Crédits photographiques

Photographie de l'image de couverture reproduite avec permissions depuis Wikimedia Commons, domaine public, accès libre.
`https://commons.wikimedia.org/wiki/File:Dark_side_of_the_moon.jpg`

Sommaire

Abbréviations

BER Base Excision Repair

gBGC GC–biased gene conversion : conversion génique biaisée vers GC

GC% Contenu en GC

VSP Very Short Patch repair : réparations courtes portées

Table des figures

Liste des tableaux

Introduction

Chez les procaryotes, le taux de guanine et cytosine (GC%) varie de 16,5 % à 75 % dans les génomes séquencés à ce jour. Une telle amplitude soulève plusieurs questions. Celles qui nous intéressent concernent les mécanismes responsables de ces variations : est-ce qu'ils sont associés à la réplication de l'ADN, à sa réparation ou à une pression de sélection sur l'usage du code génétique ? Au cours des quinzes dernières années, il a été démontré que la recombinaison homologue tend à augmenter localement le taux de GC dans les régions fortement recombinantes des génomes eucaryotes[6,17,26]. Ce mécanisme est essentiel pour la cellule : il permet de réparer les lésions de l'ADN. Au cours de la recombinaison, les brins associés au sein d'un hétéroduplex peuvent présenter des mésappariemments. Leur correction entraîne de la *conversion génique*[2] : c'est une transmission à sens unique de l'information portée par un brin vers l'autre. Ce mécanisme est biaisé vers l'introduction préférentielle des bases C et G chez les mammifères et probablement chez un grand nombre d'eucaryotes[20]. C'est un processus non adaptatif : il impacte les régions codantes et non-codantes. Il peut s'opposer à l'action de la sélection en augmentant la probabilité de fixation des allèles G et C[22,10].

En 2010, deux études simultanées[13,11] démontrent que : 1) le patron de *mutation* est universellement biaisé vers A et T chez les procaryotes, et 2) les bases G et C ont une probabilité de fixation plus élevée, probablement sous l'effet d'un processus à l'action semblable à celle de la sélection naturelle. Hildebrand *et al.* avancent que le GC% est en soi un trait soumis à la sélection, et rejettent l'hypothèse d'un biais de conversion génique biaisé vers GC chez les procaryotes. Cependant, les analyses récentes de Lassalle *et al.*[15] et de Yahara *et al.*[27] ont montré que les zones recombinantes des génomes procaryotes ont un taux de GC plus élévé que les régions non-recombinantes, une observation qui correspond précisément aux prédictions du biais de conversion vers GC (gBGC). Elles montrent également que la fixation préférentielle des allèles GC va à l'encontre de la fixation des allèles optimaux des codons, une signature caractéristique de l'action d'un processus non adaptatif tel que le gBGC. Ces observations sont compatibles avec l'hypothèse d'une conversion génique biaisée vers GC chez les procaryotes.

L'objectif de ce travail était d'obtenir une estimation expérimentale des fréquences de conversion vers GC chez la bactérie naturellement transformable *Acinetobacter baylyi ADP1*[*]. En effet, cette bactérie développe un stade de compétence pendant la phase exponentielle de

*. *A.baylyi* était appelée *A.calcoaceticus* avant 1995, mais la taxonomie du genre a fait l'objet d'une révision[8]

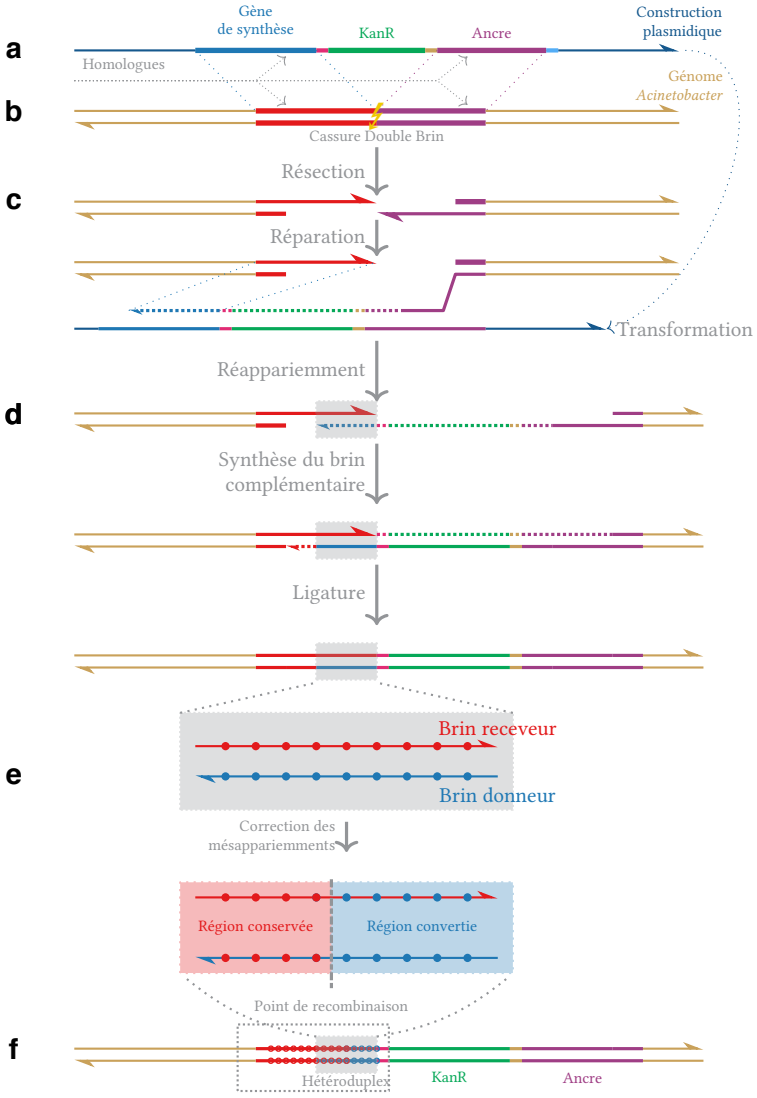


Figure 1 : **Mécanismes de conversion génique au cours de la transformation naturelle par les constructions plasmidiques**

Les constructions plasmidiques que nous avons conçues (a) comportent un gène de synthèse introduisant des mésappariemments avec la séquence sauvage receveuse, une cassette de résistance à un antibiotique (KanR) et une zone 100% homologue qui permet d'augmenter la fréquence de transformation.

Lors d'une cassure double brin (b), la région en 3' subit une résection, qui est réparée sur la base d'une matrice d'ADN homologue (c). La transformation naturelle permet de réparer ces cassures en utilisant l'ADN exogène sous forme simple brin. Le réappariemment (d) avec le brin parental conduit à la formation d'un hétéroduplex (zone grisée) entre le donneur (en bleu) et le receveur (en rouge). Les mésappariemments au sein de cet hétéroduplex sont réparés (e), du donneur vers le receveur dans la région convertie, du receveur vers le donneur dans la région conservée.

Nous avons séquencé la région encadrée (f) chez un grand nombre de recombinants.

croissance, et intègre de l'ADN exogène sous forme simple brin[1] par un mécanisme contrôlé génétiquement. Cet ADN peut être intégré dans le génome par recombinaison homologue. Nous avons utilisé ces propriétés pour forcer la recombinaison homologue à un locus neutre, en introduisant artificiellement des mésappariemments entre la séquence donneuse — nos constructions — et la séquence receveuse — le génome (voir figure 1). La réparation de ces mésappariemments donne lieu à la conversion génique d'un brin par l'autre. On distingue donc la région convertie, dont le génotype correspond à celui de l'haplotype donneur, de la région conservée, dont le génotype correspond à celui de l'haplotype receveur. Elles sont séparées par le point de recombinaison (voir figure 2). Les mésappariemments présents sur la séquence donneuse permettent 1) de localiser la position du point de recombinaison et 2) de déterminer la sens de la conversion. La réparation des mésappariemments introduits nous renseigne sur les fréquences de conversion en faveur de GC chez *A.baylyi*.

