquence SL

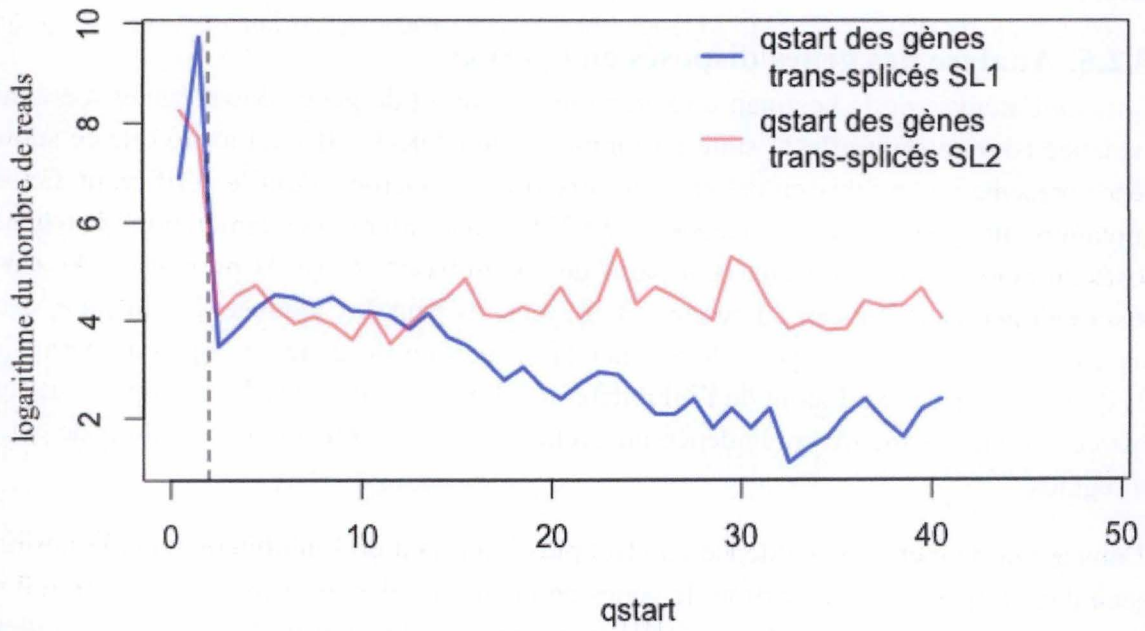


Figure 17 : Distribution du qstart, c'est-à-dire le début de l'overlap des reads blastés sur la séquence SL. La courbe bleue montre la distribution du qstart des reads identifiés comme étant SL1. La courbe rouge montre la distribution du qstart des reads identifiés comme étant SL2. La ligne verticale permet de distinguer les reads spécifiques au trans-splicing (à gauche de la ligne) de ceux qui sont aspécifiques (à droite de la ligne).

3.3. Quantification des événements de trans-splicing

Suite aux résultats de la heatmap (point 3.1 des résultats), l'analyse des données RNA-seq suggère un rôle de CDK-12 dans le trans-splicing de type SL2. C'est pourquoi un des buts de mon mémoire est de quantifier les événements de trans-splicing. Deux méthodes permettent une telle quantification : une basée sur le blast et l'autre sur un mapping. La méthode basée sur le blast est assez rapide, car elle peut être réalisée directement à partir des reads déjà alignés sur le génome lors de l'analyse transcriptomique. Bien qu'elle soit plus lente, car elle requiert plus d'étapes et repart des données brutes (reads non alignés), la technique du mapping permet de prendre potentiellement plus de données en compte et de ce fait, améliorer la confiance des résultats de l'analyse en aval.

3.3.1. Méthodes testées

a) Méthode reposant sur un blast

Dans le but de quantifier les événements de trans-splicing, une nouvelle méthode est mise au point (**figure 16, verso de la page précédente**), basée sur les résultats du RNA-seq et qui comporte 3 étapes.

Tout d'abord, le programme SAMtools est utilisé pour extraire, à partir d'un fichier d'alignement, les reads pairés dont un des reads ne mappe pas. En effet, les reads qui ne mappent pas sur le génome portent potentiellement la séquence SL. Le **tableau 1** reprend le code utilisé afin de les sélectionner et les extraire des fichiers SAM contenant les reads alignés sur le génome.

Ensuite, les reads ainsi extraits sont blastés sur les séquences SL dans le but de déterminer si le read provient d'un événement de trans-splicing SL1 ou SL2. Pour cela, la fonction blastn (nucleotide blast) du programme blast + est utilisée. Cette fonction permet d'aligner deux séquences nucléotidiques en déterminant un overlap d'une certaine taille (par défaut, l'overlap minimum est de 11 nt). Ainsi, on obtient une série de reads qui sont potentiellement hybrides (reads qui comportent la séquence SL juxtaposée à celle du pré-ARNm) qui sont filtrés par la suite afin d'obtenir une spécificité maximale. Pour cela, nous précisons que le début de l'overlap doit être égal à deux. En effet, comme le trans-splicing résulte de l'addition de la séquence SL en 5' du pré-ARNm, la position du début de l'overlap doit être égale à un. Cependant, l'adénylate ligase utilisée lors du séquençage va ajouter une adénosine à l'extrémité de certains reads. C'est pourquoi la valeur du début de l'overlap doit être égal ou inférieure à deux. Dans la suite de ce mémoire, on appellera cette notion du début de l'overlap « qstart ». La **figure 17** montre la distribution des reads blastés sur les séquences SL ainsi que ceux spécifiques au trans-splicing.

Enfin, les reads pairés sont séparés en fonction de leur événement de trans-splicing associé (SL1, SL2 ou indéterminé), identifiés grâce au blast. Pour obtenir le nombre d'événements de trans-splicing SL1 ou SL2 par gène, une "summarization" grâce à featureCounts est faite. Pour effectuer cela, les options par défaut de featureCounts sont utilisées à l'exception de :

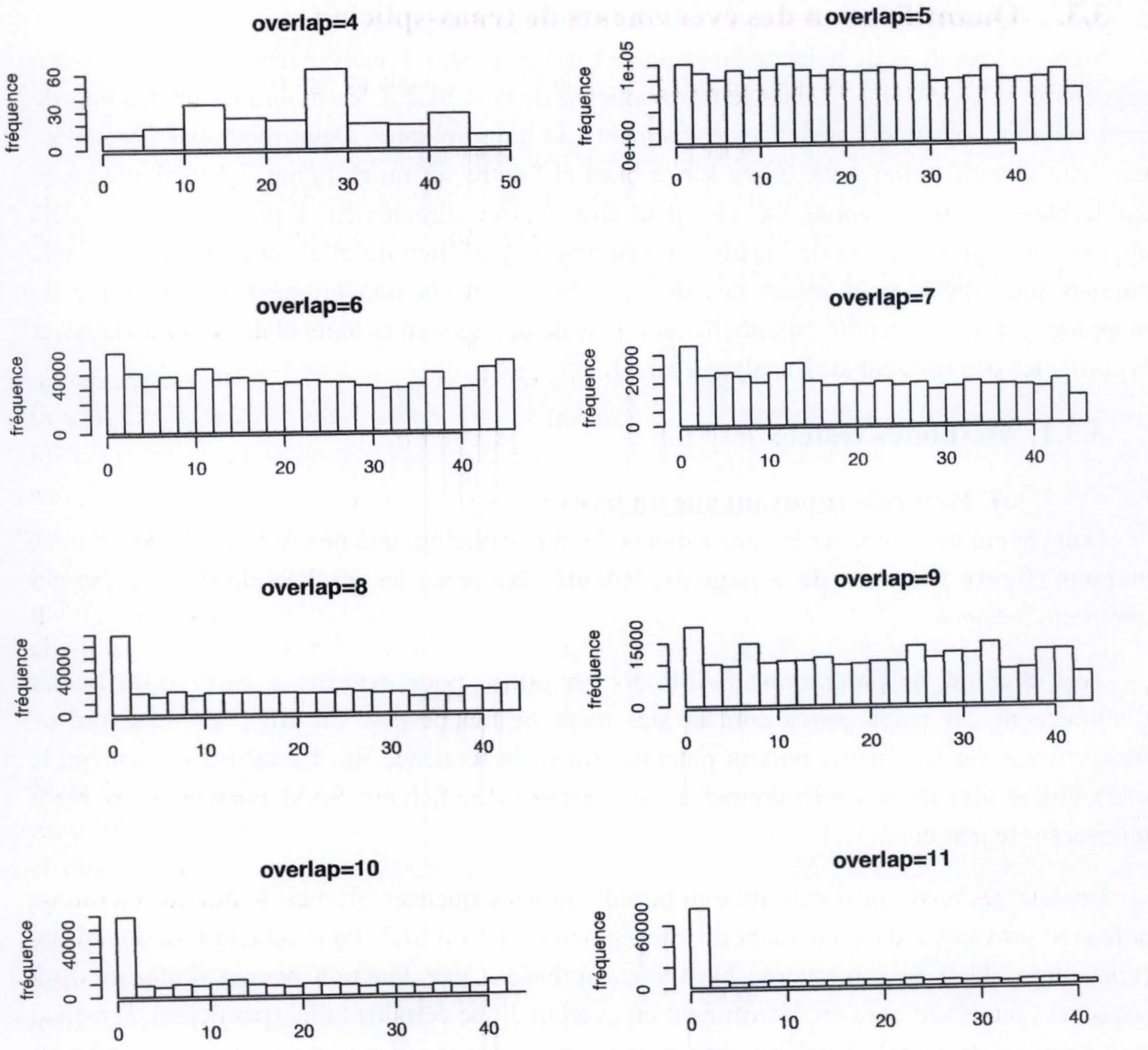


Figure 18 : Distribution du qstart, c'est-à-dire le début de l'overlap des reads blastés sur les séquences SL pour différents overlaps. A partir d'un overlap de 8 nt, la majorité des reads détectés ont un qstart assez faible.

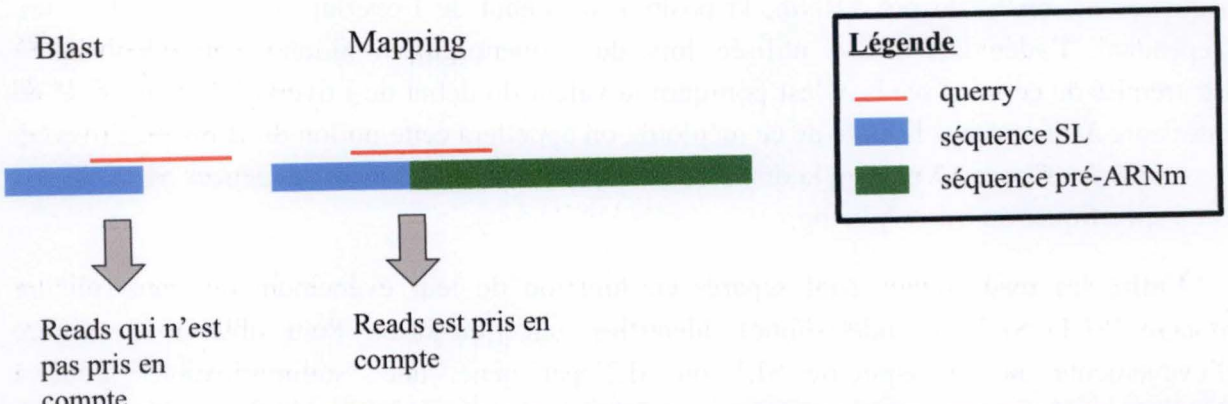


Figure 19 : Avantage du mapping sur un transcriptome trans-splicé par rapport au blast. Les reads qui sont blastés sur la séquence des splice leader (SL) doivent avoir un overlap suffisant avec la séquence SL pour être pris en compte. L'avantage du mapping sur un transcriptome trans-splicé est que l'overlap du read sur la séquence de référence est complet grâce à la séquence du pré-ARNm. De cette manière, la partie de l'overlap avec la séquence SL peut être réduite.

- *strandSpecific=2*: pour prendre en compte la spécificité du brin
- *useMetaFeatures=TRUE*: pour calculer le nombre de reads par gène et non par exon

Le fichier que génère featureCounts est présenté sous forme de table avec 32 colonnes (4 conditions * 2 réplicats * 2 souches * 2 séquences SL possibles) et 46748 lignes (nombre de gènes, y compris les transcrits non codants). Cette table est ensuite chargée dans R et rapidement analysée grâce à des barplots et boxplots.

En vue d'augmenter le nombre de reads pris en compte, la taille de l'overlap entre la query et la séquence de référence est réduite lors du blast. C'est pourquoi un blast avec un overlap minimum de 4 nt est réalisé en utilisant l'option "word size". Cependant, réduire l'overlap risque de supprimer la spécificité de la méthode. C'est dans ce cadre qu'a été réalisée une analyse de la distribution du qstart pour chaque overlap allant de 4 à 11 nt (**figure 18**). Sur cette figure, on peut constater que, pour un overlap de 8 nt, beaucoup de reads ont un qstart peu élevé. Ces résultats sont donc spécifiques. Par contre, un overlap de 4 nt présente des résultats avec des qstart aléatoires, témoignant de l'aspécificité. Ainsi, un overlap minimum de 8 nt est nécessaire pour observer une certaine spécificité. La troisième méthode consiste donc à réaliser un blast avec un overlap minimum de 8 nt

b) Méthode reposant sur un mapping

Le nombre de reads portant la séquence SL détectés par les méthodes utilisant un blast est limité par rapport à l'overlap entre le read et la séquence SL de référence. En effet, réduire cet overlap permet d'augmenter le nombre de reads pris en compte, mais cela diminue la spécificité des reads par rapport au phénomène de trans-splicing. C'est pourquoi, une autre méthode qui consiste à mapper les reads pairés sur un transcriptome trans-splicé est testée. Grâce à la technique du mapping sur un transcriptome déjà trans-splicé, les reads peuvent mapper partiellement sur la séquence SL et partiellement sur la séquence du pré-ARNm (**figure 19**). Ainsi, l'avantage de cette technique est qu'un overlap plus petit sur la séquence SL peut être pris en compte tout en conservant une spécificité élevée puisque le read entier est mappé. Cette technique a d'ailleurs été utilisée par Blumenthal *et al.* dans "a global analysis of *C. elegans* trans-splicing" (*19*).

Le concept pour générer un tel transcriptome est relativement simple. Les séquences des transcrits possédant un site de trans-splicing (TS) sont isolées (**figure 20, verso de la page suivante**). Ensuite, ces séquences sont coupées au niveau du site de trans-splicing. Enfin, les différentes séquences des splice leader (SL) sont ajoutées indépendamment de telle sorte à obtenir autant de combinaisons SL-transcrit que possible.

Cependant cette génération nécessite plusieurs étapes intermédiaires (**figure 21, verso de la page suivante**). Quatre fichiers sont téléchargés à partir de Wormbase : un fichier à propos de la position des sites TS par rapport au génome, un fichier d'annotation du transcriptome, un fichier contenant la séquence en acide nucléique du génome (fichier fasta) et un contenant la séquence fasta des SL. Tout d'abord, la position des sites TS ainsi que l'annotation du transcriptome sont utilisées pour définir l'ensemble des transcrits possédant un site TS.

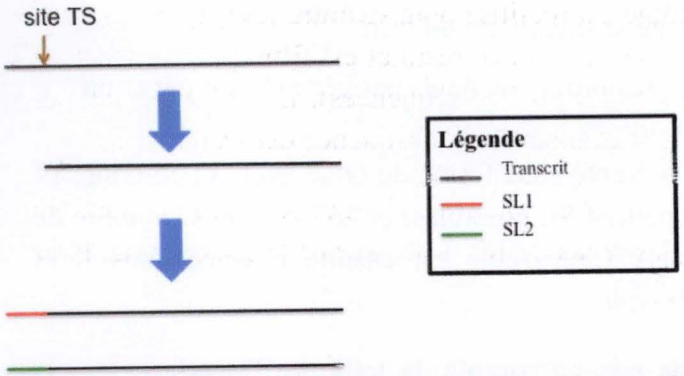


Figure 20 : Principe pour la génération d'un transcriptome virtuel déjà trans-splicé. Tout d'abord, la séquence du transcrit est coupée au niveau du site de trans-splicing (site TS). Ensuite, les différentes séquences de splice leader (SL) seront ajoutées en 5' du transcrit.

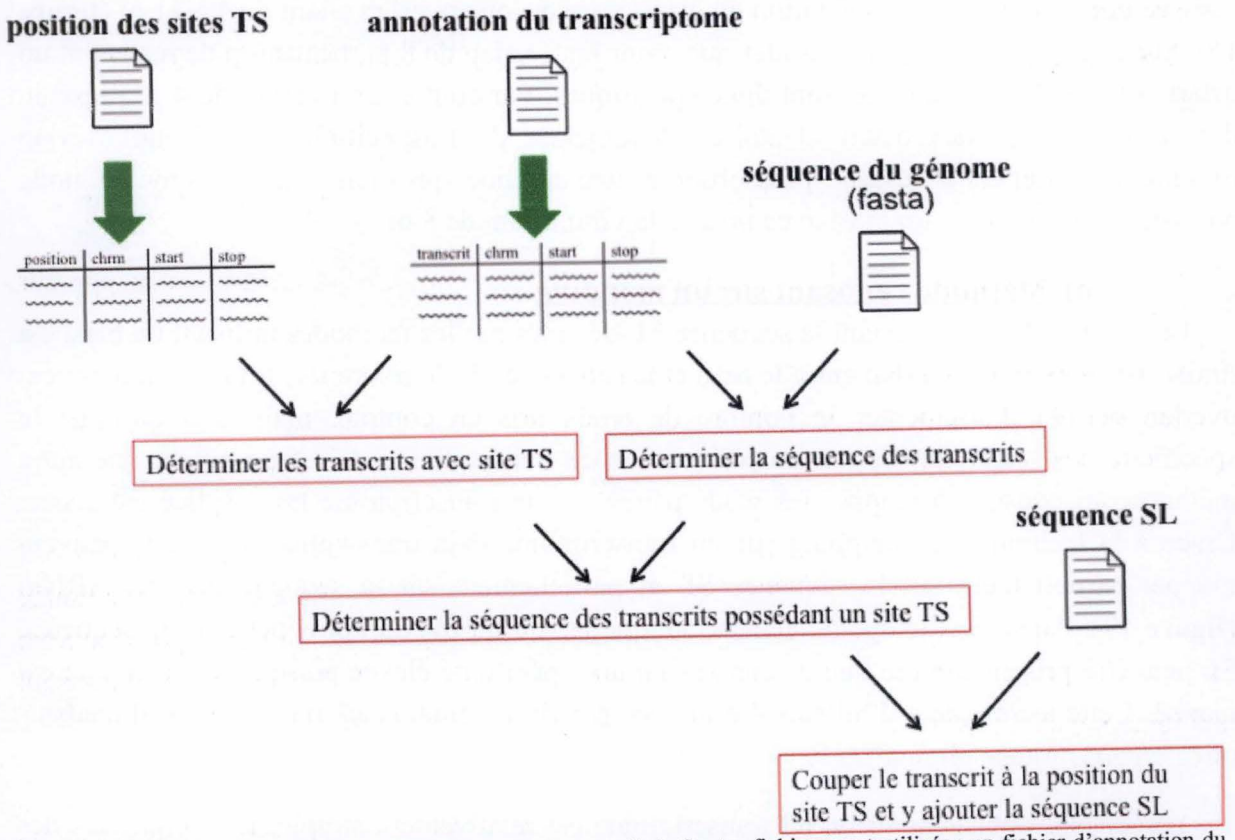


Figure 21 : Etapes pour générer un transcriptome trans-splicé. Quatre fichiers sont utilisés : un fichier d'annotation du génome avec les positions des sites de trans-splicing, une fichier d'annotation du transcriptome, un qui correspond à la séquence du génome et enfin, un avec la séquence des splice leader. A partir des deux premiers, on peut déterminer les transcrits qui possèdent un site de trans-splicing. En utilisant l'annotation du transcriptome, on peut extraire la séquence de tous les transcrits. Ensuite, les séquences des transcrits sont filtrées pour ne considérer que les séquences des transcrits qui ont un site de trans-splicing (70 %). Enfin, ces séquences sont virtuellement coupées avant d'y ajouter les séquences SL en amont.

