SOMMAIRE

PRÉAMBULE

Depuis le début des années 2000, l'amélioration des techniques de séquençage a permis l'accumulation de génomes séquencés, en particulier pour les génomes procaryotes (plus de 200.000 projets de séquençage déposés[1]). Cette période a aussi permis le développement de méthodes de génomique comparée, révolutionnant ainsi nos connaissances sur l'organisation et le contenu en gènes des génomes bactériens et permettant d'explorer les mécanismes évolutifs structurant les génomes. Aujourd'hui, le nombre et la diversité des génomes disponibles sont suffisamment importants pour aborder des questions évolutives à différents niveaux taxonomiques. C'est dans ce contexte que se situent ces travaux de thèse, où, profitant de l'abondance de données, la dynamique des génomes a été explorée à la fois au travers d'un genre bactérien, les *Streptomyces*, couvrant ainsi une histoire évolutive d'environ 400 millions d'années (McDonald and Currie, 2017), mais également au sein d'une population sympatrique de *Streptomyces* représentative d'une histoire évolutive beaucoup plus courte.

Les *Streptomyces*, modèle d'étude de ce travail, sont des organismes très étudiés par la communauté scientifique (plus de 13,000 publications référencées sur PubMed avec le mot clef "*Streptomyces*" dans le titre[2] dont 879 publications depuis le 1er janvier 2019). Cet intérêt est motivé par les nombreux phénomènes d'instabilité génétique et chromosomique observés chez ces bactéries doublés de leur capacité à synthétiser une grande diversité de métabolites secondaires présentant de nombreux intérêts industriels.

Ce travail a eu pour point d'initiation les travaux menés par Frédéric Choulet au laboratoire Dynamic à Nancy (Evolution du génome des *Streptomyces* : transfert horizontal et variabilité des extrémités chromosomiques - Thèse Université de Lorraine, 2006) où, à partir du séquençage partiel du chromosome de *Streptomyces ambofaciens* ATCC 23877 et d'application de méthodes de génomique comparée avec les génomes de 3 autres espèces de *Streptomyces*, une première vision dynamique de l'évolution de leur chromosome avait été obtenue.

Dans un premier temps, la question était de savoir si ces résultats préliminaires obtenus à partir de 3 espèces de *Streptomyces* étaient généralisables à l'ensemble des espèces de genre. Cette étape a permis d'aborder de nouvelles questions touchant à la dynamique du chromosome en considérant les distances évolutives entre les différentes espèces du genre *Streptomyces*.

Dans un second temps, la question de la dynamique du chromosome des *Streptomyces* a été abordée au niveau populationnel, c'est-à-dire entre individus sympatrique d'une même espèce.

[1] https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/
[2] https://www.ncbi.nlm.nih.gov/pubmed/?term=Streptomyces

Traiter de la dynamique des génomes met en évidence la dualité des mécanismes évolutifs : la structure d'un génome résulte du bilan net entre mécanismes conservatifs et plasticité. Dans les génomes bactériens, deux grands mécanismes d'évolution prédominent : les mutations et les transferts horizontaux. Ces deux mécanismes sont à l'origine de modifications de l'information génétique induisant des différences entre des organismes issus d'un ancêtre commun et s'opposent à la conservation de l'intégrité du génome et à la transmission de l'information génétique parentale à l'identique.

INTRODUCTION

A- Génomes bactériens : contraintes et tolérances de l'organisation génétique

L'ADN est le support universel de l'information génétique. La réplication, semi-conservative et fidèle, est un élément de préservation de l'information à l'identique, ou presque, au travers des générations. Malgré ce mécanisme universel de préservation de l'information génétique, les variations des différents paramètres qui caractérisent les génomes sont telles qu'un génome bactérien modèle n'existe pas.

1- Organisation et expression des gènes

Différents niveaux d'organisation des gènes, liés ou pas à leur expression, structurent le génome bactérien.

a- Taille et densité en gènes des génomes bactériens

La taille des génomes bactériens est extrêmement variable et ce indépendant de l'appartenance à un phylum en particulier, comme le montre la **figure 1** construite à partir de tous les génomes bactériens complets disponibles dans la base de données RefSeq (13.921 génomes). Dans cette liste, le plus petit des génomes mesure 105.760 paires de bases (pb) et correspond à un endosymbiote appartenant au phylum des Protéobactéries, *Candidatus* Hodgkinia cicadicola (Łukasik et al., 2019). À l'opposé, le plus grand génome bactérien de cette liste est *Minicystis rosea* (Garcia et al., 2014). Il appartient aussi au phylum des Protéobactéries, est 150 fois plus grand et mesure 16.040.666 pb.

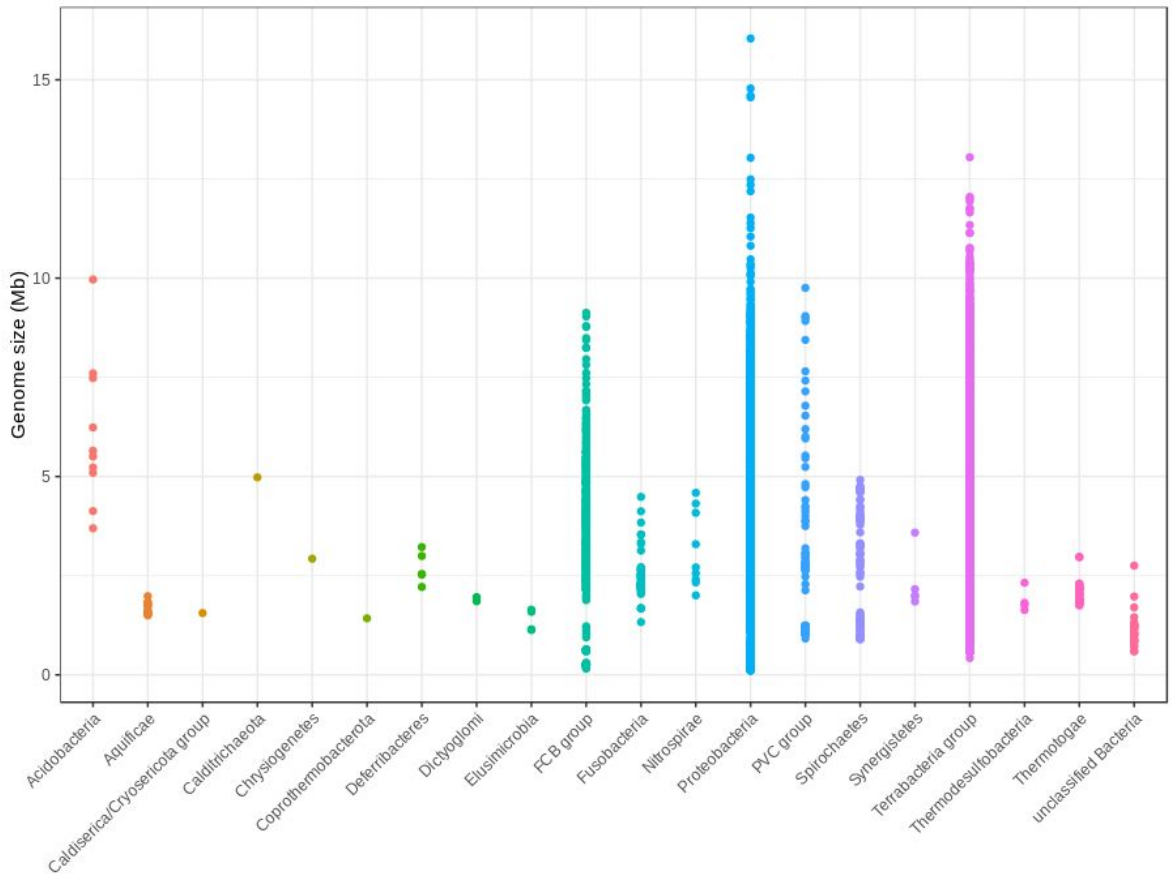

Figure 1 - Taille des génomes bactériens complets selon leur assignation aux groupes du NCBI dans la base de données NCBI RefSeq le 14/05/2019. Un groupe NCBI peut correspondre à un phylum ou à un regroupement de phyla (*ie* super phylum) comme le super phylum des Terrabactéries qui regroupe entre autres, le phylum des Actinobactéries (contient le genre *Streptomyces*) et celui des Firmicutes.

De ces observations émerge la question du génome minimal et à l'inverse du génome maximal. Les plus petits génomes correspondent à des organismes endosymbiotiques (Moran and Bennett, 2014), vivant donc dans un environnement bien défini et où des fonctions cellulaires obligatoires sont assurées par l'organisme hôte. À l'inverse, les plus grands génomes bactériens correspondent à des espèces vivant de manière autonome dans des milieux complexes et changeants, en effet dans des conditions stables, des modifications drastiques du génome sont rares (Barrick et al., 2009). Des bactéries vivant dans de tels milieux doivent être "armées" pour résister à un très grand nombre de stress différents. Cela se traduit par un plus grand nombre de gènes et donc des génomes plus grands (corrélation directe entre la taille des génomes et le nombre de gènes (Bird, 1995)) distinguant deux catégories fonctionnelles : les gènes essentiels, obligatoires au développement végétatif de la cellule quelles que soient les conditions et les gènes accessoires permettant à la cellule de s'adapter à des conditions données. En se plaçant dans une niche écologique extrêmement riche, tolérant

l'auxotrophie de la cellule pour un maximum de composés organiques, la notion de génome minimal a été abordée dès l'accès aux deux premiers génomes entièrement séquencés et assemblés en 1995 (Koonin, 2000; Maniloff, 1996) : *Haemophilus influenzae* (Fleischmann et al., 1995) et *Mycoplasma genitalium* (Fraser et al., 1995). À partir de la comparaison entre ces deux génomes, un groupe commun de 250 (A. R. Mushegian and Koonin, 1996) puis 206 (Gil et al., 2004) gènes a été extrait, largement inférieur au nombre de gènes de chaque organisme, respectivement de 1.815 et 525 gènes. De nombreux travaux ont été conduits pour construire le plus petit génome possible pour qu'un organisme puisse se développer de manière autonome. Ce génome serait composé, *a priori*, des seuls gènes obligatoires au développement d'un organisme dans des conditions idéales. En 2016, à partir du génome de *Mycoplasma mycoides*, un génome minimal de 473 gènes fut construit (Hutchison et al., 2016) avec pour objectif de construire une "cellule si simple que l'on peut déterminer les fonctions moléculaires et biologiques de chaque gène". Parmi ces 473 gènes, 149 possèdent une fonction biologique inconnue montrant l'existence de facteurs et/ou d'interactions nécessaires à la vie non caractérisés à ce jour.

D'un point de vu plus général, les bactéries symbiotiques sont caractérisées par une tendance à la réduction de leur génome (Moran and Bennett, 2014; McCutcheon and Moran, 2012). Cette observation est probablement une conséquence combinée du biais mutationnel vers les délétions communément observé chez les bactéries (Mira et al., 2001), de l'absence de sélection contre les pertes de gènes dans les environnements stables.
À l'inverse, dans des environnements complexes et changeants, une expansion des génomes induisant une accumulation de différents répertoires de gènes est observée. Les processus qui conduisent à l'expansion des génomes regroupent principalement l'intégration de gènes étrangers et l'amplification de familles de gènes (Wang et al., 2011; Tsai et al., 2018). Différents facteurs limitent la taille maximale d'un génome comme le temps de génération, la vitesse de réplication, l'énergie nécessaire à la réplication et enfin la taille physique du génome et celle de la cellule.

