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 appelés 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 des génomes d'*E. coli* et *Shigella* 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 de transcription ou comme site de reconnaissance des enzymes impliquées dans les mécanismes de la transcription.

Caractéristiques des REP et organisations en BIME

Chez *E. coli*, la taille des REP est d'environ 40 nucléotides, la classification Y, Z^1, Z^2 est basée sur leur séquence primaire. 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 séquences consensus des différentes classes 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) permettant la reconnaissance par $tnpA_{REP}$.

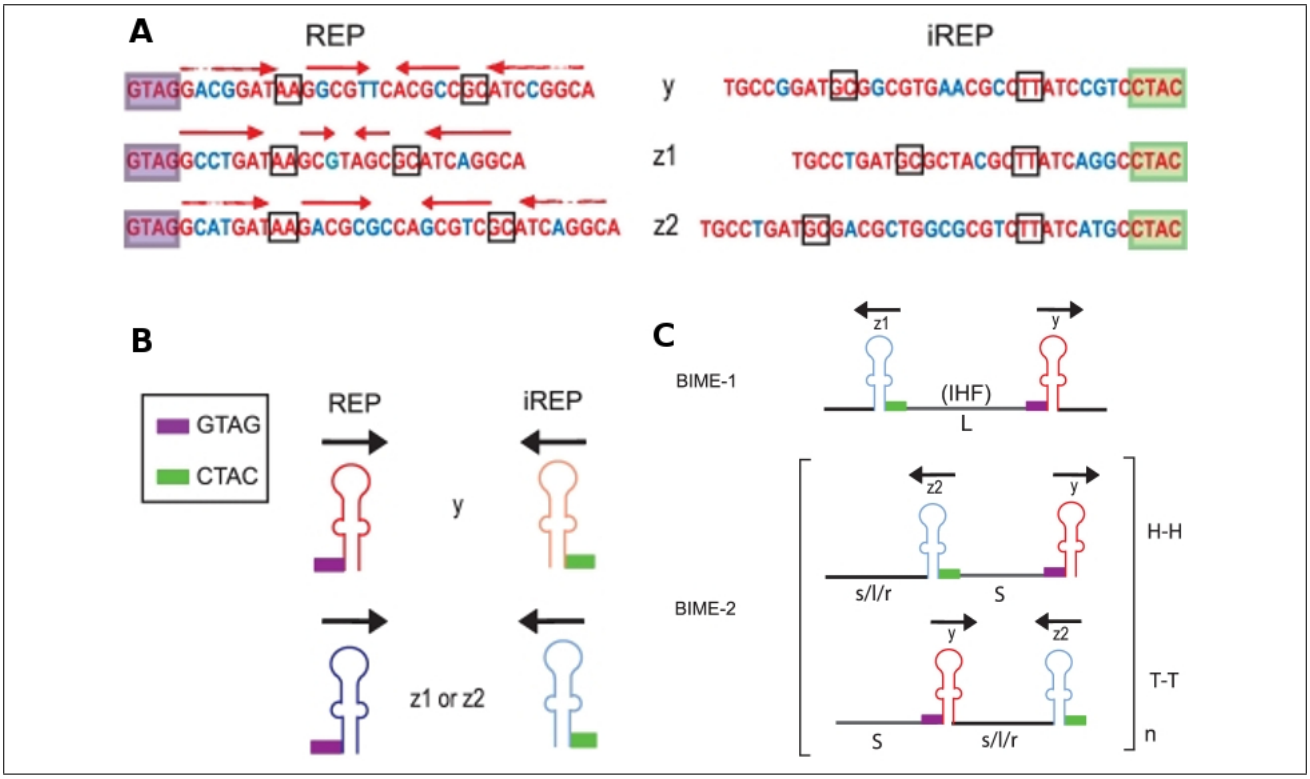


FIGURE 1 – **REP et BIME chez *Escherichia coli*. (A)** Séquences consensus Y, Z^1 et Z^2 des REP. Le tétra-nucléotide conservé GTAC est encadré en violet, le complémentaire conservé CTAC est encadré en vert, les flèches rouges situent les zones d'appariement de la tige et les positions encadrées en 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).

Une classification a été adoptée comportant 3 classes (Bachellier et al. 1997), les BIME-1 composées de REP Z^1 et Y apparaissant en paires uniques. Les BIME-2 constituées de Z^2 et de Y, apparaissant en copies multiples de cette paire. 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. 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 à quel endroit cette figure ? (Figure 2).

Propriétés associées aux REP

La littérature décrit de nombreuses fonctions hypothétiques associées aux REP au niveau structurel du génome, au niveau de l'ADN et au niveau de l'ARN. Sur un plan structurel, les REP ont été décrites comme jouant un rôle dans les événements de **recombinaisons homologues** (Kofoid et al. 2003) et 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).

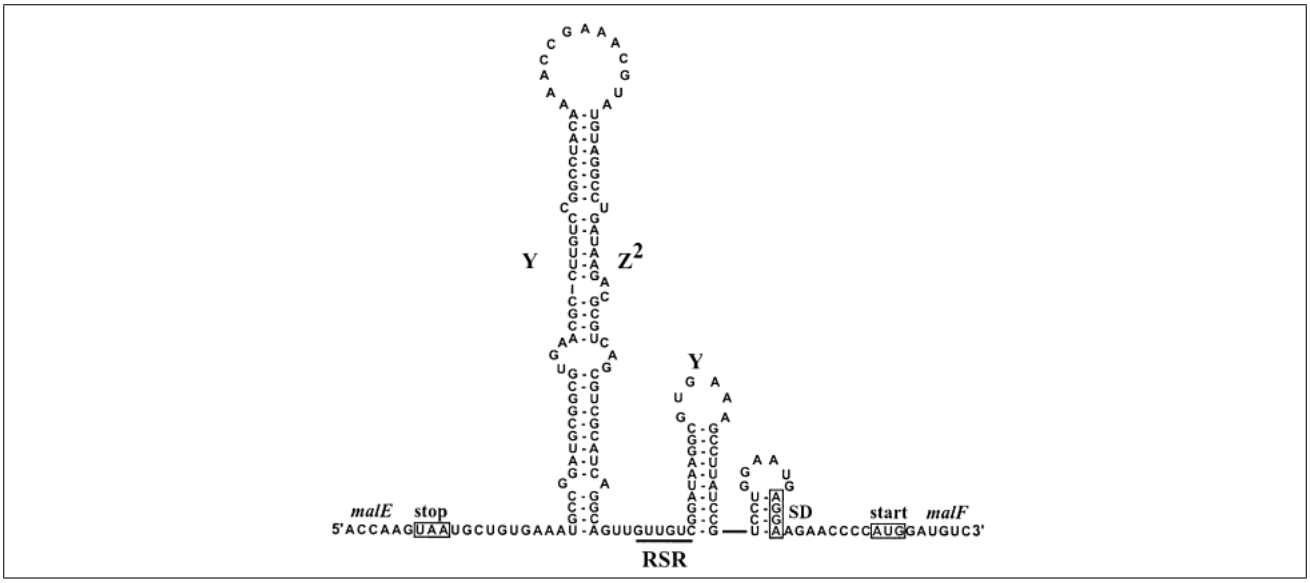


FIGURE 2 – **Structure ARN des REP au sein de l'opéron *malEFG*.** Y, Z^2 et Y indiquent la séquence des REP dans l'espace inter-génique de malE-malF. Bien que Y et Z^2 puissent former des structures tige-boucles par elles mêmes, elles s'apparient ensemble pour former une région étendue en grande partie 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).

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). Au plan de l'ARN, 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).

ARN messagers chez *E. coli*

Stabilité des ARNm

La dégradation des ARNm chez *E.coli* est réalisée par l'intervention du dégradosome. Il s'agit d'un complexe multi-enzymatique composé de quatre protéines majeures, la RNase E, la PNPase, la RhlB et l'Enolase (Figure 3).