Ensuite, le fichier d'annotation du transcriptome est réutilisé pour extraire les séquences fasta du génome correspondant aux transcrits. Troisièmement, ce résultat est filtré pour obtenir les séquences fasta des transcrits possédant un site TS (14200 séquences). Enfin, on va couper ces séquences en amont du site TS et concaténer chaque SL à la séquence des transcrits.

Suite à cette étape de génération du transcriptome (en séquence génomique), le programme Tophat2 est utilisé pour mapper les reads du RNA-seq sur ce transcriptome. Pour cela, les commandes ci-dessous sont utilisées :

- *p2:* permet de faire 2 calculs en parallèle
- *g1:* permet de spécifier que le read ne mappe qu'une fois sur le génome
- *--no-coverage-search*: permet de désactiver l'algorithme de recherche de jonction par coverage.
- *--library-type:* permet de spécifier qu'il s'agit d'un séquençage "paired end" et de prendre en compte l'orientation des reads forward-reverse.

Ensuite, grâce à featureCounts, la table des counts est générée, en utilisant les options :

- *countMultiMappingReads=T* : permet de tenir compte du fait que le read mappe sur plusieurs transcrits
- *strandSpecific=2* : pour prendre en compte la spécificité du brin
- *useMetaFeatures=F* : pour calculer le nombre de reads par transcrit
- *isPairedEnd=F* : permet de compter les fragments plutôt que les reads
- *nthreads=2* : permet de faire 2 calculs en parallèle

Ainsi, une table de 170400 lignes et 16 colonnes est obtenue où chaque colonne correspond à une condition et chaque ligne correspond à un des 14 200 transcrits possédant un site trans-splicing, trans-splicé avec une des douze isoformes de SL. Enfin, les résultats sont rapidement analysés sur R.

3.3.2. Détermination de la méthode la plus appropriée

Afin de déterminer la méthode la plus appropriée à la quantification des événements de trans-splicing, une comparaison des résultats obtenus avec les différentes techniques est réalisée. Ainsi, la puissance (capacité de détecter de vrais positifs) et la précision (proportion de vrais positifs parmi l'ensemble des positifs) sont évaluées pour chaque méthode. En effet, il ne suffit pas que la méthode détecte un nombre maximal d'événements de trans-splicing pour qu'elle soit la meilleure. Il faut également que, parmi ces événements détectés, une grande proportion soit vraie.

a) Comparaison des méthodes

Dans le but d'évaluer la puissance des différentes techniques, le nombre d'événements de trans-splicing SL1 et SL2 total et par condition est calculé. En effet, plus la méthode détecte d'événements de trans-splicing (pour autant qu'ils soient vrais), plus des différences entre conditions sont facilement mises en évidence.

La précision des méthodes est évaluée de deux manières : par le calcul de différents rapports d'événements de trans-splicing et par vérification si les gènes détectés pour être trans-splicés SL2 sont bien en position 2 ou plus des opérons.

- En ce qui concerne le rapport des événements de trans-splicing, différents rapports sont calculés. Le premier est le rapport $\frac{\text{nombre } d'\text{événements de trans-splicing SL1}}{\text{nombre } d'\text{événements de trans-splicing SL1+SL2}}$ pour les gènes en position 1 de l'opéron plus les gènes isolés. En effet, le trans-splicing SL2 est naturellement presque inexistant pour ces gènes et, par conséquent, le rapport d'une méthode précise devrait être relativement proche de 1.
  Le deuxième rapport qui devrait apporter une information similaire est le rapport $\frac{\text{nombre } d'\text{événements de trans-splicing SL1}}{\text{nombre } d'\text{événements de trans-splicing SL1+SL2}}$ pour les gènes en position 1 de l'opéron.
  Le dernier rapport évalué est le rapport $\frac{\text{nombre d'événements de trans-splicing SL2}}{\text{nombre } d'\text{événements de trans-splicing SL1+SL2}}$ pour les gènes en position 2 ou plus de l'opéron. A la différence des deux rapports précédents, celui-ci devrait être plus éloigné de 1 puisqu'en position 2 ou plus des opérons, les gènes sont majoritairement trans-splicés SL2, mais il peut y avoir du trans-splicing SL1 comme c'est le cas pour les opérons hybrides. Ce dernier ratio sera donc plus difficile à interpréter dans le contexte de la comparaison de méthodes, mais est présenté à titre indicatif.

- En ce qui concerne la cohérence entre les gènes détectés pour être trans-splicés SL2 et le fait qu'ils soient en position 2 ou plus des opérons, un graphique en nuage de points est réalisé où les gènes y sont représentés en fonction du nombre d'événements de trans-splicing SL1 et SL2. Nous allons également déterminer l'appartenance de chaque gène à un opéron et sa position grâce à la table créée précédemment (voir point 2.2.6).

b) Comparaison selon les courbes de ROC

Un moyen régulièrement utilisé pour les performances de différentes méthodes, et qui regroupe la puissance et la précision, est l'utilisation de courbes de ROC. Grâce à cette comparaison, cela va permettre de valider les conclusions précédentes à propos de la méthode à retenir pour l'analyse en aval.

Une courbe de "receiver operating characteristic" (ROC) est une courbe qui met en relation le TPR (true positive –rate, la proportion de vrais positifs par rapport au nombre total de positifs détectés) au FPR (false positive rate, la proportion de faux négatifs) d'un prédicteur binaire. En effet, toute méthode détecte un certain nombre de résultats positifs parmi lesquels certains sont pertinents, les vrais positifs, et d'autres sont des erreurs techniques, les faux positifs. Plus une méthode détecte un nombre important de vrais positifs pour un minimum de faux positifs, plus sa qualité est élevée.

Afin de pouvoir calculer le TPR et FPR, il est nécessaire de connaitre une « vérité » binaire (vrai ou faux) qui devra être prédite par un prédicteur. Dans notre cas, la vérité est l'appartenance d'un gène au groupe des gènes en aval dans les opérons (déterminée par l'annotation du *C.elegans*), tandis que les prédicteurs seront le nombre d'événements de transplicing SL2 détectés avec les différentes méthodes. En effet, les gènes situés en position deux ou plus sont majoritairement trans-splicés SL2. Ainsi, si le gène subit un nombre important d'événements de trans-splicing SL2 et que ce gène est situé en position deux ou plus d'un opéron, il est considéré comme un vrai positif. Si par contre, le gène subit des événements trans-splicés SL2, mais qu'il est situé en position 1 d'un opéron ou qu'il n'appartient pas à un opéron, il est considéré comme un faux positif.

Pour dessiner une courbe de ROC, une valeur seuil est tout d'abord délimitée afin de déterminer le TPR et le FPR. Ce seuil permet de préciser le nombre d'événements de trans-splicing SL2 souhaité pour qu'un gène soit pris en compte. Ensuite, ce processus est répété pour différents seuils. . Par conséquent, à un seuil très élevé, aucun vrai positif et aucun faux positif ne sont détectés. La courbe commence donc au point de coordonnées (0, 0). A l'inverse, pour un seuil nul, un maximum de vrais positifs et de faux positifs sont détectés.

Pour comparer différentes méthodes, leurs courbes sont représentées sur un même graphique et la valeur du point supérieur gauche de la courbe des différentes courbes est comparée, c'est-à-dire le point qui correspond à un maximum de vrais positifs détectés pour un minimum de faux positifs.

Une fois la meilleure méthode définie, une analyse du nombre d'événements de trans-splicing est réalisée

3.3.3. Analyse de la quantification des événements de trans-splicing

a) Principe

Suite à la comparaison des différentes méthodes (voir résultats, point 3.2.1), celle utilisant un blast caractérisé par un overlap minimum de 8 nt a été retenue pour la quantification des événements de trans-splicing. Grâce à cette méthode, trois analyses sur des subsets de gènes différents sont réalisées : la première analyse se fait sur l'ensemble des gènes en position 2 ou plus des opérons, la seconde se concentre sur des gènes en position 2 ou plus des opérons qui sont identifiés pour être trans-splicés SL1 et SL2 et la troisième concerne les gènes en position 2 ou plus des opérons et qui sont uniquement trans-splicés SL2. La raison d'analyser ces différents subsets est expliquée dans les résultats.

b) Adaptations

Dans le but de réaliser une quantification la plus pertinente possible, différentes adaptations ont été effectuées.

- La première adaptation consiste à regrouper les deux réplicats ensemble. Pour cela, la table obtenue grâce à la "summarization", table de 32 colonnes, est divisée en quatre tables intermédiaires, une par réplicat selon le type d'événements de trans-splicing (SL1 ou SL2). Ensuite, ces tables sont regroupées deux à deux en fonction du SL en faisant la somme des événements de trans-splicing des tables. Ainsi, deux tables de huit colonnes

sont obtenues. Enfin, ces dernières sont juxtaposées pour ne former qu'une seule table de seize colonnes grâce à la fonction cbind() disponible sur R.

- Une deuxième adaptation est de se concentrer sur les rapports $\frac{\text{nombre d'événements de trans-splicing SL2}}{\textit{nombre d'événements de trans-splicing SL1}}$ plutôt que sur le nombre d'événements de trans-splicing. Comme expliqué dans la partie résultats, cela permet une certaine normalisation par rapport aux conditions. Pour calculer ces rapports, un "pseudocount" de 1 est ajouté au numérateur et au dénominateur. Ainsi, les rapports sont calculés à partir de la formule $\frac{\text{(nombre d'événements de trans-splicing SL2+1)}}{\textit{(nombre d'événements de trans-splicing SL1+1)}}$

- La troisième adaptation est le fait de représenter ces rapports en graphiques grâce à la moyenne des logarithmes des rapports plutôt que la moyenne des rapports. En effet, le logarithme des rapports permet de rendre la distribution des rapports symétriques. Par exemple, des rapports de 0,001 et de 1000 sont considérés. La moyenne de ceux-ci est de 500,0005. Cela met en évidence que la distribution des rapports n'est pas symétrique. Par contre, la moyenne de leur logarithme revient à faire la moyenne de -3 et 3, ce qui vaut 0. Ainsi, les deux valeurs sont symétriques. Pour transformer ces rapports, la formule $\log_2$ (rapport) est appliquée.

- Ne sont pris en compte pour dessiner ce diagramme que les gènes dont la somme des reads dans les différentes conditions vaut 100 ou plus. Cette filtration des gènes pris en compte permet de ne considérer que des gènes exprimés dans nos conditions.

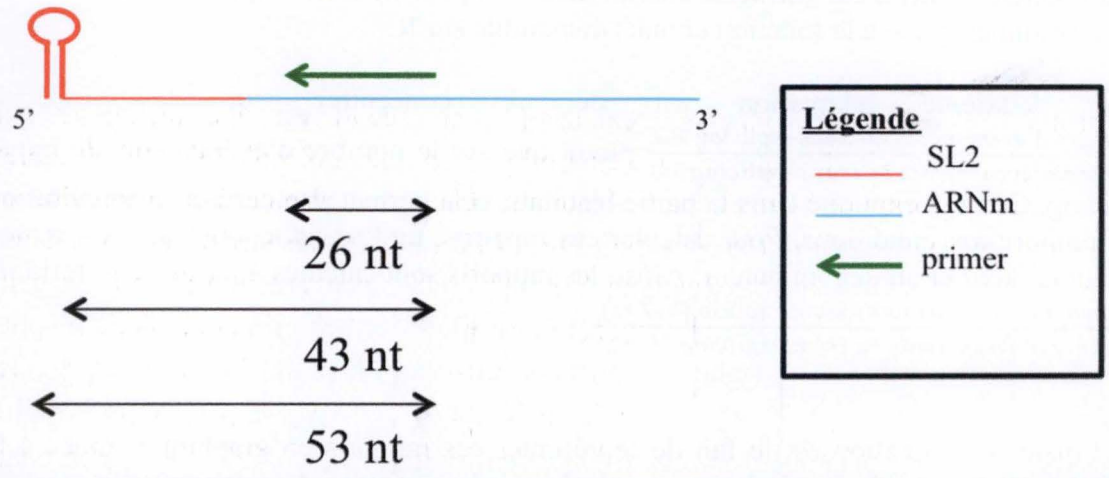


Figure 22 : Taille des bandes attendues suite au primer extension. Ainsi, le primer isolé a une longueur de 26 nucléotides (nt). Lorsque celui-ci interagit avec l'ARNm et qu'il y a un arrêt de la rétro-transcription (RT) au niveau de la structure secondaire, la longueur de l'oligonucléotide généré est de 43 nt. Par contre, lorsque la RT est complète, l'oligonucléotide a une taille de 53 nt.

3.4. Primer extension

3.4.1. Introduction à la technique du primer extension

Le primer extension (en français « extension d'amorce ») est une technique qui, historiquement, permettait de déterminer l'extrémité 5' d'une molécule d'ARN d'intérêt. Cependant, le même principe a été utilisé à d'autres fins, notamment pour mettre en évidence une éventuelle interruption de la rétro-transcription (RT). En effet, les structures secondaires ainsi que des modifications post-traductionnelles sur l'ADN peuvent interférer avec cette réaction (*46*) (*47*). Le principe du primer extension est d'utiliser des primers radioactifs pour la rétro-transcription (RT) et d'établir le profil électrophorétique après avoir chargé les produits de réaction sur un gel d'acrylamide. Dans le but de générer des primers radioactifs, de l'[γ-^{32}P]ATP est utilisé, c'est-à-dire de l'adénosine triphosphate (ATP) dont le phosphate le plus éloigné du ribose, appelé le phosphate γ, porte un isotope radioactif du phosphore. Ensuite, par l'intermédiaire d'une polynucléotide kinase (PNK), le groupement phosphate (radioactif) de l'ATP est transféré sur l'extrémité 5' du primer. La seconde étape est la réalisation de la RT. Au terme de celle-ci, des fragments de tailles différentes sont obtenus en fonction du fait que la RT soit complète ou non. Si la reverse transcriptase (RTase) ne rencontre aucun obstacle sur l'ARN, la réaction aura généré un polynucléotide ayant atteint l'extrémité 5' de l'ARN cible et, donc ayant une taille qui peut être préalablement déterminée si la séquence est connue. À l'inverse, si la RTase fait face à une structure secondaire stable ou à une modification post-transcriptionnelle, la réaction peut bloquer et aura dès lors généré des polynucléotides plus courts. L'électrophorèse sur gel d'acrylamide permet d'observer ces possibles arrêts et/ou blocages puisque plus la molécule est petite et plus celle-ci pourra facilement se faufiler à travers les mailles du filet de polyacrylamide. Donc, les petits fragments, résultant d'un arrêt de la RT, migreront davantage que ceux plus longs qui correspondent à une RT ininterrompue.

3.4.2. Gène d'intérêt et primer utilisé

Un ARN synthétique dont la séquence correspond à celle du gène *gpd-3* est utilisé pour réaliser le primer extension. En effet, le gène *gpd-3* est connu pour être en position 3 d'un opéron, est trans-splicé SL2 et a déjà été utilisé dans un autre contexte pour faire des expériences de primer extension chez *C. elegans* (*21*) (*26*). Ainsi, la séquence utilisée est 5'-GGTTTTAACCCAGTTACTCAAGggaaactgcttcaacgcatcATGACCAAGCCAAG -3' (les lettres en majuscules à l'extrémité 5' correspondent à la séquence de SL2, les lettres en minuscules qui suivent correspondent à la séquence comprise entre le site de splicing et la séquence codante de *gdp-3*, les lettres en majuscules en 3' correspondent à une partie de la séquence codante de *gdp-3*).

De plus, le primer utilisé est 5'- GGCTTGGTCATgatgcgttgaagcag -3', caractérisé par une température de melting (Tm) de 66°C et d'un pourcentage en GC, 50% (valeur déterminée grâce au logiciel Ape). Les lettres en majuscules correspondent à la séquence de SL2 et les lettres en minuscules correspondent à une partie de la séquence codante de *gdp-3*. L'intérêt d'avoir choisi ce primer est qu'il permet de distinguer les différentes tailles des bandes observées après migration (**figure 22**).

3.4.3. Marquage radioactif des primers

Un solution est préparée contenant 1 μl de primer concentré à 10 μM, 1 μl de T4 PNK (10U/μl) (Promega M40101), 1 μl de buffer T4 PNK 10X, 3 μl [γ-32P]ATP (3,000 Ci/mmol, 10mCi/ml) (Perkin NEG002A250UC) et 4 μl d'eau traitée au diéthylpyrocarbonate (DEPC). Ensuite, cette solution est incubée 10 minutes à 37°C, chauffée pendant 2 minutes à 90°C de façon à inactiver l'enzyme et brièvement centrifugée. 15 μl d'eau RNase-free sont ajoutés à la solution pour obtenir une concentration optimale du primer marqué et le résultat est stocké à -20°C.

3.4.4. Marquage radioactif de l'étalon

L'étalon utilisé s'étale de 15 nt à 200 nt. Tout d'abord, un mix de 12 oligonucléotides est préparé à partir de 12 solutions individuelles (chacune contenant un oligonucléotide concentré à 5 μM) de façon à ce que la concentration finale de chaque oligonucléotide soit de 0,42 μM. A partir de ce mix, 5 μl sont extraits et ajoutés dans un microtube avec 1 μl de T4 PNK (10U/μl) (Promega M40101), 1 μl de buffer T4 PNK 10X et 3 μl [γ-32P]ATP (3,000 Ci/mmol, 10mCi/ml) (Perkin NEG002A250UC). Ensuite, la solution est incubée 10 minutes à 37°C avant d'être chauffée 2 minutes à 90°C et centrifugée. Ensuite, un mix est préparé avec 1 μl du résultat, 9 μl de buffer TE et 10 de μl loading buffer (formamide 98%, EDTA 10mM, 0,1% bromophénol et 0,1% de cyanol xylene). Enfin, la solution finale est stockée à -20°C.