insérer débuts de conclusions ici

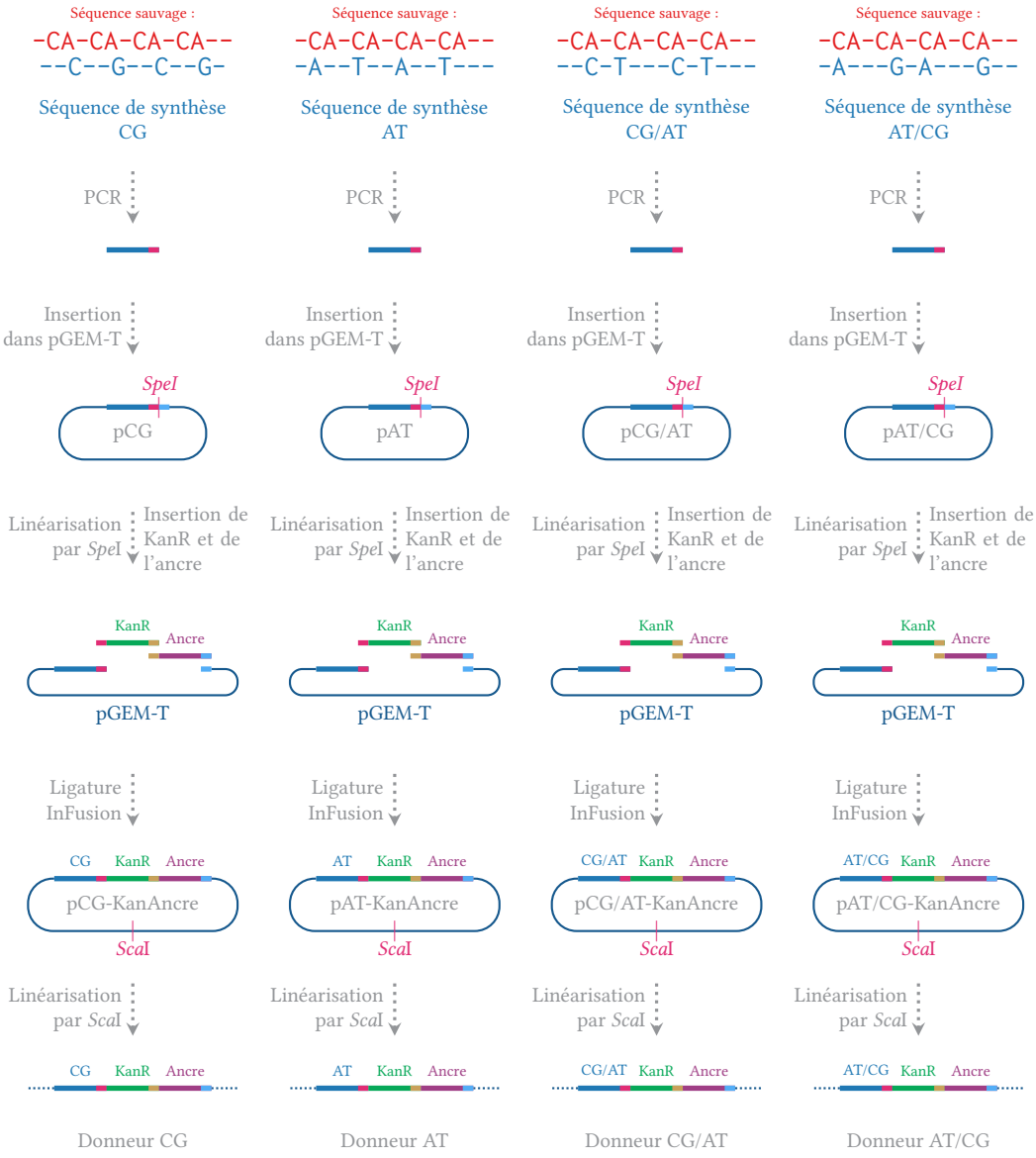


Figure 2 : **Constructions moléculaires et préparation de l'ADN recombinant**
Les gènes de synthèse sont amplifiés par PCR spécifique. Les fragments sont ensuite ligaturés dans le plasmide pGEM-T. Les plasmides obtenus sont linéarisés par digestions enzymatiques *Spe*I au niveau des sites d'insertion de la cassette de résistance à la kanamycine et de l'ancre. Ces fragments sont préalablement amplifiés par PCR en utilisant des amorces porteuses de régions 3' flottantes, complémentaires des extrémités des fragments voisins. La ligature entre les trois fragments obtenus est réalisée à l'aide du kit InFusion. Les constructions plasmidiques obtenues sont amplifiées par culture, extraites et linéarisées par digestion enzymatique *Sca*I.

1 Matériels & Méthodes

1.1 Conception des séquences de synthèse

Pour étudier la correction des mésappariemments au cours de la recombinaison homologue chez *A. baylyi*, nous avons conçu des séquences permettant d'introduire ces mésappariemments. Nous avons choisi 23 sites sur un locus neutre du génome, séparés par 30 paires de bases*. Ce locus est connu au laboratoire pour permettre de fortes efficacités de transformations. Il est localisé à la position 47184 du génome de référence. Il code putativement une 3-oxoacyl-ACP reductase, impliqué dans la synthèse des acides gras[25].

Les sites choisis sont des dinucléotides alternant une base GC avec une base AT. Le gène de synthèse GC introduit des mésappariemments GC aux sites AT (GC | AT), le gène de synthèse AT introduit des mésappariemments AT aux sites GC (AT | GC) (voir figure 2). Pour s'affranchir d'un possible effet dû à l'écart de GC% entre la séquence synthétique et la séquence sauvage, nous avons également conçu deux séquences qui introduisent en alternance un mésappariemment GC | AT et AT | GC. Le gène de synthèse CG/AT introduit d'abord un mésappariemment GC | AT, puis un mésappariemment AT | GC, le gène de synthèse AT/CG suit l'ordre inverse. (voir figure 2). Les quatre séquences ont été synthétisées par ThermoFischer (Waltham, États-Unis).

1.2 Constructions des plasmides

De façon à pouvoir sélectionner les clones recombinants au locus d'intérêt, nous avons construit les plasmides représentés dans la figure 2. Le gène synthétique est associé avec une cassette de résistance à la kanamycine, ainsi qu'une région "ancre". Cette région recombinogène est complètement homologue à la séquence en 3' du locus d'intérêt et permet d'augmenter les fréquences de transformation[5,19].

Les gènes synthétiques ont d'abord été amplifiés par PCR avec les amorces 1392 et 1393 (voir annexe 1.2), puis ligaturés dans le plasmide pGEM-T. Les plasmides obtenus ont été insérés dans la souche *E.coli* OneShot® `TOP-10` (Invitrogen, Carlsbad, États-Unis), suivant le protocole

*. Ils introduisent 3% de divergence avec la séquence sauvage. Un plus grand nombre de marqueurs diminue les fréquences de recombinaison ; un nombre inférieur diminue la qualité du signal.