Un élément clé dans la dégradation du transcrit chez *E. coli* est que celle-ci débute toujours par un clivage réalisé par une endoribonucléase (RNase E). Une fois clivés, les transcrits sont complètement dégradés par des exoribonucléases (RNase R, RNase II ou PNPase) dégradant l'ARNm par l'extrémité 3' et par des oligoribonucléases grâce à la coopération de nombreuses enzymes telles que la poly(A) polymérase (PAP) et les RNA hélicases qui facilitent l'accès aux fragments d'ARN. La RNase E possède une affinité pour les substrats possédant une ex-

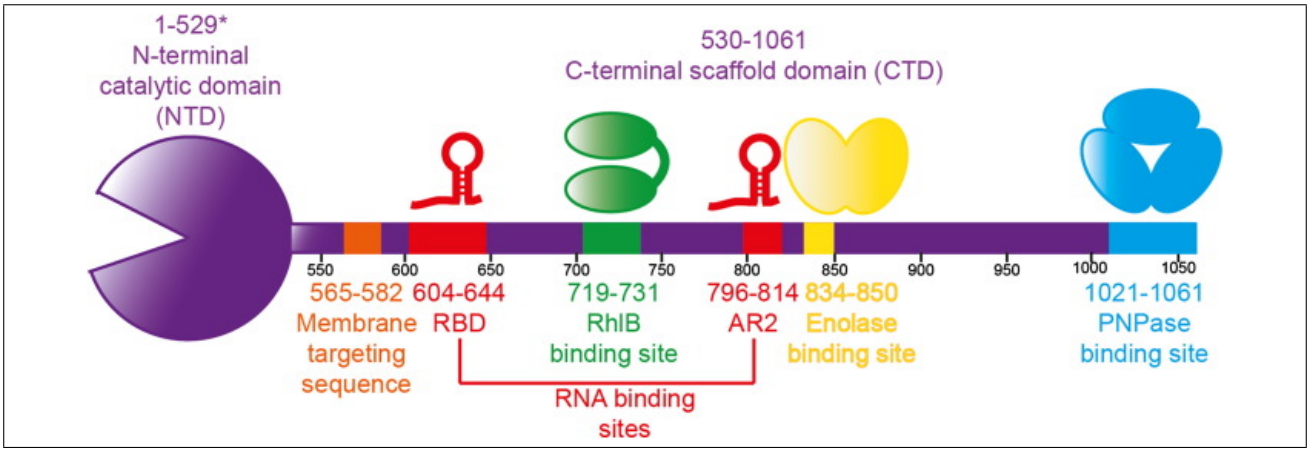


FIGURE 3 – **Structure du dégradosome.** Représentation canonique du dégradosome, la partie violette symbolise la RNase E avec le domaine catalytique à gauche, la partie verte le site de liaison de la RhlB, la jaune celui de l'Enolase et la bleue celui de la PNPase (Bandyra et al. 2013).

trémité mono-phosphate en 5'. Les transcrits primaires bactériens possèdent une extrémité 5' tri-phosphate qui les protège de la dégradation jusqu'à ce qu'ils soient déphosphorylés par l'activité des pyrophosphohydrolases. La RNAse E reconnaît alors l'extrémité 5' mono-phosphate des transcrits par contact de son domaine catalytique (Bandyra et al. 2013; Callaghan et al. 2005).

La présence de structure secondaires peut entraver le processus de dégradation initié par les exoribonucléases, la PAP intervient alors en ajoutant sur l'extrémité 3' du transcrit une séquence poly-A qui va déstabiliser la structure secondaire et permettre ainsi aux exoribonucléases de poursuivre la dégradation (Figure 4).

Terminaison de la transcription

Le mécanisme de terminaison de la transcription chez les procaryotes est gouverné par deux classes signaux de fin de transcription. Les terminateurs Rho-dépendant dont l'activité s'appuie sur la liaison de la protéine Rho à un site *rut* (Rho utilization) présent sur le transcrit associé à une interaction avec la RNA Polymérase et les terminateurs Rho-indépendants caractérisés par une structure G-C riche formant une tige-boucle suivie d'une série de résidus U. Ces terminateurs peuvent être bi-directionnels (Figure 5) (Henkin 2000; Lesnik et al. 2001).

Dans l'état actuel de la recherche beaucoup de pistes pointent vers le fait que les REP joueraient un rôle soit dans la terminaison de la transcription, soit dans le processus de dégradation des transcrits au niveau ARN CITER REFERENCE ou protéique CITER REFERENCE. Plusieurs technologies sont disponibles pour étudier l'expression des gènes, les principales sont le micro-array, le tiling-array et le RNA-Seq. Notre choix s'est porté sur le RNA-Seq car il

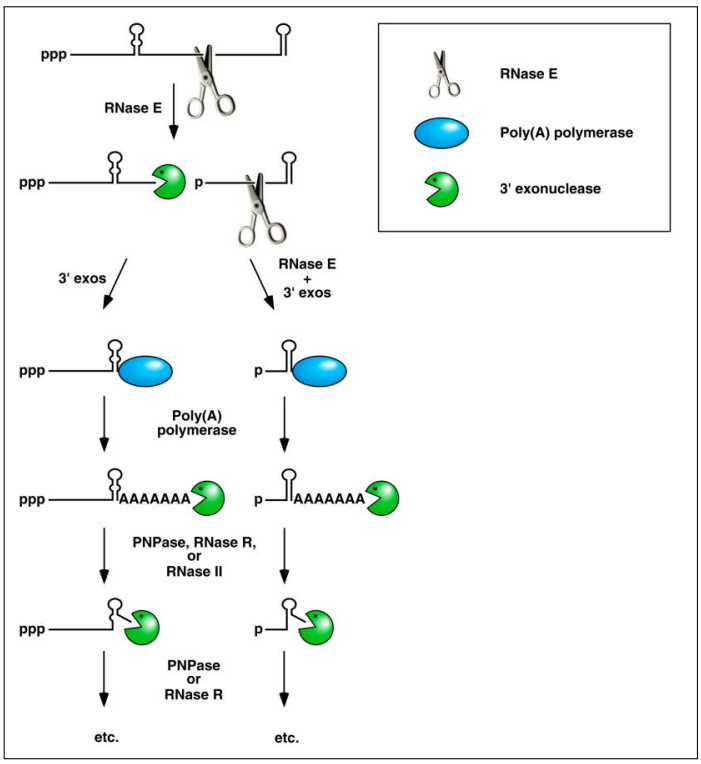


FIGURE 4 – **Facilitation de la dégradation des ARNm chez E. coli par l'intervention de polyadénylation.** Le clivage endonucléolytique par la RNase E génère de multiples fragments, dont certains possèdent à leur extrémité 3' une structure en tige-boucle. Ces fragments subissent une digestion de leur extrémité 3' par la PNPase, la RNase II et/ou la PNPase jusqu'à ce que cette structure soit rencontrée interrompant la dégradation. La PAP intervient pour déstabiliser la tige boucle par l'ajout d'une séquence poly-A en 3' autorisant la reprise de la dégradation par la PNPase et/ou la RNase R (Belasco 2010).

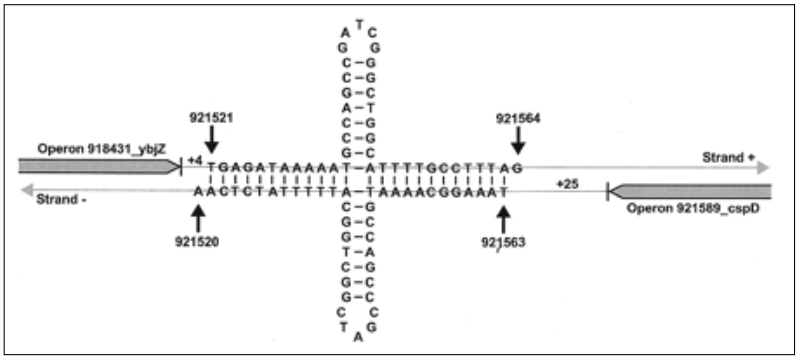


FIGURE 5 – **Terminateur Rho-indépendant bi-directionnel.** La région riche en G-C constitue la structure en tige, la boucle étant formée par les bases non appariées. A la suite de cette structure, nous observons la répétitions de T (U) caractéristique. (Lesnik et al. 2001).

présente plusieurs avantages sur les autres technologies. A la différence des arrays, cette technologie ne nécessite pas la synthèse de sondes spécifiques des espèces ou des transcrits, elle peut donc détecter de nombreux événements non attendus. Elle offre des seuils de détection beaucoup plus bas car elle ne souffre pas du bruit spécifique aux arrays ce qui améliore sa spécificité et sa sensibilité. Sa profondeur de séquençage permet de détecter des transcrits rares et de faible abondance.

Concernant le RNA-Seq, nous pouvons distinguer deux approches. L'une classique qui consiste à analyser les données d'expression de mutants par rapport à un individu sauvage (Wild-Type WT) ou une population dans des conditions classiques et une population dans des conditions perturbées. Et l'autre, alternative, où pour une même condition, nous nous intéressons à la **différence de niveau d'expression d'un gène par rapport à un autre**. Classiquement dans le cas d'un opéron, nous nous attendons à ce que les gènes qui le constituent aient un profil d'expression similaire. Cette approche est privilégiée pour notre étude puisque nous cherchons à déterminer le rôle des REP dans la régulation de la transcription. A notre connais-

sance, il n'existe pas d'approche globale sur l'implication des REP dans la transcription, nous avons développé des méthodes pour tenter de découvrir le rôle des REP en nous basant sur des données d'expressions issues d'expériences de RNA-Seq.