3.4.5. Réaction de primer extension

Des microtubes contenant différentes quantités d'ARN synthétique sont utilisés (0,3 ; 3 ; 30 et 300 ng). Les tubes sont incubés à 90°C pendant 5 minutes. Ensuite, ceux-ci sont mis sur glace pendant 1 minute avant d'être "poolés" à l'aide d'une microcentrifugeuse. En parallèle, un mix de primer extension est préparé dans un microtube de 1,5 ml. Celui-ci contient 1 μl de primer radioactif, 1,5 μl de désoxynucléotides triphosphate (dNTP) dont la concentration finale est de 150 μM, de 2 μl de buffer 5x (Invitrogen 18080-093) et 0,5 μl de rétrotranscriptase (SuperScript III) (Invitrogen 18080-093) (=100 units). 5 μl de ce mix sont ensuite ajoutés à chaque échantillon d'ARN. Le résultat est alors incubé 1h à 50°C. Par la suite, 10 μl de loading buffer sont ajoutés dans chaque tube. Enfin, les tubes sont chauffés à 90°C pendant 10 minutes et les résultats sont mis sur glace pendant une minute.

3.4.6. Migration électrophorétique

La migration se fait sur un gel d'acrylamide 15% en présence d'urée 7M et de TBE 1x. Un champ électrique de 225V est appliqué au gel pendant 10 à 20 minutes et ensuite, les échantillons sont ajoutés et migrent pendant 3 à 4h. Enfin, les résultats sont révélés par autoradiographie.

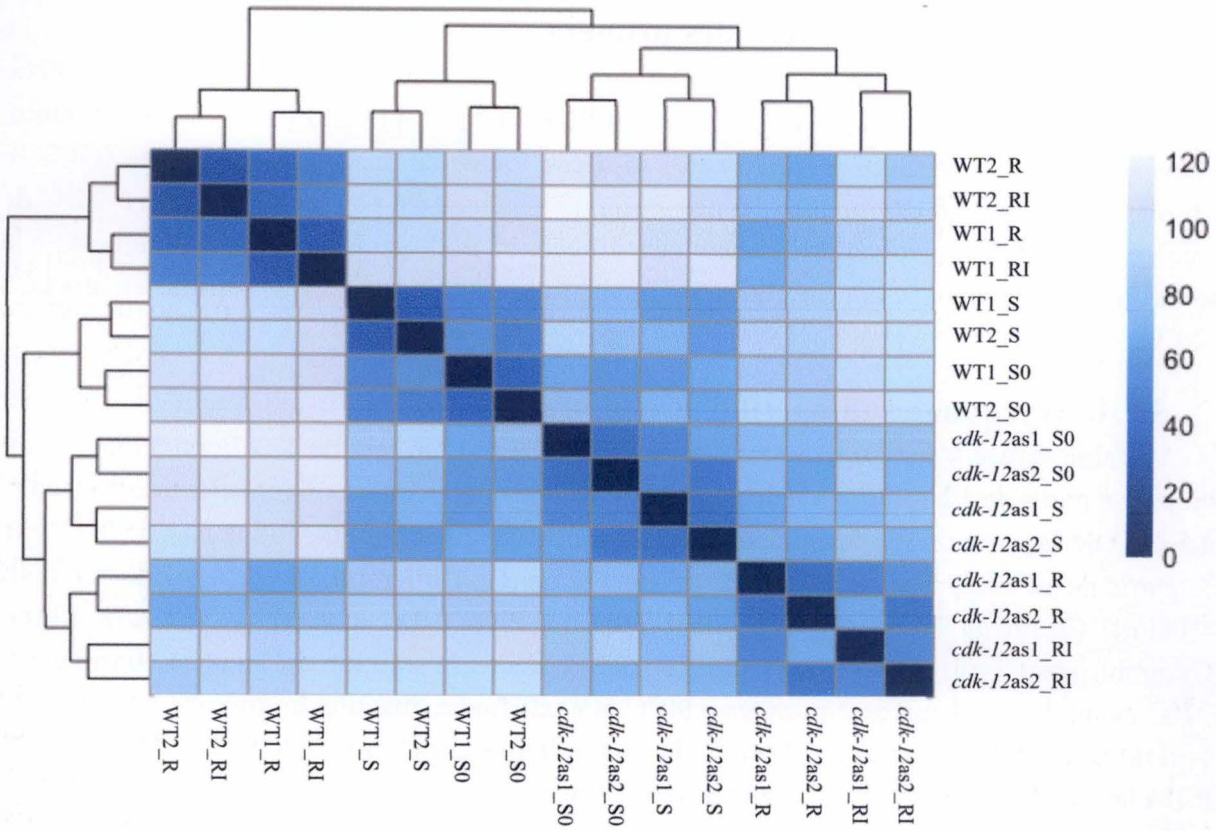


Figure 23 : Heatmap des distances des différentes conditions testées. Cette figure représente une évaluation de la différence entre les réplicats (1 – 2) et/ou les conditions (S0, S, R et RI). Le dégradé de bleu correspond aux distances euclidiennes calculées entre les échantillons.

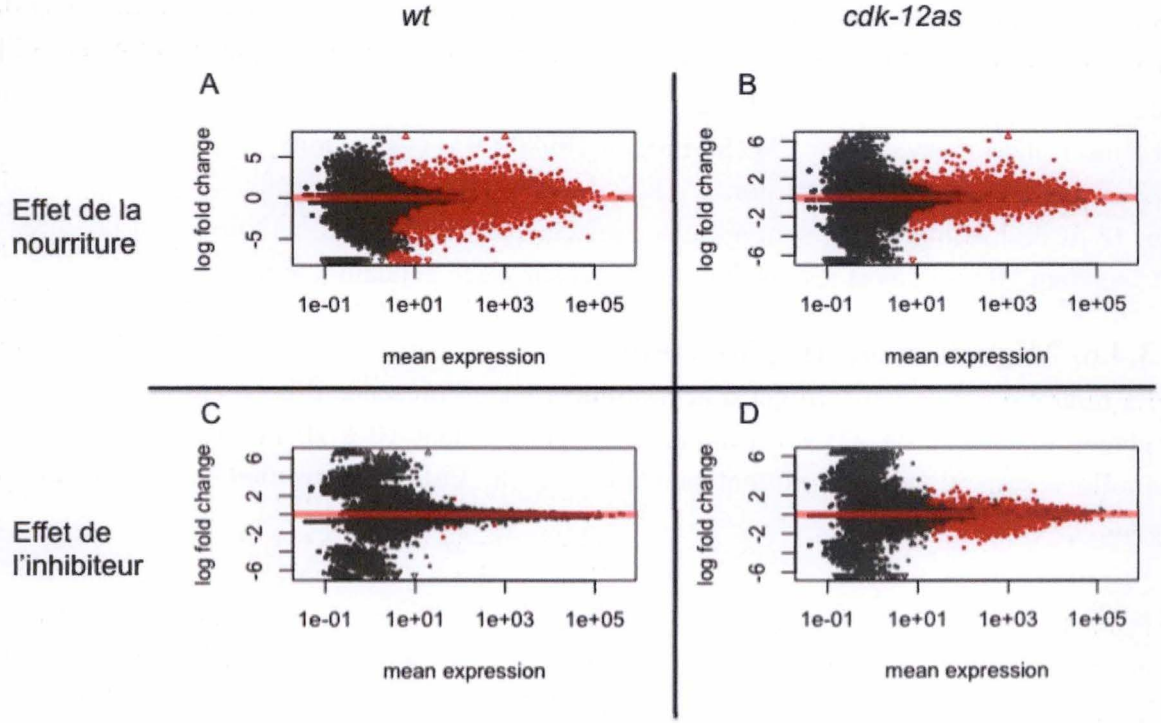


Figure 24 : MAplots illustrant le "fold change" des gènes en fonction de leur expression moyenne. Chaque gène est représenté par un point. Les points noirs signifient que le gène n'est pas exprimé différentiellement alors que les points rouges correspondent à des gènes significativement différentiellement exprimés (padj<0,1).

4. Résultats et discussion

4.1. Analyse du transcriptome

4.1.1. Relation entre conditions

Avant de commencer l'analyse du transcriptome, il est important de s'assurer de la qualité de nos données. Pour cela, les distances euclidiennes entre les réplicats et entre les conditions sont calculées et ensuite représentées grâce à une heatmap (**figure 23**). On peut constater sur cette figure que les réplicats, les souches ainsi que les conditions sont regroupés entre eux, ce qui est logique par rapport à ce que nous nous attendions. En effet, trois groupes principaux sont distinguables. Le premier regroupe les réplicats du WT en recovery (R) et recovery + inhibiteur (RI), ce qui semble normal, car l'inhibiteur chez le WT n'est pas sensé avoir d'effet sur CDK-12 et, par conséquent, l'expression génique devrait être similaire entre les deux conditions. Le deuxième groupe est composé des deux réplicats des deux souches en condition contrôle (S0) et les conditions de starvation (S). De nouveau, cela paraît logique, car les deux conditions sont caractérisées par des vers phénotypiquement similaires. Enfin, le dernier groupe est constitué des deux réplicats de la souche *cdk-12as* dans les conditions de recovery (R) et recovery + inhibiteur (RI). Le fait que la condition *cdk-12as*_RI est séparée du WT_RI est normal puisque dans le premier cas, la kinase CDK-12as est inhibée et dans le second cas, la kinase CDK-12 ne l'est pas. Par contre, il est étonnant que la condition *cdk-12as*_R soit séparée de celle WT_R. En effet, cela signifie que la condition *cdk-12as*_R est plus proche de la condition *cdk-12as*_RI plutôt que celle WT_R. L'hypothèse émise pour expliquer cela est que la mutation a un effet sur l'expression. En effet, la souche cdk-12as se caractérise par une substitution de base au niveau de la séquence génomique. Or, cela peut éventuellement avoir un impact sur le génome comme induire la méthylation d'îlot C^pG ou l'apparition de site de splicing. Ce dernier cas a été récemment rencontré par Bartkowiak (*48*).

4.1.2. Analyse statistique du transcriptome

Afin de comprendre l'impact de CDK-12 sur le transcriptome, des MAplots sont réalisés. Les Maplots sont des graphiques permettant de représenter la significativité d'une différence d'expression génique grâce à une analyse statistique réalisée, dans le cas présent, avec le programme DESeq2. Etant donné que le phénotype de l'inhibition de CDK-12, un arrêt en L1 est similaire à un arrêt en L1 en starvation, l'impact de la présence de nourriture ainsi que celui de l'inhibiteur sont évalués chez le wild-type (WT) et le *cdk-12as*.

La présence de nourriture influence fortement le transcriptome. Cette tendance s'observe à la fois chez le WT (**figure 24A**) et chez le *cdk-12as* (**figure 24B**). Par contre, quand l'effet de l'inhibiteur est analysé, de fortes différences entre les deux souches sont remarquées. En effet, la plupart des gènes ne présentent pas de différence significative chez le WT (**figure 24C**), confirmant que l'inhibiteur est spécifique à la kinase CDK-12as. Par contre, chez le *cdk-12as*, plusieurs gènes sont différentiellement exprimés (**figure 24D**), ce qui renforce le fait que l'inhibiteur est spécifique à la kinase CDK-12as. De plus, la plupart des gènes différentiellement exprimés sont situés sous l'axe des abscisses, ce qui signifie que l'inhibition de CDK-12 tend vers une downrégulation.

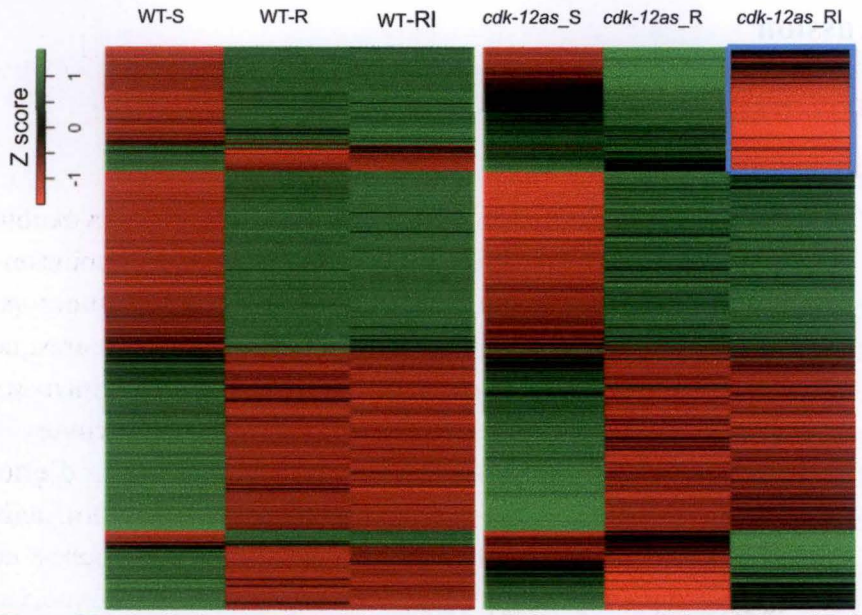


Figure 25 : Profil d'expression génique dans les différentes conditions S, R et RI. Chaque ligne représente un gène alors qu'une colonne est une condition. Pour déterminer le Z score, les données rlog de la table de summarization sont centrées et réduites. Le Z score a pour intérêt de permettre une comparaison aisée entre plusieurs conditions en tenant compte de la moyenne entre les différentes conditions et la déviation standard. Ainsi, une représentation de l'expression peut être faite. Pour une condition donnée, un Z score négatif signifie que le gène est downrégulé et la ligne correspondante est colorée en rouge. Par contre, s'il est positif, le gène est uprégulé et la ligne correspondante est colorée en vert. Ainsi, en se concentrant sur une seule ligne, on peut suivre l'évolution de l'expression d'un gène dans les différentes conditions.

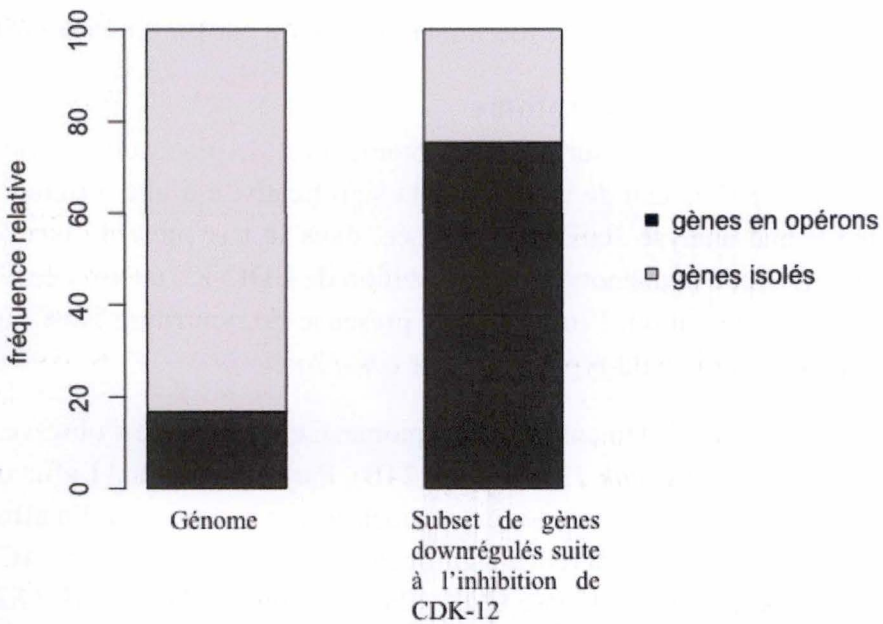


Figure 26 : Proportions des gènes en opérons dans le génome de *C. elegans* et dans le subset de gènes d'intérêt. Parmi les 20447 gènes de *C. elegans* codant pour des protéines, 3424 sont connus pour être en opéron, soit 17%. Par contre le subset de gènes downrégulés suite à l'inhibition de CDK-12 contient 1031 gènes dont 777 sont en opérons soit 75%.

4.1.3. Subsets de gènes affectés suite à l'inhibition de CDK-12

Pour représenter l'ensemble des différences d'expression génique dans les différentes conditions (S, R et RI pour le WT et le *cdk-12as*), une heatmap est réalisée (**figure 25**).

Quand les profils obtenus en S et R sont comparés, un shift prononcé du niveau d'expression des ARN messagers d'un groupe de gènes est observé. Ces données corroborent celles de Zaslaver (2011) qui a rapporté qu'un set de gènes downrégulés en starvation est rapidement uprégulés en recovery après addition de nourriture (*23*) alors qu'un autre groupe de gènes présente le comportement inverse. Les souches WT et mutante sont presque indistinguables dans ces conditions. Par contre, quand les profils sont comparés entre les conditions R (uniquement présence de nourriture) et RI (présence de nourriture et d'inhibiteur), une différence claire entre souches est observée. Chez le WT, l'ajout d'inhibiteur n'affecte quasiment pas l'expression génique. Cependant, chez le mutant *cdk-12as*, la condition recovery montre qu'un subset de gènes est downrégulé en présence de l'inhibiteur alors que ce même subset est uprégulé en absence d'inhibiteur. Un second subset de gènes, plus petit, est à l'inverse downrégulé en recovery et il est uprégulé suite à l'ajout d'inhibiteur. La **figure 24D** met en évidence que l'inhibition de CDK-12 provoque majoritairement une downrégulation. C'est pourquoi nous nous concentrerons sur l'analyse du premier subset pour réaliser l'analyse en aval. Il faut aussi noter que la majorité des gènes dont l'expression varie en cours de recovery ne sont pas affectés par l'inhibition de CDK-12, ce qui pose la question des bases moléculaires de la spécificité de l'effet.

Remarque : Pour faciliter la compréhension, uniquement les conditions S, R et RI sont présentées. En effet, les résultats de la condition S0 sont difficilement interprétables, car la condition varie des autres conditions par deux paramètres : l'absence de DMSO et le temps avant l'extraction d'ARN (0h VS 4h). Néanmoins, les résultats de cette condition sont disponibles en **annexe 1**.

4.1.4. Analyse du subset de gènes downrégulés lors de l'inhibition de CDK-12

La disposition en opérons des gènes qui sont downrégulés suite à l'inhibition de CDK-12 ainsi que leur position au sein de l'opéron sont évaluées. En effet, comme dit précédemment, le génome de *C. elegans* comporte plusieurs opérons. En outre, il a été vérifié sur base de leur annotation par C. Cassart que, parmi le subset de gènes downrégulés, la majorité de ceux-ci sont localisés dans un opéron (**figure 26**). De plus, l'effet de l'inhibiteur sur l'expression des gènes dépend de leur position dans l'opéron : une large majorité des gènes affectés sont localisés en position deux et plus alors que les gènes en position un sont épargnés (**figure 27, verso de la page suivante**).

Dans un second temps, un "gene ontology enrichment analysis" a été réalisé sur base de la heatmap, sur les gènes downrégulés suite à l'inhibition de CDK-12, dans le but de mettre en évidence les processus affectés (**figure 28, verso de la page suivante**). On retrouve un enrichissement pour les catégories fonctionnelles liées au développement et à la reproduction. Sur base de ceci, on peut proposer l'hypothèse que la lignée germinale est perturbée, mais elle n'est véritablement établie que plus tard dans le développement.

