Rapport de stage de M2

Nicolas JEANNE

11 mars 2015

Introduction

En 1982, la découverte par Higgins et al. 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 al. 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 tétranucléotide 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 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 et 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).

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

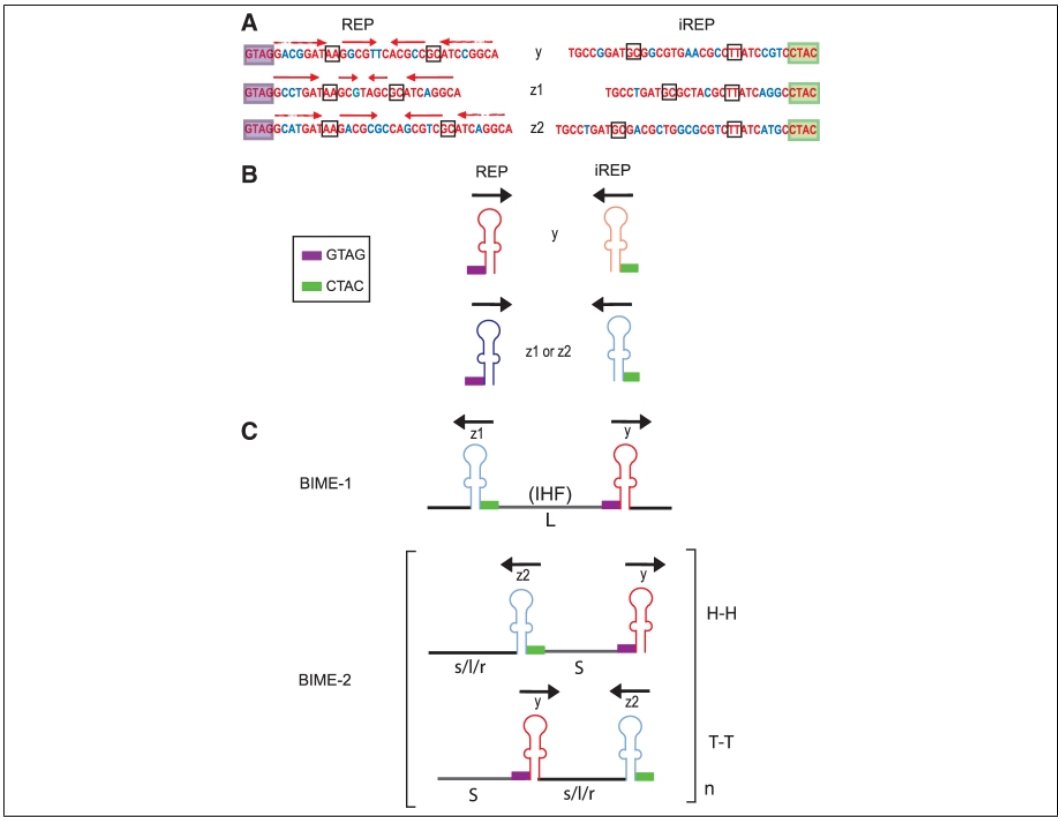

FIGURE 1 – **REP et BIME chez Escherichia coli. (A)** Séquences consensus Y, Z^1 et Z^2 des REP. Le tétra-nucléotide conservé GTAC est encadré en violet, le complémentaire conservé CTAC est encadré en vert, les flèches rouges situent les zones d'appariement de la tige et les positions encadrées en noir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

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).