Matériel & Méthodes

Matériel

Fichiers d'annotations

Le fichier General Feature Format (GFF) d'annotation du génome d'*E. coli K12* a été généré par un script Perl à partir du fichier GenBank. Les fichiers répertoriant les opérons et les promoteurs proviennent de RegulonDB. Pour les opérons, le fichier se compose de 2640 entrées dont 848 sont des opérons de plus d'un gène et parmi ces derniers, 235 ont été annotés comme ayant de fortes preuves de leur existence (expérimentalement vérifiés), les 613 autres présentent des preuves d'existence plus faibles (inférence automatique, bibliographie...). Le fichier des promoteurs contient 8580 entrées dont 6461 ont été annotées comme présentant des preuves fortes d'existence. Le fichier répertoriant les terminateurs de transcription provient de Door^{2}DB) et contient 1835 entrées, nous n'avons pas d'information quand à la manière dont ils ont été annotés. Quand aux REP et BIME, les fichiers d'annotations proviennent de l'équipe **Donner la façon dont ils ont été annotés**.

Données

Plusieurs jeux de données ont été utilisés, tous issus d'expériences RNA-Seq publiques, accessibles sur le base de données GEO (Gene Expression Omnibus) du NCBI au format Sequence Read Archive (SRA). Grâce au SRA toolkit et à la commande `fastq-dump`, elles sont décompressées au format `fastq`. Un contrôle de qualité est effectué afin d'inspecter les reads grâce au logiciel `fastqc`.

Le premier jeu de données que nous avons exploité est issu 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 important de réplicats (9) pour la condition de croissance en milieu pauvre en glucose (GSE61327_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. Elles ont été obtenues par séquençage 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 sequencing. 8 réplicats ont été validés 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.

Le second jeu de données provient d'une expérience visant à développer un algorithme pour la détection des opérons chez *E. coli K12* (Li et al. 2013). Nous nous sommes intéressés à celles provenant des cultures ayant subi un choc thermique pendant 15 minutes (HS-15min) ainsi qu'à celles provenant de culture privées de phosphore pendant 4 heures (M-P4h). Ces 2 conditions ont été retenues car elles possèdent 3 réplicats contre 2 pour toutes les autres. Les librairies ont été conçues en Single-end sequencing et brin spécifique en utilisant le kit Illumina's TruSeq Small RNA Sample Prep, puis séquencées à la fois sur Illumina HiSeq 2000 (générant des reads de 100 bases) et Illumina GA II (générant des reads de 76 bases). Aucun réplicat n'a été rejeté suite aux contrôles qualité.

Alignement des reads

Les reads ont été alignés puis mappés sur le génome d'*E. coli* NC_000913.2, qui est le génome utilisé pour annoter les REP par l'équipe, grâce au logiciel BWA. Ce logiciel propose 3 algorithmes distincts, BWA-backtrack, BWA-SW et BWA-MEM. Pour chacun de ces alignements, il est nécessaire de disposer de la séquence du génome 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) pour sa rapidité et sa précision. Il reprend les mêmes principes que BWA-SW (utilisation de la programmation dynamique pour trouver les points d'ancrage (seeds) en autorisant les mésappariements (mismatchs) et les brèches (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 l'ancrage avec des MEM, puis il réalise l'extension en prenant en compte les pénalités dues aux gaps et aux mismatchs. Les valeurs par défaut du logiciel ont été utilisées.

```
# 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 converti au format Binary Alignment Map format (BAM), puis des critères de qualité sont appliqués. Seules les séquences possédant une qualité de mapping $>$ 30, valeur d'usage commun, et n'étant pas étiquetées (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 sont ensuite triées par position génomique. Finalement, le fichier BAM trié est 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: mapping minimal, -F 2048: pas de sequence 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
```

Intégration des données

Les fichiers d'annotation 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 permettant de croiser les informations provenant de plusieurs sources de données et ainsi de rechercher des intersections, des positions proches ou déterminer une couverture de reads. Ces outils ont généré les fichiers `BED` qui serviront de référence pour la suite de l'analyse statistique.

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 6).

Méthodes

Création de la table de comptages et normalisation

Afin d'obtenir des résultats de comptage par région d'intérêt et ainsi estimer l'expression, nous avons utilisé le package Bioconductor easyRNASeq (Delhomme et al. 2012) puisqu'il permet de réaliser les opérations de comptage de façon documentée et qu'il offre la possibilité d'effectuer une normalisation par Reads Per Kilobase per Million mapped reads (RPKM). Les annotations du génome d'*E. coli* relatives aux gènes sont extraites à partir du fichier `GFF` et stockées sous forme d'une base de données. Cela a nécessité une manipulation préalable de ce

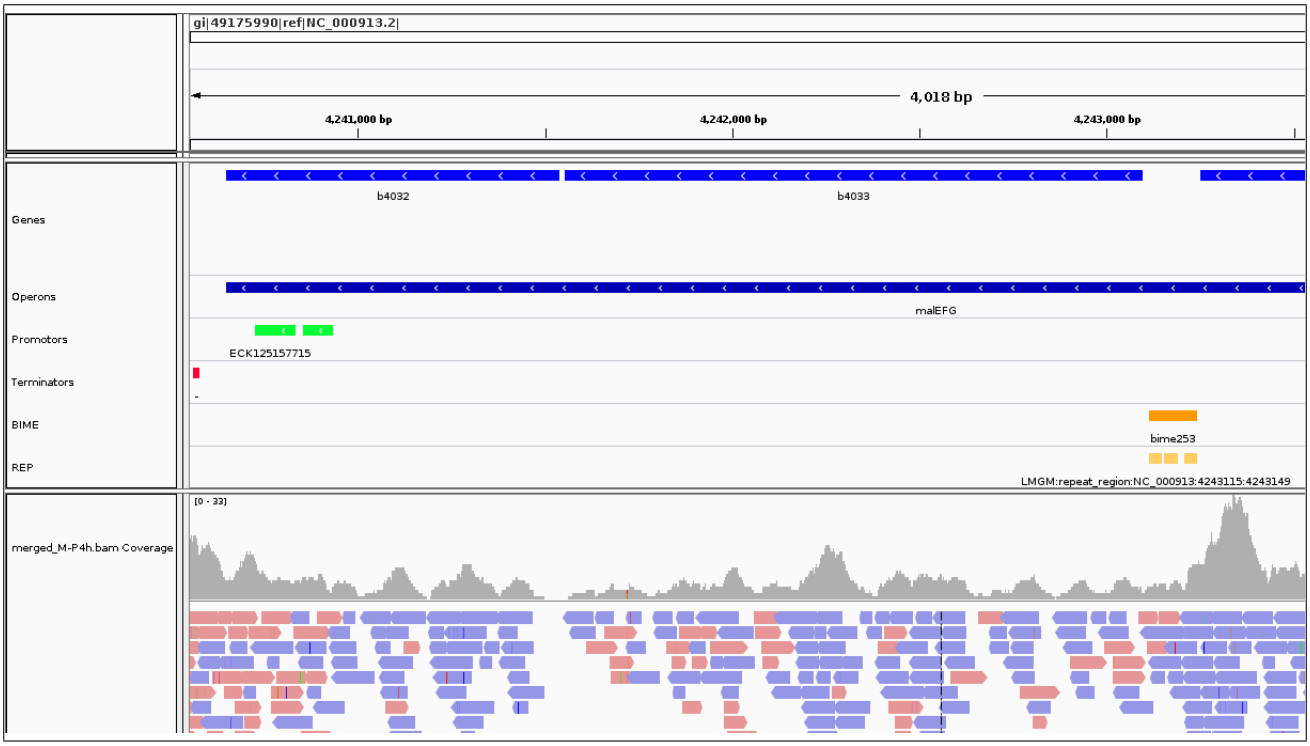


FIGURE 6 – **Visualisation du mapping de l'opéron malEFG sur IGV.** Les premières pistes représentent les positions et orientations des gènes, des opérons, la présence de promoteurs et de terminateurs, ainsi que la position des BIME et des REP qui composent les BIME. Les 2 pistes suivantes affichent 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 sur le brin direct et la couleur rouge sur le brin complémentaire.

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 gène est ensuite extraite pour un total de 4605 éléments que nous nommerons régions. La couverture par région est ensuite calculée pour chaque fichier BAM, le résultat est obtenu en réalisant l'union des positions extraites de la liste des régions 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 régions figurant en ligne et les fichiers BAM en colonnes (Figure 7).

