Rapport de stage de M2

Nicolas JEANNE

11 mars 2015

Introduction

En 1982, la découverte par Higgins de nouveaux éléments génétiques communs dans les régions intercistroniques des opérons de Escherichia coli et Salmonella typhimurium a constitué le premier pas de la recherche sur les Repeated Extragenic Palindrome (REP) (Higgins et al. 1982). En 1991, Gilson et al. ont mis en évidence l'organisation en clusters de ces REP (Gilson et al. 1991), ces clusters ont été caractérisés comme Bacterial Interspesed Mosaic Element (BIME). Chez E. coli en 1994, Bachelier et son équipe ont réussi à catégoriser les REP constituant les BIME en 2 types Y et Z, constituants 3 motifs Y, Z^1, Z^2 (Bachellier et al. 1994).

Les REP constituent une part non négligeable du génome bactérien, chez E. coli K12 ou S. typhimurium elles représentent environ 1% de celui-ci (Gilson et al. 1991). Nous les retrouvons chez de nombreux règnes bactériens, notamment chez les pathogènes humains tels que *Escherichia coli, Salmonella enterica, Neisseria meningitidis, Mycobacterium tuberculosis et Pseudomonas aeruginosa* mais également chez des pathogènes des plantes comme *Agrobacterium tumefaciens* ou chez des bactéries ubiquitaires, *Deinococcus radiodurans* ou *Pseudomonas putida* par exemple. Les travaux précédents de l'équipe ont permis l'annotation des REP au sein du génome d'E. coli et de mettre en évidence le lien existant entre la prolifération des REP et le gène $tnpA_{REP}$ (Bosc 2014; Weyder 2013), ainsi que la reconstruction des états ancêtres des REP (Bosc 2014). Le rôle exact des REP n'est pas clairement défini, des hypothèses sont avancées sur leur implication dans la régulation de l'expression des gènes, que ce soit en tant que terminateur ou comme site de reconnaissance des enzymes impliquées dans les mécanismes de la transcription.

Caractéristiques des REP et organisations en BIME

La taille des REP varie de 20 à 40 nucléotides, la classification Y, Z^1, Z^2 est basée à la fois sur la taille de la séquence consensus de la REP ainsi que sur sa structure secondaire. Par convention, une REP en orientation inversée est nommée iREP (inversed REP) (Ton-Hoang et al. 2012). Un tétra-nucléotide caractéristique de séquence GTAC est présent à l'extrémité 5' des REP, sa séquence complémentaire est CTAC en 3' pour les iREP. Les différentes classes

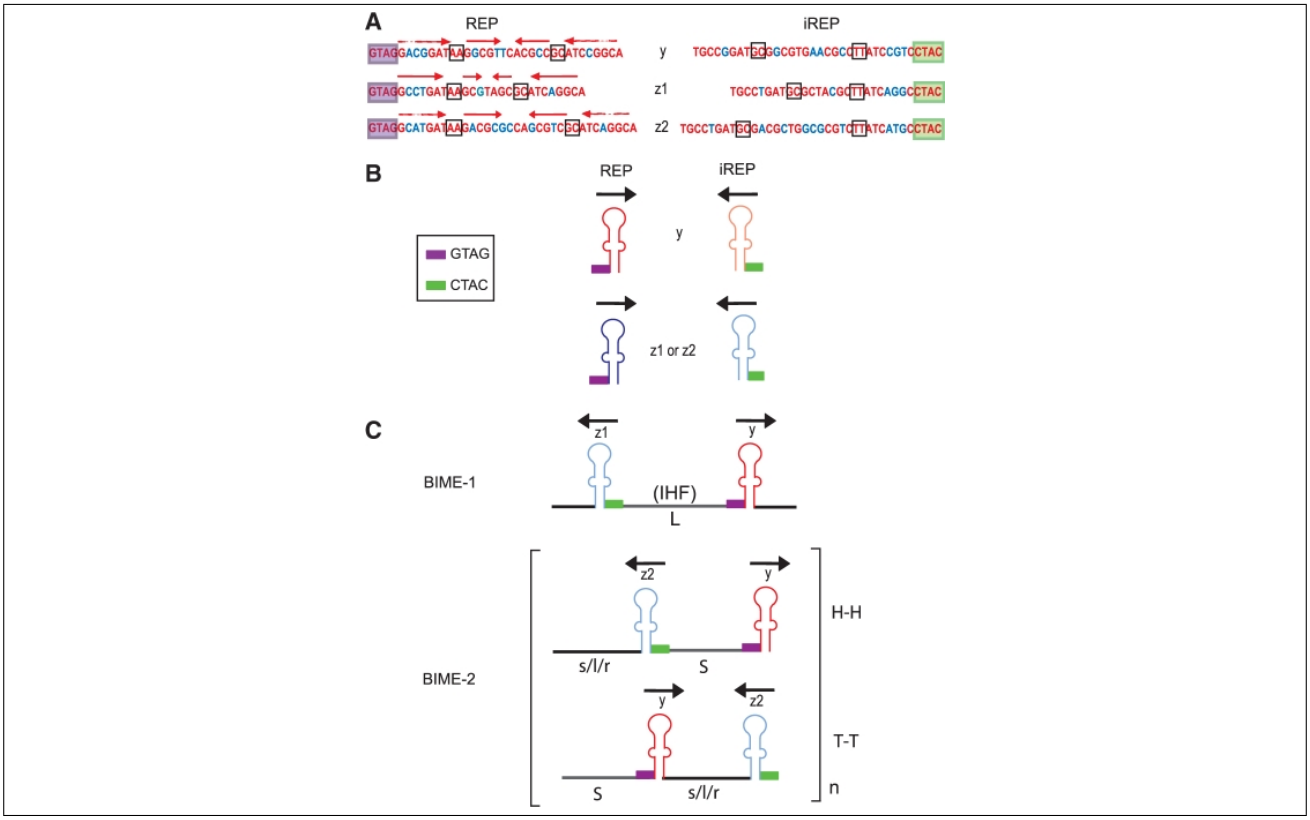


FIGURE 1 – **REP et BIME chez Escherichia coli. (A)** Séquences consensus Y, Z^1 et Z^2 des REP. Le tétra-nucléotide conservé GTAC est encadré en violet, le complémentaire conservé CTAC est encadré en vert, les flèches rouges situent les zones d'appariement de la tige et les positions encadrées en noir sont les zones de mésappariement. Les positions conservées parmi les classes de REP sont en rouge, les positions variables en bleu. **(B)** Structure secondaire des REP. Les rectangles violets et verts représentent respectivement les tétra-nucléotides conservés GTAC pour les REP et CTAC pour les iREP. Les flèches noires indiquent l'orientation des REP. **(C)** Structures des BIME-1 et BIME-2. Les BIME-1 sont composées de REP et de iREP Y et Z^1 séparées par un linker de séquence longue (L), les BIME-2 sont composées de Y et Z^2, de linker courts (S) et de séquences séparatrices s, l ou r. H-H et T-T dénotent respectivement une organisation tête à tête et queue à queue des REP. (Ton-Hoang et al. 2012).

de REP partagent des nucléotides conservés (Figure 1A). La structure secondaire des REP est caractérisée par sa forme en tige-boucle, le caractère palindromique permet la formation de la tige malgré un mésappariement situé dans la partie centrale de celle-ci (Figure 1B). Pour le génome d'E. coli K-12, 93 REP ont été répertoriées comme étant uniques sur les 605 annotées par le laboratoire, les autres sont organisées par paires sous forme de BIME. Une classification a été adoptée comportant 3 entrées, les BIME-1 composées de REP Z^1 et Y apparaissant en paires uniques dans lesquelles la REP et l'iREP sont séparées par un linker de séquence longue (L) pouvant lier l'IHF (Integration Host Factor). Les BIME-2 constituées de Z^2 et de Y, apparaissant en copies multiples de cette paire dont la REP et l'iREP sont séparées par un linker court (S) et une des trois séquences flanquantes (s, l ou r). La troisième catégorie est constituée des BIME dites atypiques qui sont des chimères de BIME-1 et BIME-2, comportant différentes combinaisons de Y, Z^1, Z^2, S, L, s, l et r. Tout comme les BIME-2, nous les retrouvons sous forme de copies multiples (Figure 1C). Les REP peuvent former des structures secondaires

avec elles-même, mais également entre elles lorsqu'elles sont organisées sous forme de BIME (Figure 2).