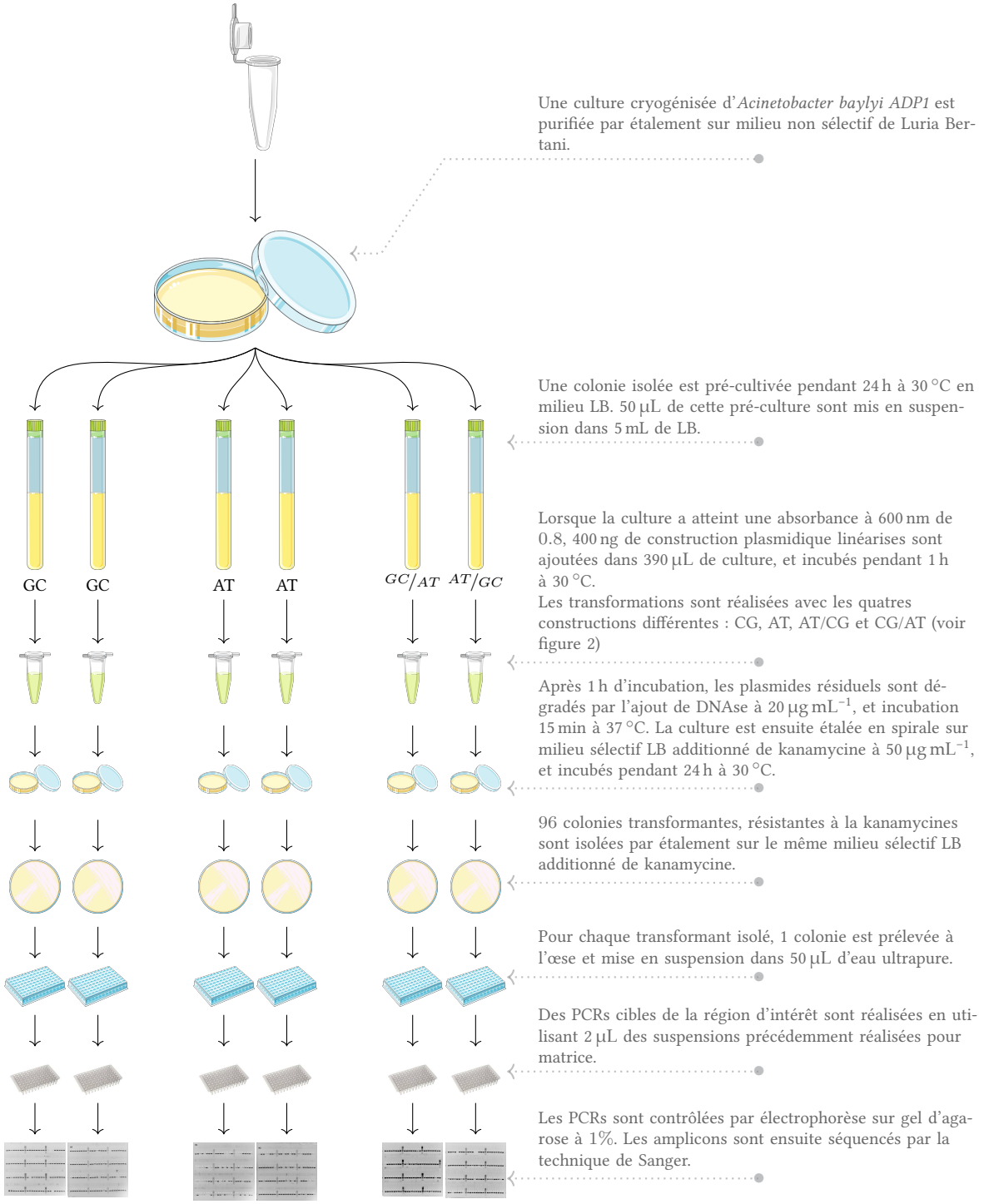


Figure 3 : **Protocole de transformation et d'obtention des amplicons des zones de recombinaison.**

du fabriquant. Le sens d'insertion du gène synthétique dans le plasmide a été vérifié par PCR avec les amorces M13R et 1392 (voir annexe 1.2), et l'absence de mutation a été vérifié par séquençage (GATC Biotech, Constance, Allemagne). Les plasmides des clones validés par séquençage ont été extraits par le kit Nucleospin-Plasmid (Macherey-Nagel, Düren, Allemagne), et linéarisés par l'enzyme *Spe*I (ThermoFischer).

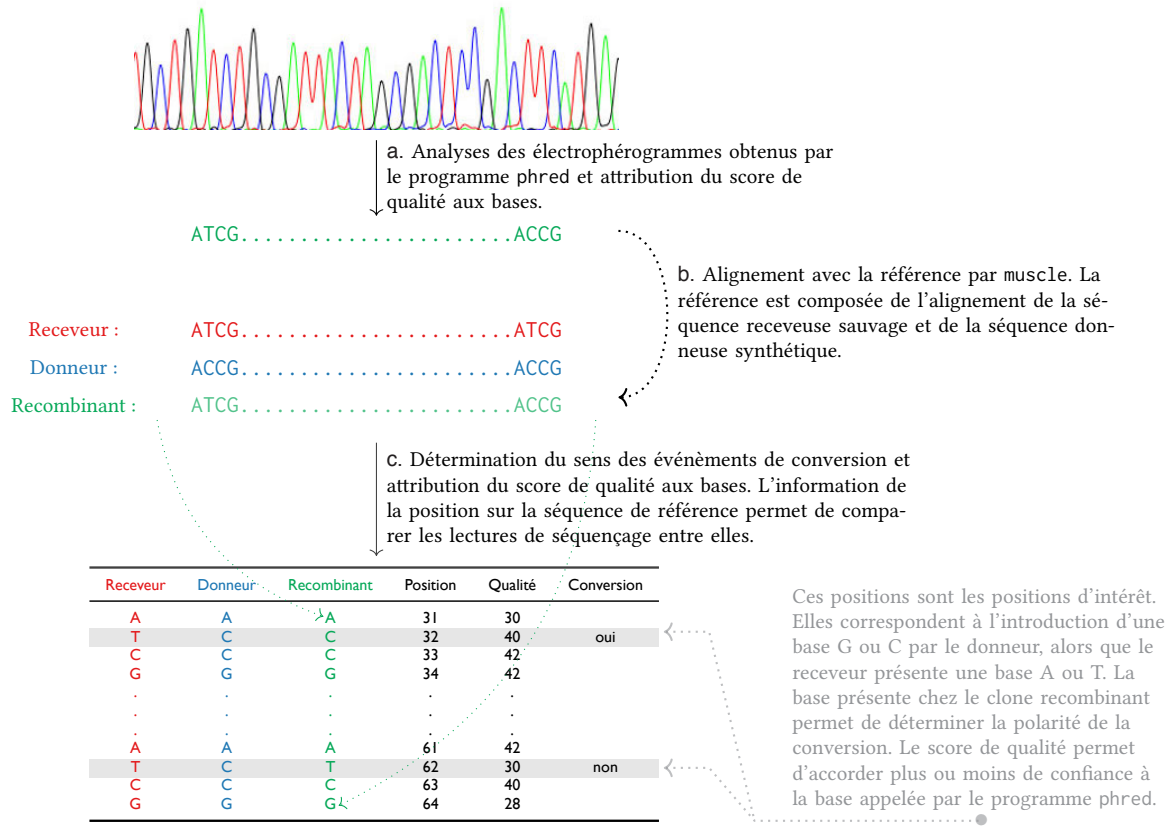
La cassette de résistance à la kanamycine *aphA*3 et l'ancre ont été respectivement amplifiés par PCR avec le couple d'amorce 1408 / 1409 et 1410 / 1411. Les deux amplicons obtenus ont été ligaturés dans le plasmide pGEM-T porteurs des gènes de synthèse. La ligature a été réalisée simultanément par le kit InFusion (Takara Clontech, Saint Germain en Layes, France). Le produit de ligature obtenu a été inséré dans la souche optimisée pour la chimio-compétence *E.coli* Stellar (Takara Clontech). Les transformants ont été séléctionnés sur milieu LB solide additionné d'ampicilline à $75\,\mu g\,mL^{-1}$ de kanamycine à $50\,\mu g\,mL^{-1}$. Les transformants ont été confirmés par PCR spécifique de l'insert avec les amorces 1393 et 1411 (voir annexe 1.2).

1.3 Transformations d'*Acinetobacter baylyi*

$1\,\mu g$ de plasmide a été extrait et linéarisé par l'enzyme *Sca*I (ThermoFisher). L'enzyme a été ensuite inactivée par incubation $10\,min$ à $70\,°C$. $390\,\mu L$ d'une culture pure d'*Acinetobacter baylyi* ADP1, avec une absorbance de $0{,}8$ à $600\,nm$ ont été incubés pendant $1\,h$ à $28\,°C$ en présence de $200\,ng$ de plasmide linéarisé. Les suspensions ont été ensuite incubées $15\,min$ à $37\,°C$ en présence de DNAse à $20\,\mu mol\,L^{-1}$ pour éliminer les plasmides résiduels. Les cellules recombinantes ont été sélectionnées par étalement en spirales (InterScience, St Nom la Bretêche, France) sur milieu LB solide additionné de kanamycine à $50\,\mu g\,mL^{-1}$ et incubées $24\,h$ à $30\,°C$. Les dénombrements ont été effectués via un compteur automatique Scan®1200 (InterScience). Les fréquences de transformations F ont été calculées de la façon suivante :

$$\begin{aligned} F &= \frac{\text{Nombre de transformants}}{\text{Nombre de cellules réceptrices totales}} \\ F &= \frac{\text{Nombres de cellules résistantes à la kanamycine}}{\text{Nombre de cellules totales}} \end{aligned} \tag{1}$$

96 clones recombinants ont été isolés et incubés $24\,h$ à $30\,°C$ sur milieu LB solide additionné de kanamycine à $50\,\mu mol\,mL^{-1}$. Une colonie isolée par clone a été mise en suspension dans $50\,\mu L$ d'H_2O ultra-pure. $2\,\mu L$ de ces suspensions ont servi de matrice pour amplifier par PCR les régions recombinantes avec la Taq polymérase haute fidélité Phusion (ThermoFischer, Waltham,



Receveur	Donneur	Recombinant	Position	Qualité	Conversion
A	A	A	31	30	
T	C	C	32	40	oui
C	C	C	33	42	
G	G	G	34	42	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
A	A	A	61	42	
T	C	T	62	30	non
C	C	C	63	40	
G	G	G	64	28	

Figure 4 : **Exemple d'analyse de la zone de recombinaison pour un clone transformant**
Les électrophérogrammes de séquençage obtenus ont été analysés en utilisant un programme permettant d'attribuer à chaque position un score de qualité (a). Les séquences obtenues ont été alignées à la référence (b), ce qui a permis d'inférer pour chaque site polymorphe le sens de la conversion du recombinant (c).