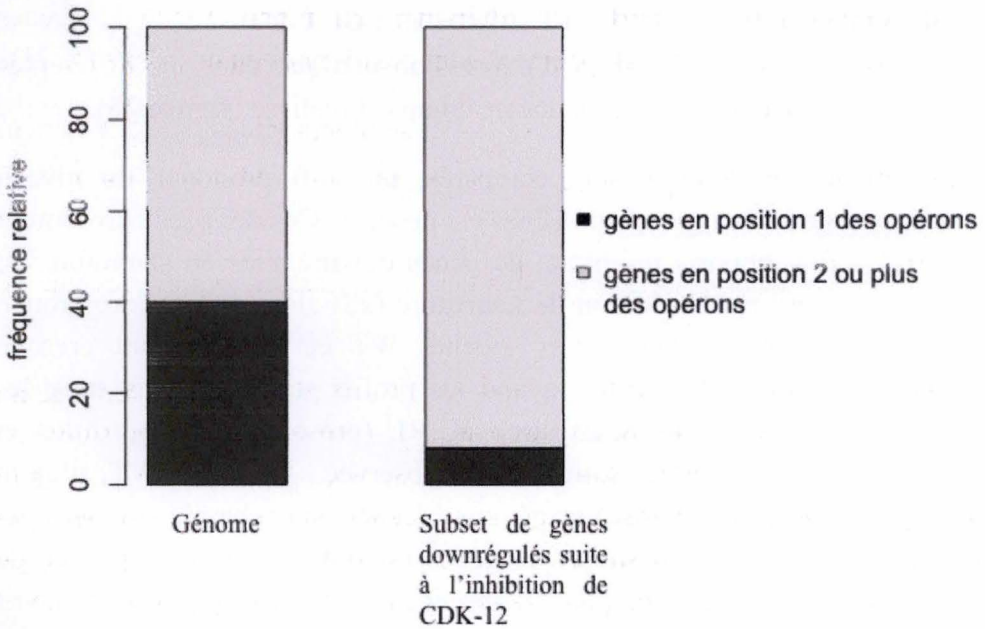


Figure 27 : Proportions des positions des gènes situés dans un opéron. Sur les 3424 gènes appartenant à un opéron, 40 % sont en position 1 de l'opéron et 60% sont en position 2 ou plus. Par contre, sur les 777 gènes du subset qui sont en opérons, moins de 10 % sont en position 1 des opérons et plus de 90 % sont en position 2 ou plus.

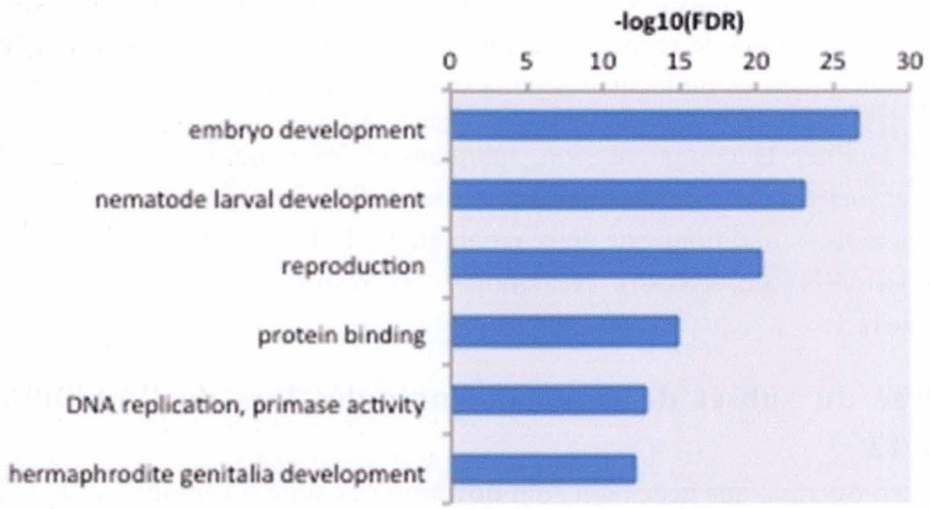


Figure 28 : Gene onthology des gènes dont l'expression est downrégulée suite à l'inhibition de CDK-12 (figure réalisée par Carlo Yague-Sanz). Les différents chiffres représentent l'inverse du logarithme de la p-value ajustée. FDR est l'acronyme de false discovery rate.

Ces résultats sont cohérents par rapport à la littérature qui rapporte que les gènes impliqués dans la croissance au cours du développement sont positionnés en opérons (*15, 23*).

En résumé, les gènes affectés par l'inhibition de CDK-12 semblent situés majoritairement en position deux et plus au sein des opérons et codent principalement pour des gènes liés au développement larvaire et à la reproduction.

Pour expliquer la downrégulation spécifique de ces gènes, différentes hypothèses ont été émises.

- Tout d'abord, en présence de l'inhibiteur, la machinerie de transcription serait bloquée au niveau des gènes en position deux ou plus des opérons, suite par exemple à un problème de polyadénylation en position un. Ainsi, la transcription de ces gènes est diminuée et *in fine*, une downrégulation est observée. Cependant, cette hypothèse n'est pas supportée par les résultats de Chromatin Immunoprecipitation (ChIP) qui ont été réalisés par C. Cassart et qui n'ont mis en évidence aucune différence significative de la distribution de la RNA pol II après l'inhibition de CDK-12as (Données du laboratoire non montrées).

- La seconde hypothèse est que la downrégulation observée résulte d'un défaut de trans-splicing SL2. En effet, le trans-splicing de type SL2 est spécifique aux gènes situés en position 2 et plus et garantit la stabilité de l'ARNm en bloquant l'accès des exonucléases. Pour tester l'hypothèse que CDK-12 affecte directement la transcription des ARN SL2, des RT-qPCR ont été réalisées dans les différentes conditions par C. Cassart (Données du laboratoire non montrées). Les résultats ne montrant pas de différence significative du niveau des précurseurs SL2, il est peu probable qu'un déficit de ces précurseurs soit la cause des problèmes d'expression observés.

- Finalement, puisque l'inhibition spécifique de CDK-12 provoque la downrégulation des gènes trans-splicés SL2, l'hypothèse que CDK-12 interviendrait non pas dans l'expression des ARN précurseurs de SL2 mais bien directement dans le mécanisme de trans-splicing SL2 a été émise. De plus, Blumenthal a récemment (2016) publié une étude où plusieurs ChIP-seq ont été réalisés et où il a mis en évidence que le pic de S2P, à l'échelle du génome, coïncide avec celui de la formation de ce complexe permettant le trans-splicing SL2 (*49*). Cela conforte notre hypothèse.

4.1.5. Effet gène spécifique de CDK-12

L'inhibition de CDK-12 n'affecte que l'expression de certains gènes, ce qui contraste avec sa distribution ubiquiste sur le génome et son rôle global présumé dans la polyadénylation et la terminaison.

Le fait que la S2P ne soit pas nécessaire pour tous les gènes est connu dans la littérature. En effet, Coudreuse a publié que, chez la levure de fission, *Schizosaccharomyces Pombe*, plus ou moins 250 gènes sont différentiellement exprimés en absence de *lsk1*. Concernant *Drosophila melanogaster*, la présence d'une kinase homologue à Ctk1 a été mise en évidence (*33*). Cette kinase, nommée Cdk12, est la principale kinase de la sérine 2 chez la drosophile. Une étude de Cdk12, basée sur une approche ARNi, a démontré que son silençage impacte

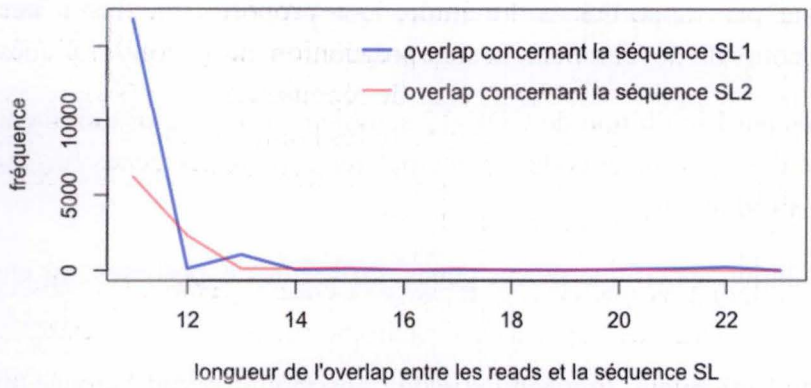


Figure 29 : Distribution de l'overlap des reads hybrides blastés sur les séquences SL. La courbe bleue représente l'overlap des reads blastés sur la séquence SL1 tandis que la courbe rouge correspond à l'overlap des reads blastés sur la séquence SL2.

l'expression d'un subset de gènes. En effet, dans cette étude, Li a proposé que, face à une augmentation du stress cellulaire, Cdk12 permettait la downrégulation de gènes impliqués dans le métabolisme et provoquait une uprégulation des gènes de réponse au stress (*12*). En ce qui concerne *Homo sapiens*, deux kinases sont connues pour être les homologues de CDK-12, à savoir CDK12 et CDK13 (*13*). Liang (2015) a démontré que le silençage de CDK12 et celui de CDK13 n'affectent que l'expression d'un subset de gènes (*50*).

Il est également intéressant de souligner que chez *C. elegans* et chez *S. pombe*, les subsets de gènes concernés regroupent des gènes de réponse à l'environnement. En effet, chez *C. elegans*, ce sont principalement des gènes impliqués dans la sortie de diapause qui sont affectés. Chez *S. pombe*, ce sont des gènes impliqués dans la différentiation sexuelle en condition de starvation. Cependant, le mécanisme moléculaire sous-jacent est probablement différent. Le laboratoire a montré que chez *S. pombe*, la S2P est nécessaire au promoteur pour la transcription du gène *ste11* qui encode le régulateur maître de la différentiation sexuelle. Chez *C. elegans*, comme suggéré plus haut, l'effet gène spécifique est potentiellement lié aux gènes disposés en opérons et à la maturation post-transcriptionnelle de ceux-ci.

4.2. Quantification des événements de trans-splicing

4.2.1. Choix d'une méthode appropriée

Dans l'objectif de tester l'hypothèse selon laquelle un défaut de trans-splicing SL2 est à l'origine de la downregulation d'un set de gènes après inhibition de CDK-12, une quantification des événements de trans-splicing dans les différentes conditions a été envisagée. Dans ce contexte, la question de l'exploitation de nos données transcriptomiques pour en retirer une information sur le trans-splicing s'est posée. C'est dans ce but que différentes méthodes ont été testées. Ces dernières sont décrites dans le matériel et méthodes, point 2.3.1. Elles diffèrent notamment par leur valeur de l'overlap minimum, c'est-à-dire la longueur de la zone de recouvrement entre le read et la séquence SL de référence (**figure 11**). Cette valeur est particulièrement importante pour les résultats de l'analyse en aval. En effet, plus un overlap est élevé, plus le nombre d'événements de trans-splicing détectés est faible (**figure 29**). Ainsi, diminuer l'overlap minimum permettrait théoriquement d'augmenter le nombre d'événements de trans-splicing détectés. Néanmoins, en réduisant l'overlap minimum, il y a un risque de perdre la spécificité par rapport au trans-splicing. Cela a déjà été discuté lors de la mise au point des différentes méthodes (voir matériel et méthodes point 2.3.1.). De plus, plus la taille de l'overlap est réduite, plus la confiance attribuée aux résultats est faible. Il est donc intéressant de déterminer une méthode qui serait un compromis entre une puissance relativement élevée, une spécificité adéquate et une confiance dans les résultats importante. C'est dans ce but que les différentes méthodes sont comparées. Cette comparaison est réalisée à l'aide de deux critères : la puissance des méthodes (capacité à détecter de vrais positifs) et leur précision (rapport de vrais positifs parmi les positifs). Dans le but d'évaluer la puissance de la méthode, le nombre d'événements de trans-splicing détectés est déterminé. En ce qui concerne la précision, celle-ci est estimée sur base de la spécificité du trans-splicing SL2 sur les gènes en aval des opérons.

Tableau 2 : Comparaison entre les quatre méthodes. nt signifie nucléotide

Base de la technique	blast						mapping	
Overlap minimum sur la séquence SL	11 nt		8 nt		4 nt		3 nt	
N^{bre} d'événements de TS détectés	**SL1**	**SL2**	**SL1**	**SL2**	**SL1**	**SL2**	**SL1**	**SL2**
	492467	122179	1270201	215524	1489571	601049	311171	173445
Ratio $\frac{SL1}{(SL1+SL2)}$ par condition, des gènes situés en position 1 des opérons et les gènes isolés	95,85 %		89,37 %		74,95 %		67,32 %	

Concernant la méthode du blast, le nombre d'événements de trans-splicing augmente quand l'overlap diminue, ce qui est en accord avec la théorie expliquée ci-dessus. Par contre, le nombre d'événements de trans-splicing détectés suite au mapping sur le transcriptome trans-splicé est moins important que les trois autres, ce qui est en désaccord avec la théorie. L'hypothèse d'une erreur de code lors de la génération du transcriptome trans-splicé a été émise. Ce problème n'a pas encore été résolu à ce jour, mais on espère résoudre cela prochainement. En conclusion, le blast utilisant un overlap minimum de quatre nucléotides est la méthode qui détecte le plus d'événements de trans-splicing. Il est également intéressant de souligner que le blast avec un overlap minimum de 8 nucléotides est, certes, moins puissant mais un nombre non négligeable d'événements de trans-splicing est détecté. Cependant, la puissance n'est pas la seule mesure à prendre en compte lors d'une comparaison de méthodes, il y a aussi la précision.

En vue d'évaluer la précision de la méthode, le fait que le trans-splicing SL1 soit majoritaire au niveau des gènes isolés ainsi que pour le premier gène de l'opéron est utilisé. Ainsi, en ce qui concerne ces gènes, le rapport entre le nombre d'événements de trans-splicing SL1 et le nombre de trans-splicing SL1 + SL2, c'est-à-dire le ratio $\frac{SL1}{(SL1+SL2)}$, devrait est proche de 1. Autrement dit, ce ratio va, de par sa valeur, permettre de déterminer la précision d'une méthode. C'est dans ce contexte que les ratios moyens de chaque condition sont calculés. Pour simplifier la comparaison entre les méthodes, une seule valeur qui correspond à la moyenne des ratios des différentes conditions est reprise dans le **tableau 2**. Les différents ratios par condition sont repris en **annexe 2**. A partir du **tableau 2**, la méthode reposant sur un blast nécessitant un overlap minimum de 11 nucléotides (nt) est la plus spécifique. En effet, le ratio $\frac{SL1}{(SL1+SL2)}$ pour les gènes qui sont soit isolés soit en position 1 des opérons est de plus ou moins 95%, ce qui paraît logique puisque le trans-splicing SL2 est presque inexistant pour ces gènes. En ce qui concerne la méthode utilisant un blast reposant sur un overlap de 4 nt, les ratios témoignent d'une méthode moins précise. La conclusion est la même pour le mapping. Par contre, la méthode basée sur un blast nécessitant un overlap minimum de 8 nt présente un $\frac{SL1}{(SL1+SL2)}$ élevé (près de 90%), ce qui signifie qu'elle est assez précise.

Un autre ratio non repris dans le tableau 2 est le rapport entre le nombre d'événements de trans-splicing SL2 et le nombre d'événements de trans-splicing SL1 + SL2 pour les gènes en position 2 ou plus des opérons. Ainsi, pour les méthodes du blast avec un overlap minimum de 11 nt, de 8 nt et de 4 nt, les ratios respectifs de 75,82 % ; 59,06 % et 60,51 % ont été obtenus. En ce qui concerne la méthode basée sur le mapping, ce ratio est de 56,81 %. Ces ratios dont la valeur est inférieure à 100 %, s'interprètent par le fait que les gènes en position 2 ou plus des opérons subissent majoritairement, mais pas exclusivement, le trans-splicing SL2.

Une autre représentation, peut-être plus visuelle, de la précision est de vérifier si les gènes en position 2 ou plus des opérons sont trans-splicés SL2. Ainsi, les vrais positifs sont les gènes trans-splicés SL2 et qui sont en position 2 ou plus des opérons. Pour réaliser cela, les gènes détectés par une méthode donnée sont représentés graphiquement en fonction de

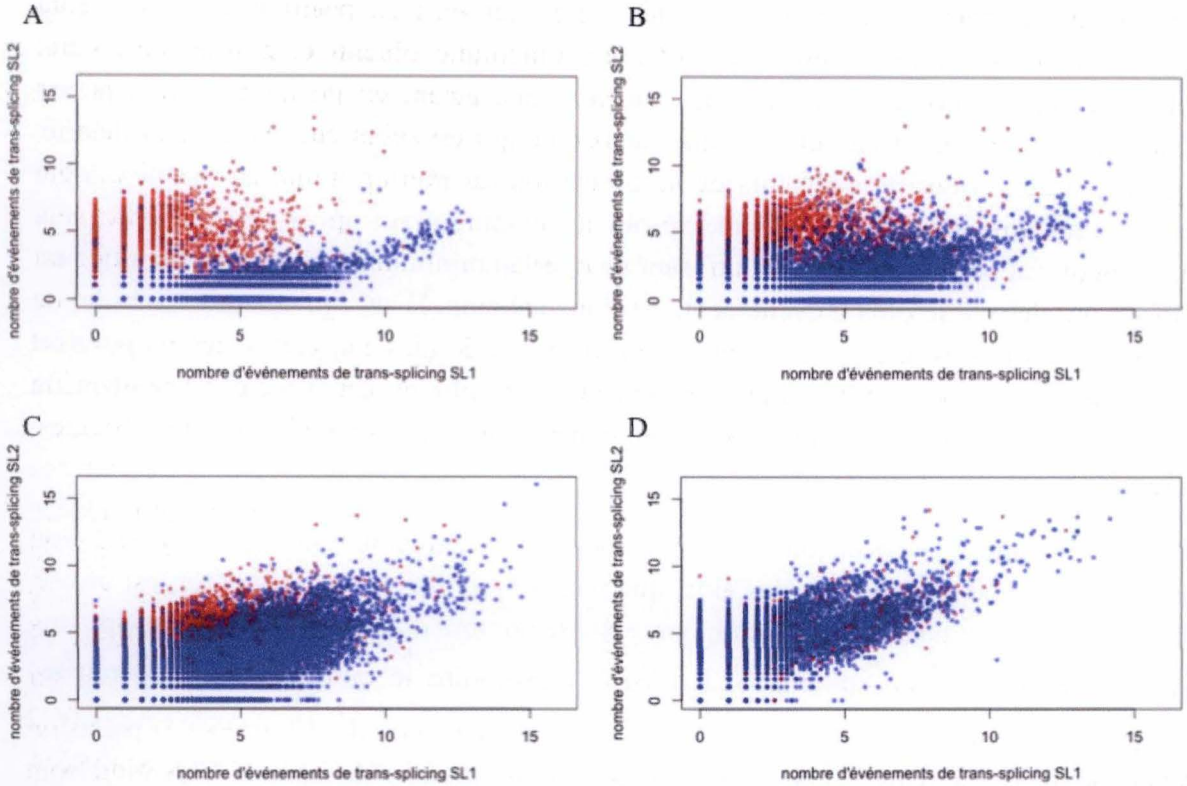


Figure 30 : Nuage de points représentant les gènes en fonction de leur nombre d'événements de trans-splicing. L'axe des abscisses est le nombre d'événements de trans-splicing SL1 tandis que l'axe des ordonnées représente le nombre d'événements de trans-splicing SL2. Chaque point représente un gène. La couleur des points nous renseigne sur la position des gènes dans le génome. La couleur rouge est attribuée aux gènes en position 2 ou plus des opérons tandis que la couleur bleue est attribuée aux gènes qui sont soit isolés soit en position 1 des opérons. (**A**) Résultat du blast avec un overlap minimum de 11 nt. (**B**) Résultat du blast avec un overlap minimum de 8 nt. (**C**) Résultat du blast avec un overlap minimum de 4 nt. (**D**) Résultat du mapping