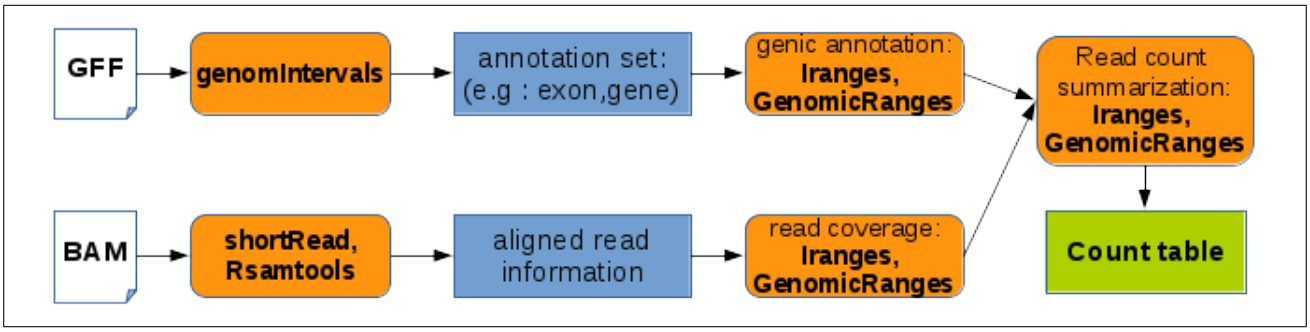


FIGURE 7 – **easyRNASeq : création d'une table de comptage.** Pour le modèle choisi sur notre analyse, les fichiers d'entrée sont en blanc, les traitements par les packages R sont colorés en orange, les données intermédiaires en bleu et la table de comptage produite est en vert.

Les résultats de comptages doivent ensuite être normalisés afin de permettre la comparaison de l'expression des gènes et des régions génomiques d'intérêt. De nombreuses méthodes de normalisation existent dont certaines introduisent un biais notamment pour la recherche de DE. La méthode du 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 conditions. 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. Notre choix s'est porté donc porté sur cette méthode (Mortazavi et al. 2008) :

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

Il faut souligner que comme nous ne nous intéressons pas à un même gène dans 2 conditions différentes mais à 2 gènes dans une même condition, un biais dû à leur composition en GC peut intervenir.

Les données sur lesquelles nous allons travailler sont issues de la table de comptage et ont subi une normalisation, afin de vérifier que l'homogénéité entre les réplicats est toujours présente, nous réalisons une analyse en composante principale.

Corrélation de profils d'expression

Le principe de cette analyse repose sur la recherche de corrélation entre le profil d'expression issu de données de RNA-Seq et un profil simulé représentant un changement de niveau d'expression, cette méthode a été mise au point pour la prédiction d'opérons dans les génomes bactériens (Fortino et al. 2014). Le concept général est de délimiter les bornes des transcrits à partir de données de couverture en s'appuyant sur un test de corrélation entre ce profil et un profil simulé de 0 et de 1. Les 0 représentant une zone sans couverture donc en dehors d'un transcrit et les 1 la zone couverte, donc le transcrit. Le cœur de leur méthode consiste à déplacer une fenêtre glissante de 100 pb parcourant base à base le génome en réalisant des tests de corrélation entre le profil d'expression réel et le profil d'expression simulé par le vecteur de même taille 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, 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 du taux de couverture sont calculées 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 1 ce qui représente un changement de 2 fois du niveau d'expression. 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 8) pour être validée. Les auteurs ont fixé un seuil de

corrélation de 0.7 et une p-value de 10^{-7} comme seuil significatif pour le test de corrélation.

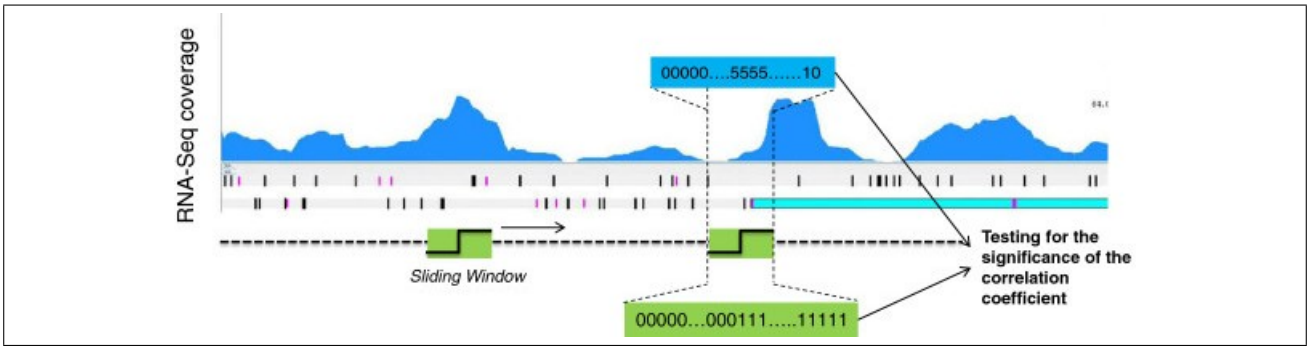


FIGURE 8 – **Recherche de corrélation sur des profils d'expression.** La fenêtre glissante (en vert) parcourt la région d'intérêt et pour chaque déplacement une corrélation est calculée entre le vecteur du profil d'expression obtenu par RNAseq (en bleu) et celui simulé par le vecteur de 0 et de 1 (en vert) (Fortino et al. 2014).

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). 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 de Poisson, en K segments, tel que :

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

où m est une partition de $[1, n]$ en r segments, le paramètre θ_r est la moyenne associée au segment r. L'objectif étant d'estimer la position des segments et le paramètre θ_r résultant de la segmentation. $M_{K,n}$ est alors l'ensemble des partitions possibles avec K le nombre maximal 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 :

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

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 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 meilleure 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.

États ancêtres et structures secondaires

Résultats

Pour réaliser nos analyses, nous nous sommes inspirés de la méthodologie employée en RNA-Seq pour l'étude de Différence d'Expression (DE). L'analyse statistique de recherche de changement d'expression liée à la présence de BIME a été menée sur le logiciel R. La première étape à consisté à examiner les données concernant les BIME et l'expression de manière générale, puis nous nous sommes intéressé aux opérons pour lesquels nous nous attendons à un niveau d'expression similaire des gènes les composant, nous avons ensuite étendu notre étude à tous les gènes en recherchant des changements de niveaux d'expression qui pourraient être liés à la présence de BIME à proximité par une approche locale (la corrélation de profils d'expression) et une approche globale (la segmentation).

Examen des données et des résultats de comptage

REP et BIME

Pour le génome d'*E. coli K-12*, 93 REP ont été répertoriées comme étant solitaires sur les 605 annotées par l'équipe. Les REP sont organisées en 287 BIME pouvant contenir une ou plusieurs d'entre elles.

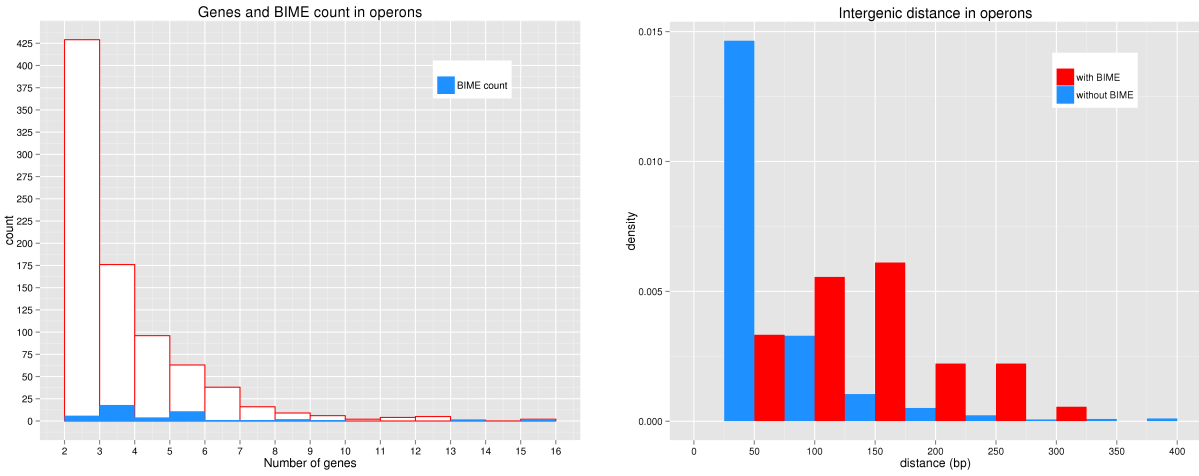
Visualisation de la couverture

Il est important de noter que la couverture le long du génome n'est pas uniforme, ni même sur les gènes, car nous observons la présence de nombreuses vallées et pics (Figure 6). Ce phénomène peut s'expliquer par plusieurs raisons techniques (Li et al. 2013). Premièrement, les **méthodes de fragmentation** des protocoles de préparation des bibliothèques amènent un biais en cassant ou dégradant certaines séquences. Le second biais possible est produit 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 (formation de structures secondaires et température de dénaturation plus élevée). Le dernier point, spécifique au séquençage Illumina, implique des **interférences spécifiques aux séquences** lors du pro-

cessus d'élongation pendant le séquençage générés par des schémas particuliers du template, tel que des répétitions de GCC ou CGG et des répétitions inversées d'une séquence de plus de 8 pb sur un même brin (ex : AAAAAACCTTGAAAAGCCAGGCTTTTCAAGGTTTTTTT), produisant des repliements du brin d'ADN et altérant l'affinité des enzymes (Nakamura et al. 2011).