États-Unis). Les amplicons ont été vérifiés par migration sur gel d'agarose à 1% et séquencés par la technique de Sanger[24] (GATC Biotech).

1.4 Alignements

Les spectrogrammes de séquençage reçus au format propriétaire `abi` (Applied Biosystem, Foster City, États-Unis) ont été analysés par le programme `phred`[9] et convertis en format universel `FASTA` (voir figure 4). Les séquences obtenues ont été alignées aux références par `muscle v3.8.31`[7]. Les références en question correspondent à la séquence sauvage et la séquence du gène de synthèse, respectivement *receveur* et *donneur* de l'évènement de recombinaison. Un programme Python[3] a été développé pour analyser les alignements obtenus. Il détermine les positions des SNPs d'intérêt dans l'alignement de référence et infère le génotype du clone séquencé. Les alignements par paire en colonne obtenus ont été analysés par `R 3.2.3`[21]. Les programmes développés sont accessibles à l'adresse `https://github.com/sam217pa/gbc-seq_mars`. Les données formattées et les fonctions d'analyse ont été assemblées dans le package `R gcbiasr` disponible à l'adresse `https://github.com/sam217pa/gbc-gcbiasr`.

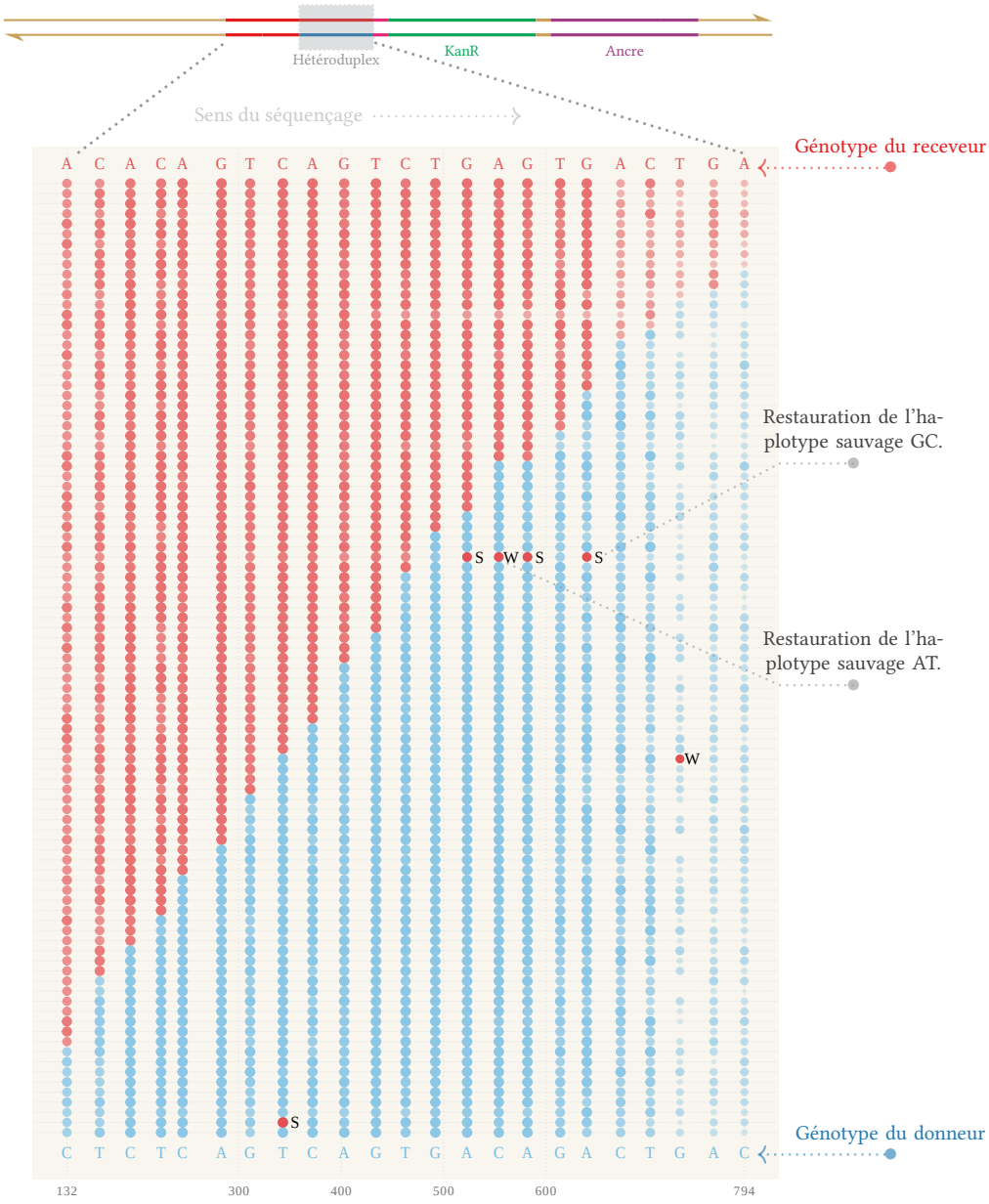


Figure 5 : **Zones de recombinaison entre un locus génomique d'*Acinetobacter baylyi* et un gène synthétique donneur d'allèles CG et AT.**
Chaque ligne horizontale représente une séquence. Les points représentent les positions des marqueurs sur les séquences. L'intensité et le diamètre des points représentent le score de qualité du site. Les points sont bleus lorsque le site est dans la région convertie : ils correspondent à l'haplotype du donneur. Ils sont rouges lorsque le site est dans la région conservée. Les séquences sont triées par longueur de région convertie. Les alternances rouge / bleu marquent la transition de l'haplotype converti à l'haplotype sauvage : le point de recombinaison est localisé entre ces deux marqueurs.

Table 1 : **Fréquences de transformation**

Construction Donneuse	Fréquences de transformation
CG	$6{,}53 \times 10^{-5}$
AT	$2{,}42 \times 10^{-5}$
AT/CG	$4{,}97 \times 10^{-5}$
CG/AT	$1{,}74 \times 10^{-4}$

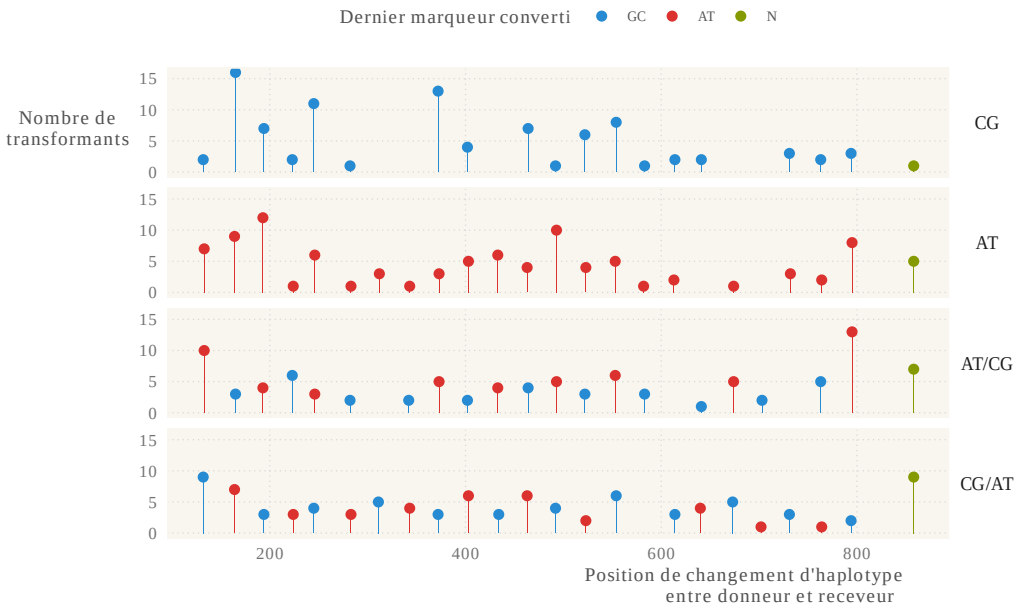


Figure 6 : **Distribution de la position du dernier marqueur converti**
Ce graphique représente en ordonnées le nombre de transformants dont le dernier marqueur converti est à la position représentée en abscisse. Les panneaux du graphique représentent les quatres constructions donneuses. La position du dernier marqueur converti indique la position du point de recombinaison. Les transformants qui ne montrent aucun marqueur converti sont indiqués par des N.