Les génomes procaryotes présentent une densité d'ADN codant élevée (*ie* proportion de la somme des taille des séquences codantes par rapport à la taille totale d'un génome en paire de bases), entre 85 et 90 % (Land et al., 2014; McCutcheon and Moran, 2012) (**figure 2**). Cela révèle une pression forte sur la taille des génomes qui s'oppose à l'accumulation de séquences d'ADN non fonctionnelles. L'augmentation de la taille du génome chez les bactéries s'accompagne donc d'une augmentation du nombre de gènes à la différence des eucaryotes.
Les rares génomes qui présentent une faible densité appartiennent à des organismes symbiotiques ou des pathogènes obligatoires. C'est la cas de *Serratia symbiotica str. Cinara cedri*, par exemple, qui est

un symbiote d'insecte et présente une densité en gènes de 38 % ou encore de *Nostoc azollae 0708* est un symbiote de plante avec une densité de 52 % (Land et al., 2015). Un autre cas bien décrit est celui de *Mycobacterium leprae*, espèce proche *Mycobacterium tuberculosis*. Le génome complet de *M. leprae* contient 3.268.203 pb (Cole et al., 2001) et 49,5 % d'ADN codant contre un génome de 4.411.532 pb et 90,8 % d'ADN codant pour *M. tuberculosis* (Cole et al., 1998). Chez *M. leprae*, 27 % du génome correspond à des pseudogènes qui ont des homologues fonctionnels chez *M. tuberculosis*. Cet exemple illustre un cas extrême d'évolution dite réductive (Cole et al., 2001) où le changement brusque d'habitat lève la pression de sélection sur de nombreux gènes favorisant leur délétion ou dégénérescence (pseudogénisation) à grande échelle. Une des rares exceptions connues à ce jour est la Cyanobactérie *Trichodesmimum erythraeum* avec seulement 63 % de densité en gènes, mais qui ne présente aucune condition de vie particulière pouvant expliquer la réduction de sa proportion en génome codant (Pfreundt et al., 2014).

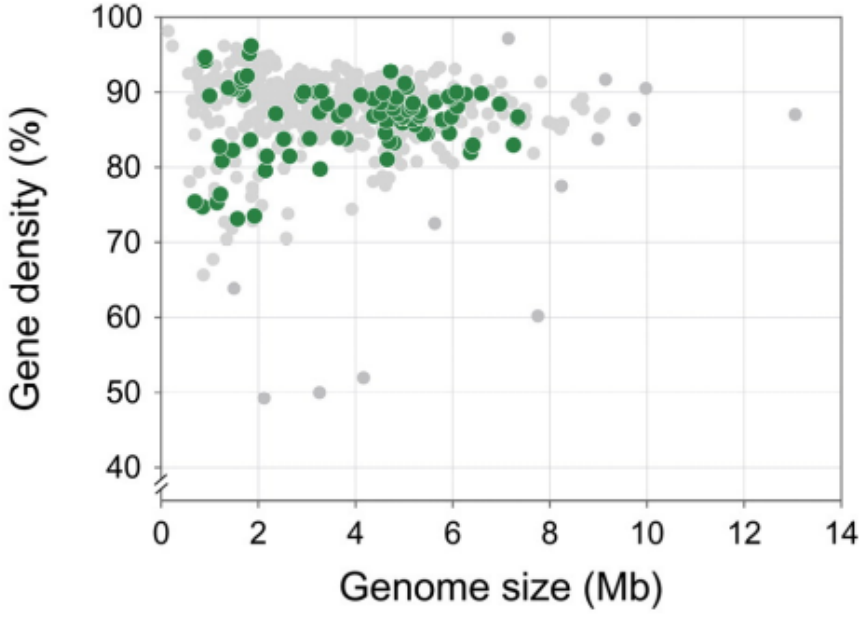

Figure 2 - Relation entre taille du génome et densité en gènes de 488 espèces bactériennes. Figure tirée de Kuo et al. (2009).

b- Le core et pan-génome

La disponibilité croissante de génomes complètement séquencés a changé notre vision de l'évolution des génomes et a révélé les flux des gènes importants entre les organismes. Le concept de stabilité d'un génome a cédé la place à une vision plus dynamique dans laquelle les génomes peuvent perdre ou gagner des gènes (Snel, 2001). Au sein d'un même groupe taxonomique, comme par exemple l'espèce, la variabilité génomique peut être très importante (Touchon et al., 2009). Le terme

pan-génome fut proposé pour représenter l'ensemble de la diversité génétique d'un ensemble d'individus (Tettelin et al., 2005), c'est-à-dire l'ensemble des gènes de tous les génomes analysés. Par opposition, les gènes partagés par tous les individus du groupe, définissent le core-génome. Dans la littérature, de nombreux autres termes sont utilisés pour décrire ces répertoires de gènes et dépendent parfois de la relation phylogénétique entre les organismes étudiés : gènes "dispensables" (Medini et al., 2005), supra-génome, gènes distribués et uniques (Hiller et al., 2007). Dans ce manuscrit, en plus des termes pan et core-génome nous allons utiliser la notion de génome accessoire qui fait référence aux gènes partagés par un sous-groupe d'espèces, autrement dit, le pan-génome est composé de l'ensemble du génome accessoire et du core-génome (**figure 3**).

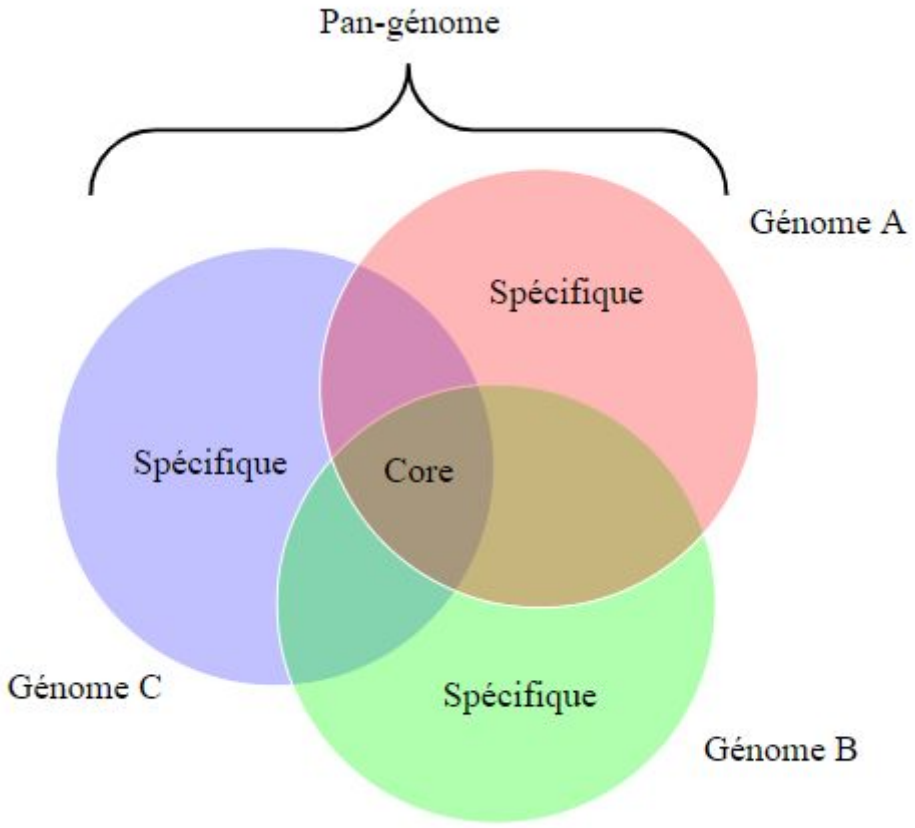

Figure 3 - Diagramme de Venn illustrant un pan-génome théorique. Chaque cercle représente le génome d'un organisme et les intersections les gènes partagés entre ces organismes. Le core-génome (intersection des 3 cercles) correspond aux gènes partagés par tous les génomes, le génome accessoire à tous les gènes qui n'appartiennent pas au core-génome.

De nombreux travaux portent sur les répertoires de gènes et se distinguent principalement par le niveau phylogénétique étudié. Une analyse des répertoires de gènes à l'échelle du règne des bactéries (utilisation de 573 génomes bactériens) indique que seuls environ 250 gènes forment le core-génome (Lapierre and Gogarten, 2009). Cette valeur est inférieure au génome minimal de 473 gènes

(Hutchison et al., 2016) construit à ce jour permettant de supposer que ces 250 gènes constituent le "squelette" de base d'un génome sur lequel se greffent d'autre gènes essentiels, mais dépendant de l'environnement des organismes et/ou ayant trop divergés pour être identifiés comme orthologues à l'échelle du règne des bactéries.

À des niveaux phylogénétiques plus restreints (intra genre, espèce, population), la comparaison génomique a permis de mettre en évidence une très grande variabilité entre les espèces, mais aussi dans certains cas l'absence de variabilité. Chez *E. coli* où deux études différentes utilisant un ensemble de 17 souches (Rasko et al., 2008) et 21 (Touchon et al., 2009) souches ont mis en évidence une grande diversité génétique chez cette espèce, où chaque nouvelle souche considérée augmente la diversité en gènes définissant un pan-génome ouvert. Le core génome représente quant à lui environ la moitié du génome d'*E. coli* (2.344 à partir de 17 souches, 1.976 depuis 21). Contrastant avec *E. coli*, l'analyse de 238 souches de *Bacillus*, réparties en 5 espèces, a révélé une très faible variabilité avec un core-génome représentant jusqu'à 98 % des gènes totaux de *B. amyloliquefaciens*.

La diversité entre les génomes ne peut se résumer à la composition en gènes, il faut également considérer l'organisation et la structure des génomes pour comprendre cette diversité.

c- Opérons et regroupements de gènes

Dès les années 50, il fut montré que les gènes ne sont pas distribués aléatoirement dans les génomes (Demerec and Hartman, 1959). Au travers des travaux de François Jacob et Jaques Monod, un premier modèle d'organisation des gènes fonctionnellement associés et co-transcrits fut proposé : l'opéron. L'opéron consiste en une séquence d'ADN regroupant plusieurs gènes dont l'expression est régulée par un même promoteur, comme l'opéron *lacZYA* mis en évidence chez *E. coli* (Jacob et al., 1960). Cette organisation est répandue chez les bactéries avec en moyenne 3 à 4 gènes par opéron (Zheng et al., 2002). Les gènes d'un opéron codent souvent pour des protéines qui interagissent physiquement (Huynen et al., 2000) ou qui possèdent des fonctions complémentaires (Rogozin et al., 2002). L'opéron n'est pas le seul procédé de co-régulation retenu par l'évolution; de nombreux gènes distribués le long des génomes bactériens sont co-régulés en *trans*. Le nombre et la taille des opérons diminuent avec la taille du génome, et cela est corrélé avec l'augmentation de la proportion de gènes de régulation chez les bactéries possédant un grand génome (Nuñez et al., 2013; Stover et al., 2000).