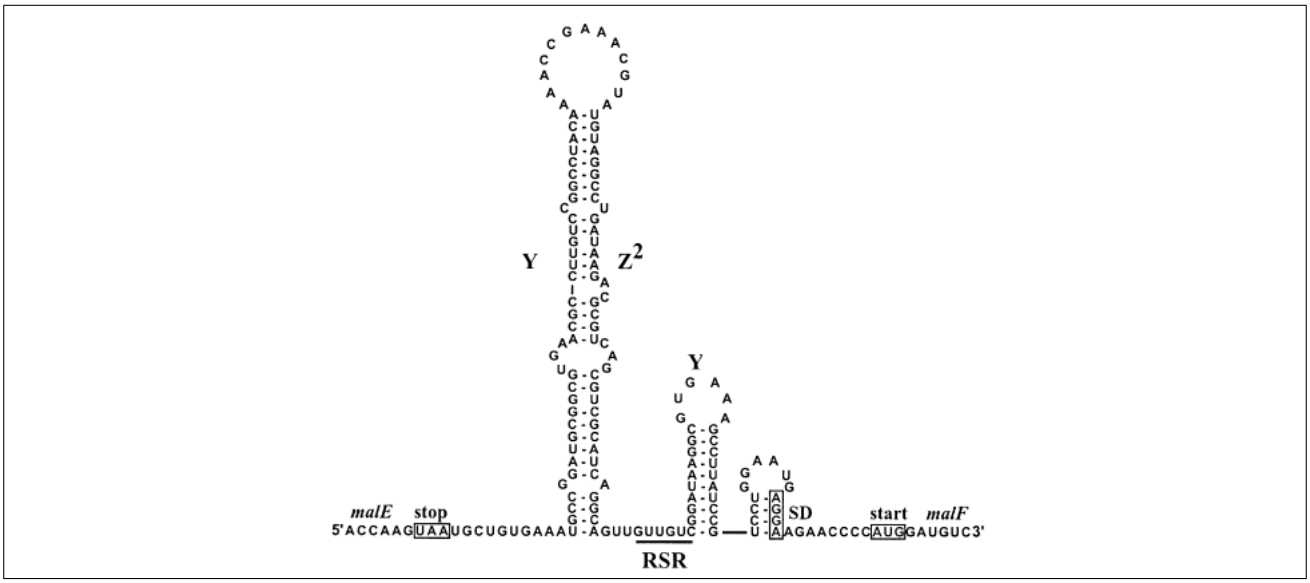


FIGURE 2 – **Structure ARN des REP au sein de l'opéron malEFG.** Y, Z^2 et Y indiquent la séquence des REP dans l'espace inter-génique de malE-malF. Bien que Y et Z^2 puissent former des structures tige-boucles par elles mêmes, elles s'apparient ensemble pour former une région étendue en grande partie à double brin (70% des nucléotides sont appariés). La séquence affichée provient du génome d'E. coli K12. La région REP-stabilized RNA (RSR) indique l'extrémité 3′ du messager malE mature, qui s'étend de 3 à 9 nucléotides depuis la base de la tige-boucle formée par Y et Z^2. Les codons STOP de malE et START de malF sont encadrés. SD représente la séquence Shine–Dalgarno nécessaire à l'initiation de la traduction de malF. (Khemici and Carpousis 2004).

Propriétés associées aux REP

La littérature décrit de nombreuses fonctions associées aux REP, mais certaines d'entre elles restent encore peu étudiées. Les REP ont été décrite comme jouant un rôle dans les événements de recombinaisons homologues (Kofoid et al. 2003). Les BIME ont été décrites comme des sites privilégiés pour l'insertion de séquences d'ADN mobiles comme certaines familles d'IS (Insertion Sequence) (Bachellier et al. 1997; Choi et al. 2003; Clément et al. 1999; Tobes and Pareja 2005). Lorsqu'elles sont transcrites, les REP joueraient un rôle dans la stabilisation de l'ARNm grâce à leur structure en tige-boucle (Aguena et al. 2009; Espéli et al. 2001; Khemici and Carpousis 2004; Newbury et al. 1987), la terminaison de la transcription (Gilson et al. 1986) et le contrôle de la traduction (Stern et al. 1988). Au niveau de l'ADN, les REP sont capables de lier plusieurs facteurs protéiques tels que l'ADN Gyrase (Espéli and Boccard 1997) et l'ADN polymérase (Gilson et al. 1990). Plus spécifiquement, la BIME-1 peut lier l'IHF sur son linker (Boccard and Prentki 1993) qui peut être notamment responsable de l'initiation de la transcription et d'événements de recombinaisons sites spécifiques (Goosen et al. 1995).

Transcription chez E. coli

Stabilité des ARN

Matériel & Méthodes

Données

Les données que nous avons exploité sont issues des expériences d'évolution adaptatives en laboratoire visant à découvrir l'émergence de mutations clés permettant la croissance rapide d'E. coli K-12 MG1655 sur un medium pauvre en glucose (LaCroix et al. 2014). Ces données ont été choisies car elles proviennent d'expériences de RNA-Seq comportant un nombre non négligeable de réplicats (9) pour la condition de croissance en milieu pauvre en glucose (ALE) et 2 réplicats pour le Wild Type (WT) , mais nous n'avons exploité que la condition ALE, le nombre de réplicats du WT étant faible. Il s'agit de données Next Generation Sequencing (NGS) publiques, accessibles sur le base de données GEO (Gene Expression Omnibus) du NCBI au format Sequence Read Archive (SRA), séquencées sur Illumina MiSeq à partir d'ARN total extrait des cultures d'E. coli et rétro-transcrit en cDNA. La librairie a été conçue en Paired-end **et brin spécifique en utilisant la méthode dUTP** (Levin et al. 2010) pas sûr du tout vu la tête des données sur IGV.... Grâce au SRA toolkit et à la commande `fastq-dump`, elles sont décompressées au format `fastq`. Un contrôle de qualité a été effectué afin d'inspecter les séquences grâce au logiciel `fastqc`. 8 ALE ont été validés puisque disposant d'une qualité de séquence par base supérieure à 30 pour des reads de 62 pb, seul le fichier `SRR1573441.fastq` a été rejeté car la longueur des reads allait de 35 à 502 pb avec des scores de qualités très variables.

Alignement des reads

Les reads ont été alignés grâce au logiciel BWA sur le génome d'E. coli NC_000913.2 qui est le génome utilisé pour annoter les REP au laboratoire. Le logiciel BWA propose 3 algorithmes distincts, BWA-backtrack, BWA-SW et BWA-MEM. Pour chacun de ces alignements, il est nécessaire de disposer d'une séquence de référence indexée, obtenue par la commande `bwa index NC_000913.2.fasta`. L'algorithme que nous avons sélectionné est le MEM (Maximal

Exact Matches) car il s'agit du plus récemment développé. Il reprend les mêmes principes que BWA-SW (utilisation de la programmation dynamique pour trouver les seeds en autorisant les mismatchs et les gaps. Il n'étend les alignements des seeds que lorsque ceux-ci ont peu d'occurrences sur le génome de référence, cela permet de diminuer le temps d'alignement en éliminant les extensions des séquences très répétées) mais en utilisant le seeding avec des MEM, puis il réalise l'extension avec les autorisations de mismatchs et de gaps.

```
# Alignement avec l'algorithme MEM de BWA
bwa mem ref.fasta file.fastq > aln.sam
```

Le fichier d'alignement généré est au format Sequence Alignment Map format (SAM), afin de poursuivre l'analyse il doit être convertit au format Binary Alignment Map format (BAM), puis des critères de qualité sont appliqués. Seules les séquences possédant une qualité de mapping > 30 et n'étant pas taggées comme alignement chimérique sont conservées. Cette opération a aussi le mérite de compresser l'information et ainsi de gagner en espace de stockage. Les séquences vont ensuite être triées par position génomique.