2 Résultats

Nous avons transformé une supension d'*Acinetobacter* par des constructions dont l'intégration dans le génome par recombinaison homologue entraîne la réparation des mésappariemments, un mécanisme qui est biaisé vers l'introduction des bases CG chez les eucaryotes.

Les fréquences de transformations obtenues sont représentées dans le tableau 1 page 5. Le type de construction donneuse n'a pas d'influence sur l'efficacité de la transformation. Les fréquences de l'ordre de 1×10^{-5} ont permis d'obtenir un grand nombre de recombinants.

La figure 5 représente le détail des zones de recombinaison obtenues avec une construction donneuse alternant CG et AT. Les zones de recombinaisons des clones transformés par les donneurs CG, AT et AT/CG sont détaillées en annexe 1.4. En moyenne, 393 ± 228 nucléotides ont été transférés et intégrés dans le génome. La position du dernier marqueur converti indique que le point de recombinaison (voir figure 1 et 5) se situe entre les 30 bases le séparant du premier marqueur conservé. Nous avons donc estimé la position du point de recombinaison par celle du dernier marqueur converti, dont la distribution est représentée dans la figure 6. De façon surprenante, la distribution du point de recombinaison est très variable entre les transformants.

Nous nous sommes intéressés à deux paramètres permettant d'estimer les fréquences de conversion en faveur de GC chez *A. baylyi* : la position du point de recombinaison, et la correction ponctuelle des mésappariemments.

2.1 Comparaison de la longueur des régions converties

Selon l'hypothèse gBGC, la région convertie devrait être plus longue lorsque la construction donneuse induit des réparations vers CG que lorsqu'elle induit des réparations vers AT. La différence entre la longueur moyenne de région convertie par les donneurs respectivement CG et AT n'est pas significative (test de Wilcoxon, probabilité critique $= 0{,}31$) (voir figure 6). De la même façon, la différence entre la longueur de région convertie par le donneur AT/CG et celle par le donneur CG/AT n'est pas significative (test de Wilcoxon, probabilité critique $= 0{,}22$). Globalement, le type de construction donneuse n'explique pas la variabilité de la longueur de région convertie (test de Kruskal-Wallis, probabilité critique $= 0{,}10$). Le type de donneur n'a pas d'influence sur la longueur de région convertie.

Table 2 : **Dénombrements des derniers marqueurs avant le point de recombinaison.**

ADN synthétique donneur	Dernier marqueur converti	
	AT	CG
AT/CG	55	33
CG/AT	37	50
Total	92	83

Table 3 : **Dénombrement des cas de restauration**

ADN synthétique donneur	Nucléotide Restauré	
	AT	CG
AT	-	4
CG	0	-
AT/CG	3	2
CG/AT	1	4
Total	4	10

2.2 Distribution du dernier marqueur converti

Les constructions alternant AT et CG permettent de déterminer si le point de recombinaison se situe plus souvent après un marqueur introduisant une conversion vers CG qu'après un marqueur introduisant une conversion vers AT (voir table 2). Le dernier marqueur converti est AT dans 92 transformants ; il est CG dans 83 transformants. Cet écart n'est pas significatif (test du χ^2 d'homogénéité, probabilité critique $= 0{,}49$) (voir table 2).

Il semblerait que lorsque le premier marqueur donneur est AT, le dernier marqueur converti est plus souvent AT que GC. De la même façon, lorsque le premier marqueur donneur est GC, le dernier marqueur converti est plus souvent GC que AT.

2.3 Restaurations de l'haplotype sauvage

Certains transformants montrent des régions de conversions qui alternent entre l'allèle sauvage receveur et l'allèle donneur. Ces alternances ponctuelles affectent de 1 à 3 marqueurs consécutifs (voir figure 5 et figure 12 en annexe 1.4 page iv). Nous avons confirmé qu'il s'agissait bien de restaurations de l'allèle sauvage de deux façons. 1) Expérimentalement, nous avons séquencé une sous-population de 30 clones issus d'un isolat séquencé en premier lieu. Tous montrent la même alternance au même site (voir figure 11 en annexe 1.5). 2) Analytiquement, le score de qualité moyen des sites montrant des restaurations de l'allèle sauvage permet de s'affranchir d'une possible erreur de séquençage : celle-ci se traduit généralement par un indice de qualité plus faible au site concerné. Le score de qualité moyen est de $49{,}36$ aux sites restaurés, contre $52{,}79$ aux sites non-restaurés. La différence entre les deux n'est pas significative (test de Wilcoxon, probabilité critique $= 0{,}94$). Les marqueurs correspondant à des restaurations de l'haplotype sauvage ne sont donc pas des erreurs de séquençage, et correspondent à un signal biologique.

Sur les 14 cas de restaurations de l'haplotype sauvage, 4 sont des restaurations de l'allèle AT, 10 sont des restaurations de l'allèle CG. Cet écart n'est pas significatif (voir la table 3). L'écart n'est pas statistiquement significatif (test du χ^2 d'homogénéité, probabilité critique $= 0{,}11$).

2.4 Estimation des fréquences de conversion en faveur de GC

Nous avons estimé les fréquences de conversion en faveur de GC par la différence entre le taux de GC des marqueurs des clones recombinants et celui du receveur. En l'absence d'un biais, il est estimé à 0 sur l'ensemble des transformants séquencés.

Nous avons échantillonné 80 séquences par type de construction donneuse, sélectionné uniquement les positions des marqueurs, mesuré la différence entre le taux de GC du receveur et le taux de GC du clone recombinant et moyenné cette différence sur l'ensemble des clones. En repétant cette mesure 1×10^4 fois, la différence moyenne entre le taux de GC des recombinants et le taux de GC des receveur est de $1{,}13\%$.

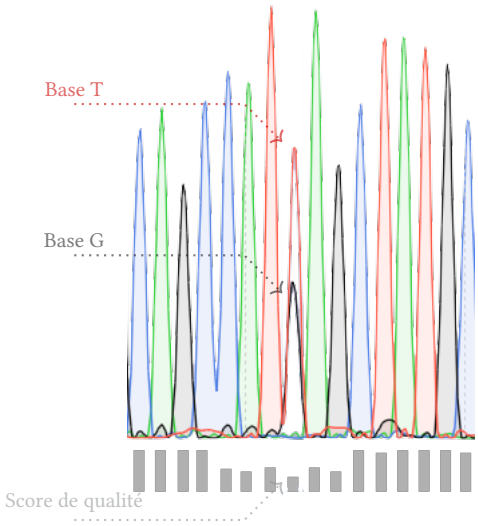


Figure 7 : **Exemple de marqueur montrant des traces de contaminations**
Cet électrophérogramme montre les bases autour du marqueur à la position 200. Dans une région de qualité moyenne élevée (bases en 5' et en 3'), le marqueur présente une trace de contamination par une autre base. La base déterminée est la base T mais une base G est présente dans la population d'amplicon séquencée.