Opérons

Chez *E. coli K12*, 1792 opérons sont formés d'un seul gène et sur les 848 autres opérons formés par plus d'un gène, 36 contiennent au moins une BIME. Nous avons étudié la répartition de la composition en nombre de gènes de cette catégorisation et nous avons visualisé laquelle contient le plus de BIME (Figure 9(a)).



(a) (b)

FIGURE 9 – **(a) Distribution du nombre de gènes par opérons et répartition des BIME.** L'échelle des abscisses indique le nombre de gènes dans l'opéron. Les histogrammes blancs représentent le compte du nombre d'opérons, les histogrammes bleus le compte des BIME au sein des opérons. La grande majorité des opérons sont constitués de 2 gènes mais peuvent aller jusqu'à 16 gènes. Sur les 36 BIME présentes dans les opérons, 18 sont dans des opérons de 3 gènes, 11 dans ceux de 5 gènes et 6 dans ceux de 2 gènes. **(b) Distances inter-géniques dans les opérons avec et sans BIME.** L'échelle des ordonnées représente la densité car l'écart entre les valeurs est trop important pour une visualisation correcte de l'histogramme avec les comptes. En bleu absence de BIME dans la RIG, en rouge présence de BIME. Pour les données en absence de BIME, nous observons une valeur modale dans les régions inter-géniques de 50 pb, alors qu'en présence de BIME, la valeur modale se situe dans les régions inter-géniques de 150 pb.

Nous avons ensuite établi un indice d'expression, I_{expr}, permettant de mesurer l'écart entre le niveau d'expression de deux gènes consécutifs dans un opéron. La numérotation des gènes se faisant dans le sens de la transcription :

$$I_{expr} = (gene2 - gene1)/(gene2 + gene1) \quad I_{expr} \in [-1, 1]$$

Cet indice I_{exp} permet de comparer la variation entre les couples de gènes quelque soit leur niveau d'expression. Si le premier gène est plus exprimé que le second, la valeur de I_{expr} est négative sinon elle devient positive. Cet indice nous a permis de nous intéresser aux niveaux

d'expression des gènes des opérons en observant trois facteurs de la Région Inter-Génique (RIG), la présence d'une BIME, la présence de promoteurs et/ou de terminateurs de transcription et la taille de cette région. Les annotations des promoteurs et des terminateurs proviennent respectivement de RegulonDB et Door^{2}DB et sont issus à la fois de validations expérimentales et de prédictions. Pour la taille de cette région nous avons étudié la distance inter-génique dans les opérons en présence ou absence de BIME (Figure 9(b)). La taille des RIG ne contenant pas de BIME possède une valeur modale de 50 pb alors qu'en leur absence celle-ci se situe à 150 pb. Nous avons fixé le seuil de la taille de la RIG à 150 pb.

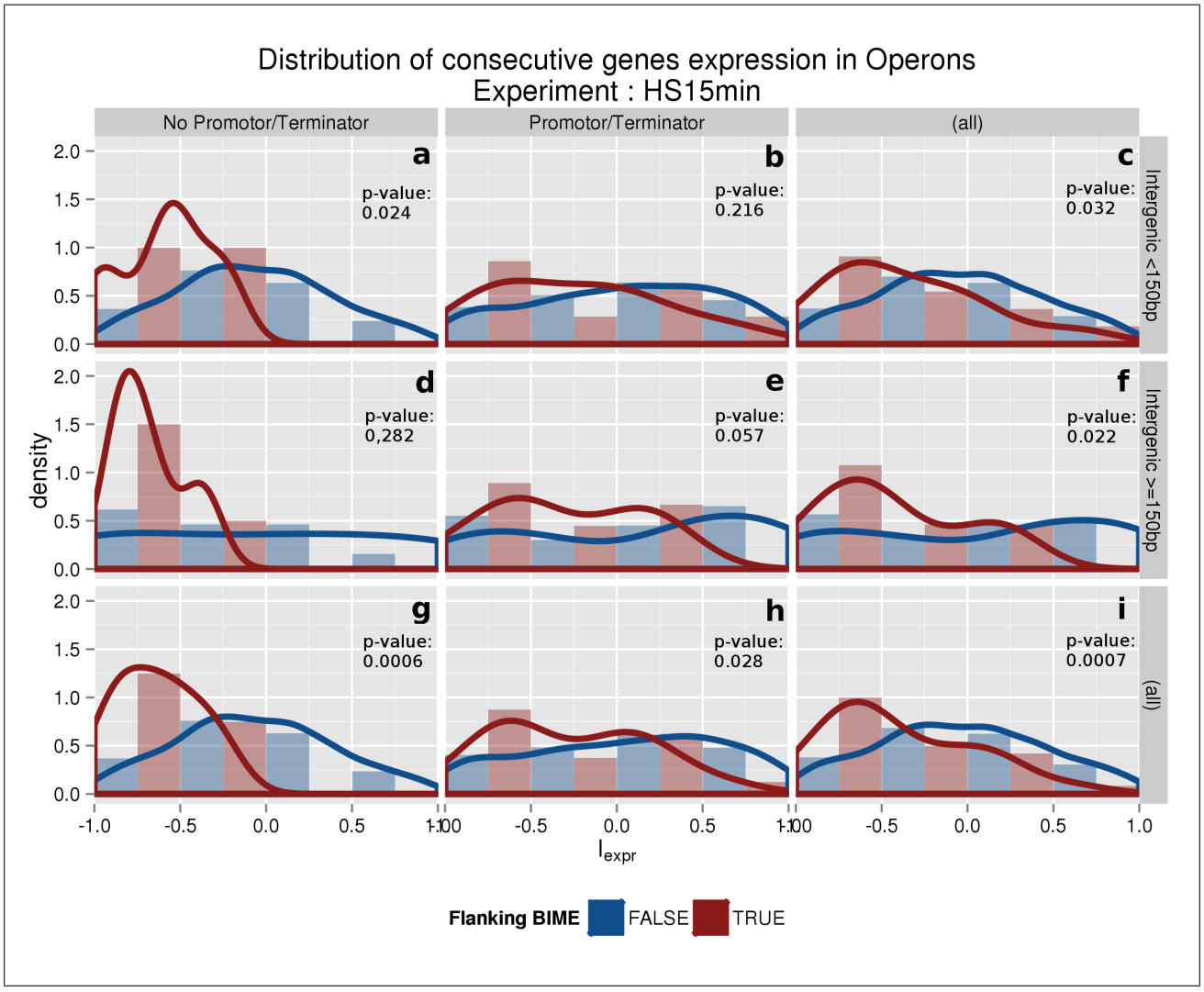


FIGURE 10 – **Niveau d'expression des gènes consécutifs dans les opérons.** Les données proviennent de l'expérience de choc thermique pendant 15 minutes (HS-15min). L'I_{expr} en abscisses est l'indice d'expression du gène 1 par rapport au gène 2 $(gene2 - gene1)/(gene2 + gene1)$, borné sur [-1,1]. Les valeurs négatives indiquent un niveau d'expression plus élevé du gène 1, les valeurs positives l'inverse. Les éléments de couleur rouge se réfèrent aux gènes flanquant une BIME, ceux en bleu aux gènes sans BIME entre eux. Les histogrammes indiquent la densité pour chacune des valeurs de l'I_{expr}. Les courbes de densité montrent la tendance de l'I_{expr} par rapport à la présence/absence de BIME, la comparaison entre les 2 distributions est réalisée à l'aide d'un test de Wilcoxon dont la p-valeur est affichée dans chaque graphique. Les graphiques sont décomposées sur 2 critères, les Élements Transcriptionnels, tels que les promoteurs et les terminateurs de transcription (ET) dans la RIG et la taille de la RIG (<150 pb ou >= 150 pb).