Ainsi chez *Streptomyces coelicolor* A3(2), 965 gènes, soit environ 12 % des gènes, codent pour des protéines régulatrices (Bentley et al., 2002).

Chez les *Streptomyces*, les gènes impliqués dans la synthèse de métabolites spécialisés sont regroupés pour former des regroupements (cette organisation n'implique pas obligatoirement une unique unité de transcription et donc un opéron) (Rudd and Hopwood, 1980; Rudd and Hopwood, 1979). Un

regroupement comprend outre les gènes de biosynthèse, des gènes de régulation, d'export et/ou de résistance (notamment lorsque le composé produit est toxique, comme les antibiotiques). Un génome de Streptomyces peut contenir de nombreux regroupements de biosynthèse. Ainsi chez *S. coelicolor* A3(2) au moins 29 regroupements ont été identifiés, 37 chez *S. avermitilis* (Liu et al., 2013). En considérant 3 espèces de *Streptomyces* : *S. coelicolor*, *S. avermitilis* et *S. griseus* des regroupements de métabolisme secondaire sont conservés. Par exemple les regroupements codant pour l'hopanoïde et la desferrioxamine sont conservés et sont arrangés de manières similaires (taille, orientation et localisation chromosomique) entre ces 3 espèces (**figure 4**).

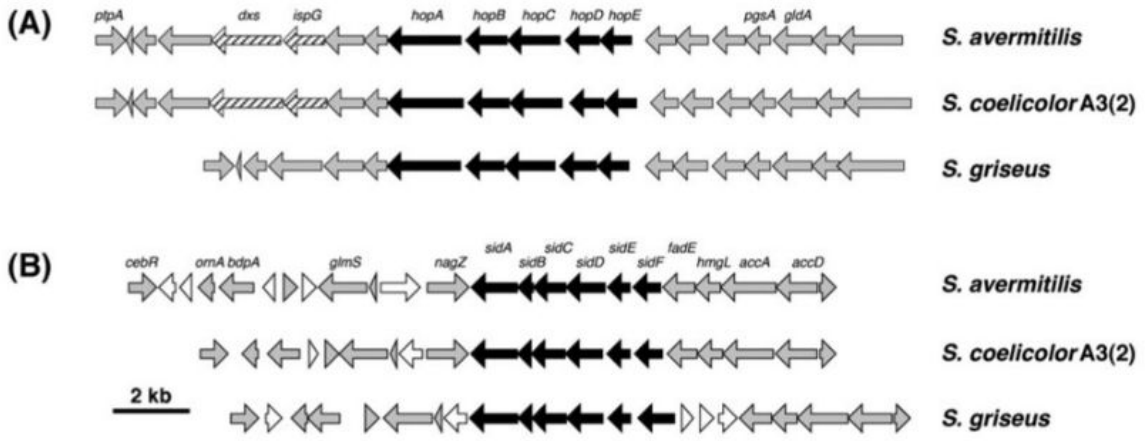

Figure 4 - Régions conservées entre *S. coelicolor*, *S. avermitilis* et *S. griseus* contenant le regroupement biosynthétique de l'hopanoïde (**A**) et la desferrioxamine (**B**). Figure tirée de Nett et al. (2009).

A l'inverse, des regroupements de métabolismes secondaires sont spécifiques à certaines espèces et localisés au sein de régions conservées entre ces 3 espèces comme le regroupement de biosynthèse de l'oligomycine, spécifique de *S. avermitilis*. Cette organisation suppose une adaptation fonctionnelle rapide de l'organisme probablement effectuée au travers de l'acquisition de la cassette de gènes du regroupement par transfert horizontal.

d- Impact de l'organisation en opérons et en regroupements sur la synténie

Le séquençage des 2 premiers génomes bactériens, *Haemophilus influenzae* (Fleischmann et al., 1995) et *Mycoplasma genitalium* (Fraser et al., 1995), a permis d'aborder la question de l'évolution de l'organisation des génomes. De cette première approche, il est apparu que l'ordre des gènes n'est pas conservé entre des bactéries distantes phylogénétiquement (Arcady R. Mushegian and Koonin, 1996), et ce même pour les gènes organisés en opéron, illustré par le cas de l'opéron tryptophane (*trp*) décrit chez *E. coli* (Bertrand et al., 1976) où les différents gènes sont conservés entre les espèces comparées, mais organisés différemment (insertion de gènes, ordre relatif, orientation, "perte" de l'organisation

en opéron) révélant de nombreux événements de réarrangements même au sein de la structure d'un opéron (Dandekar et al., 1998). Ces remaniements fréquents de l'ordre des gènes auraient favorisé le regroupement de gènes impliqués dans une même fonction et plus particulièrement les gènes codant des protéines interagissant physiquement (Dandekar et al., 1998).

En 2006, une étude portant sur un nombre plus conséquent de génomes bactériens (126) a permis d'analyser l'évolution de l'ordre des gènes en fonction de la distance phylogénétique et de distinguer différents comportements selon les clades de bactéries considérées (Rocha, 2006). Un indice, le GOC (Gene Order Conservation) a été défini comme la proportion du nombre total de paires d'orthologues entre deux génomes par rapport au nombre total d'orthologues entre ces deux génomes (Rocha, 2006). Le GOC correspond à la fréquence relative de réarrangements du matériel génétique partagé (transféré verticalement depuis l'ancêtre commun) entre deux espèces en s'affranchissant du transfert horizontal. Le GOC varie donc entre 1 (conservation parfaite de l'ordre des gènes) et 0 (aucune conservation). Une valeur élevée de GOC peut correspondre à deux situations, soit l'organisation des génomes comparés est stable soit ces deux génomes ont divergés récemment. Pour distinguer ces situations, le GOC a été modélisé en fonction de la distance phylogénétique séparant les espèces considérées pour les gènes isolés ou en opérons. Ainsi, la présence d'opérons impacte sur l'évolution de la valeur de GOC avec la distance phylogénétique; l'organisation en opéron tend à biaiser à la hausse la valeur de GOC, l'organisation en opéron étant plus conservée. De plus, il est apparu que le nombre de réarrangements endogènes, perturbant l'ordre des gènes ne serait pas aussi important qu'il est apparu à travers les premières études de génomique comparée. Il a également pu être montré qu'il n'existe pas de corrélation directe entre instabilité génomique et mode de vie de l'organisme. Cependant, certains organismes apparaissent plus instables que d'autres et en particulier les *Streptomyces*.

e- Impact de la localisation chromosomique

La localisation chromosomique d'un gène est un paramètre important de son expression. Parmi de nombreux gènes étudiés, le gène *spoIIR* de *Bacillus subtilis* illustre parfaitement cet effet. Le gène *spoIIR* est exprimé dans les pré-spores nouvellement formées (Khvorova et al., 2000). L'activation de l'expression de *spoIIR* intervient immédiatement après la formation du septum distinguant la cellule mère de la pré-spore. La région du chromosome proche de l'origine de réplication est localisée dans la pré-spore, région où est situé le gène *spoIIR*. Cette localisation chromosomique assure l'expression précoce de *spoIIR* durant la sporulation qui en est un acteur majeur. Pour tester le possible impact de la localisation chromosomique du gène *spoIIR* à proximité de l'origine de réplication, la séquence a été transloquée à différents endroits du chromosome. Le chromosome de *B. subtilis* étant circulaire, les positions sont exprimées en degré (°) où l'origine de réplication est à 0° et la localisation native de

spoIIR à 324°. La translocation a été effectuée à 3 endroits : de l'autre côté de l'origine de réplication (24°), aux environs du terminus de réplication (190°) et à mi-distance entre origine et terminus de réplication (283°). Pour chaque localisation du gène *spoIIR*, l'efficacité de la sporulation par rapport à la souche sauvage. Seul les translocations aux positions 283° et 190° ont révélé une perte d'efficacité de la sporulation (respectivement 80 et 94 % de perte). La translocation de ce gène dans une région distante de l'origine de réplication mène donc à une sporulation déficiente chez *Bacillus subtilis* (Khvorova et al., 2000) suggérant une régulation par la position chromosomique de ce gène qui impact à la fois sur le taux et le moment d'expression de *spoIIR*.

Chez *E. coli* l'ordre des gènes le long de l'axe origine - terminus de réplication est corrélé avec l'expression temporelle des gènes durant la croissance bactérienne, les gènes impliqués dans les premières et dernières étapes de la croissance bactérienne sont ainsi retrouvés respectivement aux environs de l'origine et de la terminaison de la réplication (Sobetzko et al., 2012). Cette organisation des gènes semble répandue, au moins, chez les γ-protéobactéries suggérant un lien avec la réplication.

Cependant, chez les *Streptomyces* (ou plus globalement chez les *Actinobactéries*), l'effet de la localisation chromosomique des gènes est peu décrit, et ce, malgré l'utilisation importante de ces bactéries dans la synthèse de molécules bioactives. Des travaux ont identifié une protéine spécifique aux *Actinobactéries* (sIHF) responsable de la compaction de l'ADN et qui semble avoir un impact sur le taux d'expression global des gènes (Swiercz et al., 2013). D'autres travaux ont identifié une protéine "histone-like" (DdbA) chez les *Streptomyces* capable de coupler des changements de conformation de l'ADN et d'expression en réponse à des stress . Cette complexité dans l'organisation du nucléoïde des *Streptomyces* mène à des variations d'expression selon les régions chromosomique, causant un effet de localisation chromosomique sur l'expression des gènes (Aldridge et al., 2013). Plus récemment, l'impact de localisation chromosomique sur l'expression génique a été testée par l'analyse de l'expression d'un regroupement de gènes de biosynthèse d'un métabolite spécialisé (expression hétérologue[3]) chez *S. albus* J1074 (Bilyk et al., 2017). Le taux d'expression varie d'un facteur 8 selon la localisation chromosomique, montrant bien un impact de la localisation chromosomique chez *Streptomyces*.

d - L'effet dose

Un impact de la localisation chromosomique particulier a été mis en évidence chez les bactéries à croissance rapide : l'effet dose. Cet effet a d'abord été révélé chez *E. coli* où durant la phase exponentielle de croissance, la réplication bidirectionnelle commençant à l'origine de réplication génère un gradient important de gènes : plusieurs cycles de réplication du chromosome peuvent être

[3] Expression d'un gène ou d'un fragment de gène dans un organisme hôte, qui ne possède pas naturellement le gène ou son fragment.

initiés en même temps avant la séparation des chromosomes répliqués et la division cellulaire. Par conséquent les gènes proches de l'origine de réplication sont présents en un nombre de copies plus important à cet instant que les gènes situés plus loin sur le chromosome pouvant augmenter l'expression de ces gènes, c'est l'effet dose associé à la réplication (Couturier and Rocha, 2006).

Lors de la phase exponentielle, si un ou plusieurs gènes, dont l'expression est limitante pour la croissance bactérienne, leur position sur le chromosome devient un paramètre soumis à une forte sélection. Cet effet a, par exemple, été montré chez *Lactococcus lactis* où la création d'inversions artificielles sur le chromosome, perturbant la symétrie des réplichores, a révélé deux régions particulièrement contraintes où la modification de l'organisation des gènes n'est pas tolérée : aux environs de l'origine de réplication et à l'opposé du chromosome vers le terminus de réplication (Campo et al., 2004). Ainsi, pour les génomes montrant une organisation façonnée par l'effet dose, les réarrangements asymétriques auraient des effets plus marqués et seraient, par conséquent, contre sélectionnés. L'effet dose serait donc un frein à la plasticité génomique.