Finalement, le fichier BAM trié doit être indexé pour être visualisable sur un Genome Browser. Les outils utilisés sont compris dans la suite des samtools.

```
# Conversion du SAM en BAM et application des filtres
# (-q 30 : score de mapping minimal, -F 2048 : suppression des
# sequences chimeriques).
samtools view -Sbh -q 30 -F 2048 aln.sam > aln.bam

# Tri en fonction des positions genomiques
samtools sort aln.bam aln_sorted

# Indexation du fichier d'alignement
samtools index aln_sorted.bam
```

Pour les besoins ultérieurs de l'analyse, les réplicats d'une même condition sont fusionnés en un seul fichier.

```
# Fusion des replicats.
samtools merge merged.bam aln_sorted_1.bam \
aln_sorted_2.bam aln_sorted_3.bam
```

Préparation des fichiers de référence

Le fichier General Feature Format (GFF) d'annotation du génome a été généré par un script Perl à partir du fichier GenBank. Les fichiers répertoriant les opérons et les promoteurs (source RegulonDB), les terminateurs de transcription (source Door^{2}DB) ainsi que les REP et BIME (source laboratoire) ont été transformés au format Browser Extensible Data (BED) grâce à des scripts Python. Ces changements de format permettent de travailler aisément avec la suite de logiciels BEDtools pour la recherche d'intersections, de positions proches ou de couverture de reads. Ces outils ont généré les fichiers `BED` nécessaires à la suite de l'analyse tel que celui des positions génomiques des opérons contenant des REP, celui des REP, celui des gènes contenus dans des opérons bordant une REP, celui des BIME et celui de la couverture sur les régions contenant des BIME.

Visualisation du mapping

L'alignement des reads et le mapping sur le génome de référence de E. coli sont visualisés grâce au genome browser IGV (Robinson et al. 2011; Thorvaldsdóttir et al. 2013) (Figure 3).

Il est important de noter que la couverture au long du génome n'est pas uniforme, ni même sur les gènes, car nous observons la présence de nombreuses vallées et pics. Ce phénomène s'explique par plusieurs raisons techniques (Li et al. 2013). Premièrement, les méthodes de fragmentation des protocoles de préparation des bibliothèques amenant un biais en cassant ou dégradant certaines séquences. Le second biais possible est amené par le Random Priming lors de l'étape de rétro-transcription pouvant préférentiellement transcrire certaines séquences. Troisième point, les ligases peuvent lier préférentiellement les adaptateurs à certaines séquences. Quatrième point, l'amplification de la PCR est bien connue pour introduire des biais dépendant de la proportion en GC des séquences. Le dernier point, étudié sur le séquençage Illumina, implique des interférences spécifiques aux séquences lors du processus d'élongation pendant le séquençage générés par des schémas particuliers du template produisant des repliements du brin d'ADN et altérant l'affinité des enzymes (Nakamura et al. 2011).

Analyse statistique des données d'expression

Pour réaliser nos analyses, nous nous sommes inspirés de la méthodologie employée en RNA-Seq pour l'étude de Différence d'Expression (DE). Il est important de noter qu'une différence importante existe avec notre approche, car nous ne nous intéressons pas à une DE pour un même gène dans différentes conditions mais plutôt à la *DE entre deux gènes pour lesquels ont*

FIGURE 3 – **Visualisation du mapping de l'opéron malEFG sur IGV.** Les premières pistes représentent les positions et orientations des gènes, des opérons, la présence de promoteurs et de terminateurs, ainsi que la position des BIME et des REP qui composent les BIME. Les 2 pistes suivantes affichent la couverture des fichiers BAM fusionnés de la région visualisée (histogramme gris) et l'alignement des reads (flèches pleines rouges et bleues). La couleur bleue sur cette piste indique un alignement en anti-sens et la couleur rouge, un alignement en sens.

pourrait supposer un profil d'expression similaire dans une même condition, comme c'est le cas dans les opérons par exemple. L'analyse statistique de recherche de changement d'expression liée à la présence de BIME a été menée sur le logiciel R avec 3 méthodes différentes.

Création de la table de comptages et normalisation

Afin d'obtenir des résultats de comptage par transcrit et ainsi estimer l'expression, nous avons utilisé le package Bioconductor easyRNASeq (Delhomme et al. 2012). Les annotations du génome d'E. coli touchant aux transcrits sont extraites à partir du fichier `GFF` et stockées sous forme d'une base de données. Cela a nécessité une manipulation préalable de ce fichier, en effet les opérons récupérés sur RegulonDB sont composés à la fois de gènes dont les transcrits sont annotés ARNm, ARNt et ARNr. Seul les gènes dont le transcrit est annoté ARNm est pris en compte par le package easyRNASeq, donc pour ne pas avoir d'erreur dans l'analyse, nous avons transformé les annotations ARNt et ARNr en ARNm. La liste des transcrits par gènes est ensuite extraites pour un total de 4605 éléments. La couverture par transcrit est ensuite calculée pour chaque fichier BAM, le résultat est obtenu en réalisant l'union des positions

extraites de la liste des transcrits et des positions des reads extraites des fichiers BAM qui auront été préalablement transformées au format Genomic Ranges (GRanges) (Lawrence et al. 2013). Une table de comptage est alors produite, les transcrits figurant en ligne et les fichiers BAM en colonnes (Figure 4).

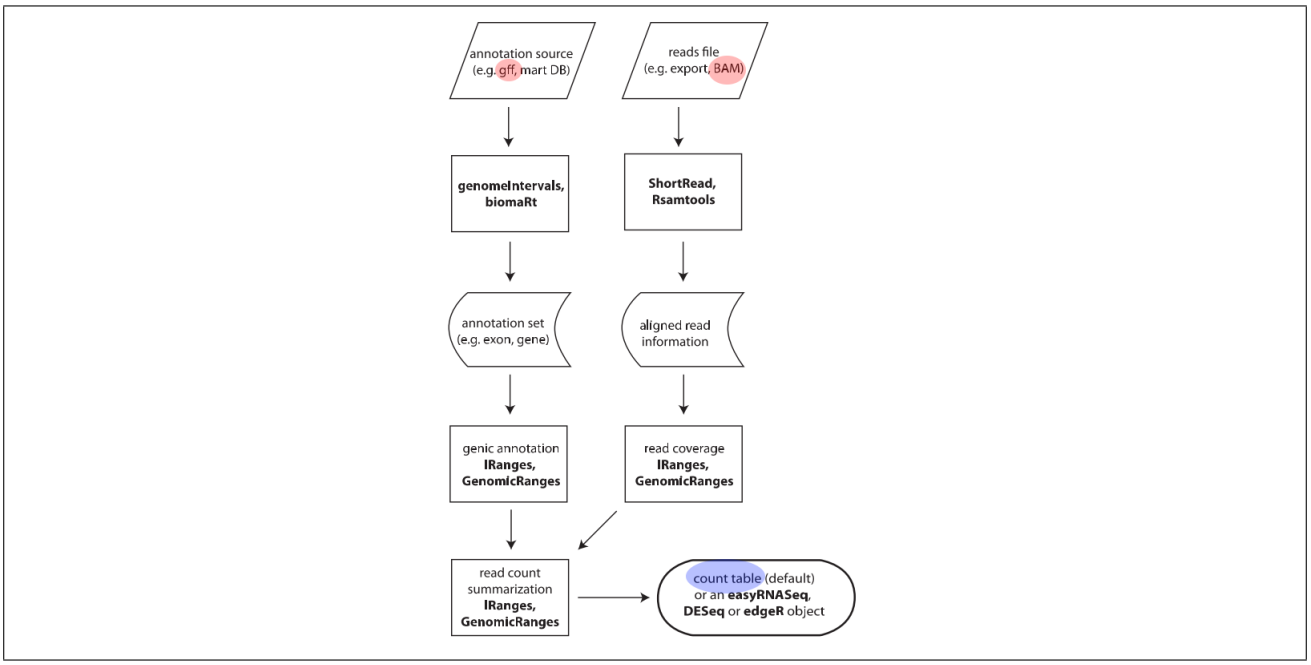


FIGURE 4 – **easyRNASeq : création d'une table de comptage.** Processus de création de la table de comptage grâce à l'union des intervalles génomiques. Les formats d'entrées de notre analyse sont surlignés en rouge, le format de sortie en bleu (Delhomme et al. 2012).