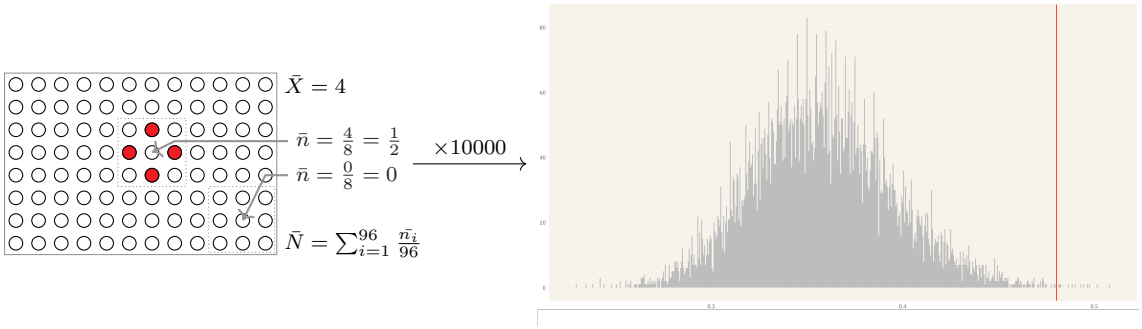


Figure 8 : **Des contaminations dues au hasard ?**
Par plaque de 96 puits, nous avons déterminé X le nombre de puits dont la séquence montre des traces de pics secondaires (voir figure 7) et mesuré $\bar{n}$ la moyenne du nombre de puits voisins contaminés. $\bar{N}$ est la moyenne des 96 $\bar{n}$ obtenus par plaque. Nous avons simulé 1×10^4 plaques avec X puits contaminés répartis aléatoirement, mesuré $\bar{N}$ et comparé la valeur expérimentale de $\bar{N}$ (trait vertical rouge) avec la distribution des 1×10^4 $\bar{N}$ (en gris) Seules $78/10000$ plaques simulées montrent un $\bar{N}$ supérieur à la valeur expérimentale : la répartition des séquences contaminées dans les plaques ne peut pas être attribuée au hasard.

3 Discussions

Nous avons séquencé un grand nombre de clones d'*A. baylyi* recombinants à un locus choisi de façon à introduire la correction des mésappariemments, un processus qui est biaisé vers l'introduction des bases C et G chez certains eucaryotes.

Des erreurs de séquençage sans conséquences ?

Certains marqueurs sont d'une qualité inférieure au reste des marqueurs sur la lecture. C'est la marque de pics secondaires sur les électrophérogrammes (voir figure 7). Des pics secondaires apparaissent quand la population d'amplicon séquencée n'est pas homogène au site considéré. De façon surprenante, les deux pics présents à un marqueur donné correspondent toujours soit à la base sauvage, soit à la base synthétique introduite. Cette donnée peut être interprétée de deux façons.

1) S'il s'agit de la marque d'un signal biologique, la colonie dont la zone de recombinaison a été séquencée montre une hétérogénéité au marqueur considéré. Autrement dit, une part de la population séquencée a converti la base, l'autre partie l'a conservée. Nous avons séquencé à nouveau 31 isolats issus d'un clone séquencé en premier lieu qui montrait des pics secondaires. Aucun des sous-clones ne montrent l'allèle correspondant au pic secondaire (voir figure 11).

2) S'il s'agit d'un erreur de séquençage, la population d'amplicon séquencée est hétérogène à cause de contaminations entre les puits des plaques, qui ont pu avoir lieu au cours du séquençage ou au cours des PCRs. Dans ce cas, la répartition dans la plaque des puits dont la séquence montre des traces de contamination ne devrait pas être déterminée que par le hasard. Nous avons montré par simulation qu'elles ne l'étaient pas (voir figure 8). Les séquences montrant des pics secondaires sont plus souvent voisines avec une autre séquence affectée que si elles étaient réparties aléatoirement. En conséquence, nous avons filtré de façon très stringente les sites de faible qualité.

3.1 Une distribution des points de recombinaison surprenante

L'hypothèse gBGC prédit que la région convertie devrait être plus longue lorsque la conversion introduit des bases GC que lorsqu'elle introduit des bases AT. Nous n'avons pas détecté de différences significatives. En fait, la distribution du point de recombinaison est assez uniforme,

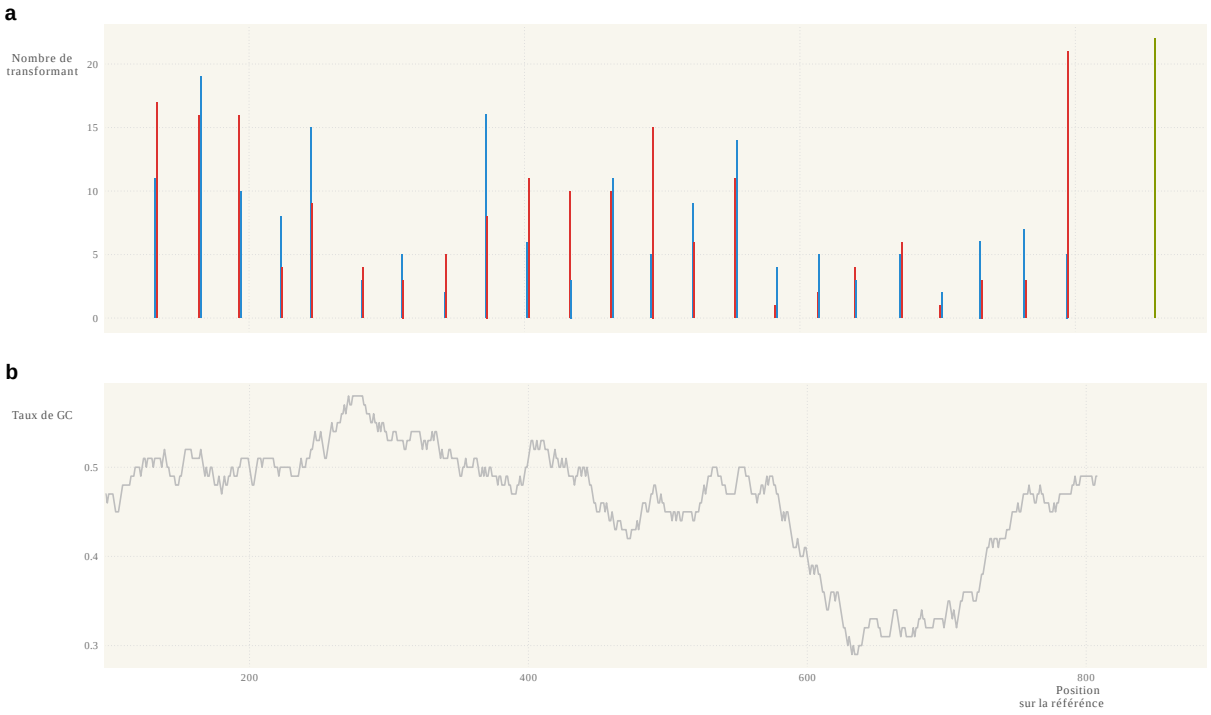


Figure 9 : Taux de GC
TODO légender

contrairement à ce qui est décrit par Yáñez-Cuna *et al.*[28]. Chez un modèle de *Rhizobium etli*, la distribution de la longueur des régions converties est bimodale, avec un pic à respectivement 120 et 600 bp. Comment peut-on expliquer ces variations ?

Une influence du taux de GC local ?

La figure 6 montre un point froid de recombinaison dans la région entre 600 et 800 bp, environ 200 paires de bases après l'origine de l'hétéroduplex. Nous avons supposé que le taux de GC local diminuait localement l'efficacité de la recombinaison : un taux de GC plus faible diminue le nombre de liaisons hydrogène entre les brins et pourrait donc diminuer les probabilités d'appariemment avec un brin d'ADN exogène. En effet, lors de la recherche par la protéine RecA d'une matrice homologue permettant de réparer la lésion, la reconnaissance est pûrement basée sur l'appariemment Watson-Crick entre les brins[16]. Nous avons donc représenté le taux de GC moyen par fenêtre de 50bp au long de la séquence sauvage (voir figure 9b).

Le point froid de recombinaison (fig.9a) semble bien correspondre à un taux de GC local plus faible (fig.9b). Un taux de GC plus faible pourrait donc conduire au rejet du brin homologue, et diminuer les fréquences de recombinaison localement. C'est une observation qui va à l'encontre de la théorie du gBGC : si un taux d'AT élevé implique une faible fréquence de recombinaison, il diminue les probabilités d'introduire des bases GC par conversion génique. À l'inverse, les zones riches en GC feraient plus souvent l'objet de conversion, ce qui conduirait à l'effacement progressif du pic local de GC : c'est le paradoxe des points chauds de recombinaison[4].

Lieb *et al.*[18] ont montré chez *E. coli* que les enzymes du VSP (Very Short Patch repair) répairaient un mésappariemment dans un sens spécifique en fonction des bases au voisinage immédiat du mésappariemment.

TODO terminer analyses des dinucléotides.

TODO mal dit “Si ce genre de mécanisme” est à l'œuvre chez *A. baylyi*, les marqueurs AT et les marqueurs GC de nos constructions ne sont pas rigoureusement dans le même contexte nucléotidique. Pour palier à cet éventuel biais, il faudrait transformer à nouveau un clone ayant converti tous les marqueurs par la séquence sauvage : chaque évènement de conversion serait alors rigoureusement dans le même contexte.