L'analyse menée sur les données d'expression des bactéries soumises à un choc thermique pendant 15 minutes (HS-15min) est représentative des deux autres jeux de données (résultats non montrés). La comparaison des distributions en présence et absence de BIME est effectuée à

l'aide d'un test de Wilcoxon car toutes les distributions ne suivent pas une loi normale (test de Shapiro), l'hypothèse nulle étant la similitude des distributions. L'absence de BIME montre une répartition symétrique de l'expression des deux gènes centrée sur une valeur d'I_{expr} à zéro sauf en présence de promoteurs/terminateurs dans la RIG, la tendance étant vers une expression plus importante du gène 2 (Figure 10 - (b,e,h)). En s'intéressant dans un premier temps au RIG de moins de 150 pb, en présence de promoteur/terminateur, le test statistique ne montre pas de différence entre les deux distributions (Figure 10 - (b) ; p-valeur : 0.216), ce qui n'est pas le cas en leur absence ou avec toutes les valeurs (Figure 10 - (a-c) ; p-valeurs : 0.024, 0.032). Pour les RIG de plus de 150 pb, la tendance vers un niveau d'expression supérieur du gène 1 est plus marquée avec un décalage de la courbe en présence de BIME (rouge) vers la gauche mais la significativité n'est pas avérée sauf lorsque le facteur promoteur/terminateur n'est plus pris en compte (Figure 10 - (f) ; p-valeur : 0.022). En étudiant maintenant le facteur promoteur/terminateur, nous remarquons en leur absence une tendance à l'expression plus importante du gène 1 qui devient significative en ne s'intéressant plus à la taille de la RIG (Figure 10 - (g) ; p-valeur : 0.0006). En leur présence, nous notons que la courbe marquant l'absence de BIME (bleue) tend vers les valeurs positives, donc une expression plus importante du gène 2, ceci s'explique par la présence de promoteurs dans la RIG (Figure 10 - (b,e,h) ; p-valeurs : 0.216, 0.057, 0.028). Le décalage de la courbe des BIME (rouge) vers les valeurs négatives est visible de façon plus ou moins marquée dans tous les cas de figure, cet effet est peut être dû à rôle de terminateur de transcription joué par les BIME ou à un effet de stabilisation de la partie 5' du transcrit favorisé par la présence de la BIME.

En approfondissant l'étude de ses données et en nous plaçant dans le cas de la (Figure 10 - (i)), nous nous intéressons à la distribution de l'I_{expr} en n'ayant que pour critère la présence ou l'absence de BIME et la relation existant entre l'absence d'ET et la présence de chacun d'entre eux. En étudiant d'abord les cas où les BIME sont absentes (Figure 11-(a)), nous observons un effet des promoteurs avec une expression plus importante du gène 2 et l'effet inverse pour la courbe des terminateurs de transcription. Ces deux effets sont visibles de façon plus modérée sur la courbe des promoteurs et terminateurs. Quand à la courbe des RIG sans ET, elle est symétrique et centrée sur sur une valeur négative mais proche de zéro. Des tests de Wilcoxon ont été réalisés entre les valeurs d'I_{expr} sans ET (courbes bleues) dans la RIG et chacun des trois autres cas de figure (Tableau 1). Ces résultats nous indiquent qu'il existe une différence significative entre la distribution sans ET et les trois autres (p-valeurs < 0.05), résultat attendu puisque les ET ont un effet sur la transcription des gènes. En revanche dans le cas où les BIME sont présentes (Figure 11-(b)), uniquement la distribution de l'I_{expr} contenant des promoteurs se différencie de manière significative (p-valeurs : 0.034) de celle sans ET. Ce résultat nous conforte dans l'hypothèse que les BIME joueraient un rôle similaire aux terminateurs de transcription ou de stabilisation de la partie 5' du transcrit face au dégradosome.

Afin d'approfondir ces hypothèses, nous avons étudié plus en détail le niveau d'expression des gènes dans les 36 opérons contenant des BIME.

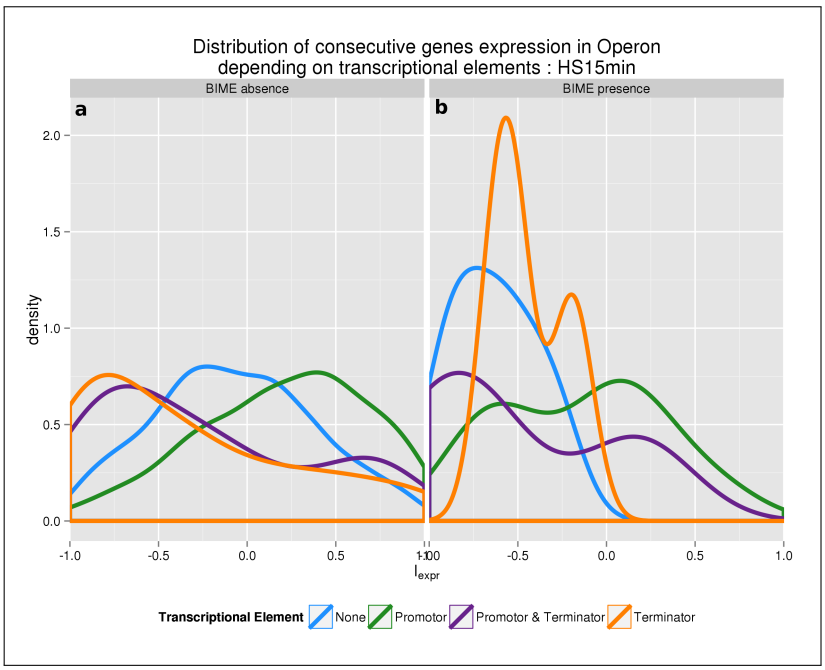


FIGURE 11 – **Niveau d'expression des gènes consécutifs dans les opérons en fonction des éléments transcriptionnels (ET).** L'I_{expr} en abscisses est l'indice d'expression du gène 1 par rapport au gène 2 $(gene2 - gene1)/(gene2 + gene1)$, borné sur [-1,1]. Les valeurs négatives indiquent un niveau d'expression plus élevé du gène 1, les valeurs positives l'inverse. Les courbes de densité indiquent la distribution de l'I_{expr} en fonction des ET présent dans la RIG. Les courbes bleues indiquent l'absence de promoteur ou de terminateur, les vertes la présence de promoteur uniquement, les oranges la présence de terminateurs uniquement et les violettes la présence à la fois de promoteurs et de terminateurs. **(a)** présence de BIME dans la RIG, **(b)** absence de BIME dans la RIG.

Wilcoxon test p-valeurs	Élément transcriptionnel		
	Promoteur	Terminateur	Promoteur & Terminateur
Absence de BIME	$5.48\ e^{-14}$	$7.3\ e^{-5}$	0.0091
Présence de BIME	0.034	0.375	0.921

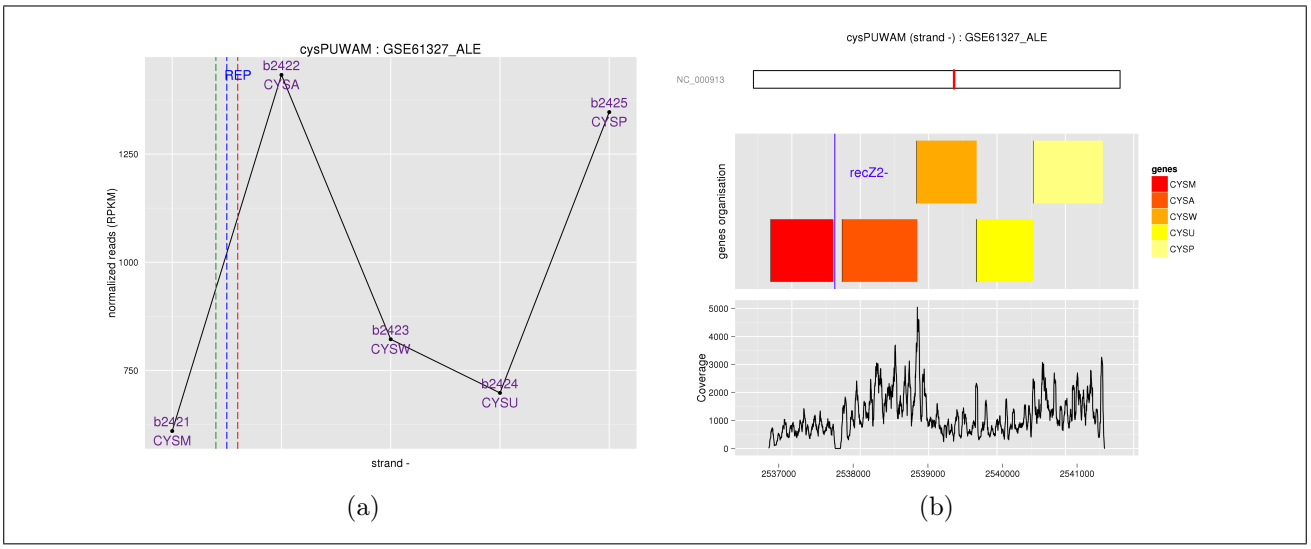
TABLE 1 – **P-valeurs des tests de Wilcoxon des I_{expr} en présence des éléments transcriptionnels comparé à leur absence.** Les p-valeurs sont présentées en présence et en absence de BIME et renvoient aux cas de la Figure 11-(i).