Les résultats de comptages doivent ensuite être normalisés afin de permettre la comparaison de l'expression des gènes et des régions génomiques d'intérêt. Notre choix s'est porté sur la méthode du Reads Per Kilobase per Million mapped reads (RPKM) (Mortazavi et al. 2008) :

$$RPKM = Nb.\ reads\ transcrit\ *\ \frac{1000\ bases\ *\ 10^6}{Nb.\ total\ reads\ *\ Taille\ du\ transcrit}$$

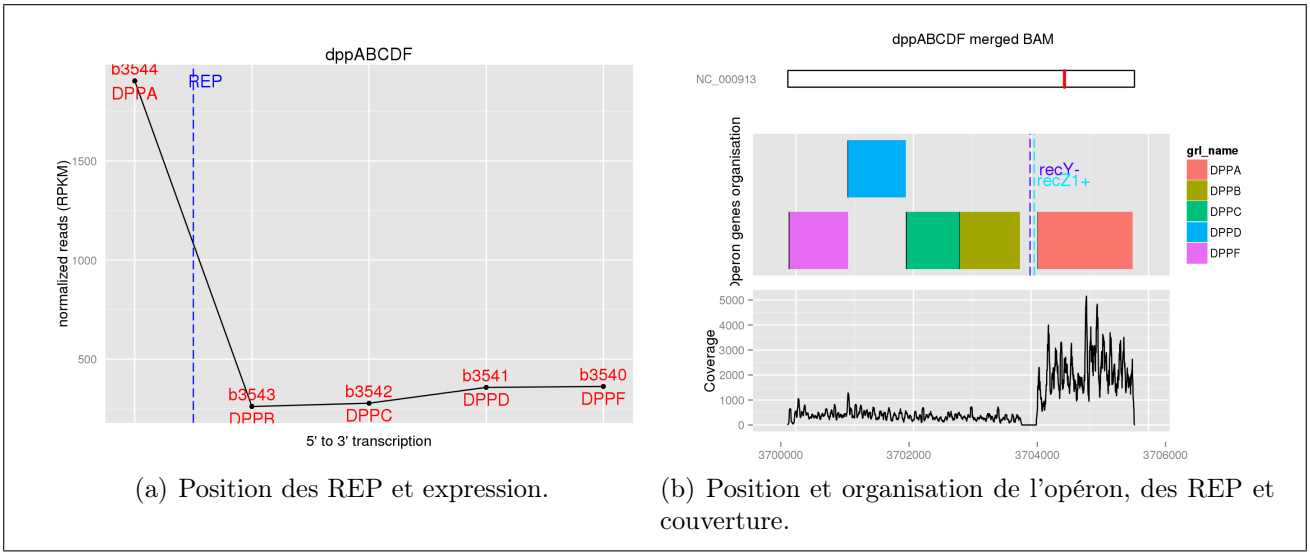
Le RPKM reflète la concentration molaire du transcrit en normalisant par la longueur du brin d'ARN et le nombre de reads de la bibliothèque. Cette normalisation est soumise à critique à juste titre (Dillies et al. 2013) car elle induit un biais de lors d'une analyse de DE dans le cas de gènes fortement exprimés dans un condition par rapport aux autres. Comme nous ne nous situons pas dans le cadre d'une analyse différentielle sur plusieurs conditions, mais que nous comparons des réplicats d'une même condition, nous pouvons appliquer cette normalisation.

Une analyse en composante principale est réalisée sur ces données normalisées afin de vérifier l'homogénéité des réplicats et les taux d'expressions moyens des gènes et des BIME sont visualisés sous forme de BoxPlot.

Différence d'expression dans les opérons contenant des BIME

L'hypothèse privilégiée ici est que les gènes appartenant à un opéron vont être exprimés à un taux similaire, la question qui se pose est de savoir si la présence d'une BIME entre 2 gènes d'un opéron va avoir un impact sur la transcription d'un des gènes. Dans ce cadre, les opérons contenant des BIME sont sélectionnés et l'expression des 2 gènes de l'opéron entourant la BIME recueillie si au moins un des deux gènes a une couverture > 10. Si l'expérience contient au moins 5 réplicats, un test non paramétrique de rangs de Wilcoxon est effectué dont l'hypothèse nulle est qu'il n'existe pas de différence d'expression entre les 2 gènes. La p-value significative étant fixée à 0.01. Dans le cas où le nombre de réplicats est inférieur à 5, un test de Student est réalisé avec la même p-value significative.

Pour les gènes dont le test est significatif, deux représentations graphiques sont générées (Figure 5). La 1ère est un schéma décrivant les taux d'expression des gènes de l'opéron ainsi que la position des REP formant la BIME (5(a)). La 2nde est une représentation de la couverture sur l'opéron par rapport à l'organisation génomique de celui ci ainsi que la catégorisation des REP composant la BIME (5(b)). Deux fichiers au format `CSV` sont créés, l'un pour les gènes où nous observons un différence d'expression significative, le second pour les autres cas.



(a) Position des REP et expression.

(b) Position et organisation de l'opéron, des REP et couverture.

FIGURE 5 – **Résultats de l'étude de l'expression des gènes dans les opérons contenant des BIME. (a)** Le taux d'expression de chaque gène de l'opéron est représenté en ordonnée, l'organisation des gènes de l'opéron est schématisée sur l'axe des abscisses dans le sens 5' $\mapsto$ 3'. La position de la ou les REP composant la BIME est schématisée par la ligne blue verticale. **(b)** La position de l'opéron sur le génome est indiquée par la barre rouge sur l'idéogramme de la partie supérieure. La partie médiane représente l'organisation des gènes de l'opéron dans le sens 5' $\mapsto$ 3' ainsi que le positionnement et la classe des REP composant la BIME. La partie inférieure montre la couverture par rapport à l'organisation de la partie médiane.

Analyse par corrélation de profils d'expression

Nous avons créé un outil réalisant un test de corrélation entre les profils d'expression des régions contenant des BIME et un profil modèle de changement d'expression en nous inspirant d'une technique mise au point pour la prédiction d'opérons dans les génomes bactériens (Fortino et al. 2014). Pour cela, il a d'abord été nécessaire de délimiter nos régions d'intérêt. Celles-ci se modélisent par la présence du 1er gène, de la 1ère région inter-génique, de la BIME, de la 2nde région inter-génique et du 2nd gène. Une fois ces régions extraites, le calcul de la couverture base par base a été réalisé à l'aide des BEDtools :

```
# Couverture base par base.
bedtools coverage -abam merged.bam -b regionOfInterest.bed \
-d > unsorted_cov_perBase.bed

# Tri en fonction de la position genomique
# puis de la position des bases dans chaque transcrit
sort -k2 -k7 -n unsorted_cov_perBase.bed | uniq > \
cov_perBase_strandToFix.bed

# Remplacement des '.' par des '*' dans la colonne
# des brins pour l'utilisation sous R
awk 'BEGIN{OFS = "\t"} {gsub(/\./,"*",$6); print }' \
cov_perBase_strandToFix.bed > cov_perBase.bed
```