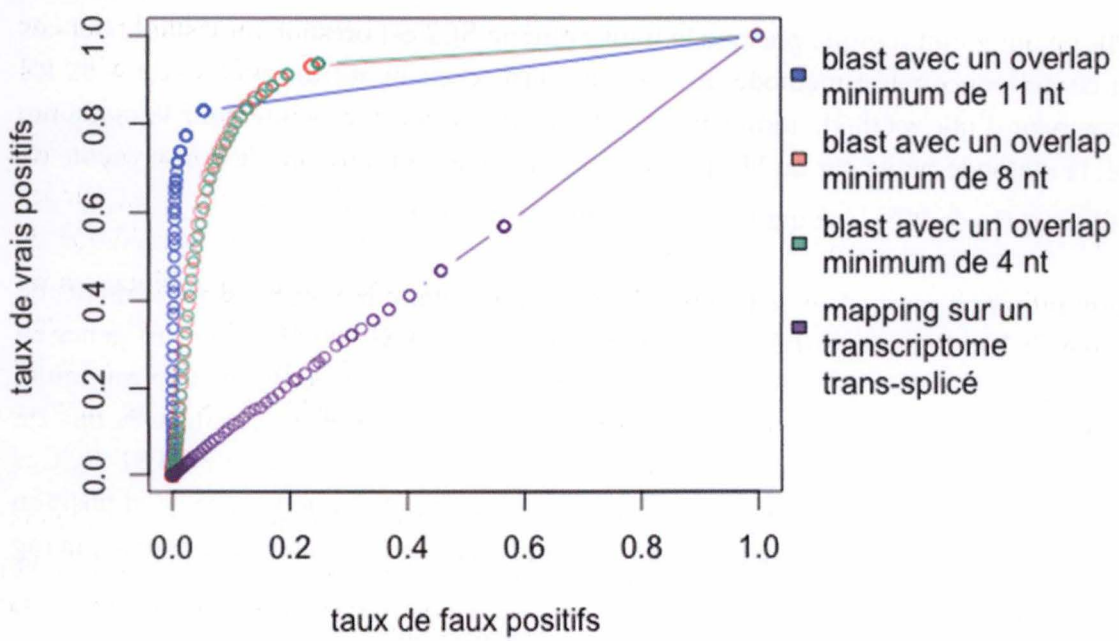


Figure 31 : Courbes de ROC des différentes méthodes. Les différentes couleurs permettent de différencier les différentes méthodes utilisées.

leur événement de trans-splicing. De plus, les gènes qui sont en position 2 ou plus sont identifiés et colorés d'une couleur spécifique. Le graphique obtenu comporte donc deux informations : une information sur le fait que les gènes soient en position 2 ou plus des opérons et une à propos du phénomène de trans-splicing que subissent les gènes. C'est de cette manière que les résultats obtenus par le blast avec un overlap minimum de 11 nt sont présentés à la **figure 30A**. Ainsi, les gènes en position 2 ou plus des opérons sont majoritairement corrélés à un taux de trans-splicing SL2 élevé tandis que les gènes qui sont soit isolés soit en position 1 des opérons sont davantage associés à un taux de trans-splicing SL1 important. La conclusion de ce graphe est donc que la position des gènes en opérons (position 2 ou plus ou non) détermine globalement la prévalence d'un type de trans-splicing sur l'autre. Ceci valide notre méthode puisqu'on sait que le trans-splicing de type SL2 est spécifique aux gènes en aval dans les opérons. Par la suite, les résultats obtenus à partir des autres méthodes sont analysés. Le blast avec un overlap de 8 nt présente la même tendance qu'avec 11 nt (**figure 30B**). Par contre, la diminution de l'overlap à 4 nt (**figure 30C**) altère la discrimination entre les événements SL1 et SL2. Ces résultats sont en accord avec la **figure 18** qui met en évidence que les méthodes basées sur le blast avec un overlap de 8 et 11 donnent des résultats plus spécifiques qu'avec un overlap de 4 nt. Le résultat du mapping sur le transcriptome trans-splicé est très inattendu puisque aucune discimination entre SL1 et SL2 n'est visible (**figure 30D**). Comme nous l'avons dit, les causes de cet échec sont analysées. En conclusion, les résultats obtenus par les blasts et un overlap minimum de 11 ou 8 sont ceux qui sont les plus en accord avec ce qui est connu du trans-splicing chez *C. elegans*.

Un troisième moyen de comparer différentes méthodes est de réaliser des courbes de ROC. Ces courbes permettent d'évaluer la capacité d'une méthode à prédire une vérité en mettant en relation sa spécificité avec sa sensibilité grâce à l'évaluation du taux de vrais positifs et de faux positifs. La meilleure méthode est généralement celle maximisant le taux de vrais positifs tout en minimisant les faux positifs. Dans le cas présent, la vérité à prédire est que le gène soit trans-splicé SL2 sur base du fait qu'il soit situé en position 2 ou plus des opérons et d'un paramètre "seuil" variable. La **figure 31** montre ces courbes pour les différentes méthodes utilisées. Ces courbes confirment que les techniques reposant sur les blasts de 11 et de 8 nt sont les plus appropriées. Les résultats obtenus avec la méthode du mapping semblent aléatoire, car la courbe est sur la diagonale. Ces graphiques nous renseignent sur la puissance de la méthode en se concentrant uniquement sur le taux de vrais positifs. Il est intéressant de remarquer que le taux de vrais positifs le plus important est celui relatif au blast avec un overlap de 4 nt, technique qui a été considérée comme la plus puissante.

En conclusion, la technique utilisant un blast avec un overlap minimum de 8 nt est retenue pour la quantification des événements de trans-splicing. En effet, cette technique semble un compromis entre d'une part, la puissance du blast avec un overlap de 4 nt et d'autre part, la précision du blast avec un overlap de 11 nt.

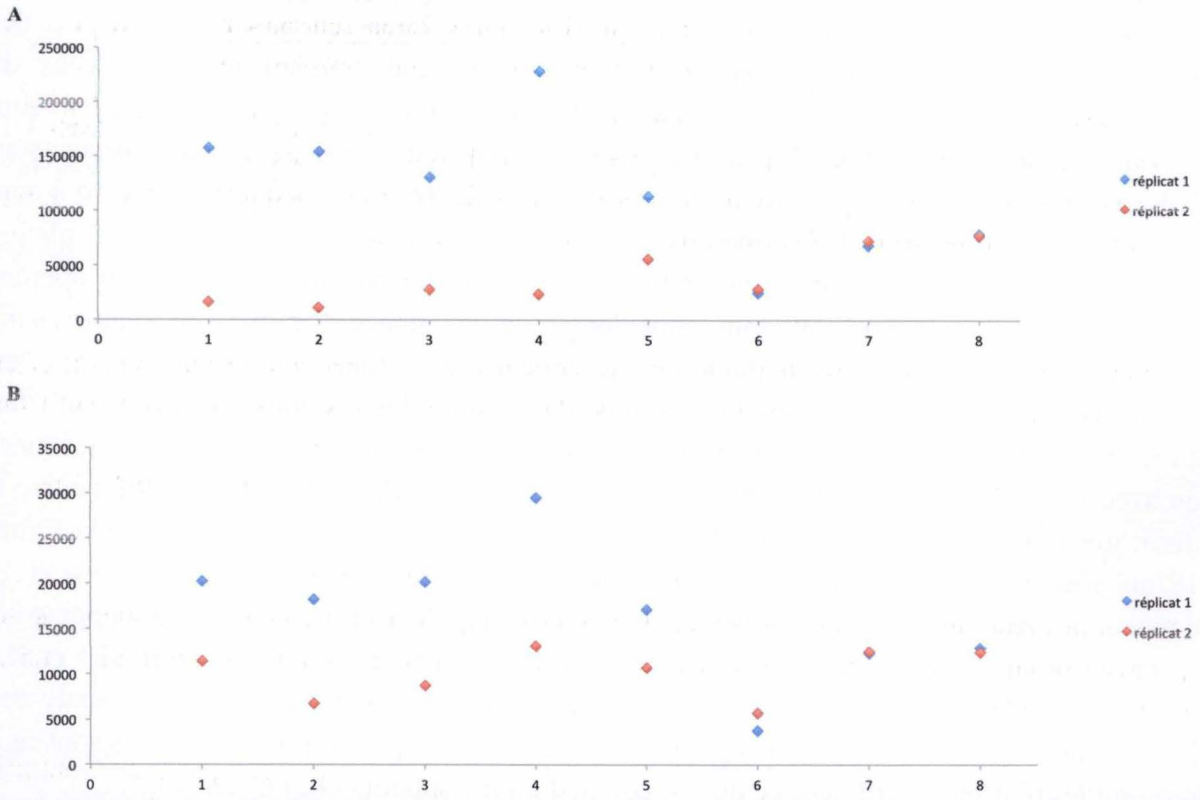


Figure 32 : Nombre d'événements de trans-splicing détectés avec notre méthode. L'axe des abscisses permet d'identifier les 8 conditions tandis que celui des ordonnées permet de quantifier le nombre d'événements de trans-splicing détectés. Les couleurs indiquent le réplicat utilisé. (**A**) Nombre d'événements de trans-splicing SL1. (**B**) Nombre d'événements de trans-splicing SL2.

4.2.2. Analyse des événements de trans-splicing

- Dans cette analyse, le rapport entre le nombre des événements de trans-splicing SL2 et celui de SL1, c'est-à-dire le ratio $\frac{\text{nombre d'événements de trans-splicing SL2}}{nombre\ d'événements\ de\ trans-splicing\ SL1}$, est utilisé plutôt que les événements de trans-splicing SL1 ou SL2. La raison est que les différentes conditions présentent une certaine variabilité résiduelle quant au nombre d'événements de trans-splicing (**figure 32**). Cette variabilité peut être à l'origine d'une mauvaise interprétation des résultats. Par contre, le calcul des ratios $\frac{\text{nombre d'événements de trans-splicing SL2}}{nombre\ d'événements\ de\ trans-splicing\ SL1}$ permet de normaliser les conditions. Notons cependant que ce choix d'utiliser les ratios $\frac{\text{nombre d'événements de trans-splicing SL2}}{nombre\ d'événements\ de\ trans-splicing\ SL1}$ plutôt que les événements de trans-splicing va influencer les conclusions de l'analyse. En effet, un ratio qui varie peut être dû à une variation du nombre d'événements de trans-splicing SL2 ou SL1.
- Pour augmenter le nombre d'événements de trans-splicing détectés, la somme de ceux-ci dans les deux réplicats et par condition est effectuée.

- En ce qui concerne la représentation graphique des ratios, le logarithme des ratios est calculé. En effet, les ratios présentent une distribution asymétrique. La transformation logarithmique permet d'obtenir une distribution symétrique. Cela est expliqué dans le matériel et méthodes.

- Une précision est à apporter par rapport aux différents graphiques. Les gènes qui sont pris en compte lors de la quantification résultent d'une filtration. Comme expliqué dans le matériel et méthodes, celle-ci permet de ne considérer que les gènes qui sont exprimés dans nos conditions.

  Remarque : ne sont présentés ici que les résultats obtenus pour les conditions S, R et RI, car un éventuel effet dans la S0 est difficilement interprétable comme discuté plus haut. Cependant, les résultats avec S0 sont montrés en **annexe 3**.

a) Analyse des gènes en position 2 ou plus des opérons

Dans un premier temps, les ratios $\frac{trans-splicing\ SL2}{trans-splicing\ SL1}$ de l'ensemble des gènes en position 2 ou plus des opérons sont analysés et leur logarithme est représenté sous forme de boxplot (**figure 33A, verso de la page suivante**). Tout d'abord, les conditions de starvation (S) et recovery (R), c'est-à-dire présence de nourriture, sont comparées pour le WT et le mutant *cdk-12*as. La différence observée entre ces deux conditions est similaire chez les deux souches : une augmentation du ratio $\frac{trans-splicing\ SL2}{trans-splicing\ SL1}$ est observée en recovery. Donc, la présence de nourriture a une influence sur le trans-splicing. Par contre, pour mettre en évidence l'effet de l'inhibiteur, les conditions R (présence de nourriture) et RI (présence de nourriture + inhibiteur) sont comparées. Ainsi, l'ajout d'inhibiteur engendre une augmentation du ratio $\frac{trans-splicing\ SL2}{trans-splicing\ SL1}$ chez le WT alors que cet ajout provoque une diminution du ratio dans la souche *cdk-12*as. Cette diminution est faible, mais va dans le sens de notre hypothèse selon laquelle CDK-12 a un impact sur le trans-splicing. Concernant l'effet chez le WT, celui-ci s'interprète comme étant un artéfact. En effet, aucune différence phénotypique entre les vers du groupe WT_R et WT_RI n'est observée. De plus, il n'y a pas de différence

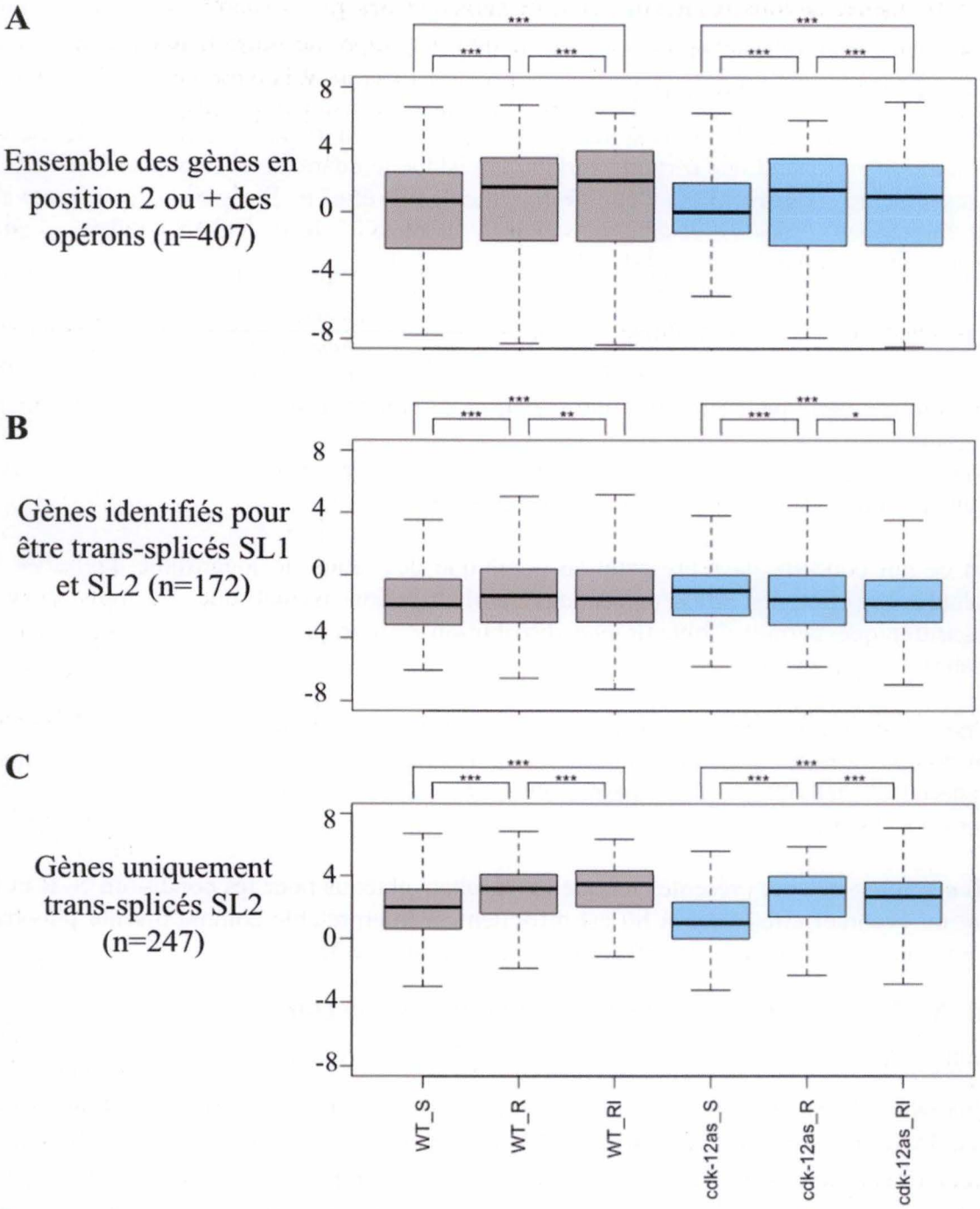


Figure 33 : Comparaison de la distribution des ratios $\frac{nombre\ d'événements\ de\ trans-splicing\ SL2}{nombre\ d'événements\ de\ trans-splicing\ SL1}$ de différents subsets de gènes pour les conditions S, R et RI. Les différents subsets sont (**A**) l'ensemble des gènes en position 2 ou plus des opérons, (**B**) gènes en position 2 ou plus des opérons et qui sont identifiés pour être trans-splicés SL1 et SL2 et (**C**) gènes qui sont uniquement trans-splicés SL2. Les couleurs nous renseignent sur la souche utilisée : le WT en gris et la souche *cdk-12as* en bleu clair. Pour chacun des subsets, un test de Wilcoxon est effectué. Les astérisques permettent de se rendre compte de la significativité des différences : * signifie que la différence est significative (p-value<0.05), ** signifie que celle-ci est hautement significative (p-value<0.01) et *** signifie que la différence est très hautement significative (p-value<0.001).

transcriptionnelle. Notons également que, la moyenne des ratios (barre rouge) est positive dans les différentes conditions, ce qui signifie que le nombre de trans-splicing SL2 dans les opérons est plus important que celui du SL1. Enfin, un test de Wilcoxon permet de souligner la significativité de ces différences (n=407). Grâce à ce test statistique, différentes conditions ont été testées. Ainsi, la p-value des conditions WT_S et WT_R ; WT_R et WT_RI ; *cdk-12as*_S et *cdk-12as*_R ; *cdk-12as*_R et *cdk-12as*_RI ; *cdk-12as*_S et *cdk-12as*_RI est inférieure à 2,2 e^{-16}. La seule p-value qui diffère est celle entre les conditions WT_S et WT_RI et est de 5,89 e^{-10}.

Les conclusions de cette première analyse sont tout d'abord, que le trans-splicing SL2 semble être favorisé en recovery plutôt qu'en starvation et ensuite, que l'ajout d'inhibiteur impacte le trans-splicing de type SL2 pour les gènes en position deux ou plus des opérons.

b) Analyse des gènes identifiés pour être trans-splicés SL1/SL2

Dans une étude du trans-splicing en 2011, Blumenthal *et al.* ont identifié près de 702 gènes comme pouvant être trans-splicés SL1 et SL2. Puisque ces gènes sont trans-splicés SL1 et SL2, ces gènes sont supposés être en opérons hybrides (opéron caractérisé par le fait que les gènes peuvent être trans-splicés SL2 et SL1). Le but de l'analyse ci-dessous est de ne considérer que ces gènes qui sont en position 2 ou plus des opérons hybrides et de se concentrer sur l'évolution de leur ratio suite à l'inhibition de CDK-12. L'intérêt de cela est que, puisque les gènes sont trans-splicés SL1 et SL2, ceux-ci peuvent réagir différemment à l'inhibition de CDK-12.