Étude des opérons, différence d'expression en présence de BIME

L'hypothèse privilégiée ici est que les gènes appartenant à un opéron vont être exprimés à un niveau similaire, la question qui se pose est de savoir si la présence d'une BIME entre deux gènes d'un opéron va avoir un impact sur la transcription ou la dégradation d'un des gènes. Dans ce cadre, les opérons contenant des BIME sont sélectionnés et l'expression des deux gènes de l'opéron flanquant la BIME prise en compte si au moins un des deux gènes a une couverture > 10, cette valeur est choisie car elle représente le premier quartile de nos données, supprimant les gènes les moins exprimés. De plus, cela nous permet d'éliminer des variations qui seraient jugées importantes pour de petites valeurs (e.g : 2 et 8 qui implique un facteur 4 pour le changement d'expression). Nous avons vérifié par un test de Shapiro que nos valeurs de comptage ne suivaient pas une loi Normale, ce qui nous a conduit à choisir d'appliquer un test non paramétrique de rangs de Wilcoxon si l'expérience contient au moins 5 réplicats et un test de Student si l'expérience contient moins de réplicats. L'hypothèse nulle de ces tests est qu'il n'existe pas de DE entre les 2 gènes. La p-value significative étant fixée à 0.01 pour nous permettre d'être plus stricts sans pour autant limiter les résultats. Nous n'avons pas jugé utile d'appliquer de correction à ces tests car le nombre d'opérons contenant des BIME se limite à

36.

Pour les gènes dont le test est significatif, deux représentations graphiques sont générées. La 1ère est un schéma décrivant les niveaux d'expression normalisé des gènes de l'opéron ainsi que la position relative des REP formant la BIME (Figure 12(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 (Figure 12(b)).



(a) (b)

FIGURE 12 – **Résultats de l'étude de l'expression des gènes dans les opérons contenant des BIME. (a)** Le taux d'expression normalisé de chaque gène de l'opéron est représenté en ordonnées, l'organisation des gènes de l'opéron est schématisée sur l'axe des abscisses dans le sens 5' $\mapsto$ 3', l'orientation du brin est notée sur l'axe des abscisses. La position de la ou les REP composant la BIME est schématisée par la ligne bleue verticale. En tenant compte de l'orientation du brin, si au moins un promoteur ou un terminateur de transcription est présent dans la RIG contenant la BIME, ceux-ci sont représentés respectivement par une ligne en pointillés verticale verte ou rouge. **(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' (le sens du brin est précisé dans le titre) ainsi que le positionnement et la classe des REP composant la BIME. La partie inférieure montre la couverture issue des fichiers BAM fusionnés de l'expérience par rapport à l'organisation de la partie médiane.

Nous avons testé cette méthode sur nos trois jeux de données qui ont produit les résultats du Tableau 2.

Jeux de données	$\frac{Nb.\ operons\ avec\ genes\ DE}{Nb.\ d'operons\ exprimes}$	Nb. BIME	Nb. Terminateurs	Nb. Transporteurs ABC
GSE61327_ALE	17/32	18	6	7
HS-15min	8/20	8	1	4
M-P4h	12/30	12	4	5

TABLE 2 – **Résultats de l'étude de différence d'expression des gènes des opérons flanquant une BIME.** La 2ème colonne indique le nombre d'opérons contenant une 2 gènes flanquant une BIME et pour lesquels nous observons une différence de niveau d'expression par rapport au nombre d'opérons dont les 2 gènes flanquant la BIME sont exprimés (au moins un des 2 possède une couverture > 10). La 3ème colonne indique le nombre de BIME concernées, un opéron pouvant en contenir plusieurs. La 4ème le nombre d'opérons contenant un terminateur de transcription dans la RIG avec la BIME pour les gènes DE. La dernière colonne indique le nombre de transporteurs ABC parmi les opérons dont les gènes sont DE.

Approche locale, corrélation de profils d'expression

Nous avons développé une méthode réalisant un test de corrélation entre les profils d'expression des régions contenant des BIME et un profil modèle de changement d'expression en