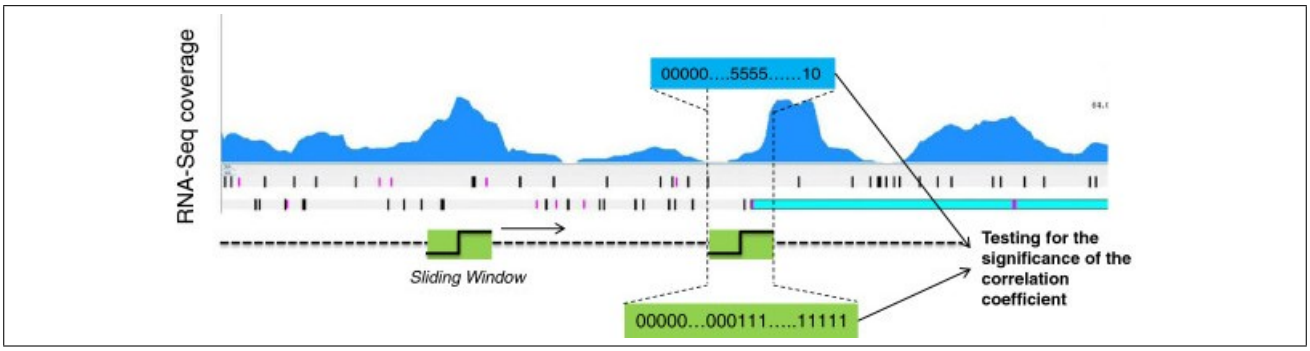


FIGURE 6 – **Recherche de corrélation sur des profils d'expression.** La fenêtre glissante (en vert) parcourt la région d'intérêt et pour chaque déplacement une corrélation est calculée entre le vecteur du profil d'expression obtenu par RNAseq (en bleu) et celui simulé par le vecteur de 0 et de 1 (en vert) (Fortino et al. 2014).

Le cœur de cette méthode consiste déplacer une fenêtre glissante parcourant base à base la région d'intérêt définie ci-dessus en réalisant des tests de corrélation entre le profil d'expression réel et un profil d'expression simulé par un vecteur contenant un nombre égal de 0 et de 1 (si l'on cherche une croissance d'expression en sens ou une décroissance d'expression en anti-sens,

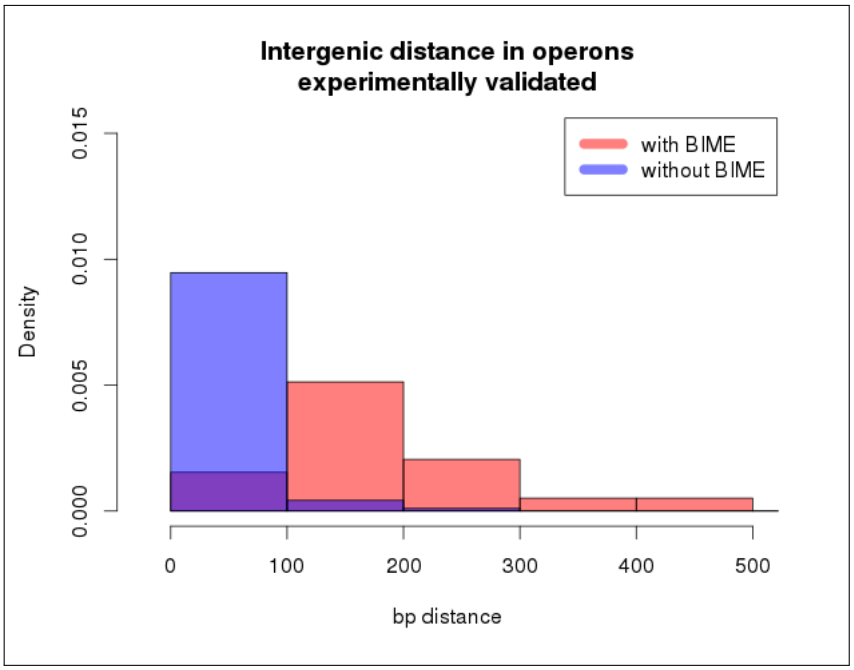


FIGURE 7 – **Distances inter-géniques dans les opérons avec et sans BIME.** En bleu absence de BIME dans la région inter-génique, en rouge présence de BIME. Nous observons une croissance de 2 à 3 fois de la taille de cette région lors de la présence de BIME.

e.g : 000111), ou de 1 et de 0 (dans le cas d'une décroissance en sens et d'une croissance en anti-sens, e.g : 111000). Les moyennes d'expression sont récupérées sur les parties gauche et droite de la fenêtre pour dans un premier temps filtrer les données. Le Log_2 du rapport de ces moyennes doit être supérieur à un seuil, si ce 1er filtre est passé, la corrélation entre le profil réel et celui simulé est calculée et doit être supérieure à un seuil avec une p-valeur significative (Figure 6). Les seuils que nous avons fixés sont les suivants :

- fenêtre glissante de 300 bases
- au moins un des deux gènes possède une couverture > 10
- $log_2(\frac{\overline{couverture\ droite} + 1}{\overline{couverture\ gauche} + 1}) \geq 1$ pour un profil $\_\_\_|\hat{}\hat{}\hat{}$
- $log_2(\frac{\overline{couverture\ gauche} + 1}{\overline{couverture\ droite} + 1}) \geq 1$ pour un profil $\hat{}\hat{}\hat{}|\_\_\_$
- une corrélation > 0.7
- une p-valeur du test de corrélation $< 10^7$

Le choix de la taille de la fenêtre glissante a été motivé par la taille des régions inter-géniques contenant des BIME dans les opérons . L'idée étant de rechercher le changement d'expression en ayant des informations sur la couverture moyenne des gènes entourant la BIME. La Figure 7 nous indique que les majorité des régions contenant des REP/BIME se situent entre 200 et 300 pb. Comme la BIME se situe dans la majorité des cas proche d'un des deux gènes nous avons opté pour une fenêtre de taille 300 pb.

Sur un ensemble de positions consécutives dont les corrélations sont significatives et sont situées sur l'espace génomique de la BIME (étendu de 40 pb de chaque côté), celle dont la cor-

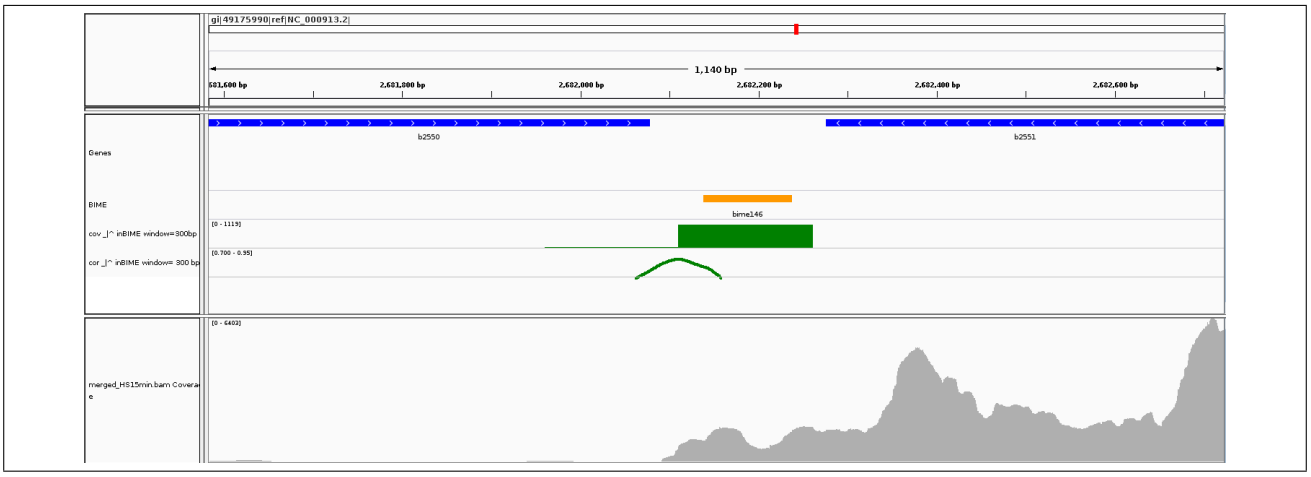


FIGURE 8 – **Visualisation des changements d'expression obtenus par la méthode de corrélation des profils.** La position de la BIME est représentée en orange, l'histogramme à 2 colonnes en vert montre les profils d'expression moyens de chaque moitié de la fenêtre et la courbe en points verts indique l'évolution de la corrélation sur cette zone. Dans cet exemple, ce changement d'expression peut se traduire par une augmentation en sens ou une diminution en anti-sens.