Plus récemment, en analysant le taux d'expression d'un gène rapporteur inséré à différents dans le chromosome d'*E. coli*, il a été mis en évidence que son taux d'expression peut varier d'un facteur 300 selon sa position sur le chromosome (Bryant et al., 2014). Cet effet n'est que partiellement dû à l'effet dose décrit précédemment : le taux d'expression a été analysé en fonction du nombre de copies du gène rapporteur. Cela a révélé que l'effet dose ne peut être responsable que d'un facteur 1,4 dans la variation du taux d'expression des gènes. La variation du taux d'expression d'un gène selon sa position chromosomique ne semble donc que partiellement causée par la présence de ce gène en multi-copies générées par les génomes partiellement répliqués, renforçant l'importance de la position chromosomique des gènes sur leur taux d'expression.

2- Organisation du nucléoïde

Le nucléoïde correspond à la région dans les cellules procaryotes dans laquelle se trouve le matériel génétique. La taille d'un chromosome bactérien étiré comme celui d'*E. coli* est de l'ordre du millimètre, soit environ 1.000 fois plus grand que celle d'une bactérie. L'ADN doit donc être compacté afin d'être contenu dans une cellule tout en étant compatible avec de nombreux processus cellulaires comme la réplication, la transcription, la réparation, la recombinaison et l'intégration. Chez les bactéries, cette contrainte est d'autant plus importante que ces processus sont simultanés.

a- Organisation en chromosome et plasmide (unités de réplication)

Le génome bactérien comprend la plus souvent un unique chromosome, circulaire ou linéaire (Badrinarayanan et al., 2015), complété selon les organismes d'un ou plusieurs plasmides (sous forme circulaire et/ou linéaire). Le premier chromosome linéaire bactérien a été identifié chez *Borrelia*

burgdorferi (Baril et al., 1989) puis chez d'autres genres bactériens comme *Streptomyces* (Lin et al., 1993), *Saccharopolyspora* (Reeves et al., 1998) et *Agrobacterium* (Allardet-Servent et al., 1993). Les relations phylogénétiques (**figure 5**) entre ces bactéries suggèrent au moins 3 apparitions indépendantes de la linéarité du chromosome (Volff and Altenbuchner, 2000). Chez *Streptomyces*, l'origine de réplication est localisée dans la région centrale du chromosome, définissant deux bras chromosomiques où des protéines se liant covalemment aux extrémités chromosomiques génèrent une topologie fermée du chromosome (Tsai et al., 2012; Yang and Losick, 2001). Chez *Borrelia* les extrémités chromosomiques sont organisées en tiges-boucles (Tourand et al., 2003). Chez *Agrobacterium*, l'organisation des extrémités chromosomiques est comparable à celle de *Borrelia* (Goodner et al., 2001). Tout comme chez les *Streptomyces* une position centrale de l'origine de réplication a été mise en évidence chez *Borrelia* (Picardeau et al., 1999) et *Agrobacterium* (Goodner et al., 2001), suggérant une pression de sélection sur le positionnement centrale de l'origine de réplication comparable à l'emplacement diamétralement opposé des sites de terminaison et de réplication dans les chromosomes circulaires, permettant la progression simultanée de deux fourches de réplication.

Chez *Streptomyces* et *Borrelia,* des mutants circulaires ont pu être générés en laboratoire (Chen et al., 2010; Ferdows et al., 1996). Les chromosomes de *Streptomyces* artificiellement circularisés présentent une instabilité génétique égale ou supérieure à la forme linéaire (Chen et al., 2010; Lin and Chen, 1997). Une des explications possibles à l'instabilité de la topologie circulaire serait le manque de domaines associés à la terminaison de la réplication, nécessaire à l'intégrité du chromosome.

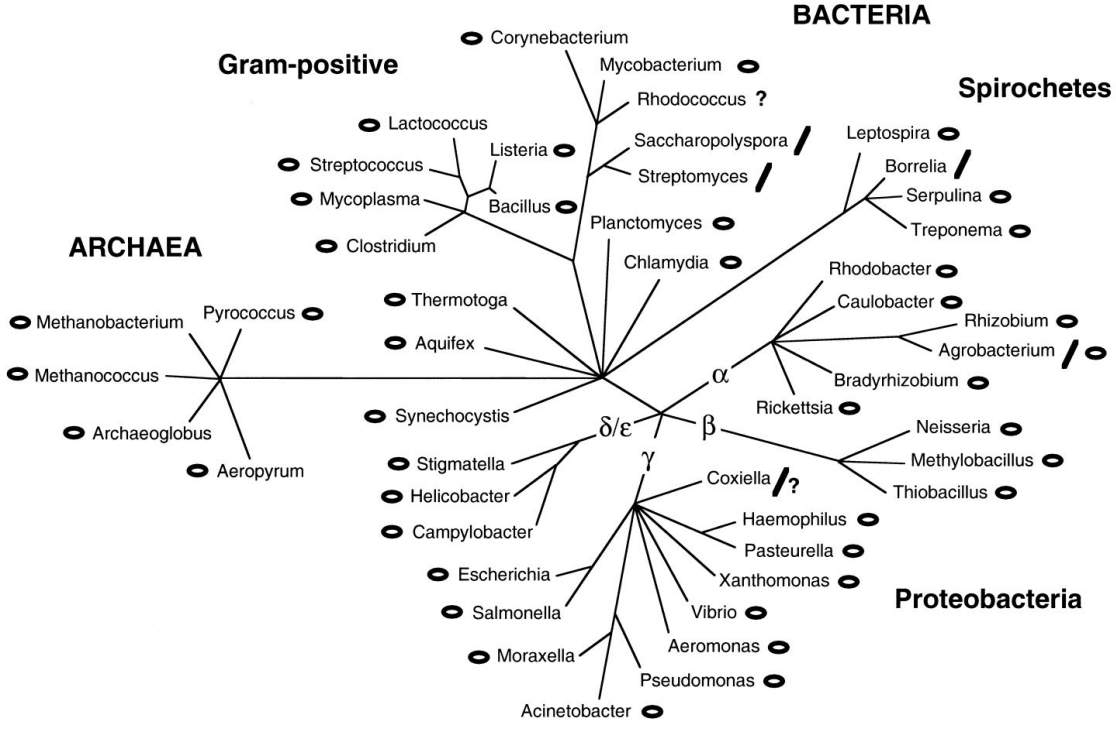

Figure 5 - Distribution des chromosomes circulaires et linéaires chez les procaryotes. Les cercles symbolisent les chromosomes circulaires, les barres les chromosomes linéaires. Figure tirée de Volff and Altenbuchner (2000)

b- Organisation en macro-domaines

Les chromosomes bactériens sont organisés en régions structurées de l'ordre du Megabase appelées macro-domaines. Ce type d'organisation fut tout d'abord suggérée, par des résultats d'hybridation *in situ* en fluorescence (FISH (Langer-Safer et al., 1982)) chez *Escherichia coli* (Niki et al., 2000) avec la mise en évidence d'une région contenant l'origine de réplication (domaine Ori) et d'une autre contenant le terminus de réplication (domaine Ter). Ces deux macro-domaines suivent une chorégraphie au cours du cycle cellulaire qui souligne le caractère fonctionnel de ces régions. Dans un second temps, toujours chez *E. coli*, il a été mis en évidence que des loci au sein d'un même macro-domaine recombinent entre eux plus fréquemment que des loci localisés dans différents macro-domaines (Valens et al., 2004). Quatre macro-domaines ont ainsi été identifiés chez *Escherichia coli* : Ori, Ter, Left et Right ainsi que deux régions moins structurées bordant le domaine Ori (**figure 6**). De plus, les séquences contenues dans les macro-domaines sont moins mobiles que les séquences des régions peu structurées (Espeli et al., 2008).

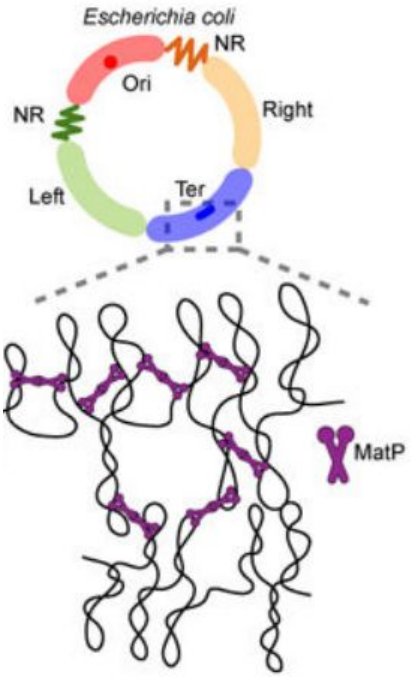

Figure 6 - Organisation en macro-domaines du chromosome d'*Escherichia coli* : Ori, Ter, Left et Right et les deux régions non structurées (NR). La protéine MatP (violet) structure le macro-domaine Ter. Figure tirée de Badrinarayanan et al. (2015).

L'organisation du macro-domaine Ter chez *Escherichia coli* dépend de la présence d'un motif répété de 13 pb : *matS* (Mercier et al., 2008). Ces motifs sont le site de fixation de la protéine MatP, sous forme de dimère, qui pontent deux sites *matS* distants (par association de dimères MatP) et compactent le macro-domaine (**figure 6**). Des modes de compaction comparables sont pressentis pour les autres macro-domaines d'*Escherichia coli,* mais n'ont pas été identifiés à ce jour.

L'utilisation récente de méthodes (3C, 5C, Hi-C) permettant la construction de carte de contact à l'échelle du chromosome (Le and Laub, 2014) a permis d'entrevoir une organisation en macro-domaines chez d'autres espèces bactériennes. Ces approches infèrent la structure du chromosome *in vivo* en capturant des interactions distantes (réalisées par l'intermédiaire d'acteurs *trans*). La cartographie de ces interactions est réalisée grâce au séquençage massif (**figure 7**).

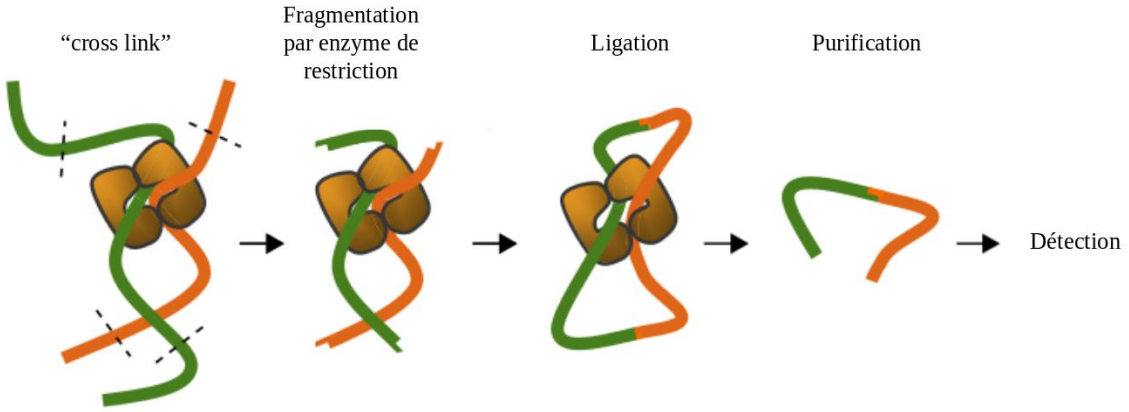

Figure 7 - Schémas du protocole des techniques de capture de conformation du chromosome. L'étape de "cross link" correspond à un traitement (ajout de formaldéhyde dans le cas de la 3C par exemple) "gelant" la conformation du chromosome. Les différences principales entre les 3 méthodes, 3C, 5C et Hi-C se situent au niveau de l'étape de détection, utilisant différentes techniques de séquençages.