3.2 Des restaurations de l'haplotype sauvage inattendues

Au sein des régions converties, 14 cas de restauration de l'allèle sauvage ont été détectés. Quels mécanismes peuvent les expliquer ? Ce ne sont probablement pas des mutations spontanées : bien que la recombinaison soit un processus mutagène en soi[23,12], les restaurations sont retrouvées spécifiquement aux positions des marqueurs, et correspondent précisément à l'allèle sauvage. Une mutation spontanée pourrait introduire aléatoirement l'une des quatre bases. S'il s'agit de l'action de la machinerie de correction des mésappariemments, ces cas sont inattendus : pourquoi la machinerie de correction des mésappariemments restaurerait spécifiquement ces marqueurs dans une région qui présente un mésappariemment toutes les 30 paires de bases ?

Dans le contexte de la conversion génique biaisée vers GC, est-ce que ces cas de restaurations pourraient expliquer un biais ? Chez les eucaryotes, la correction des mésappariemments dans les cellules en mitose est effectuée par la voie du BER (Base Excision Repair). Cette voie excise spécifiquement une base mésappariée en détectant les malformations qu'elle occasionne dans la structure de la double-hélice, puis la remplace par la base complémentaire à celle du brin opposé[14]. Cette voie a été considérée comme l'un des moteurs possibles du gBGC chez les mammifères[6], mais elle a été rejetée chez la levure[17]. Chez cette dernière, le gBGC est associé aux régions converties les plus longues et dans lesquelles tous les allèles sont convertis depuis le même haplotype donneur.

Est-ce que le BER, s'il est actif chez les procaryotes, pourrait conduire à de la conversion biaisée vers GC ? L'étude des régions converties chez des mutants d'*A. baylyi* délétés pour les fonctions clés de la correction des mésappariemments (ΔMutS, ΔMutH, ΔMutL) devrait permettre de répondre à cette question.

Ces cas de restaurations de l'haplotype sauvage ajoutent une incertitude sur la nature des premiers marqueurs conservés (en rouge sur la figure 5). En effet, ces marqueurs peuvent être la résultante de trois choses : 1) ils ne font pas partie de l'hétéroduplex, et ne sont pas soumis à de la conversion, 2) ils font partie de l'hétéroduplex, mais la conversion génique conserve l'allèle sauvage ou 3) ils ont été d'abord convertis par l'allèle donneur, puis restaurés. Dans tous les cas, la base séquencée correspond à l'allèle sauvage du marqueur.

3.3 Comment augmenter la puissance du test ?

Pour détecter une différence de l'ordre de 30 paires de bases entre longueurs de régions converties correspondant à l'introduction d'un marqueur supplémentaire, en supposant qu'elles sont normalement distribuées, avec l'écart type observé dans nos résultats (227), avec une puissance de $0{,}8$ et un niveau de confiance de $0{,}05$, il faudrait une taille d'échantillon de 717 (test t) par groupe.

Cette taille d'échantillon n'est pas envisageable en utilisant les techniques actuelles : le séquençage par la technique de Sanger implique un coût prohibitif pour une telle taille d'échantillon. Nous avons étudié différents procédés permettant de séquencer les régions converties à haut débit. Les séquençage à haut débit permettrait de réduire les coûts de séquençage. Il nécessite néanmoins de prendre en compte les facteurs suivants. 1) La taille des lectures est de l'ordre de 2×300 paires de bases par lecture, soit un total de 600pb, inférieure à la taille du locus considéré. 2) Lors de la construction de la librairie de séquence, les amplicons sont mélangés les uns aux autres. Il faut pouvoir les discriminer les uns des autres facilement, de façon à individualiser les évènements de recombinaison. 3) La divergence très faible des amplicons peut introduire des artéfacts lors de la capture des clusters par la caméra de séquençage.

Conclusion

Références

[1] **Chen, I. et D. Dubnau**. 2004. DNA uptake during bacterial transformation. Nat Rev Micro **2** :241–249.

[2] **Chen, J.-M., D. N. Cooper, N. Chuzhanova, C. Férec, et G. P. Patrinos**. 2007. Gene conversion : mechanisms, evolution and human disease. Nat Rev Genet **8** :762–775.

[3] **Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, et M. J. L. d. Hoon**. 2009. Biopython : freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25** :1422–1423.

[4] **Coop, G. et S. R. Myers**. 2007. Live Hot, Die Young : Transmission Distortion in Recombination Hotspots. PLoS Genet **3** :e35.

[5] **de Vries, J. et W. Wackernagel**. 2002. Integration of foreign DNA during natural transformation of *Acinetobacter sp.* by homology-facilitated illegitimate recombination. Proceedings of the National Academy of Sciences **99** :2094–2099.

[6] **Duret, L. et N. Galtier**. 2009. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annual Review of Genomics and Human Genetics **10** :285–311.

[7] **Edgar, R. C.** 2004. MUSCLE : multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. **32** :1792–1797.

[8] **Euzéby, J. P.** 1997. List of Bacterial Names with Standing in Nomenclature : a folder available on the Internet. International Journal of Systematic and Evolutionary Microbiology **47** :590–592.

[9] **Ewing, B., L. Hillier, M. C. Wendl, et P. Green**. 1998. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. Genome Res. **8** :175–185.

[10] **Galtier, N., L. Duret, S. Glémin, et V. Ranwez**. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends in Genetics **25** :1–5.

[11] **Hershberg, R. et D. A. Petrov**. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet .

[12] **Hicks, W. M., M. Kim, et J. E. Haber**. 2010. Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. Science **329** :82–85.

[13] **Hildebrand, F., A. Meyer, et A. Eyre-Walker**. 2010. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet **6** :e1001107.

[14] **Krokan, H. E. et M. Bjørås**. 2013. Base Excision Repair. Cold Spring Harb Perspect Biol **5** :a012583.

[15] **Lassalle, F., S. Périan, T. Bataillon, X. Nesme, L. Duret, et V. Daubin**. 2015. GC-Content Evolution in Bacterial Genomes : The Biased Gene Conversion Hypothesis Expands. PLOS Genetics **11** :e1004941.

[16] **Lee, J. Y., T. Terakawa, Z. Qi, J. B. Steinfeld, S. Redding, Y. Kwon, W. A. Gaines, W. Zhao, P. Sung, et E. C. Greene**. 2015. Base triplet stepping by the Rad51/RecA family of recombinases. Science **349** :977–981.

[17] **Lesecque, Y., D. Mouchiroud, et L. Duret**. 2013. GC-Biased Gene Conversion in Yeast Is Specifically Associated with Crossovers : Molecular Mechanisms and Evolutionary Significance. Molecular Biology and Evolution **30** :1409–1419.

[18] **Lieb, M.** 1985. Recombination in the λ repressor gene : evidence that very short patch (VSP) mismatch correction restores a specific sequence. Mol Gen Genet **199** :465–470.

[19] **Meier, P. et W. Wackernagel**. 2003. Mechanisms of homology-facilitated illegitimate recombination for foreign DNA acquisition in transformable Pseudomonas stutzeri. Molecular Microbiology **48** :1107–1118.

[20] **Pessia, E., A. Popa, S. Mousset, C. Rezvoy, L. Duret, et G. A. B. Marais**. 2012. Evidence for Widespread GC-biased Gene Conversion in Eukaryotes. Gen Biol Evol **4** :675–682.

[21] **R Core Team**. 2015. R : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[22] **Ratnakumar, A., S. Mousset, S. Glemin, J. Berglund, N. Galtier, L. Duret, et M. T. Webster**. 2010. Detecting positive selection within genomes : the problem of biased gene conversion. Philosophical Transactions of the Royal Society B : Biological Sciences **365** :2571–2580.

[23] **Rodgers, K. et M. McVey**. 2016. Error-Prone Repair of DNA Double-Strand Breaks. Journal of Cellular Physiology **231** :15–24.

[24] **Sanger, F., S. Nicklen, et A. R. Coulson**. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A **74** :5463–5467.

[25] **Vallenet, D., E. Belda, A. Calteau, S. Cruveiller, S. Engelen, A. Lajus, F. L. Fèvre, C. Longin, D. Mornico, D. Roche, Z. Rouy, G. Salvignol, C. Scarpelli, A. A. T. Smith, M. Weiman, et C. Médigue**. 2013. MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucl. Acids Res. **41** :D636–D647.