A nouveau, les ratios sont calculés et la distribution de leur logarithme est illustrée sous forme de boxplot (**figure 33B**). Sur base de cette figure, les conditions de starvation (S) et les conditions de recovery (R) sont comparées chez le WT et le mutant *cdk-12*as. Ainsi, la présence de nourriture a un impact sur le ratio des gènes en position 2 ou plus des opérons hybrides. Il est cependant observé que le ratio n'évolue pas de la même façon selon la souche WT ou *cdk-12as*. Pour expliquer cela, l'hypothèse est que la mutation a un impact sur le trans-splicing des gènes en position 2 ou plus des opérons hybrides. Lorsque les conditions recovery (uniquement présence de nourriture) et recovery + inhibiteur (présence de nourriture et inhibiteur), c'est-à-dire R et RI, sont comparées, une diminution du ratio est observée pour les deux souches. Deux hypothèses sont émises pour expliquer ces résultats : soit les diminutions observées chez les deux souches ont une origine commune soit elles ont une origine différente. En effet, un artéfact n'est pas à exclure quant à l'origine de ces diminutions. Cependant, l'analyse précédente (point 3.2.2.a) suggère un artéfact uniquement pour le WT. Cela pourrait également être le cas ici. En vue d'analyser la significativité de ces différences, un test de Wilcoxon est réalisé sur ces valeurs. Le nombre de gènes analysés par ce test est 172. Ainsi, la significativité des différences entre la condition avec uniquement de la nourriture (R) et celle avec de la nourriture plus de l'inhibiteur (RI) est différente selon la souche. En effet, la différence entre la condition R et RI chez le mutant *cdk-12as* est significative (p-value =0,0248). Par contre, en ce qui concerne le WT, cette différence est hautement significative (p-value = 0,00235).

De plus, une analyse sur l'expression de ces gènes effectuée par C. Yague-Sanz révèle que leur expression n'est pas significativement perturbée. Ainsi, le fait que les gènes soient exprimés de manière similaire, mais que leur ratio $\frac{\text{nombre d'événements de trans-splicing SL2}}{\textit{nombre d'événements de trans-splicing SL1}}$ diminue, suggère que, suite à l'inhibition de CDK-12, le trans-splicing SL1 pourrait compenser, du moins en partie, la diminution du trans-splicing SL2.

Le but de cette analyse était de mettre en évidence l'évolution du ratio $\frac{\textit{trans-splicing SL2}}{\textit{trans-splicing SL1}}$ suite à l'inhibition de CDK-12. Les conclusions de cette analyse sont d'une part, que la présence de nourriture influence le ratio $\frac{\textit{trans-splicing SL2}}{\textit{trans-splicing SL1}}$ et d'autre part, que l'inhibition de CDK-12 a potentiellement un impact sur ce ratio.

c) Analyse des gènes uniquement trans-splicés SL2

La première analyse réalisée concerne l'ensemble des gènes situés en position 2 ou plus des opérons. La deuxième a étudié les gènes en positions 2 ou plus des opérons hybrides. Dans cette troisième partie, la question est de savoir quel est l'effet observé pour les gènes des opérons qui ne sont pas hybrides et qui sont donc exclusivement trans-splicés SL2. Pour répondre à celle-ci, une soustraction des gènes en position 2 ou plus des opérons hybrides est faite à partir de l'ensemble des gènes en position 2 ou plus des opérons.

L'analyse en aval de ce subset de gènes est la même que précédemment, c'est-à-dire qu'un boxplot est généré à partir des logarithmes des ratios $\frac{\textit{trans-splicing SL2}}{\textit{trans-splicing SL1}}$ (**figure 33C, verso de la page précédente**). Sur base de ce graphique, les conditions de starvation (S) et recovery (R) sont comparées. Une augmentation du ratio en recovery par rapport à la starvation est clairement observée pour les deux souches : WT et *cdk-12as*. En ce qui concerne l'effet de l'inhibiteur, les conditions recovery (uniquement présence de nourriture) et recovery + inhibiteur (présence de nourriture et inhibiteur), c'est-à-dire R et RI, sont comparées. Une augmentation du ratio $\frac{\textit{trans-splicing SL2}}{\textit{trans-splicing SL1}}$ suite à l'ajout d'inhibiteur est observée chez le WT. L'hypothèse qu'un artéfact soit à l'origine de celle-ci est à nouveau émise. Par contre, concernant le mutant *cdk-12as*, une diminution du ratio $\frac{\textit{trans-splicing SL2}}{\textit{trans-splicing SL1}}$ est observée. Comme expliqué précédemment, l'hypothèse de l'artéfact n'est pas à exclure mais il se pourrait que l'ajout d'inhibiteur ait un impact sur ce ratio. Suite à ces résultats, le même test statistique que précédemment a été réalisé sur ces valeurs. Grâce à ce test, toutes les différences sont identifiées comme étant très hautement significatives avec à chaque fois une p-value inférieure à $2e^{-16}$.

Les conclusions de cette analyse sont donc que la présence de nourriture influence positivement le ratio $\frac{\textit{trans-splicing SL2}}{\textit{trans-splicing SL1}}$ et que l'inhibition de CDK-12 peut potentiellement influencer le ratio $\frac{\textit{trans-splicing SL2}}{\textit{trans-splicing SL1}}$

d) Analyse transversale des résultats obtenus

L'intérêt des différentes approches développées ci-desssus est d'analyser différents subsets de gènes. En effet, la première analyse prend en compte l'ensemble des gènes de *C. elegans* qui sont en position 2 ou plus des opérons. La seconde se concentre sur les gènes en position 2 ou plus des opérons hybrides, c'est-à-dire des gènes trans-splicés SL1 et SL2. La dernière permet de ne considérer que les gènes trans-splicés SL2. Une rapide analyse transversale des différents boxplots permet de souligner plusieurs différences.

- Une différence au niveau de l'ordre de grandeur des ratios $\frac{trans-splicing\ SL2}{trans-splicing\ SL1}$ entre les différents subsets de gènes analysés est observée. En effet, en ce qui concerne le subset des gènes en position 2 ou plus des opérons, les médianes et, par conséquent les ratios $\frac{trans-splicing\ SL2}{trans-splicing\ SL1}$, sont plus ou moins nulles. Cela signifie que pour la majorité des gènes, il y a autant de trans-splicing SL1 que SL2. Par contre, concernant le subset de gènes en position 2 ou plus des opérons hybrides, les médianes sont négatives. Le fait que la majorité des ratios $\frac{trans-splicing\ SL2}{trans-splicing\ SL1}$ soient négatifs signifie qu'il y a davantage de trans-splicing SL1 que SL2 pour ces gènes. A l'inverse, quand on se concentre sur l'analyse des gènes uniquement trans-splicés SL2, les médianes, et par conséquent les ratios $\frac{trans-splicing\ SL2}{trans-splicing\ SL1}$, sont positifs. Cela s'explique par le fait que ces gènes sont majoritairement trans-splicés SL2.

- L'effet de l'inhibition de CDK-12 sur le trans-splicing n'est pas différent selon le subset de gènes. En effet, quand les conditions de recovery (R) et recovery + inhibiteur (RI) sont comparées pour les différents subsets de gènes, la différence entre ces deux conditions est la même quel que soit le subset analysé. Par contre, les différences entre les conditions de starvation (S) et celles de recovery (R) varient avec le subset analysé. En effet, celles-ci sont relativement faibles pour les gènes en 2 ou plus des opérons hybrides. Ces différences sont un peu plus marquées lorsque l'ensemble des gènes en position 2 ou plus des opérons sont considérés et le sont encore davantage quand sont pris en compte uniquement les gènes en position 2 ou plus des opérons trans-splicés SL2. Cela coïncide avec l'étude de Baugh selon laquelle les gènes en opérons sont davantage transcrits en recovery plutôt qu'en starvation (*16*). Or, la majorité des gènes en aval des opérons sont trans-splicés SL2. Il semble donc normal que le ratio SL2/SL1 augmente en recovery.

- Il est également important de noter que les différents subsets de gènes ne sont pas de la même taille. Or, la p-value calculée lors du test statistique et qui a permis de souligner la significativité des différences est influencée par ce nombre d'échantillons analysés (n). En effet, plus le n augmente, plus la p-value est faible et plus le résultat du test est significatif.

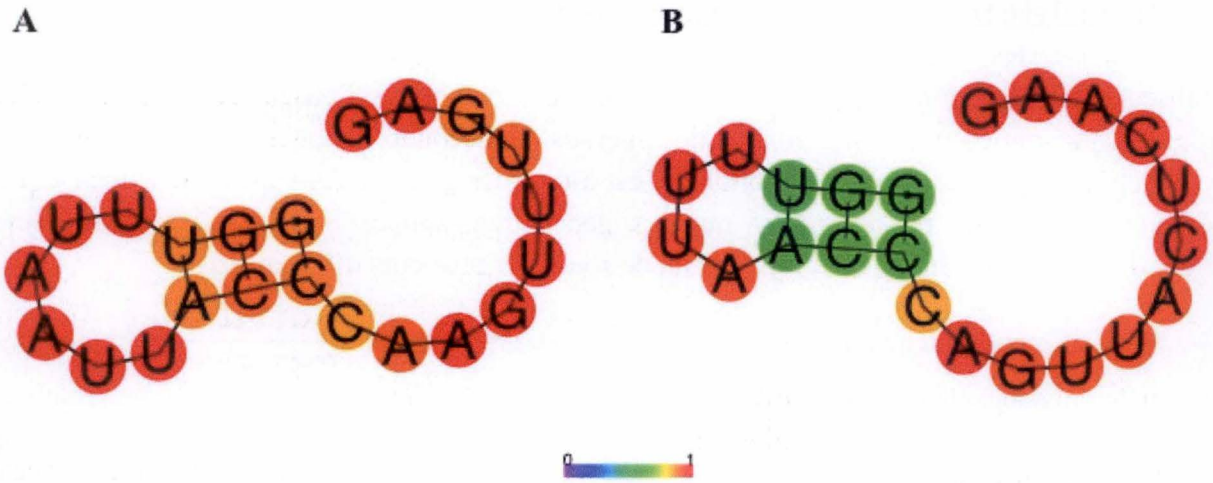


Figure 34 : Prédiction de la structure secondaire des splice leader. Les couleurs nous renseignent sur la probabilité que les bases soient pairées. (**A**) Structure du splice leader 1 (SL1). (**B**) Structure du splice leader 2 (SL2).

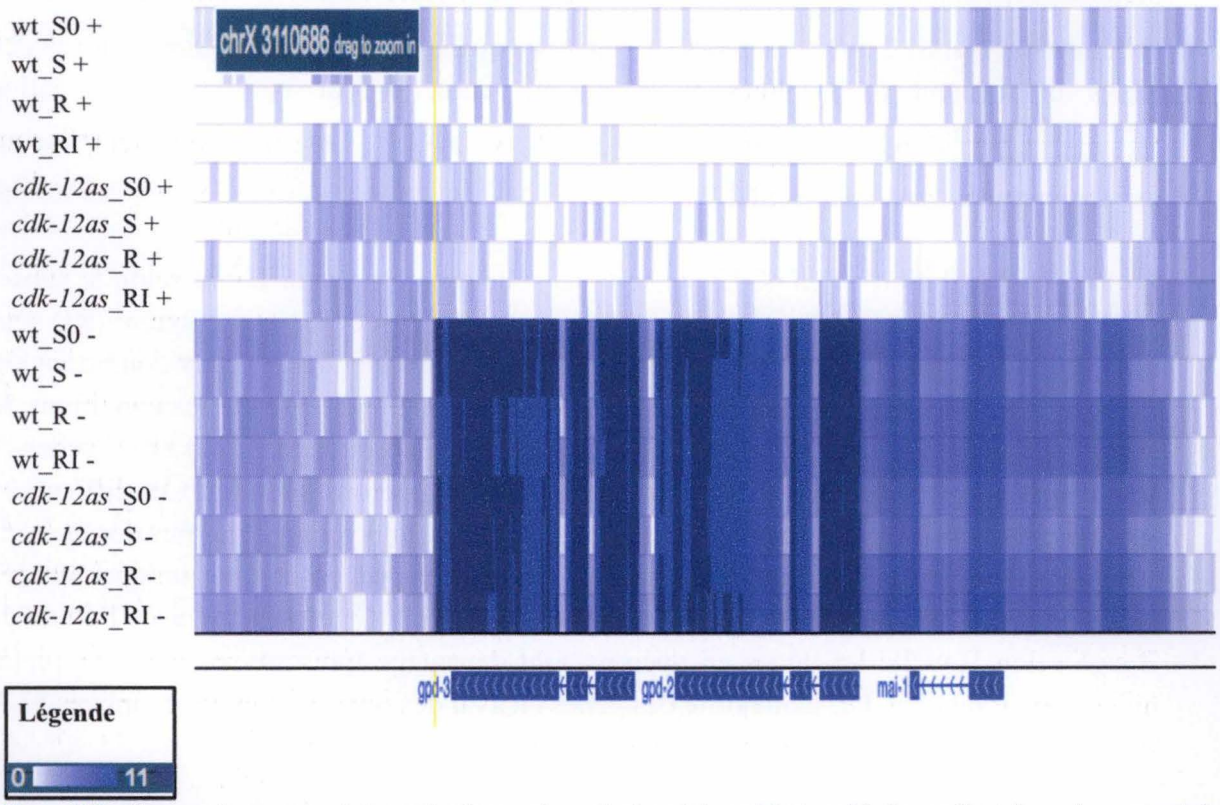


Figure 35 : *Genome browser* centré sur l'opéron qui contient *mai-1*, *gpd-2* et *gpd-3*. Le gradient de couleur permet de visualiser le nombre de reads qui mappent à ces positions du génome de *C. elegans*. Sur cette figure, on voit que le gène *gdp-3* est en position 3 d'un opéron. De plus, cet opéron est positionné sur le brin négatif. Grâce au dégradé de couleur bleu, on voit que l'opéron est exprimé uniformément dans les différentes conditions.

4.3. Détermination expérimentale d'un arrêt de rétro-transcription dans la structure du splice leader par primer extension

Il a été mentionné précédemment que plus la valeur de l'overlap entre le read et la séquence SL augmente, plus le nombre d'événements de trans-splicing détectés diminue (**figure 29**). Pour expliquer cela, on peut proposer qu'une structure secondaire au sein des séquences des splice leader (SL) interfère avec la rétro-transcription, rendant moins probable des rétro-transcrits longs. Le fait que la rétro-transcription soit sensible à des structures secondaires a été rapporté pour les ARN de transfert (ARNt), les ARN ribosomiaux (ARNr) ou encore chez les rétrovirus (*46*). Concernant *C. elegans*, plusieurs études mentionnent que les SL possèdent une structure secondaire (*22, 26*) (*29*) (**figure 34**).

Dans le but de tester l'hypothèse d'un potentiel arrêt de la rétro-transcription (RT) à cause d'une structure secondaire du splice leader (SL), la technique du primer extension a été mise en œuvre. Cette technique permet d'identifier, grâce à l'utilisation de primers radioactifs, des polynucléotides de tailles différentes ayant été générés au cours d'une réaction de rétro-transcription (RT). Pour réaliser cette expérience, un gène d'intérêt (*gpd-3*) a été choisi, car il est décrit dans la littérature comme étant en position 3 d'un opéron et est exprimé dans les différentes conditions utilisées dans cette étude (**figure 35**).

Un primer extension a été réalisé avec l'aide d'Olivier Finet. Il utilise comme template un ARN synthétique de 57 nt, dont la séquence correspond à celle de *gpd-3* après trans-splicing. Le résultat est illustré à la **figure 36 (verso de la page suivante)**. Sept conditions sont testées. Le puits 0 contient l'étalon, les puits 1 jusque 4 contiennent une quantité croissante d'ARN (0,3 ; 3 ; 30 et 300 ng). Le puits 5 est un contrôle négatif sans ARN. Les puits 6 et 7 sont d'autres contrôles et contiennent le primer, de l'ARN (respectivement 0,3 et 300 ng), mais pas de rétrotranscriptase. L'analyse du gel est ensuite réalisée. La **figure 22**, présentée précédemment, permet de faciliter l'interprétation. Différentes bandes sont observées. Premièrement, il y a une bande à 80 nt dépendante de la quantité d'ARN et de la présence de la rétrotranscriptase. Cette bande est absente des puits 5, 6 et 7 où la réaction de RT n'a pas lieu. La présence de cette bande dont la taille est supérieure au template ne peut s'expliquer que par le fait que l'enzyme reviendrait sur ses pas comme illustré à la **figure 37 (verso de la page suivante)**. Une deuxième bande qui est observée a une taille d'environ 55 nt, ce qui s'approche de la taille prédite (53 nt) dans le cas d'une RT complète et son intensité augmente avec la quantité d'ARN. Etrangement, cette bande ne disparaît pas totalement en absence d'ARN synthétique (puits 5) ou de rétrotranscriptase (puits 6), bien qu'elle soit fortement atténuée. Troisièmement, la bande ayant une taille de 26 nt correspond au primer non incorporé et diminue logiquement proportionnellement à l'augmentation des autres bandes. Quatrièmement, une fine bande ayant une taille de plus ou moins 30 nt est visible notamment dans les puits 2 et 3 et correspond sans doute à un arrêt mineur de la rétro-transcription au niveau de la jonction entre la séquence du SL2 et celle de *gpd-3*.

Cette expérience de primer extension révèle la difficulté d'obtenir des résultats de RT non ambigus à l'extrémité d'un ARN trans-splicé. Aucune évidence d'un arrêt complet au pied de l'hairpin n'est présente, mais les bandes à 30 nt et 80 nt témoignent d'un comportement inattendu qui pourrait interférer avec l'analyse des reads en RNA-seq.