Chez *Caulobacter crescentus* par exemple, le chromosome est organisé en larges domaines appelés CID, pour Chromosomal Interaction Domains, d'une taille moyenne de 166 kb (Le et al., 2013). De plus, la bordure des CID correspond à la position de gènes fortement exprimés comme ceux codant les protéines ribosomales.

c- Les domaines topologiques

Des boucles de super-enroulement, appelées plectonèmes, ont été caractérisées par microscopie électronique chez *Escherichia coli* (Kavenoff and Ryder, 1976). Ce surenroulement négatif du chromosome est principalement dû à l'activité des topoisomérases, mais aussi à celle des activités des polymérases d'ADN et d'ARN. Ces régions sur-enroulées sur elles-même sont attachées à leur base

par des protéines générant une boucle isolée d'ADN (Stavans and Oppenheim, 2006) (**figure 8**). Chez *Escherichia coli* environ 400 plectonèmes d'environ 10 kb ont été identifiés (Postow et al., 2004).
Les limites entre les boucles sur-enroulées sont dynamiques et peuvent dépendre à la fois des protéines de liaison à l'ADN et de l'expression des gènes. Comme pour les macro-domaines, l'expression génique semble également jouer un rôle majeur dans l'établissement de domaines topologiques. Les protéines de liaison à l'ADN peuvent relier des loci distants, isolant topologiquement l'ADN formant la boucle et empêchant la propagation du plectonème. Les domaines sur-enroulés varient au cours du cycle cellulaire, impactant l'expression des gènes localisés dans ces régions, cependant, les limites des domaines bordent toujours les gènes les plus fortement exprimés (Wright et al., 2007). Des travaux sur *Caulobacter* ont montré que l'inhibition de la transcription abolit les domaines sur-enroulés autour des gènes fortement exprimés. À l'inverse, l'insertion d'un gène fortement exprimé (*rsaA*) est suffisant pour générer une nouvelle frontière de CID (Le et al., 2013). Les loci fortement transcrits constituent des bornes empêchant la diffusion du domaine topologique, bien que le mécanisme sous-jacent précis ne soit pas clair (Badrinarayanan et al., 2015).

Les relations entre les différents niveaux de compaction du chromosome bactérien n'est pas clairement défini à ce jour, mais une relation hiérarchique entre les échelles peut être envisagée. Les macro-domaines, de l'ordre du mégabase, définissent des régions chromosomiques précises spatialement proches favorisant les interactions au sein d'un même macro-domaine. Un macro-domaine contient des domaines sur-enroulés, de l'ordre de la dizaine de kilobases, aux frontières plus diffuses et variables au cours du cycle cellulaire à l'exception des gènes très fortement exprimés qui apparaissent toujours en dehors des plectonèmes.

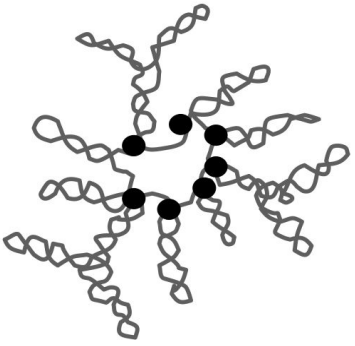

Figure 8 - Schéma de l'organisation en domaines sur-enroulés d'un chromosome bactérien circulaire. Chaque domaine est topologiquement isolé et chaque point noir correspond à une protéine maintenant l'organisation en plectonème. Figure tirée de Stavans and Oppenheim (2006)

d- Les protéines associées au nucléoïde

L'organisation du chromosome est fortement impactée par des protéines interagissant avec l'ADN, le maintien des domaines sur-enroulés décrit au point précédent en est un premier exemple. Une classe particulière de ces protéines, les protéines associées au nucléoïde (NAP) peuvent se lier de manière non-spécifique au chromosome et impacter localement l'organisation de l'ADN et même sur la transcription (Dillon and Dorman, 2010).

Un autre type de protéine capable d'interagir avec l'ADN, les H-NS (pour Histone-like Nucleoid-Structuring) sont capables de créer des "ponts" ADN-protéine-ADN, rapprochant physiquement des loci distants dans la séquence d'ADN. Ces protéines se lient à des centaines de sites dans le chromosome, mais présentent une préférence pour les sites riches en AT (Grainger et al., 2005; Kahramanoglou et al., 2011). En se liant à l'ADN, ces protéines peuvent s'associer aux sites d'interactions d'ARN polymérase ou de facteurs de transcription régulant l'expression des gènes associés (Müller et al., 2006). Les H-NS sont impliqués dans la régulation de l'expression de gènes horizontalement acquis, généralement plus riches en AT que le génome hôte et donc préférentiellement ciblés par les H-NS (Singh et al., 2014).

Les protéines HU sont un autre exemple de NAP. Elles sont trouvées dans la plupart des bactéries et sont capables, assemblées en octamères, d'enrouler l'ADN de manière similaire aux histones (Ali Azam et al., 1999). Les protéines HU ne présentent pas de spécificité de liaison à l'ADN. Chez *E. coli* il a été estimé que les protéines HU interagissent avec environ 10 % du génome (Prieto et al., 2012) et introduisent des courbures dans la topologie du chromosome (Swinger et al., 2003) qui stabilisent les boucles sur-enroulées (Le et al., 2013).

L'abondance des NAPs varie fortement selon l'étape de croissance de la cellule et des conditions environnementales permettant d'ajuster la compaction du génome selon les besoins de l'organisme. Il existence une très grande diversité chez les NAPs (Dillon and Dorman, 2010) et elles sont capables d'influer sur l'organisation des chromosomes bactériens, mais aussi sur l'expression des gènes.

e- Autres facteurs

L'ADN bactérien ne diffuse pas dans toute la cellule, il est confiné dans une petite fraction de la cellule, le nucléoïde (Valkenburg and Woldringh, 1984). De plus, le volume cellulaire est encombré (Milo, 2013) par de nombreuses autres molécules (**figure 9**).

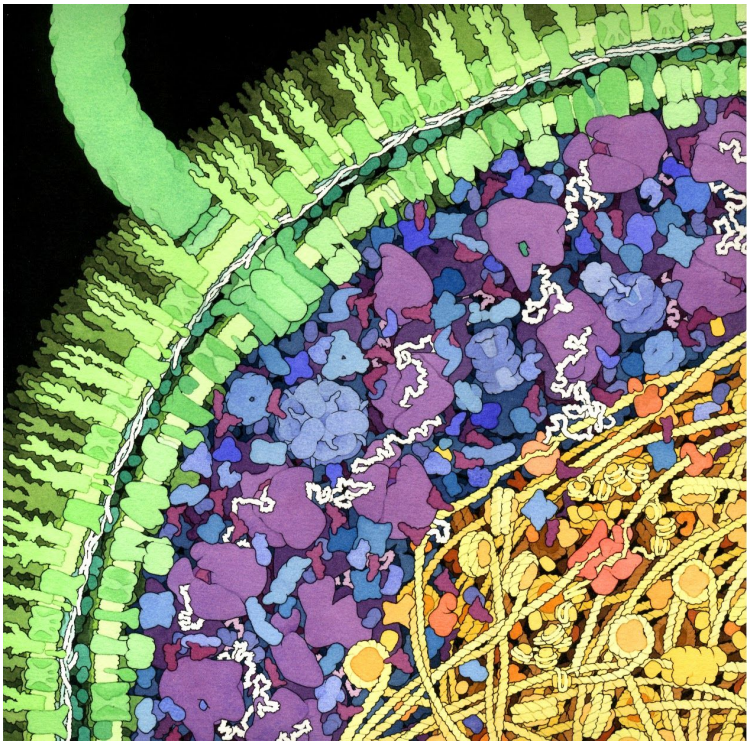

Figure 9 - Vue d'artiste d'une coupe transversale d'une section d'une cellule d'*E. coli* illustrant l'encombrement cellulaire. La membrane est représentée en vert, la région cytoplasmique en bleu et violet. Les grandes molécules bleues correspondent aux ribosomes, les molécules marron en forme de "L" aux ARNt, les brins blancs aux ARNm et les enzymes sont en bleu. La région du nucléoïde est en orange et jaune : l'ADN circulaire, en jaune, est enroulé autour de protéines HU. Illustration proposée par David S. Goodsell[4] .

Cet encombrement moléculaire serait en faveur de la compaction de l'ADN (de Vries, 2010). À l'inverse, Jun and Mulder (2006) ont proposé un modèle dans lequel les domaines sur-enroulés du chromosome se repousseraient et favoriseraient la dispersion du chromosome dans la cellule. La transertion, c'est-à-dire l'insertion de protéines dans la membrane cellulaire durant une traduction co-transcriptionnelle (*ie* lorsque des ribosomes traduisent un ARNm qui est lui-même en cours de synthèse), associe l'ADN à la membrane cellulaire (Libby et al., 2012). En observant l'impact de l'inhibition de la transcription ou de la traduction sur le nucléoïde chez *E. coli,* il a été montré que la transertion agit comme une force de dispersion du chromosome (Bakshi et al., 2012).

[4] https://mgl.scripps.edu/people/goodsell/illustration/public/

B- Plasticité et dynamique des génomes bactériens

Les mutations génétiques et/ou chromosomiques apparues spontanément sont le matériau de base de la sélection naturelle. Les mutations ponctuelles peuvent correspondre à la substitution, l'insertion ou la délétion d'une unique base ou de plusieurs bases. La conséquence d'une mutation ponctuelle peut être silencieuse ou à l'inverse, dans le cas de mutation non-sens par exemple, avoir des conséquences phénotypiques pour le gène concerné. Les réarrangements chromosomiques sont des événements génétiques impactant la structure d'un ou plusieurs chromosomes; les inversions, délétions, duplications et translocations affectent la structure chromosomique. La recombinaison qui est à l'oeuvre dans ces réarrangements favorise également l'incorporation de matériel exogène. C'est le transfert horizontal, qui permet l'acquisition rapide de gènes 'prêts à l'emploi'. Cette section a pour but de décrire les différents événements capables de modifier l'information et/ou l'organisation d'un génome bactérien.