modifiant la technique mise au point pour la prédiction d'opérons dans les génomes bactériens (Fortino et al. 2014) décrite dans la partie Corrélation de profils d'expression. Elle a été adaptée pour nous permettre de localiser le point de cassure dans la RIG contenant une BIME. Au lieu de faire une recherche sur le génome complet, nous avons privilégié une approche locale en ciblant des régions d'intérêt qu'il a fallu au préalable délimiter. Celles-ci se modélisent par la présence du 1er gène, de la 1ère RIG, de la BIME, de la 2nde RIG 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
```

Les seuils que nous avons utilisés sont différents de ceux de la méthode originelle et nous avons introduit un filtre supplémentaire quand à la couverture minimale d'un des deux gènes :

- 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 ___|^^^
- $log_2(\frac{\overline{couverture\ gauche} + 1}{\overline{couverture\ droite} + 1}) \geq 1$ pour un profil ^^^|___
- 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 plusieurs raisons, suite à des essais nous sommes arrivés aux conclusions qu'une fenêtre trop petite (100 pb) présente une sensibilité trop importante aux variations locales dues au manque d'uniformité de la couverture (voir Visualisation de la couverture), alors qu'une fenêtre de taille trop grande (500 pb) produit une perte de sensibilité en lissant les couvertures de chaque moitié de la fenêtre. L'idée étant de rechercher le changement d'expression en ayant des informations sur la couverture moyenne des gènes entourant la BIME, les informations d'annotation (Figure 9(b)) nous indiquent que la majorité des régions inter-géniques contenant des BIME mesurent dans les 150 pb. Comme la BIME se situe dans la plupart des cas proche d'un des deux gènes nous avons opté pour une

fenêtre de taille 300 pb pour couvrir à la fois un des gènes et la RIG. Les seuils de corrélation, de p-valeur ainsi que de niveau de changement d'expression ont été repris de la publication.

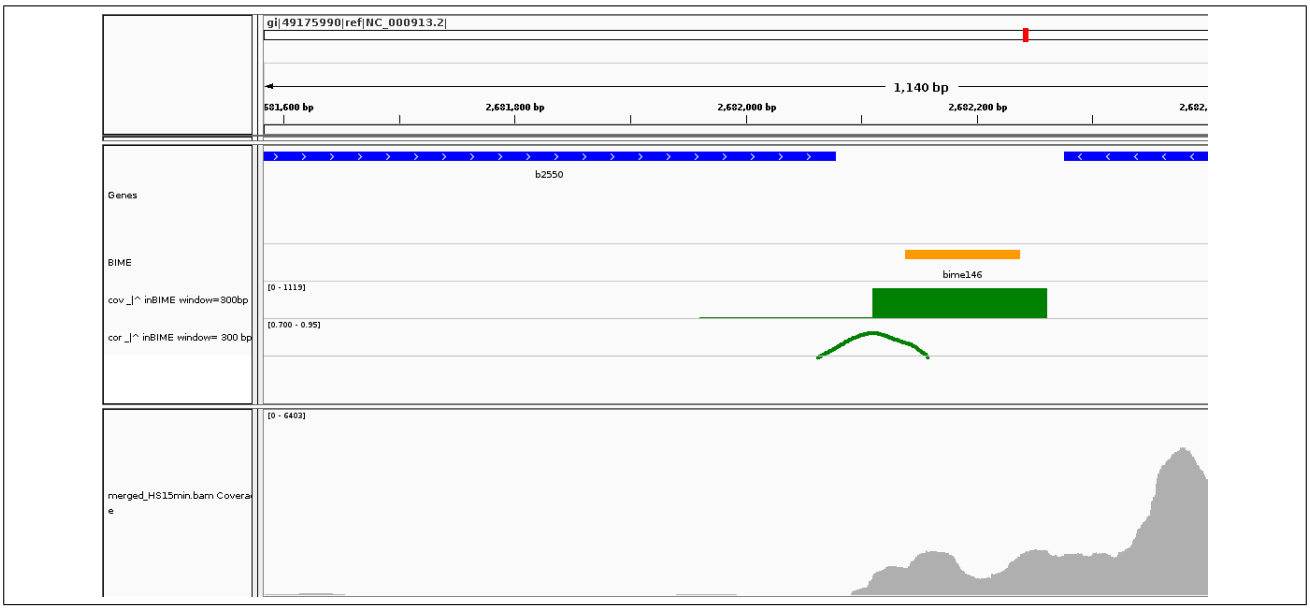


FIGURE 13 – **Visualisation des changements de couverture obtenus par la méthode de corrélation des profils.** La position de la BIME est représentée en orange, le diagramme de barres à 2 colonnes en vert montre les profils d'expression moyens de chaque moitié de la fenêtre et la courbe en points verts 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.

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 corrélation 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 de diagramme de barre représentant les couvertures moyennes des deux parties de la fenêtre, le second représentant l'évolution de la corrélation sur la zone (Figure 13). 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 14). Finalement, un fichier au format `CSV` recueille toutes les informations de l'analyse.

Approche globale, segmentation

En reprenant la méthode décrite dans la partie Segmentation, 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 RIG. Dans notre étude, nous nous intéressons aux positions de ces points de cassures pour nos régions d'intérêt, définis de la même manière que précédemment, 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 RIG et si les deux gènes appartiennent à un opéron, un test de statistique sur la différence d'expression est réalisé avec la même méthodologie que pour la partie **??**.

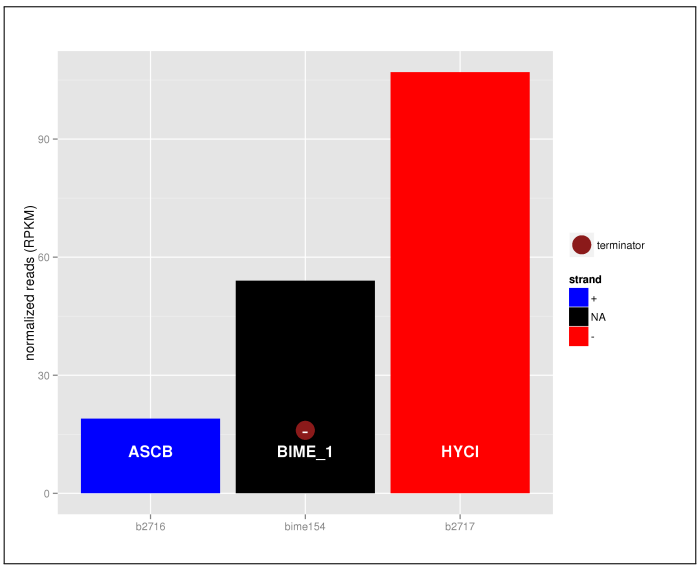


FIGURE 14 – **Résultat de corrélation de profils.** Les niveaux d'expression des gènes encadrant la BIME et de cette dernière sont représentés par les histogrammes. La couleur de l'histogramme indique le sens de transcription de l'élément, bleu pour le brin sens, rouge pour l'anti-sens et noir lorsque aucun brin est défini. La présence d'éléments de régulation dans la RIG 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.

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 15). Deux fichiers au format `CSV` sont générés pour recueillir les informations de la segmentation, le 1er pour les gènes sur le même brin, le 2nd pour les gènes sur des brins opposés.

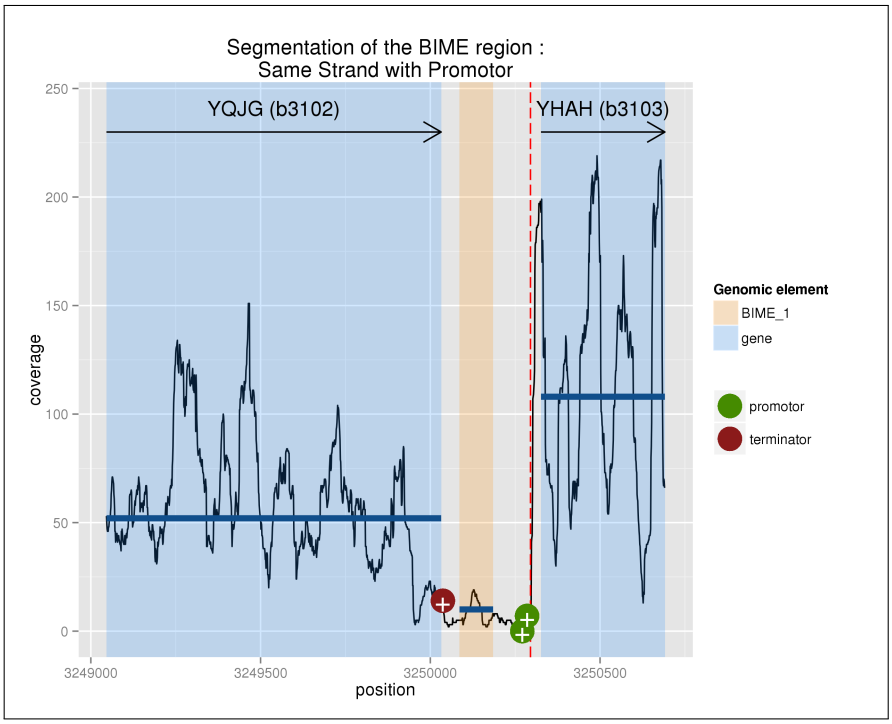


FIGURE 15 – **Résultat de segmentation pour** K_{max}=4**.** Les gènes sont symbolisés par les zones bleues, leur sens de transcription par les flèches noires. La BIME est représentée par la zone orange et sa classe est précisée dans la légende (Genomic element). Les promoteurs sont représentés par des points verts et les terminateurs par des points rouges, 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 en dehors de la BIME après les promoteurs.

Discussion

Glossaire

alignement chimérique Alignement d'un read qui ne peut pas être représenté comme un alignement continu. 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. 8

BAM Binary Alignment Map format. 8

BED Browser Extensible Data. 9

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. 9

DE Différence d'Expression. 14

ET Élements Transcriptionnels, tels que les promoteurs et les terminateurs de transcription. 16

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. 10

GFF General Feature Format. 7

opéron unité d'ADN fonctionnelle regroupant des gènes sous le contrôle d'un signal moléculaire régulateur. Les gènes sont transcrits en ARN messager ensemble et concourent à la réalisation d'une même fonction physiologique. Ainsi, soit tous les gènes d'un opéron sont transcrits tous ensemble, soit aucun n'est transcrit puisqu'ils sont tous sous le contrôle du même régulateur. 5

Paired-end sequencing 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. 8

PDP Pruned Dynamic Programming. 12

reads Séquence nucléotidique issue d'un séquençage NGS. 7

REP Repeated Extragenic Palindrome. 1

RIG Région Inter-Génique. 16

RNA-Seq Technique de séquençage haut-débit pour l'étude de l'expression de l'ARN. Un échantillon d'ARN est rétro-transcrit puis amplifié par PCR, le cDNA est séquencé sur un séquenceur haut-débit. Le compte des reads produits mappé sur un transcrit représente son abondance. 4

RPKM Reads Per Kilobase per Million mapped reads. 9

SAM Sequence Alignment Map format. 8

Single-end sequencing Technique de séquençage haut débit la plus simple consistant à ne réaliser le séquençage que depuis une extrémité du template.. 8

SRA Sequence Read Archive. 7

tiling-array Cette technique diffère des micro-arrays traditionnels par la nature des sondes, au lieu de disposer de sondes pour des séquences de gènes connues ou prédits, elle sonde des séquences connues pour être disposées dans des régions contiguës et ainsi détecter la présence ou l'absence des transcrits dans ces régions. 4

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.

K. J. Bandyra, M. Bouvier, A. J. Carpousis, and B. F. Luisi. The social fabric of the RNA degradosome. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1829(6-7) : 514–522, 2013. ISSN 18749399. doi : 10.1016/j.bbagrm.2013.02.011. URL http://dx.doi.org/10.1016/j.bbagrm.2013.02.011.

J. G. Belasco. All things must pass : contrasts and commonalities in eukaryotic and bacterial mRNA decay. *Nature reviews. Molecular cell biology*, 11(7) :467–78, July 2010. ISSN 1471-0080. doi : 10.1038/nrm2917. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3145457&tool=pmcentrez&rendertype=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.

A. J. Callaghan, M. J. Marcaida, J. a. Stead, K. J. McDowall, W. G. Scott, and B. F. Luisi. Structure of Escherichia coli RNase E catalytic domain and implications for RNA turnover. *Nature*, 437(7062) :1187–1191, 2005. ISSN 0028-0836. doi : 10.1038/nature04084.

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.

T. M. Henkin. Transcription termination control in bacteria. *Current Opinion in Microbiology*, 3(2) :149–153, 2000. ISSN 13695274. doi : 10.1016/S1369-5274(00)00067-9.

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.

E. a. Lesnik, R. Sampath, H. B. Levene, T. J. Henderson, J. a. McNeil, and D. J. Ecker. Prediction of rho-independent transcriptional terminators in Escherichia coli. *Nucleic acids research*, 29(17) :3583–3594, 2001. ISSN 1362-4962. doi : 10.1093/nar/29.17.3583.

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.