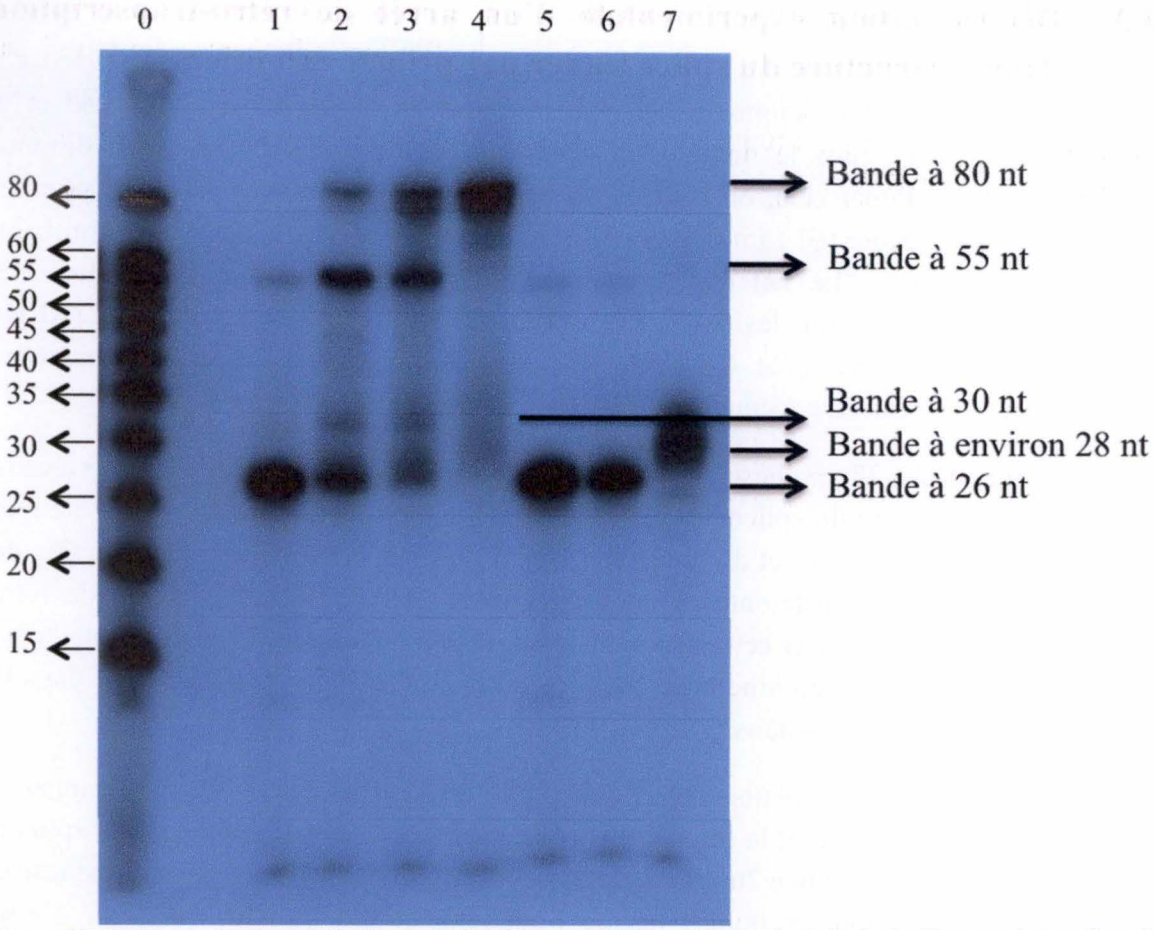


Figure 36 : Primer extension réalisé à partir d'ARNm synthétique. Le premier puits (puits 0) est un étalon allant de 15 à 80 nt. Les puits 1, 2, 3 et 4 contiennent des concentrations croissantes d'ARN : respectivement 0,3 ; 3 ; 30 et 300 ng. Le puits 5 est un contrôle négatif où l'ARNm est remplacé par de l'eau. Les puits 6 et 7 constituent également des contrôles négatifs : ce sont des puits similaires aux puits 4 et 7 en absence de rétrotranscriptase.

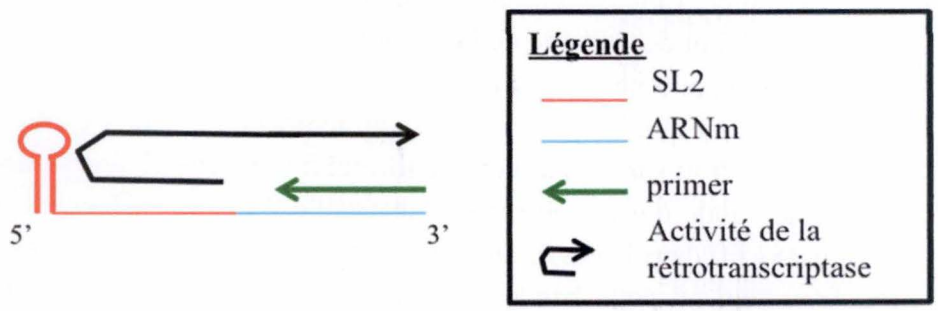


Figure 37 : Réaction potentielle de la rétrotranscriptase face à une structure secondaire. La réaction de rétro-transcription se fait en direction de l'extrémité portant la séquence SL (extrémité 5') grâce à une rétrotranscriptase. Lorsque celle-ci rencontre la structure secondaire, la rétrotranscriptase reviendrait sur ces pas.

Très récemment, il a été rapporté qu'un traitement à la pyrophosphatase est nécessaire pour améliorer le séquençage des ARN de type SL. En effet, ce traitement a été utilisé par Maxwell pour séquencer et distinguer les séquences SL (*51*). Cette donnée souligne encore que la prudence doit être de mise dans l'analyse de reads (ou de rétro-transcription) aux extrémités trans-splicées.

5. Conclusion et perspectives

5.1. Conclusion

Compte tenu de la nature de mon travail, nous avons opté pour un agencement combinant résultats et discussion. Nous commencerons ici par un brève conclusion avant d'évoquer les perspectives du travail.

Le but de mon mémoire était d'étudier le rôle *in vivo* de CDK-12, la kinase de la sérine 2 du CTD, chez l'organisme modèle *C. elegans*. Pour cela, une souche de ver *cdk-12as* qui permet une inhibition complète et réversible de la kinase a été utilisée. Les travaux antérieurs à mon arrivée ont révélé que le phénotype résultant de l'inhibition de CDK-12 est un arrêt au premier stade larvaire indépendamment de la présence de nourriture. Une étude du transcriptome par RNA-seq dans ces conditions a mis en évidence que CDK-12 est requise pour l'induction en présence de nourriture d'un sous-groupe de gènes. Des analyses complémentaires ont révélé que ceux-ci sont majoritairement localisés en position 2 ou plus des opérons et que leur fonction est apparentée au développement larvaire. Par la suite, le mécanisme moléculaire par lequel l'inhibition de CDK-12 provoque la downrégulation de ces gènes a été investigué. La particularité des gènes en position 2 ou plus des opérons chez *C. elegans* est un trans-splicing de type SL2. Comme développé dans l'introduction, l'ajout d'un ARN SL2 est couplé à la polyadénylation du premier messager de l'opéron et les complexes CPSF et CstF impliqués contiennent des sous-unités qui lient le CTD phosphorylé sur la sérine 2.

C'est pourquoi, pour expliquer la downrégulation des gènes en position 2 ou plus des opérons suite à l'inhibition de CDK-12, l'hypothèse que CDK-12 serait impliquée dans le phénomène de trans-splicing SL2 a été émise. Elle tient au fait qu'une diminution du niveau de S2P après inhibition de CDK-12 pourrait diminuer le recrutement du complexe CstF et donc le trans-splicing SL2. En l'absence de leader SL2, les ARN messagers seraient beaucoup plus accessibles aux exonucléases, ce qui pourrait expliquer la forte diminution de leur niveau d'expression.

Des analyses moléculaires sont en cours au laboratoire pour tester l'hypothèse décrite ci-dessus. Lors de ce mémoire, je me suis attelé à tenter de quantifier les événements de trans-splicing en présence ou non d'inhibiteur, grâce à des méthodes reposant sur un blast des reads correspondant aux extrémités des messagers. L'analyse des ratios $\frac{\text{nombre d'événements de trans-splicing SL2}}{\textit{nombre d'événements de trans-splicing SL1}}$ révèle que le phénomène de trans-splicing est perturbé suite à l'ajout d'inhibiteur. Cependant, une variation de ce ratio ne permet pas de déterminer si c'est le trans-splicing SL2 qui est affecté ou le trans-splicing de type SL1. Cette question reste donc à être investiguée. Grâce à cette analyse, il a également été démontré que la présence de nourriture influence de manière plus marquée le ratio $\frac{\text{nombre d'événements de trans-splicing SL2}}{\textit{nombre d'événements de trans-splicing SL1}}$.

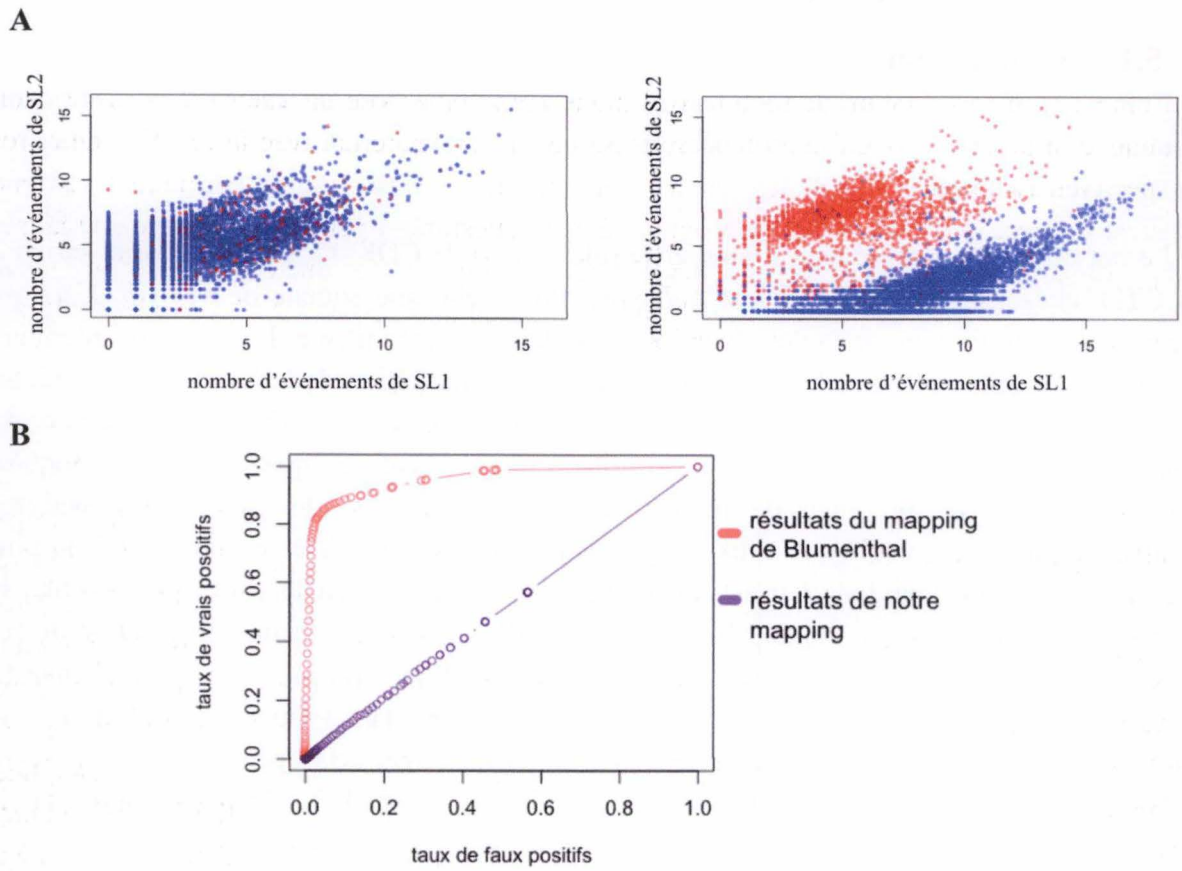


Figure 38 : Comparaison de la méthode du mapping réalisée au cours de cette étude avec la méthode du mapping effectuée par Blumenthal. (A) Comparaison des nuages de points représentant les gènes en fonction de leur nombre d'événements de trans-splicing, pour notre mapping (graphique de gauche) et celui de Blumenthal (graphique de droite). L'axe des abscisses est le nombre d'événements de trans-splicing SL1 tandis que l'axe des ordonnées représente le nombre d'événements de trans-splicing SL2. Chaque point représente un gène. La couleur des points nous renseigne sur la position des gènes dans le génome. La couleur rouge est attribuée aux gènes en position 2 ou plus des opérons tandis que la couleur bleue est attribuée aux gènes qui sont soit isolés soit en position 1 des opérons. **(B)** Courbes de ROC obtenues pour les deux méthodes.

5.2. Perspectives

5.2.1. Quantité d'ARN utilisée lors du primer extension

Des manipulations de primer extension ont été réalisées en vue de tester l'hypothèse selon laquelle la structure secondaire en 5' des séquences SL interfèrerait avec la réaction de rétro-transcription (RT). Cependant, les résultats obtenus jusqu'à présent ne permettent pas une conclusion claire. Afin d'essayer de répondre à cette question, il peut être envisagé de réaliser des primer extensions à partir de quantités d'ARN plus importantes que celles utilisées dans le cadre de cette étude ou en présence de phosphatase ou en utilisant de l'ARN extrait plutôt que synthétique.

5.2.2. Perfectionnement de la méthode du mapping pour quantifier les événements de trans-splicing

Afin de quantifier les événements de trans-splicing dans nos conditions grâce à une méthode basée sur un mapping, un transcriptome virtuellement trans-splicé a été généré. Cependant, les résultats obtenus avec la méthode du mapping semblent aléatoires. De plus, Blumenthal a déjà utilisé cette méthode pour quantifier des événements de trans-splicing dans un autre contexte avec des résultats probants (*19*). La **figure 38** permet de comparer ces deux mapping sur base de leur précision et de leur courbe de ROC. Ainsi, il a été suggéré que notre méthode comporte une erreur au niveau de la génération du transcriptome trans-splicé. Identifier cette erreur en vue de pouvoir appliquer un mapping pour quantifier les événements de trans-splicing pourrait peut-être permettre d'investir le rôle de CDK-12 dans le phénomène de trans-splicing SL2.

5.2.3. Distinction entre gènes trans-splicés SL1 et SL2 en vue de déterminer l'impact de l'inhibition de CDK-12 sur un type de trans-splicing plutôt que l'autre

Le fait de réaliser notre étude sur les rapports entre les événements de trans-splicing SL2 et SL1 nous empêche toute conclusion à propos du type de trans-splicing affecté. Pour résoudre cela, il faudrait pouvoir distinguer les deux types de splice leader (SL).

Dans une autre étude, Maxwell *et al.* ont réalisé un RNA-seq où les extractions d'ARN ont été préalablement traitées à l'aide d'une réaction de pyrophosphatase en milieu acide. De cette manière, la pyrophosphatase va agir sur la coiffe pour *in fine* rendre possible le séquençage du splice leader et ainsi identifier les gènes trans-splicés SL1 et SL2.

Il pourrait dès lors être envisagé de refaire le séquençage des conditions de recovery et recovery + inhibiteur en présence de Tobacco Acid Pyrophosphatase afin de permettre une meilleure quantification des événements de trans-splicing.

5.2.4. Etude d'un éventuel rôle de CDK-12 dans la méthylation des histones

La heatmap d'expression permet de mettre en évidence plusieurs groupes de gènes qui sont différentiellement exprimés (p-value < 0,05) suite à l'inhibition de CDK-12. Un de ces subsets est enrichi en gènes en position deux ou plus des opérons. Pour expliquer pourquoi CDK-12 agit spécifiquement sur ces gènes, l'hypothèse que CDK-12 est impliquée dans le trans-splicing SL2 a été émise. Cependant, une question qui se pose est de savoir ce qu'il en est des gènes affectés par l'inhibition de CDK-12 et qui ne sont pas en position deux ou plus des opérons. Pour expliquer cela, on peut suggérer qu'il s'agit soit d'un effet indirect soit d'un effet direct. Parmi les rôles potentiels de CDK-12, on peut proposer que celle-ci est impliquée dans la méthylation des histones pour plusieurs raisons.

- Il est connu de la littérature qu'il existe un lien direct entre la S2P et la triméthylation de la lysine 36 de l'histone 3 (H3K36me3). En effet, Bowman a mis en évidence un lien entre CDK-12, S2P et le niveau de H3K36me3 au niveau de la lignée germinale de *C. elegans* (*3*).
- En outre, cela est en accord avec les résultats du "gene onthology". En effet, les résultats du "gene onthology" suggère que la lignée germinale est affectée suite à l'inhibition de CDK-12. Une éventuelle perturbation de la méthylation des histones au niveau de la lignée germinale suite à l'inhibition de CDK-12 permettrait d'expliquer ce résultat.

Dès lors, il pourrait être intéressant d'investiguer le rôle de CDK-12 dans le méthylation des histones à l'échelle du génome.

6. Bibliographie

1. D. Eick, M. Geyer, The RNA polymerase II carboxy-terminal domain (CTD) code. *Chemical reviews* **113**, 8456 (Nov 13, 2013).
2. J. P. Hsin, J. L. Manley, The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & development* **26**, 2119 (Oct 1, 2012).
3. E. A. Bowman, C. R. Bowman, J. H. Ahn, W. G. Kelly, Phosphorylation of RNA polymerase II is independent of P-TEFb in the C. elegans germline. *Development* **140**, 3703 (Sep, 2013).
4. S. Egloff, M. Dienstbier, S. Murphy, Updating the RNA polymerase CTD code: adding gene-specific layers. *Trends in genetics : TIG* **28**, 333 (Jul, 2012).
5. M. Heidemann, D. Eick, Tyrosine-1 and threonine-4 phosphorylation marks complete the RNA polymerase II CTD phospho-code. *RNA biology* **9**, 1144 (Sep, 2012).
6. T. Kamenski, S. Heilmeier, A. Meinhart, P. Cramer, Structure and mechanism of RNA polymerase II CTD phosphatases. *Molecular cell* **15**, 399 (Aug 13, 2004).
7. J. Drogat, D. Hermand, Gene-specific requirement of RNA polymerase II CTD phosphorylation. *Molecular microbiology* **84**, 995 (Jun, 2012).
8. C. Jeronimo, P. Collin, F. Robert, The RNA Polymerase II CTD: The Increasing Complexity of a Low-Complexity Protein Domain. *Journal of molecular biology* **428**, 2607 (Jun 19, 2016).
9. B. Palancade, O. Bensaude, Investigating RNA polymerase II carboxyl-terminal domain (CTD) phosphorylation. *European journal of biochemistry* **270**, 3859 (Oct, 2003).
10. D. Coudreuse *et al.*, A gene-specific requirement of RNA polymerase II CTD phosphorylation for sexual differentiation in S. pombe. *Current biology : CB* **20**, 1053 (Jun 22, 2010).
11. C. Cassart, J. Drogat, V. Migeot, D. Hermand, Distinct requirement of RNA polymerase II CTD phosphorylations in budding and fission yeast. *Transcription* **3**, 231 (Sep-Oct, 2012).
12. X. Li, N. Chatterjee, K. Spirohn, M. Boutros, D. Bohmann, Cdk12 Is A Gene-Selective RNA Polymerase II Kinase That Regulates a Subset of the Transcriptome, Including Nrf2 Target Genes. *Scientific reports* **6**, 21455 (Feb 25, 2016).
13. D. Blazek *et al.*, The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes & development* **25**, 2158 (Oct 15, 2011).
14. O. Sin, H. Michels, E. A. Nollen, Genetic screens in Caenorhabditis elegans models for neurodegenerative diseases. *Biochimica et biophysica acta* **1842**, 1951 (Oct, 2014).
15. A. K. Corsi, B. Wightman, M. Chalfie, A Transparent window into biology: A primer on Caenorhabditis elegans. *WormBook : the online review of C. elegans biology*, 1 (Jun 18, 2015).
16. L. R. Baugh, J. Demodena, P. W. Sternberg, RNA Pol II accumulates at promoters of growth genes during developmental arrest. *Science* **324**, 92 (Apr 3, 2009).
17. C. Murgatroyd, D. Spengler, Histone tales: echoes from the past, prospects for the future. *Genome biology* **11**, 105 (2010).
18. T. Blumenthal, P. Davis, A. Garrido-Lecca, Operon and non-operon gene clusters in the C. elegans genome. *WormBook : the online review of C. elegans biology*, 1 (Apr 28, 2015).
19. M. A. Allen, L. W. Hillier, R. H. Waterston, T. Blumenthal, A global analysis of C. elegans trans-splicing. *Genome research* **21**, 255 (Feb, 2011).
20. D. A. Zorio, N. N. Cheng, T. Blumenthal, J. Spieth, Operons as a common form of chromosomal organization in C. elegans. *Nature* **372**, 270 (Nov 17, 1994).