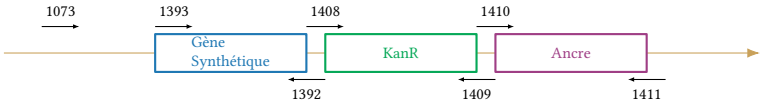
[26] **Williams, A. L., G. Genovese, T. Dyer, N. Altemose, K. Truax, G. Jun, N. Patterson, S. R. Myers, J. E. Curran, R. Duggirala, J. Blangero, D. Reich, et M. Przeworski**. 2015. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. eLife **4**.

[27] **Yahara, K., X. Didelot, K. A. Jolley, I. Kobayashi, M. C. J. Maiden, S. K. Sheppard, et D. Falush.** 2016. The Landscape of Realized Homologous Recombination in Pathogenic Bacteria. Mol Biol Evol **33** :456–471.

[28] **Yáñez-Cuna, F. O., M. Castellanos, et D. Romero.** 2015. Biased gene conversion in *Rhizobium etli* is caused by preferential double strand breaks on one of the recombining homologs. Journal of Bacteriology p. JB.00768–15.

1 Annexes

1.2 Amorces utilisées



Cible	Amorce	Séquence	Tm (°C)
Génome	1073	CAGGCTGACGTGATTGTTCA	56.6
Construction	1392	AAGGTGGAAGAGAAGGAGGC	58.7
	1393	GCGAGGAGGAAAGCAAAGAG	58.2
	1408	CACCTTAATCACTAGTTAGACATCTAAATCTAGGTAC	61.50
	1409	GGTAAAGTCAGAGGAGAGGATGAGGAGGCAGATTG	68.66
	1410	TCCTCTGACTTTACCAACAAC	48.04
	1411	AGGCGGCCGCACTAGCTTTCTGAGGGGAACGATCA	71.62
Plasmide pGEM-T	M13R	GAGGAAACAGCTATGAC	47.8
	M13F	GTAAAACGACGGCCAGT	53.9

Figure 10 : **Liste des amorces utilisées**

1.3 Carte des constructions donneuses

1.4 Traces de conversions

La figure 12 page 12 montre l'alignement de toutes les positions de marqueurs lorsque le donneur est CG, AT, CG/AT et AT/CG. L'interprétation des figures est détaillée dans la figure 5 en page 5.

1.5 Confirmations des restaurations des haplotypes sauvages

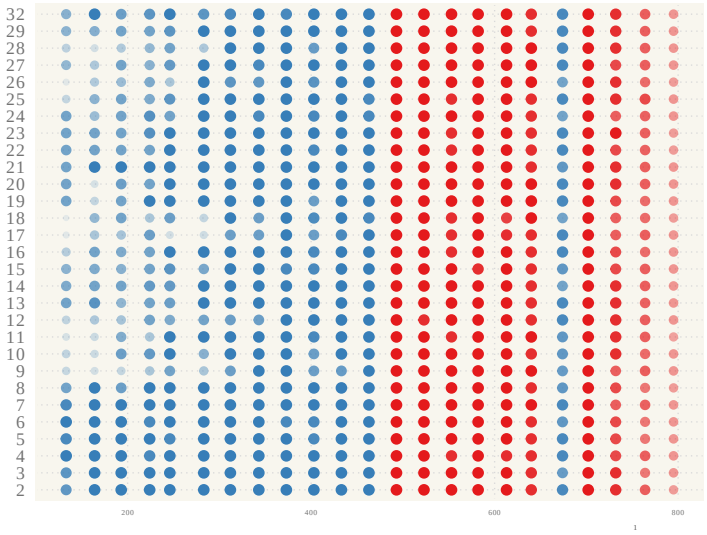


Figure 11 : **Confirmation d'une restauration d'haplotype sauvage**

Pour s'affranchir d'une possible erreur de séquençage sur les sites montrant des restaurations de l'haplotype sauvage dans la région convertie, nous avons sequencé la zone de recombinaison de 31 clones issus du clone séquencé en premier lieu. Nous avons également séquencé à nouveau la colonie "mère", de façon à comparer les sous-populations avec la population mère séquencée ; c'est le clone séquencé 32.

Les 31 clones séquencés montrent tous la même alternance au même marqueur que le clone 32. Nous avons ainsi confirmé que cette restauration de l'haplotype sauvage était bien due à une correction des mésappariemments dans la population mère.

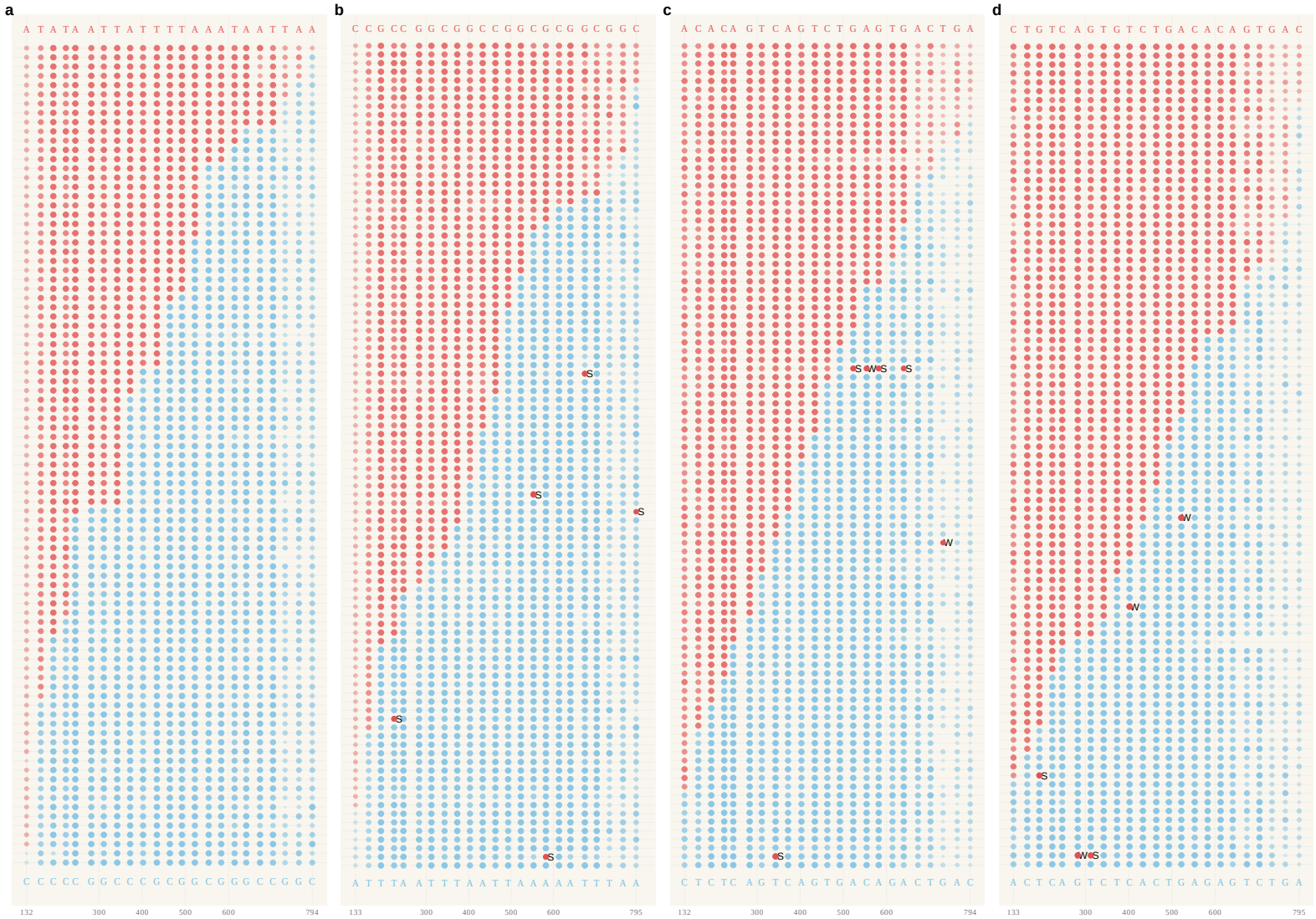


Figure 12 : Pellentesque dapibus suscipit ligula. Donec posuere augue in quam. Aliquam feugiat tellus ut neque. Nulla facilisis, risus a rhoncus fermentum, tellus tellus lacinia purus, et dictum nunc justo sit amet elit.