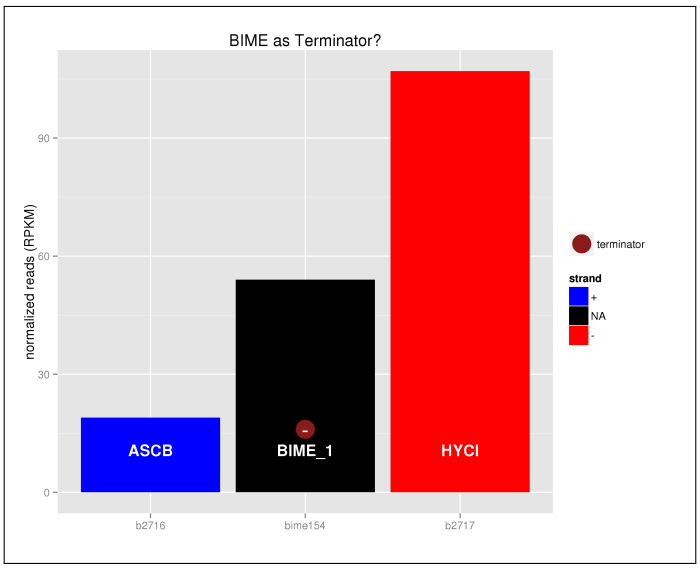


FIGURE 9 – **Résultat de corrélation de profils.** Les niveaux d'expression des gènes encadrant la BIME et de cette dernière sont représentés par les histogrammes. La couleur de l'histogramme indique le sens de transcription de l'élément, bleu pour le brin sens, rouge pour l'anti-sens et noir lorsque aucun brin est défini. La présence d'éléments de régulation dans la région inter-génique est représentée par des ronds de couleur verte pour les promoteurs et rouge pour les terminateurs avec un symbole '+' ou '-' pour indiquer le brin de cet élément. La représentation est schématique et ne donne pas d'information sur la position exacte de ces éléments.

rélation est la plus élevée sera définie comme position de changement d'expression. Deux types de fichiers sont générés au format `bedgraph` pour une visualisation sur un Genome Browser, le premier sous forme d'histogramme représentant les couvertures moyennes des deux parties de la fenêtre, le second représentant l'évolution de la corrélation sur la zone (Figure 8). Une visualisation des niveaux d'expression des 2 gènes et de la BIME est également générée sous forme d'histogrammes avec les informations de sens et du type de la BIME (Figure 9). Finalement, un fichier au format `CSV` recueille toutes les informations de l'analyse.

Analyse par segmentation

A la différence de la méthode précédente, la segmentation ne requière pas l'utilisation de profils. La méthode de segmentation que nous utilisons est issue du package R Segmentor3IsBack (Cleynen et al. 2014). Ce package a l'avantage d'avoir été développé pour s'adapter à la loi Négative Binomiale (NB) qui est une loi statistique appropriée aux données d'expression de RNAseq pour lesquelles la variance n'est pas stabilisée par rapport à la moyenne, phénomène communément appelé overdispersion. Le but de cette méthode est de rechercher, sur une zone d'intérêt, des points de changements abrupts dans la couverture en utilisant l'algorithme Pruned Dynamic Programming (PDP) (Rigaill 2010). La segmentation se fonde sur le partitionnement d'un signal de n points, la couverture de notre région d'intérêt, compris dans l'ensemble $\{y_t\}_{t=1,\dots,n}$, suivant une distribution NB, en K segments, tel que :

$$Y_t \sim NB(\theta_r, \phi) \quad \text{si } t \in r \quad \text{et } r \in m$$

où m est une partition de $[1, n]$ en r segments, le paramètre θ_r est la probabilité de succès associée au segment r et ϕ le paramètre de dispersion qui est commun à tous les segments. L'objectif étant d'estimer le point de cassure ou la position des segments et le paramètre θ_r résultant tous les deux de la segmentation. $M_{K,n}$ est alors l'ensemble des partitions possibles avec K le nombre minimal de partitions demandé et n la taille de notre région. L'algorithme tente de choisir la partition $M_{K,n}$ avec la perte γ minimale. Cette perte est calculée par la négative log-likelihood du modèle. La fonction de calcul du coût est définie comme telle :

$$c(r, \theta) = \sum_{i \in r} \gamma(y_i, \theta)$$

et dont le coût optimal sera :

$$c(r) = min_\theta\{c(r, \theta)\}$$

cela permettant de récupérer la segmentation optimale $M_{K,n}$ et son coût $C_{K,n}$. L'algorithme itératif PDP intervient ensuite et est basé sur la minimisation de la fonction de coût $C_{k,t}$ décomposée de la façon suivante :

$$C_{k,t} = \min_{\{k-1<\tau<t\}} \{C_{k-1,\tau} + \min_\theta [c([\tau + 1, t], \theta)]\} \tag{1}$$

où θ est le paramètre de coût du dernier segment directement lié au calcul de perte γ. La spécificité de cet algorithme est qu'il s'appuie sur la comparaison de candidats pour la position du dernier point de cassure notée τ à travers les permutations des minimisations de (1) et avec