1- Les mutations ponctuelles

L'ADN subit spontanément des dégâts et certaines bases sont plus susceptibles aux substitutions, notamment les cytosines, sujettes à la désamination (Duncan and Miller, 1980) et les guanines, sensibles à l'oxydation (Michaels and Miller, 1992). Il a aussi été montré que les événements de désamination sont plus fréquents lorsque l'ADN est sous forme simple brin comme durant la réplication et la transcription (Frank and Lobry, 1999). Sans mécanisme de réparation de l'ADN, la dégradation naturelle de l'ADN empêcherait le maintien et la transmission de l'information génétique. Chez *Salmonella typhimurium*, l'analyse du génome de mutants déficients pour les systèmes de réparation des mésappariements a révélé après 5.000 générations un fort biais mutationnel, favorisant certains types de substitution. Ainsi, sans mécanisme de réparation de l'ADN, le taux en G+C des bactéries diminuerait rapidement (Lind and Andersson, 2008).

Les substitutions non-synonymes entraînent un changement dans l'acide aminé codé. Selon les similarités entre les propriétés physico-chimiques des acide aminés échangés, on distingue les mutations conservatives où les caractéristiques physico-chimiques de l'acide aminé sont conservées, comme le caractère hydrophobe d'un acide aminé, des mutations non-conservatives.

Les mutations non-sens sont un cas particulier des mutations non-synonymes : elles modifient un codon codant un acide aminé en un codon stop, menant de fait à l'interruption de la phase ouverte de lecture. A l'inverse, une mutation dans un codon stop peut mener à une extension en 3' de la phase ouverte de lecture en supprimant le codon stop originel.

Ainsi, une simple substitution peut entraîner un changement important au niveau de la protéine codée, changement qui peut s'avérer délétère pour l'organisme (dégradation ou altération de la fonction) ou à

l'inverse présenter une forme avantageuse. Ce type de mutation est l'un des moteurs du polymorphisme (*ie* variations de la séquence nucléotidique d'un locus donné) observé dans une population bactérienne.

La fréquence des variants d'un même locus tend à être modulée par deux mécanismes distincts, la sélection naturelle et la dérive génétique. La sélection naturelle consiste en la pression de sélection imposée par un environnement sur la diversité génétique d'une population. Autrement dit, la sélection naturelle favorise le maintien d'un allèle conférant un avantage reproductif (ou fitness) à un individu, réduisant par conséquent la proportion d'allèles ne conférant pas d'amélioration du fitness. La dérive génétique correspond à la sélection de différents variants alléliques d'une population par des phénomènes aléatoires indépendamment de l'impact sur le fitness des organismes.

Ce type de mutations est un des mécanismes majeur de l'évolution des espèces permettant une innovation permanente du matériel génétique d'une population, mais n'impacte que peu sur le contenu en gènes d'une population sur des temps évolutifs "courts", car ces phénomènes sont cumulatifs et lents.

D'autres mécanismes sont capables de modifier fortement le contenu et l'organisation d'un génome en considérant un nombre d'événements mutationnels restreint, les réarrangements chromosomiques. La section suivante est dédiée à la présentation de ces mécanismes et de leurs conséquences sur l'évolution des génomes.

2- Les réarrangements chromosomiques

Les réarrangements chromosomiques sont des modifications de l'organisation chromosomique comme des pertes ou gains de gènes ainsi que des translocations. Ils résultent de la réparation incorrecte de cassures double brin. Chez les bactéries, deux grands mécanismes coexistent pour réparer les cassures double brin de l'ADN, la recombinaison homologue et la recombinaison illégitime.

a- Les cassures double brin de l'ADN

Les cassures double brin dans une molécule d'ADN (ou DSB pour Double Strand Break) sont les dommages les plus délétères que peuvent subir les cellules *in vivo* (Lindahl, 1993). Un DSB a pour conséquence de générer une discontinuité dans l'ADN et entraîne la mort de la cellule si la cassure n'est pas réparée. L'origine des DSB est diverse, ils peuvent être d'origine physique comme les radiations (Ward, 1994) ou chimique, résultant du métabolisme endogène ou produit par d'autres organismes (Wyrobek, 2005). Dans le cas d'une exposition artificielle de l'ADN à un agent mutagène, on parle de mutagenèse.

Quelle que soit l'origine d'un DSB, sa réparation est primordiale pour conserver l'intégrité du génome et la survie de la cellule. Deux stratégies différentes sont utilisées par les bactéries pour réparer les DSB, la recombinaison homologue (RH) et la recombinaison illégitime. Ce sont les conséquences de ces mécanismes de réparation des DSBs qui peuvent entraîner des réarrangements chromosomiques.

b- Impact de la recombinaison homologue sur les réarrangements chromosomiques

La recombinaison homologue est un des processus majeurs de réparation de l'ADN chez les bactéries qui consiste en l'échange d'ADN entre deux régions présentant une forte similarité de séquences nucléotidiques. La RH peut réparer les DSB (Szostak et al., 1983) ou relancer la réplication de l'ADN interrompue par une lésion durant la réplication (Kogoma, 1997; Kowalczykowski, 2000; Mehta and Haber, 2014).

La réparation d'un DSB par RH peut avoir 3 conséquences différentes (**figure 10**) : la conversion génique (**figure 10 A**), le "Single Strand Annealing" (SSA) (**figure 10 B**) et le "Break Induced Replication" ((BIR) **figure 10 C**) (Svendsen and Harper, 2010). Le choix du processus de réparation par la cellule dépend du type de cassure de l'ADN.

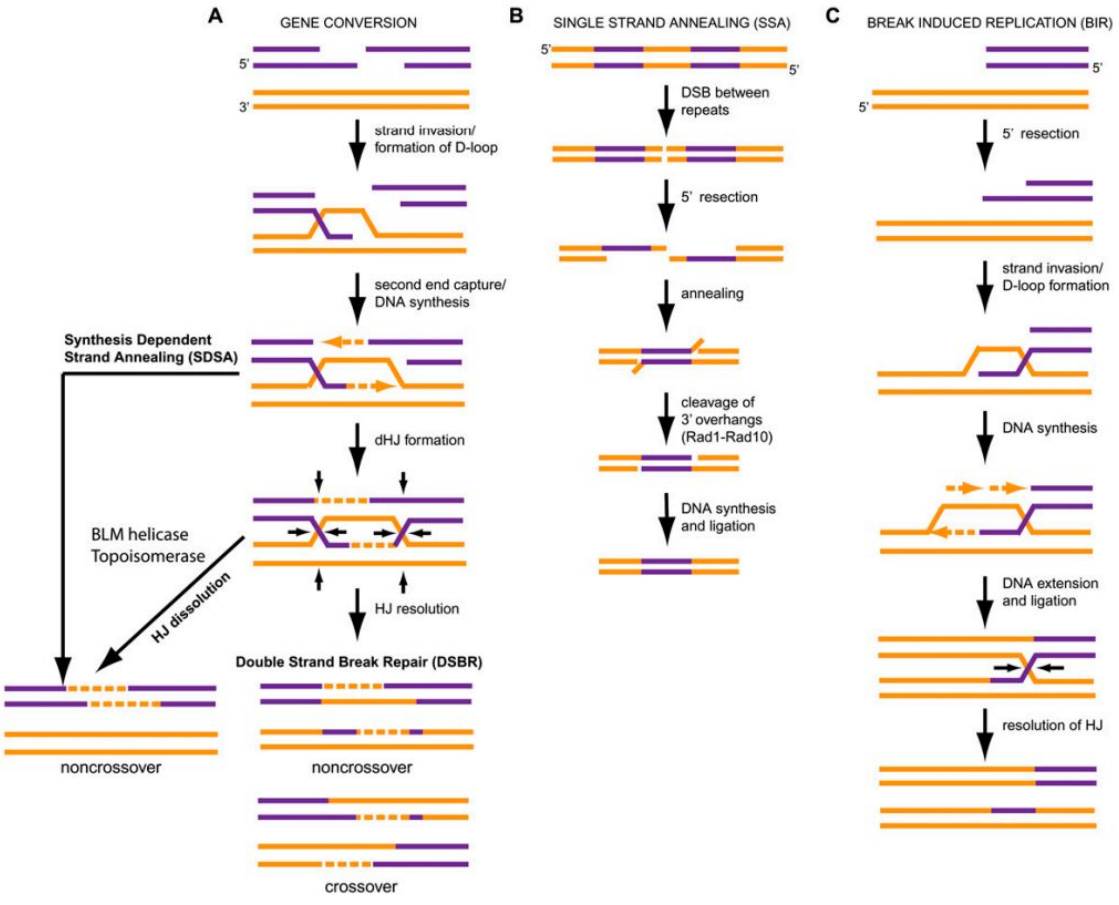

Figure 10 - Schémas des différents processus de réparation des DSB par recombinaison homologue. **A** : la

conversion génique avec ses 3 produits, le "Double Strand Break Repair" (DSBR), la dissolution de la jonction Holliday (HJ dissolution) et la "Synthesis Dependent Strand Annealing" (SDSA); **B** : la Single Strand Annealing, lorsque la cassure intervient entre deux séquences répétées; **C** : la "Break Induced Replication" lorsqu'une seule extrémité est disponible. Figure tirée de Svendsen and Harper (2010).

La recombinaison homologue est un mécanisme de réparation fidèle, en effet, la réparation d'un DSB nécessite la présence d'une séquence homologue intacte. Cependant, la recombinaison homologue peut être à l'origine de réarrangements chromosomiques, en générant des événements de recombinaison qui peuvent être délétères comme bénéfiques pour la cellule. La RH aboutit à 2 types de produits, qui correspondent soit à un transfert unidirectionnel d'information d'une double hélice d'ADN vers l'autre soit d'un échange des segments d'ADN situés de part et d'autre de la cassure (crossing over), générant de l'ADN recombiné (**figure 11**). Pour les séquences répétées en tandem, ils conduisent à une variation du nombre de copies de ces séquences (**figure 11 1.**). Cette configuration peut aussi mener à une recombinaison intramoléculaire, délétant la séquence présente entre les 2 régions dupliquées (**figure 11 2.**). Les crossing-over entre séquences répétées dispersées engendrent quant à eux des délétions ou des inversions (**figure 11 3.**) quand ces séquences sont situées sur le même chromosome la recombinaison peut entraîner des translocations (**figure 11 4.** et **5.**).

Parmi les mécanismes de RH, la voie Single Strand Annealing (SSA) se distingue des autres par son mécanisme qui génère des délétions lors de la réparation de DSB entre des répétitions homologues (**figure 10 B**). Ce mode de réparation des DSB apparaît comme un "dernier recours" et semble utilisé par une cellule que si les autres voies de réparation ne sont pas disponibles. Pour les autres mécanismes nécessitant une autre molécule d'ADN, l'événement de recombinaison peut s'effectuer entre 2 loci non homologues entraînant des réarrangements génomiques. Par exemple, chez *E. coli* une homologie de séquence parfaite sur un segment de 20 pb est suffisante pour initier une recombinaison homologue (Watt et al., 1985).

Cet effet de la recombinaison homologue contraste avec son rôle dans la réparation de l'ADN et dans le maintien de la stabilité du génome, mais peut présenter un avantage pour l'intégration de matériel exogène. Néanmoins, ces réarrangements génétiques restent rares, suggérant que des mécanismes contrôlent les événements de recombinaison délétères.