21. X. Y. Huang, D. Hirsh, A second trans-spliced RNA leader sequence in the nematode Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 8640 (Nov, 1989).
22. D. Evans *et al.*, A complex containing CstF-64 and the SL2 snRNP connects mRNA 3' end formation and trans-splicing in C. elegans operons. *Genes & development* **15**, 2562 (Oct 1, 2001).
23. A. Zaslaver, L. R. Baugh, P. W. Sternberg, Metazoan operons accelerate recovery from growth-arrested states. *Cell* **145**, 981 (Jun 10, 2011).
24. G. B. Danks *et al.*, Trans-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Molecular biology and evolution* **32**, 585 (Mar, 2015).
25. T. Blumenthal, Trans-splicing and operons in C. elegans. *WormBook : the online review of C. elegans biology*, 1 (Nov 20, 2012).
26. D. Evans, T. Blumenthal, trans splicing of polycistronic Caenorhabditis elegans pre-mRNAs: analysis of the SL2 RNA. *Molecular and cellular biology* **20**, 6659 (Sep, 2000).
27. T. Blumenthal, Trans-splicing and operons. *WormBook : the online review of C. elegans biology*, 1 (Jun 25, 2005).
28. J. H. Graber, J. Salisbury, L. N. Hutchins, T. Blumenthal, C. elegans sequences that control trans-splicing and operon pre-mRNA processing. *RNA* **13**, 1409 (Sep, 2007).
29. E. L. Lasda, T. Blumenthal, Trans-splicing. *Wiley interdisciplinary reviews. RNA* **2**, 417 (May-Jun, 2011).
30. P. Huang *et al.*, Identification and analysis of internal promoters in Caenorhabditis elegans operons. *Genome research* **17**, 1478 (Oct, 2007).
31. Q. Lei *et al.*, Evolutionary Insights into RNA trans-Splicing in Vertebrates. *Genome biology and evolution* **8**, 562 (Mar 10, 2016).
32. R. Chila, F. Guffanti, G. Damia, Role and therapeutic potential of CDK12 in human cancers. *Cancer treatment reviews* **50**, 83 (Sep 14, 2016).
33. B. Bartkowiak *et al.*, CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes & development* **24**, 2303 (Oct 15, 2010).
34. S. X. Andrew Fire, Mary K. Montgomery, Steven A. Kostas, Samuel E. Driver & Craig C. Mello, Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806 (1998).
35. J. S. Kim *et al.*, A chemical-genetic strategy reveals distinct temporal requirements for SAD-1 kinase in neuronal polarization and synapse formation. *Neural development* **3**, 23 (Sep 22, 2008).
36. P. J. Alaimo, M. A. Shogren-Knaak, K. M. Shokat, Chemical genetic approaches for the elucidation of signaling pathways. *Current opinion in chemical biology* **5**, 360 (Aug, 2001).
37. J. Gregan *et al.*, Construction of conditional analog-sensitive kinase alleles in the fission yeast Schizosaccharomyces pombe. *Nature protocols* **2**, 2996 (2007).
38. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57 (Jan, 2009).
39. D. Kim *et al.*, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36 (Apr 25, 2013).
40. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (Aug 15, 2009).
41. C. Camacho *et al.*, BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421 (Dec 15, 2009).

42. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923 (Apr 1, 2014).
43. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).
44. S. Raithel *et al.*, Inferential considerations for low-count RNA-seq transcripts: a case study on the dominant prairie grass Andropogon gerardii. *BMC genomics* **17**, 140 (Feb 27, 2016).
45. T. Stiernagle, Maintenance of C. elegans. *WormBook : the online review of C. elegans biology*, 1 (Feb 11, 2006).
46. B. Wittig, S. Wittig, Reverse transcription of tRNA. *Nucleic acids research* **5**, 1165 (Apr, 1978).
47. Y. Motorin, S. Muller, I. Behm-Ansmant, C. Branlant, Identification of modified residues in RNAs by reverse transcription-based methods. *Methods in enzymology* **425**, 21 (2007).
48. B. Bartkowiak, C. Yan, A. L. Greenleaf, Engineering an analog-sensitive CDK12 cell line using CRISPR/Cas. *Biochimica et biophysica acta* **1849**, 1179 (Sep, 2015).
49. A. Garrido-Lecca, T. Saldi, T. Blumenthal, Localization of RNAPII and 3' end formation factor CstF subunits on C. elegans genes and operons. *Transcription* **7**, 96 (May 26, 2016).
50. K. Liang *et al.*, Characterization of human cyclin-dependent kinase 12 (CDK12) and CDK13 complexes in C-terminal domain phosphorylation, gene transcription, and RNA processing. *Molecular and cellular biology* **35**, 928 (Mar, 2015).
51. C. S. Maxwell, I. Antoshechkin, N. Kurhanewicz, J. A. Belsky, L. R. Baugh, Nutritional control of mRNA isoform expression during developmental arrest and recovery in C. elegans. *Genome research* **22**, 1920 (Oct, 2012).

7. Annexes

Annexe 1 : Résultats avec condition S0

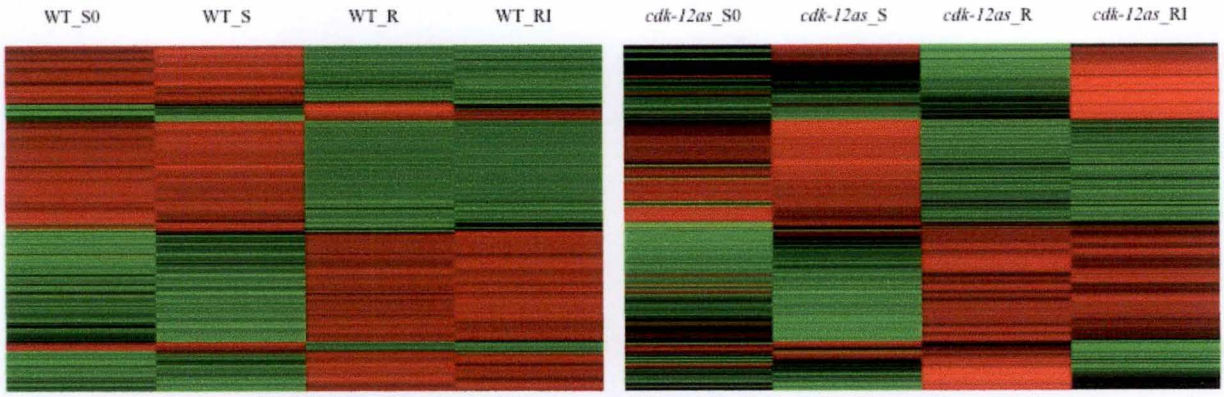


Figure : Profil d'expression génique dans les différentes conditions S0, S, R et RI. Chaque ligne représente un gène alors qu'une colonne est une condition. Pour déterminer le Z score, les données rlog de la table de summarization sont centrées et réduites. Le Z score a pour intérêt de permettre une comparaison aisée entre plusieurs conditions en tenant compte de la moyenne entre les différentes conditions et la déviation standard. Ainsi, une représentation de l'expression peut être faite. Pour une condition donnée, un Z score négatif signifie que le gène est downrégulé et la ligne correspondante est colorée en rouge. Par contre, s'il est positif, le gène est uprégulé et la ligne correspondante est colorée en vert. Ainsi, en se concentrant sur une seule ligne, on peut suivre l'évolution de l'expression d'un gène dans les différentes conditions.

Annexe 2 : table de comparaison détaillée des différentes méthodes

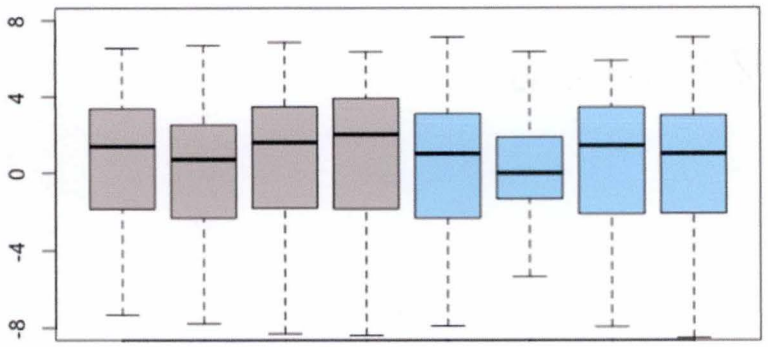
Base de la méthode	Blast des reads hybrides sur les séquences SL en spécifiant un overlap minimum de 11 nt		Blast des reads hybrides sur les séquences SL en spécifiant un overlap minimum de 8 nt	
N^bre d'événements de TS détectés	**SL1** 492467	**SL2** 122179	**SL1** 1270201	**SL2** 215524
N^bre d'événements de TS détectés par condition		**SL1** / **SL2**		**SL1** / **SL2**
	WT_S0	33711 / 9628	**WT_S0**	88624 / 15804
	WT_S	35182 / 7102,5	**WT_S**	82613,5 / 12467
	WT_R	27475 / 9682,5	**WT_R**	79195 / 14436,5
	WT_RI	55487,5 / 14456	**WT_RI**	126676 / 21335,5
	***cdk-12as*_S0**	19921 / 6727	***cdk-12as*_S0**	84519,5 / 13930
	***cdk-12as*_S**	8091,5 / 2430	***cdk-12as*_S**	26946 / 4749,5
	***cdk-12as*_R**	40327 / 5726,5	***cdk-12as*_R**	70822 / 12405
	***cdk-12as*_RI**	27038 / 5337	***cdk-12as*_RI**	77904,5 / 12634,5
Ratio $\frac{SL2}{(SL1+SL2)}$ par condition, des gènes situés en position 2 ou plus des opérons		**Ratio**		**Ratio**
	WT_S0	82,23 %	**WT_S0**	67,87 %
	WT_S	68,27 %	**WT_S**	62,34 %
	WT_R	85,77 %	**WT_R**	61,64 %
	WT_RI	75,60 %	**WT_RI**	69,13 %
	***cdk-12as*_S0**	81,44 %	***cdk-12as*_S0**	52,22 %
	***cdk-12as*_S**	75,23 %	***cdk-12as*_S**	50,59 %
	***cdk-12as*_R**	65,88 %	***cdk-12as*_R**	55,20 %
	***cdk-12as*_RI**	72,10 %	***cdk-12as*_RI**	52,66 %
Ratio $\frac{SL1}{(SL1+SL2)}$ par condition, des gènes situés en position 1 des opérons		**Ratio**		**Ratio**
	WT_S0	90,29 %	**WT_S0**	87,31 %
	WT_S	95,76 %	**WT_S**	92,85 %
	WT_R	90,20 %	**WT_R**	90,04 %
	WT_RI	96,76 %	**WT_RI**	88,68 %
	***cdk-12as*_S0**	94,75 %	***cdk-12as*_S0**	91,23 %
	***cdk-12as*_S**	97,33 %	***cdk-12as*_S**	92,44 %
	***cdk-12as*_R**	97,72 %	***cdk-12as*_R**	92,72 %
	***cdk-12as*_RI**	96,57 %	***cdk-12as*_RI**	92,50 %
Ratio $\frac{SL1}{(SL1+SL2)}$ par condition, des gènes situés en position 1 des opérons et les gènes isolés		**Ratio**		**Ratio**
	WT_S0	93,58 %	**WT_S0**	84,09 %
	WT_S	97,19 %	**WT_S**	83,30 %
	WT_R	92,89 %	**WT_R**	94,77 %
	WT_RI	96,76 %	**WT_RI**	91,59 %
	***cdk-12as*_S0**	94,75 %	***cdk-12as*_S0**	94,69 %
	***cdk-12as*_S**	97,33 %	***cdk-12as*_S**	96,85 %
	***cdk-12as*_R**	97,72 %	***cdk-12as*_R**	93,40 %
	***cdk-12as*_RI**	96,51 %	***cdk-12as*_RI**	94,66 %

Base de la méthode	Blast des reads hybrides sur les séquences SL en spécifiant un overlap minimum de 4 nt		Mapping sur un transcriptome trans-splicé		
N^{bre} d'événements de TS détectés	**SL1** 1489571	**SL2** 601049	**SL1** 311171	**SL2** 173445	
N^{bre} d'événements de TS détectés par condition		**SL1**	**SL2**		**SL1**	**SL2**
	WT_S0	116045,5	65832	**WT_S0**	20996	8845,5
	WT_S	91908,5	29635,5	**WT_S**	18807,5	9043,5
	WT_R	89735,5	32745,5	**WT_R**	23661,5	9500,5
	WT_RI	138732	44346	**WT_RI**	26435,5	16092
	***cdk-12as*_0**	104185	44935	***cdk-12as*_0**	17537	10742
	***cdk-12as*_S**	31186,5	13205	***cdk-12as*_S**	16638	8171,5
	***cdk-12as*_R**	81832	31825	***cdk-12as*_R**	15436,5	12736
	***cdk-12as*_RI**	91595	38000,5	***cdk-12as*_RI**	16073,5	11591
Ratio $\frac{SL2}{(SL1+SL2)}$ par condition, des gènes situés en position 2 ou plus des opérons		**Ratio**			**Ratio**	
	WT_S0	68,52 %		**WT_S0**	57,59 %	
	WT_S	61,50 %		**WT_S**	58,06 %	
	WT_R	62,38 %		**WT_R**	46,43 %	
	WT_RI	66,92 %		**WT_RI**	64,50 %	
	***cdk-12as*_0**	55,66 %		***cdk-12as*_0**	47,43 %	
	***cdk-12as*_S**	55,42 %		***cdk-12as*_S**	45,41 %	
	***cdk-12as*_R**	57,86 %		***cdk-12as*_R**	63,39 %	
	***cdk-12as*_RI**	56,29 %		***cdk-12as*_RI**	71,63 %	
Ratio $\frac{SL1}{(SL1+SL2)}$, par condition, des gènes situés en position 1 des opérons		**Ratio**			**Ratio**	
	WT_S0	65,23 %		**WT_S0**	68,37 %	
	WT_S	79,90 %		**WT_S**	55,77 %	
	WT_R	81,96 %		**WT_R**	71,76 %	
	WT_RI	78,01 %		**WT_RI**	47,47 %	
	***cdk-12as*_0**	80,94 %		***cdk-12as*_0**	68,58 %	
	***cdk-12as*_S**	86,48 %		***cdk-12as*_S**	70,64 %	
	***cdk-12as*_R**	83,23 %		***cdk-12as*_R**	56,20 %	
	***cdk-12as*_RI**	81,57 %		***cdk-12as*_RI**	63,33 %	
Ratio $\frac{SL1}{(SL1+SL2)}$, par condition, des gènes situés en position 1 des opérons et les gènes isolés		**Ratio**			**Ratio**	
	WT_S0	60,15 %		**WT_S0**	46,65 %	
	WT_S	68,63 %		**WT_S**	1,1502 %	
	WT_R	74,01 %		**WT_R**	63,33 %	
	WT_RI	72,96 %		**WT_RI**	61,20 %	
	***cdk-12as*_0**	72,26 %		***cdk-12as*_0**	63,36 %	
	***cdk-12as*_S**	89,55 %		***cdk-12as*_S**	86,07 %	
	***cdk-12as*_R**	82,48 %		***cdk-12as*_R**	48,92 %	
	***cdk-12as*_RI**	79,53 %		***cdk-12as*_RI**	54,02 %	

Annexe 3 :

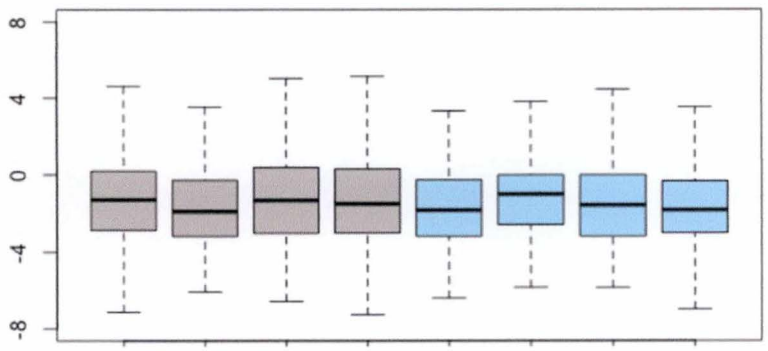
A

Ensemble des gènes en position 2 ou + des opérons (n=407)



B

Gènes identifiés pour être trans-splicés SL1 et SL2 (n=172)



C

Gènes uniquement trans-splicés SL2 (n=247)

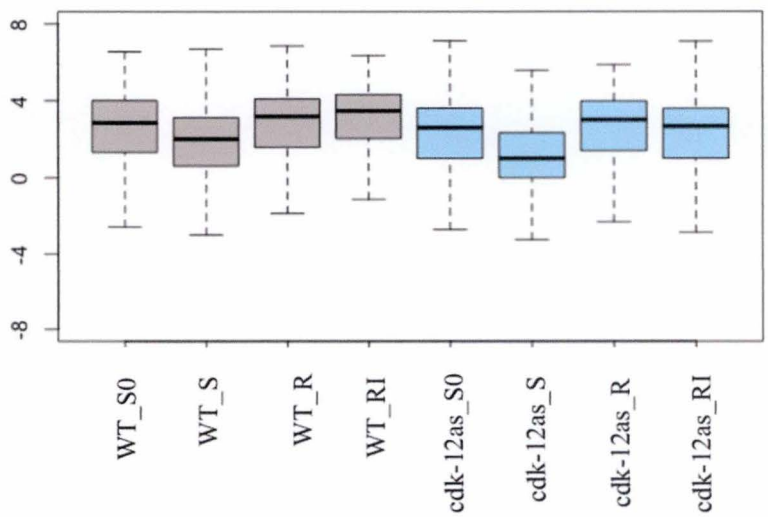


Figure : Comparaison de la distribution des ratios $\frac{nombre\ d'événements\ de\ trans-splicing\ SL2}{nombre\ d'événements\ de\ trans-splicing\ SL1}$ de différents subsets de gènes pour les conditions S0, S, R et RI. Les différents subsets sont (**A**) l'ensemble des gènes en position 2 ou plus des opérons, (**B**) gènes en position 2 ou plus des opérons et qui sont identifiés pour être trans-splicés SL1 et SL2 et (**C**) gènes qui sont uniquement trans-splicés SL2. Les couleurs nous renseignent sur la souche utilisée : le WT en gris et la souche *cdk-12as* en bleu clair.