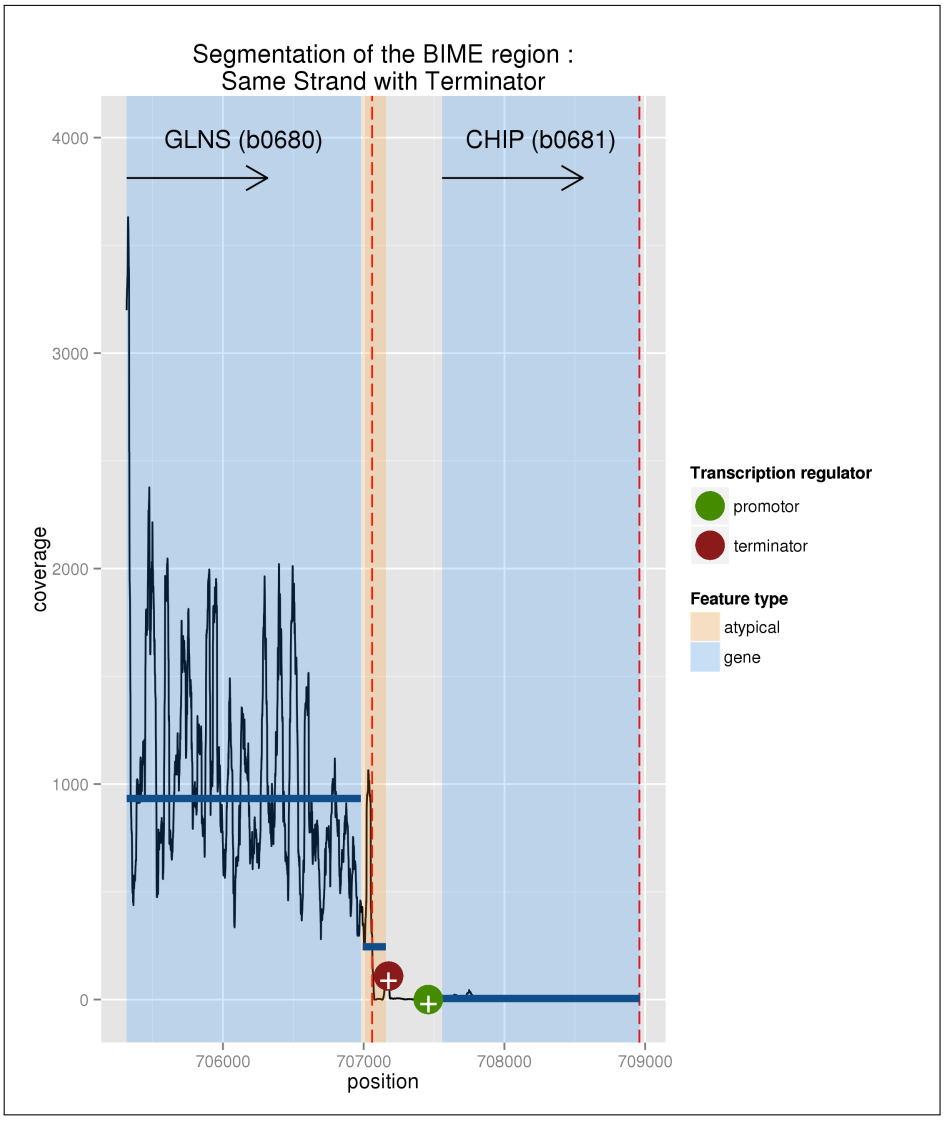


FIGURE 10 – **Résultat de segmentation pour K_{max}=4.** Les gènes sont symbolisés par les zones bleues, leur sens de transcription par les flèches noires. La BIME est représentée par la zone orange et sa classe est précisée dans la légende (Genomic element). Les promoteurs sont représentés par des points rouges et les terminateurs par des points verts, leur sens affichés par les symboles '+' ou '-' sur ces points. La courbe noire représente la couverture sur la région et les barres bleues horizontales indiquent les couvertures moyennes des éléments génomiques. Les points de cassure dans la couverture, déterminés par la segmentation, sont matérialisés par des lignes rouges verticales en pointillés. Ici, 2 segments sont représentés avec un point de cassure situé sur la BIME.

l'introduction de la fonction :

$$H_{k,t}(\theta) = \underset{\{k-1<\tau<t\}}{min}\{C_{k-1,\tau} + c([\tau+1,t],\theta)\}$$

qui est le coût de la meilleur partition en k régions jusqu'à t, le paramètre du dernier segment étant θ. $C_{k,t}$ est alors obtenu comme le $min_\theta\{H_{k,t}(\theta)\}$. Pour chaque itération k, l'algorithme travaille sur une liste de candidats pour les derniers points de cassure. Pour chaque élément τ et chaque valeur t, il met à jour un ensemble $S^\tau_{k,t}$ contenant les paramètres θ pour lequel ce candidat est optimal. Si cet ensemble $S^\tau_{k,t}$ est vide, le candidat est supprimé autorisant un élagage et une diminution de la complexité de l'algorithme.

Au final, l'utilisation de ce package produit un découpage de la région d'intérêt en K segments, K étant fixé par l'utilisateur, dont les limites sont définies par les positions de leurs

points de cassures. Nous avons fixé le paramètre K à 4 segments maximum donc 3 points de cassure de façon à vérifier la présence éventuelle de plusieurs de ces points sur la région inter-génique. Dans notre étude, nous nous intéressons aux positions de ces points de cassures pour nos régions d'intérêt, lorsqu'au moins un des deux gènes possède une couverture supérieure à 10. Les résultats sont croisés avec la présence de promoteurs ou de terminateurs dans la région inter-génique et si les deux gènes appartiennent à un opéron, un test de significativité de la différence d'expression est réalisé avec la même méthodologie que pour la partie Différence d'expression dans les opérons contenant des BIME.

L'analyse renvoie une représentation graphique de la couverture de ces régions avec les positionnements des gènes et de la BIME, ainsi que des promoteurs et terminateurs éventuels. Une classification est faite en fonction du sens des gènes et de l'impact des régulateurs de transcription sur la couverture (Figure 10). Deux fichiers au format `CSV` sont générés pour recueillir les informations de la segmentation, le 1$^{\text{er}}$ pour les gènes sur le même brin, le 2$^{\text{nd}}$ pour les gènes sur des brins opposés.

États ancêtres et structures secondaires

Résultats

Discussion

Glossaire

alignement chimérique L'alignement d'un read ne peut pas être représenté comme un alignement linéaire. Un alignement chimérique est représenté comme un ensemble d'alignements, par exemple lorsqu'une partie d'un read est mappé à un locus du génome et la suite à un autre locus. 6

BAM Binary Alignment Map format. 6

BED Browser Extensible Data. 7

BIME Bacterial Interspesed Mosaic Element. 1

couverture Appelé également profondeur de séquençage, correspond au nombre de reads alignés sur une région génomique. Dans le cas du RNAseq, la couverture fournit une information sur le taux d'expression d'un élément génomique. 7

DE Différence d'Expression. 7

Genomic Ranges Format de stockage d'informations pour les éléments génomiques sous R. L'information minimale requise est le chromosome, les positions de départ et de fin, le sens du brin. Ces champs peuvent être suivis de méta-datas où d'autres informations libres peuvent être enregistrées. 9

GFF General Feature Format. 7

loi Négative Binomiale Si une expérience consiste en une série de tirages indépendants avec une probabilité de succès p et une probabilité d'échec complémentaire, celle-ci se poursuit jusqu'à l'obtention de n succès, la variable aléatoire représentant le nombre d'échecs avant l'obtention des n succès suit une loi négative binomiale. Les paramètres de cette loi sont n le nombre de succès attendus et p la probabilité d'un succès. 14

NGS Next Generation Sequencing. 5

Paired-end Technique de séquençage haut débit consistant à réaliser les amplifications d'un fragment d'ADN en marquant l'extrémité 5' par un tag n° 1 et l'extrémité 3' par un tag n° 2. La distance entre les 2 tags est connue et fixe (négative ou jusqu'à 500 pb). Ceci permet lors de l'assemblage, de séquences de 35 pb par exemple, d'associer le read 1 et le read 2 grâce à la distance séparant les 2 et cela même si la séquence intermédiaire est inconnue. Si la distance est négative, il est possible d'obtenir des reads chevauchants de longueur plus importante que les 35 pb. 5

PDP Pruned Dynamic Programming. 14

reads Séquence nucléotidique issue d'un séquençage NGS. 5

REP Repeated Extragenic Palindrome. 1

RPKM Reads Per Kilobase per Million mapped reads. 9

SAM Sequence Alignment Map format. 6

SRA Sequence Read Archive. 5

Bibliographie

M. Aguena, G. M. Ferreira, and B. Spira. Stability of the pstS transcript of Escherichia coli. *Archives of Microbiology*, 191 :105–112, 2009. ISSN 03028933. doi : 10.1007/s00203-008-0433-z.

S. Bachellier, W. Saurin, D. Perrin, M. Hofnung, and E. Gilson. Structural and functional diversity among bacterial interspersed mosaic elements (BIMEs). *Molecular Microbiology*, 12 :61–70, 1994. ISSN 0950382X. doi : 10.1111/j.1365-2958.1994.tb00995.x.

S. Bachellier, J. M. Clément, M. Hofnung, and E. Gilson. Bacterial interspersed mosaic elements (BIMEs) are a major source of sequence polymorphism in Escherichia coli intergenic regions including specific associations with a new insertion sequence. *Genetics*, 145(3) :551–62, Mar. 1997. ISSN 0016-6731. URL /pmc/articles/PMC1207841/?report=abstract.

F. Boccard and P. Prentki. Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3' end of transcription units. *The EMBO journal*, 12 (13) :5019–27, Dec. 1993. ISSN 0261-4189. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=413762&tool=pmcentrez&rendertype=abstract.

J. Bosc. Etude de la dynamique des éléments palindromique répétées (REP) chez l ' espèce Escherichia coli par une méthode de reconstruction des états ancêtres . Technical report, 2014.

S. Choi, S. Ohta, and E. Ohtsubo. A novel IS element, IS621, of the IS110/IS492 family transposes to a specific site in repetitive extragenic palindromic sequences in Escherichia coli. *Journal of bacteriology*, 185(16) :4891–900, Aug. 2003. ISSN 0021-9193. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=166490&tool=pmcentrez&rendertype=abstract.

J. M. Clément, C. Wilde, S. Bachellier, P. Lambert, and M. Hofnung. IS1397 is active for transposition into the chromosome of Escherichia coli K-12 and inserts specifically into palindromic units of bacterial interspersed mosaic elements. *Journal of Bacteriology*, 181(22) :6929–6936, 1999. ISSN 00219193.