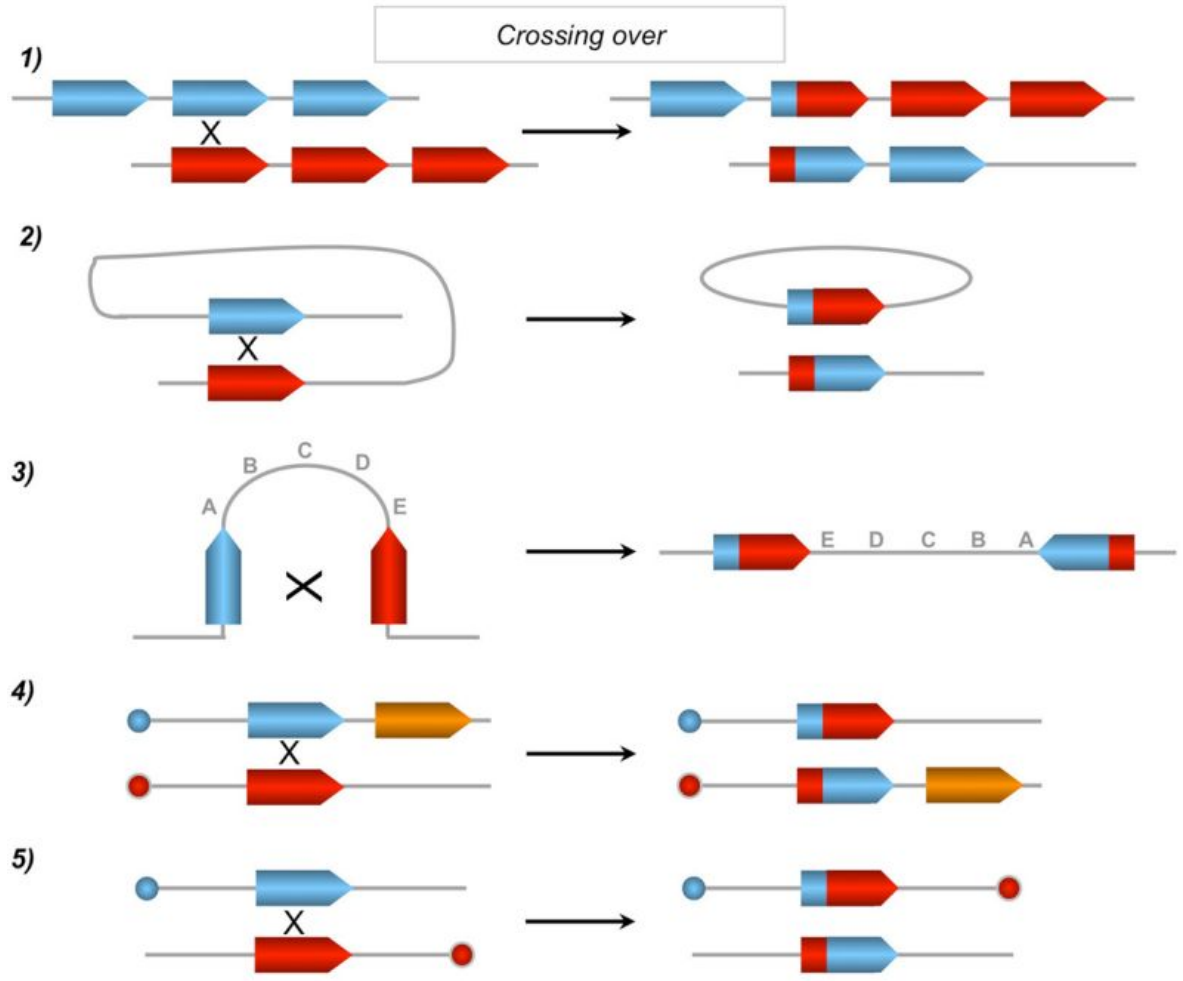

Figure 11 - Schémas des types d'ADN recombiné pouvant être obtenu lors de la réparation d'un DSB par RH.
Réarrangements chromosomiques résultant d'un crossing over (CO) entre séquences répétées. (1) Entre des séquences homologues sur deux chromosomes ou après un échange inégal de chromatides soeurs sur le même chromosome, entraînant l'amplification d'une molécule et la délétion de l'autre. (2) CO intramoléculaire entre deux séquences homologues dans une orientation directe, entraînant l'excision de la séquence intermédiaire. (3) CO intramoléculaire entre deux séquences homologues dans une orientation inversée, entraînant l'inversion du fragment interne. (4) et (5) CO interchromosomique, en fonction de l'orientation des séquences homologues par rapport à leurs centromères (cercles bleus ou rouges); ce processus génère une translocation (4) ou un chromosome dicentrique et acentrique (5). Figure tirée de Guirouilh-Barbat et al. (2014).

c- Impact de la recombinaison illégitime sur les réarrangements chromosomiques

Chez les bactéries, durant la phase stationnaire ou dans certaines cellules ne contenant qu'une seule copie du génome, comme les spores, la réparation d'une DSB ne peut être effectué selon un mécanisme de recombinaison homologue et fait intervenir la recombinaison illégitime qui ne nécessite pas de matrice homologue. Cependant, contrairement à la recombinaison homologue, ces mécanismes sont mutagènes.

Plusieurs mécanismes de recombinaisons illégitimes ont été identifiés et ont été distingués entre la jonction d'extrémités non-homologues classique (ou C-NHEJ pour classical Non-Homologous End-Joining (Moore and Haber, 1996; Seol et al., 2018)) qui est le mécanisme principal de recombinaison illégitime et d'autres mécanismes regroupés sous le nom de jonction d'extrémités non-homologues alternatifs (Frit et al., 2014) (ou A-NHEJ pour Alternative Non-Homologous End Joining).

La C-NHEJ diffère de la recombinaison homologue par sa capacité à réparer les DSB en liant les extrémités de l'ADN, en utilisant très peu (1 à 4 nucléotides) ou pas de complémentarité de séquences (Seol et al., 2018). Ce mécanisme a d'abord été identifié chez les eucaryotes (chez *S. cerevisiae)* (Moore and Haber, 1996). Chez les bactéries, la C-NHEJ fait intervenir deux acteurs principaux, Ku et LigD comme cela a été démontré pour la première fois chez *Mycobacterium tuberculosis* (Weller et al., 2002), (Della et al., 2004).

D'autres mécanismes de recombinaison illégitime ont été identifiés chez les bactéries. Par exemple, *E. coli* ne possède pas de C-NHEJ (Chayot et al., 2010) et il a longtemps été admis que la réparation des DSB reposait sur la seule recombinaison homologue (Wilson et al., 2003). Différents mécanismes indépendants de la voie C-NHEJ ont depuis été mis en évidence, regroupés sous le nom de jonction d'extrémités non-homologues alternatifs (Frit et al., 2014) (ou A-NHEJ pour Alternative Non-Homologous End Joining). Ces voies incluent, par exemple, un mécanisme de réparations des DSB basé sur la présence d'une micro-homologie de séquence au niveau d'un DSB, la voie Microhomology-Mediated End Joining (MMEJ) (Sfeir and Symington, 2015; Truong et al., 2013).

La NHEJ (désigne à la fois les voies C et A-NHEJ) est par définition infidèle et génère des insertions ou des délétions au niveau des sites de ligation modifiant la structure du génome et induit également des réarrangements par fusion de séquences issues de loci non homologues (Bahmed et al., 2011). De plus, si plus de 2 ruptures se produisent simultanément, une jonction incorrecte de ces DSB par C-NHEJ peut également produire des translocations et des réarrangements chromosomiques (Richardson and Jasin, 2000).

3- Le transfert horizontal et intégration de matériel exogène

Le transfert horizontal constitue une force majeure dans la dynamique d'évolution des génomes bactériens. Par exemple, les travaux de Nakamura (Nakamura et al., 2004) ont révélé que 14 % des gènes d'une collection de 116 génomes bactériens ont été soumis à un transfert horizontal récent. L'acquisition de fonctions nouvelles peut faciliter l'adaptation à un nouvel écosystème. Elle permet ainsi d'améliorer la valeur sélective de l'organisme aussi appelée *fitness*. La capacité d'acquisition de

gènes par transfert horizontal varie selon les organismes et est modulée par les conditions environnementales.

Il existe 3 mécanismes du transfert horizontal (**figure 12**) : (i) la transformation naturelle définie comme étant l'acquisition par la bactérie d'ADN se trouvant sous forme libre dans l'environnement (Avery et al., 1944; Griffith, 1928), (ii) la transduction bactérienne qui consiste en un transfert de matériel génétique d'une bactérie donneuse à une receveuse par l'intermédiaire d'un bactériophage (Zinder and Lederberg, 1952), et (iii) la conjugaison bactérienne qui nécessite un contact physique entre deux bactéries après formation d'un pore de conjugaison entre les deux cellules (Lederberg and Tatum, 1946).

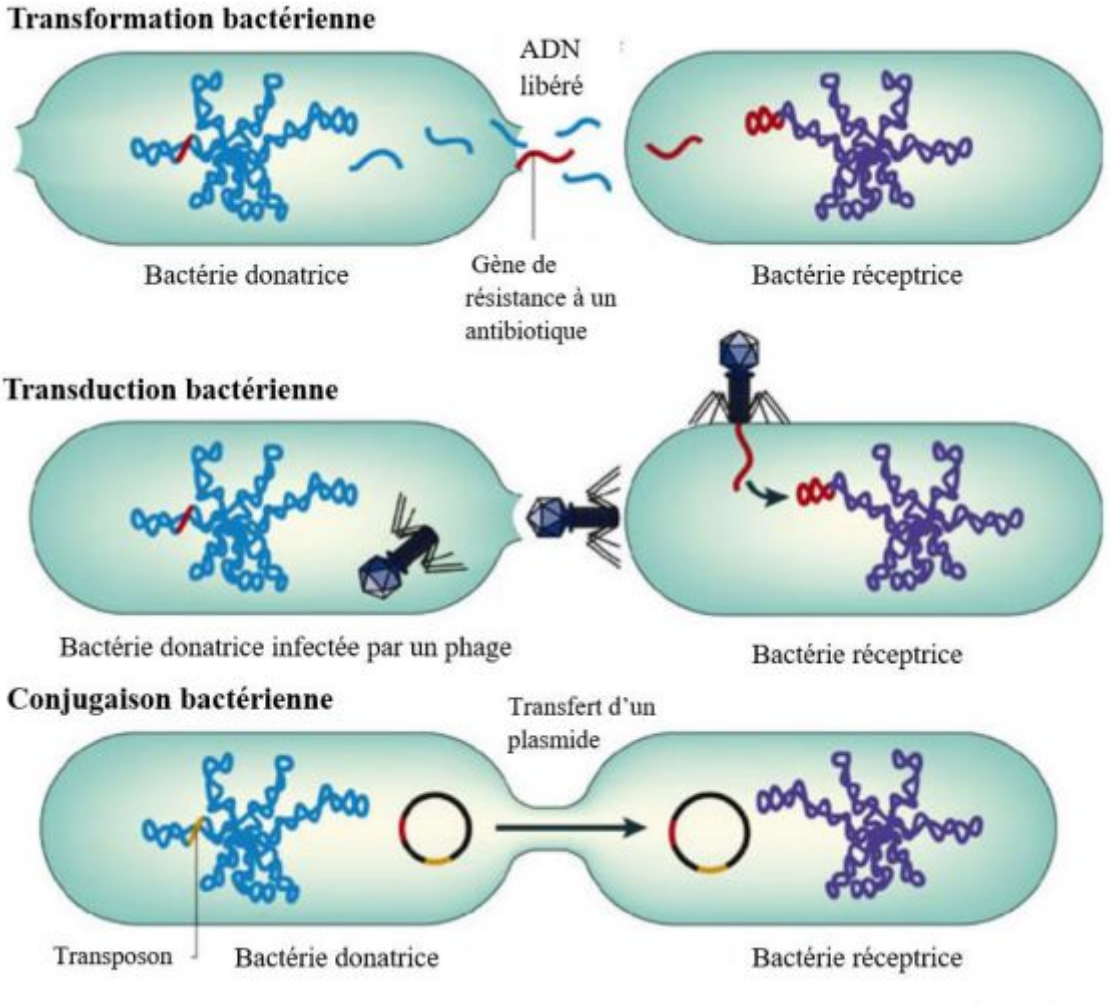

Figure 12 - Représentation schématique des 3 mécanismes majeurs de transfert horizontal impliquant le transfert d'un gène de résistance à un antibiotique. Figure tirée de Furuya and Lowy (2006)

Chez les *Streptomyces*, la conjugaison semble être le mécanisme de transfert prépondérant. En effet, de nombreux éléments conjugatifs réplicatifs et intégratifs (ou ICE pour Integrative and Conjugative

Element) ont été identifiés dans de nombreux génomes de *Streptomyces* (Bibb et al., 1981; Cohen et al., 1985; Pernodet et al., 1984; Ghinet et al., 2011).