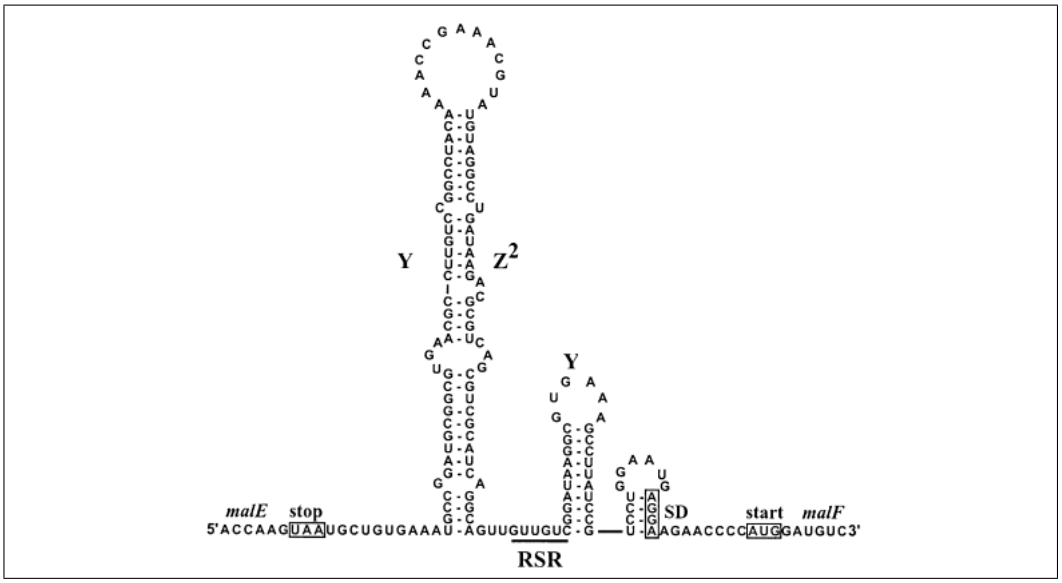

FIGURE 2 – **Structure ARN des REP au sein de l'operon malEFG.** Y, Z^2 et Y indiquent la séquence des REP dans l'espace intergénique de malE-malF. Bien que Y et Z^2 puissent former des structures tige-boucles par elles mêmes, elles s'apparient ensemble pour former une région étendue en grande partie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

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) . 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). 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`. Les 2 WT et 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 binaire `Binary Alignment Map format (BAM)`, 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 de gagner de l'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
# de qualite.
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_noDup.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 (source RegulonDB), les REP et BIME (source laboratoire) ainsi que les terminateurs de transcription (source DOOR) 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 des fichiers BED nécessaire 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.

Analyse statistique

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 pourrait supposer un profil d'expression similaire*, comme c'est le cas dans les opérons par exemple. L'analyse statistique a été menée sur le logiciel R.

Création de la table de comptages par gène

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 du fichier GFF, 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) (Figure 3) (Lawrence et al. 2013). Une table de comptage est alors produite, les transcrits figurant en ligne et les fichiers BAM en colonnes .

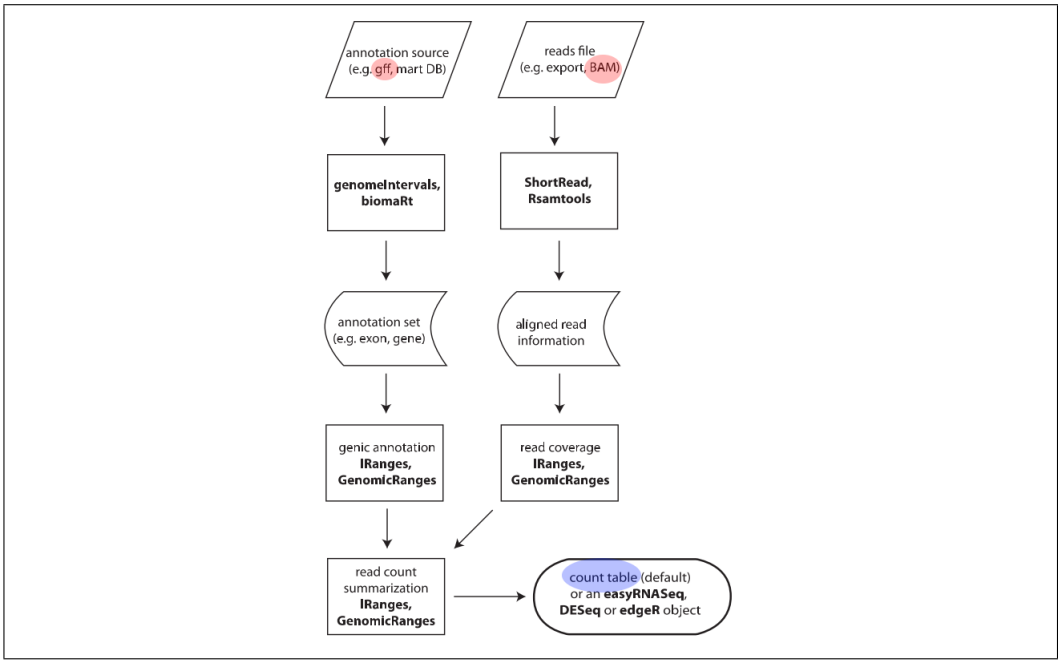

FIGURE 3 – **easyRNASeq : création d'une table de comptage.** Processus de création de la table de comptage grâce à l'union des intervalles génomiques. Les formats d'entrées de notre analyse sont surlignés en rouge, le format de sortie en bleu. (Delhomme et al. 2012).

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

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. 8

GFF General Feature Format. 7

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

reads Séquence nucléotidique issue d'un séquençage NGS. 5

REP Repeated Extragenic Palindrome. 1

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.

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.

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.

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. 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.

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.

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.