A. Cleynen, M. Koskas, E. Lebarbier, G. Rigaill, and S. Robin. Segmentor3IsBack : an R package for the fast and exact segmentation of Seq-data. *Algorithms for molecular biology : AMB*, 9(1) :6, Jan. 2014. ISSN 1748-7188. doi : 10.1186/1748-7188-9-6. URL http://www.almob.org/content/9/1/6http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3977952&tool=pmcentrez&rendertype=abstract.

N. Delhomme, I. Padioleau, E. E. Furlong, and L. M. Steinmetz. easyRNASeq : A bioconductor package for processing RNA-Seq data. *Bioinformatics*, 28(19) :2532–2533, 2012. ISSN 13674803. doi : 10.1093/bioinformatics/bts477.

M. A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, N. S. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, and F. Jaffrézic. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6) :671–683, Nov. 2013. ISSN 14675463. doi : 10.1093/bib/bbs046. URL http://www.ncbi.nlm.nih.gov/pubmed/22988256.

O. Espéli and F. Boccard. In vivo cleavage of Escherichia coli BIME-2 repeats by DNA gyrase : genetic characterization of the target and identification of the cut site. *Molecular microbiology*, 26 :767–777, 1997. ISSN 0950-382X.

O. Espéli, L. Moulin, and F. Boccard. Transcription attenuation associated with bacterial repetitive extragenic BIME elements. *Journal of molecular biology*, 314(3) :375–86, Nov. 2001. ISSN 0022-2836. doi : 10.1006/jmbi.2001.5150. URL http://www.sciencedirect.com/science/article/pii/S0022283601951502.

V. Fortino, O.-P. Smolander, P. Auvinen, R. Tagliaferri, and D. Greco. Transcriptome dynamics-based operon prediction in prokaryotes. *BMC bioinformatics*, 15 :145, Jan. 2014. ISSN 1471-2105. doi : 10.1186/1471-2105-15-145. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4235196&tool=pmcentrez&rendertype=abstract.

E. Gilson, J. Rousset, J. Clément, and M. Hofnung. A subfamily of E. coli palindromic units implicated in transcription termination ? *Annales de l'Institut Pasteur / Microbiologie*, 137 (1) :259–270, July 1986. ISSN 07692609. doi : 10.1016/S0769-2609(86)80116-8. URL http://www.sciencedirect.com/science/article/pii/S0769260986801168.

E. Gilson, D. Perrin, and M. Hofnung. DNA polymerase I and a protein complex bind specifically to E. coli palindromic unit highly repetitive DNA : implications for bacterial chromosome organization. *Nucleic acids research*, 18(13) :3941–3952, 1990. ISSN 0305-1048.

E. Gilson, W. Saurin, D. Perrin, S. Bachellier, and M. Hofnung. Palindromic units are part of a new bacterial interspersed mosaic element (BIME). *Nucleic acids research*, 19(7) :1375–1383, 1991. ISSN 03051048.

N. Goosen, P. V. D. Putte, and P. Van De Putte. The regulation of transcription initiation by integration host factor. *Molecular Microbiology*, 16 :1–7, 1995. ISSN 00219258. doi : 10.1111/j.1365-2958.1995.tb02386.x. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7961996.

C. F. Higgins, G. F.-L. Ames, W. M. Barnes, J. M. Clement, and M. Hofnung. A novel intercistronic regulatory element of prokaryotic operons. *Nature*, 298(5876) :760–762, Aug. 1982. ISSN 0028-0836. doi : 10.1038/298760a0.

V. Khemici and A. J. Carpousis. The RNA degradosome and poly(A) polymerase of Escherichia coli are required in vivo for the degradation of small mRNA decay intermediates containing REP-stabilizers. *Molecular Microbiology*, 51 :777–790, 2004. ISSN 0950382X. doi : 10.1046/j.1365-2958.2003.03862.x.

E. Kofoid, U. Bergthorsson, E. S. Slechta, and J. R. Roth. Formation of an F′ plasmid by recombination between imperfectly repeated chromosomal Rep sequences : A closer look at an old friend (F′128 pro lac). *Journal of Bacteriology*, 185(2) :660–663, 2003. ISSN 00219193. doi : 10.1128/JB.185.2.660-663.2003.

R. a. LaCroix, T. E. Sandberg, E. J. O'Brien, J. Utrilla, a. Ebrahim, G. I. Guzman, R. Szubin, B. O. Palsson, and a. M. Feist. Use of Adaptive Laboratory Evolution To Discover Key Mutations Enabling Rapid Growth of Escherichia coli K-12 MG1655 on Glucose Minimal Medium. *Applied and Environmental Microbiology*, 81(1) :17–30, 2014. ISSN 0099-2240. doi : 10.1128/AEM.02246-14. URL http://aem.asm.org/cgi/doi/10.1128/AEM.02246-14.

M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8) :1–10, 2013. ISSN 1553734X. doi : 10.1371/journal.pcbi.1003118.

J. Levin, M. Yassour, and X. Adiconis. Comprehensive comparative analysis of strand specific RNA sequencing methods. *... methods*, 7(9) :709–715, 2010. doi : 10.1038/nmeth.1491. Comprehensive. URL http://www.nature.com/nmeth/journal/v7/n9/abs/nmeth.1491.html.

S. Li, X. Dong, and Z. Su. Directional RNA-seq reveals highly complex condition-dependent transcriptomes in E. coli K12 through accurate full-length transcripts assembling. *BMC genomics*, 14(1) :520, 2013. ISSN 1471-2164. doi : 10.1186/1471-2164-14-520. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3734233&tool=pmcentrez&rendertype=abstract.

A. Mortazavi, B. a. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7) :621–628, 2008. ISSN 1548-7091. doi : 10.1038/nmeth.1226.

K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, M. Altaf-Ul-Amin, N. Ogasawara, and S. Kanaya. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13), 2011. ISSN 03051048. doi : 10.1093/nar/gkr344.

S. F. Newbury, N. H. Smith, E. C. Robinson, I. D. Hiles, and C. F. Higgins. Stabilization of translationally active mRNA by prokaryotic REP sequences. *Cell*, 48 :297–310, 1987. ISSN 00928674. doi : 10.1016/0092-8674(87)90433-8.

G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *eprint arXiv :1004.0887*, page 9, 2010. URL http://arxiv.org/abs/1004.0887.

J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature biotechnology*, 29(1) :24–6, Jan. 2011. ISSN 1546-1696. doi : 10.1038/nbt.1754. URL http://dx.doi.org/10.1038/nbt.1754.

M. J. Stern, E. Prossnitz, and G. F. Ames. Role of the intercistronic region in post-transcriptional control of gene expression in the histidine transport operon of Salmonella typhimurium : involvement of REP sequences. *Molecular microbiology*, 2 :141–152, 1988. ISSN 0950382X.

H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov. Integrative Genomics Viewer (IGV) : high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2) :178–92, Mar. 2013. ISSN 1477-4054. doi : 10.1093/bib/bbs017. URL http://bib.oxfordjournals.org/content/14/2/178.full?keytype=ref&%2520ijkey=qTgjFwbRBAzRZWC.

R. Tobes and E. Pareja. Repetitive extragenic palindromic sequences in the Pseudomonas syringae pv. tomato DC3000 genome : extragenic signals for genome reannotation. *Research in microbiology*, 156(3) :424–33, Apr. 2005. ISSN 0923-2508. doi : 10.1016/j.resmic.2004.10.014. URL http://www.sciencedirect.com/science/article/pii/S092325080400289X.

B. Ton-Hoang, P. Siguier, Y. Quentin, S. Onillon, B. Marty, G. Fichant, and M. Chandler. Structuring the bacterial genome : Y1-transposases associated with REP-BIME sequences. *Nucleic acids research*, 40(8) :3596–609, Apr. 2012. ISSN 1362-4962. doi : 10.1093/nar/gkr1198. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3333891&tool=pmcentrez&rendertype=abstract.

M. Weyder. Étude de la dynamique de la prolifération des éléments REP chez Escherichia et Shigella par une approche bioinformatique. Technical report, 2013.