Par ailleurs, des événements de transduction n'ont été que très rarement décrits chez *Streptomyces* (Burke et al., 2001; Stuttard, 1982). La transformation naturelle chez *Streptomyces* semble impossible, probablement en raison d'une forte imperméabilité de la paroi.

Pour qu'un gène horizontalement transféré se maintienne dans la descendance de l'organisme receveur, le gène doit généralement conférer un avantage sélectif au receveur (ou à lui-même dans le cas d'un élément génétique "égoïste" (Werren, 2011; Werren et al., 1988)) et la recherche s'est d'abord concentrée sur ce type de gène. Dans un second temps, il a été montré que nombreux gènes identifiés comme transférés ont un effet neutre sur l'organisme hôte (Gogarten and Townsend, 2005). Une règle pour les gènes transférés semble donc être de d'abord, ne pas nuire à l'hôte. Ces gènes, intégrés avec succès chez un receveur, sont souvent exprimés à de faibles niveaux et ne participent pas au métabolisme (Park and Zhang, 2012).

En plus de l'effet des gènes transférés sur l'organisme hôte, la distance phylogénétique entre organismes donneurs et receveurs semble avoir un impact important. En effet, plus les organisme et receveur sont proches phylogénétiquement, plus ils possèdent une grande compatibilité génétique (cohérence dans le taux en GC ou biais de codon par exemple). Cette compatibilité génétique entre donneur et receveur permet l'intégration de matériel génétiques par recombinaison homologue (Lawrence and Retchless, 2009). Par exemple, il a été montré que des souches d'*E. coli* et *Shigella* échangent majoritairement entre souches très proches plutôt qu'avec les souches plus distantes phylogénétiquement (Skippington and Ragan, 2012). Cependant, d'autres travaux suggèrent que l'échange de matériel génétique est dirigé principalement par l'écologie des souches et non pas par la phylogénie (Smillie et al., 2011).

Selon le mécanisme de transfert horizontal considéré, la proximité phylogénétique entre les organismes donneur et receveur peut être plus ou moins importante. La transformation naturelle, qui correspond à la récupération d'ADN exogène présent dans l'environnement ne dépend généralement pas de la parenté entre les organismes (Johnston et al., 2014). La conjugaison nécessite un contact physique entre donneur et receveur, le matériel génétique est transféré via un pilus de conjugaison et peut s'effectuer entre des espèces distantes phylogénétiquement comme par exemple entre la bactérie *Agrobacterium* et une plante (Kyndt et al., 2015). La transduction, où le transfert de matériel génétique est médié par un phage, nécessite la présence de récepteurs spécifiques à la surface des cellules infectées. Par exemple, le phage lambda n'infecte que des cellules d'*E. coli* qui présentent le récepteur Mal à leur surface (Randall-Hazelbauer and Schwartz, 1973).

Certaines bactéries présentent des mécanismes capables de limiter le transfert et le maintien d'ADN étranger. Chez *E. coli* par exemple, l'expression d'un couple de gènes (*traS* et *traT*) inhibe la fixation du pilus nécessaire à la mise en place de la conjugaison (Achtman et al., 1980), c'est l'exclusion de surface. Un autre exemple est le système CRISPR-Cas9 (pour Clustered Regularly Interspaced Short Palindromic Repeats et CRISPR associated protein 9 (Barrangou et al., 2007)) qui limite la réinfection d'une bactérie par un phage et par conséquent, le mécanisme de transduction. Suite à une infection par un bactériophage, les bactéries intègrent des dérivés des séquences génomiques du phage. Cela confère aux bactéries une "mémoire" de l'infection et lui permet de mieux s'en défendre en cas de réinfection. À la suite d'une infection par un phage, des séquences d'ADN du virus sont intégrées au sein de ces séquences CRISPR. Lors d'une nouvelle infection par ce même phage, une interaction entre l'ADN viral conservé dans les séquences CRISPR et la protéine Cas9 est capable de reconnaître et de cliver l'ADN du phage empêchant toute contamination par ce phage (Barrangou et al., 2007).

La restriction-modification (RM) est le mécanisme de défense contre le matériel exogène le plus répandu chez les bactéries (Makarova et al., 2013; Oliveira et al., 2014). Le système RM est composé d'une méthyltransférase capable de méthyler une séquence spécifique et d'une enzyme de restriction (activité endonucléase) qui clive l'ADN non méthylé au site de reconnaissance de l'enzyme (Mruk and Kobayashi, 2014). Cependant, les petits plasmides peuvent naturellement contourner cette restriction, car la séquence qu'ils présentent n'est pas suffisamment étendue pour présenter ces marqueurs.

C- Comment fonder une approche de génomique comparée robuste ?

1 - Evolution des méthodes de séquençage

Dans les années 1970, la technologie Sanger (Sanger et al., 1977b) a permis de séquencer les premiers organismes (le premier génome séquencé fut celui du bactériophage φX174 (Sanger et al., 1977a)) et d'obtenir le premier assemblage complet d'un génome bactérien, *Haemophilus Influenzae Rd*, en 1995 (Fleischmann et al., 1995). Cette technologie a été développée jusque dans les années 1990 et a constitué la "première génération" des technologies de séquençage. Dans les années 2000, une nouvelle technologie de séquençage nommée Next Generation Sequencing (NGS) s'est généralisée. Cette nouvelle technologie a permis l'augmentation de la quantité de données séquencées ainsi que la réduction du coût de séquençage. Le terme NGS fait référence à plusieurs techniques de séquençages , mais elles sont toutes basées sur une parallélisation massive des réactions biochimiques permettant la lecture de chacun des nucléotides composant l'ADN ou l'ARN. De plus, la rupture technologique avec le séquençage Sanger est principalement due à la disparition de l'étape d'électrophorèse (qui permettait la séparation et l'identification des acides nucléiques) qui ralentissait et augmentait le coût des séquençages.

Dans le contexte de ces travaux de thèse, les génomes utilisés sont issus de technologies de séquençages différentes, du premier génome de *Streptomyces* entièrement séquencé et assemblé en 2002 (*S. coelicolor* A3(2), technologie Sanger (Bentley et al., 2002)), à nos jours.

Les méthodes utilisées, ainsi que les objectifs des équipes ayant séquencées ces génomes peuvent avoir un impact sur la qualité des séquences. Avoir un indicateur sur la qualité de chacune d'elle est donc primordiale avant d'aborder des problématiques de génomique comparée.

2- Estimer la qualité des séquences génomiques disponibles

Une séquence génomique est une hypothèse. En effet, malgré les avancées technologiques, il est encore impossible de lire parfaitement et sans interruption une séquence d'une extrémité à l'autre. Pour contourner cette contrainte, les génomes sont fragmentés en fragments plus petits, inégaux en taille, et avec une fidélité imparfaite (*ie* insertion de nucléotide ne correspondant pas à la séquence biologique). Ces fragments sont ensuite assemblés lors de l'assemblage *de novo* (Nagarajan and Pop, 2013). De ces étapes résulte une séquence supposée correspondre à la réalité biologique. Des choix techniques sont faits à chaque étape du processus de séquençage impactant la qualité de la séquence finale.

Un grand nombre de génomes de *Streptomyces* sont assemblés à partir des technologies 454, Illumina, PacBio et nanopore. Les méthodes 454 et Illumina génèrent des lectures courtes (généralement

inférieurs à 450 pb) pouvant mener à des assemblages incomplets tandis que PacBio génére des lectures plus longues, de l'ordre du kilobase et le séquençage nanopore peut générer des lectures mesurant jusqu'à 2 Mégabases. Chez *Streptomyces* les NRPS (Non-Ribosomal Peptide Synthase) et les PKS (PolyKetide Synthase) sont de grandes protéines modulaires constituées de nombreux domaines répétés les rendant d'autant plus compliqués à assembler à partir de lectures courtes. Le cas de *S. sp.* Mg1 illustre parfaitement l'impact que peut avoir une technique de séquençage sur la séquence finale obtenue. Ce génome a été séquencé et assemblé par une même équipe selon 2 méthodes différentes (Hoefler et al., 2013) : PacBio et 454/Illumina. L'assemblage généré par la méthode PacBio est environ 20 % plus grand que celui généré par la méthode 454/Illumina (séquence de 8,705,754 pb et 7,260,368 pb, respectivement) soit un écart d'environ 1,3 Mb. De plus, des régions de l'assemblage issu de 454/Illumina sont absentes de l'assemblage issu de PacBio et réciproquement, révélant que l'assemblage 454/Illumina n'est pas un simple sous-groupe de l'assemblage PacBio.

La qualité des assemblages est souvent motivée par son utilisation et dans le cas des *Streptomyces*, le principal moteur est la découverte de nouveaux de regroupements de biosynthèse de métabolites secondaires. Simultanément au développement des techniques de séquençage, les outils de mining des génomes ont aussi connu un fort développement et sont de plus en plus tolérants sur l'état d'assemblage des génomes, avec par exemple, antiSMASH (Blin et al., 2019) qui est devenu un standard dans la recherche de nouveaux regroupements de gènes du métabolisme secondaire. L'utilisation quasi-systématique d'antiSMASH a fortement enrichi les cibles pour l'identification de molécules bioactives comme l'illustre le cas de *S. roseochromogenes* où antiSMASH prédit 43 nouveaux regroupements de biosynthèse métabolites secondaires, en plus de celui de la clorobiocine déjà mis en évidence expérimentalement (Rückert et al., 2014) à partir d'un niveau d'assemblage faible (433 contigs). Paradoxalement, la capacité d'antiSMASH à prédire des regroupements de métabolites secondaires à partir de génomes fragmentés comme complets n'encourage pas à un gain de qualité dans les nouveaux génomes de *Streptomyces* séquencés dans le but d'identifier de nouveaux regroupements de métabolites secondaires.

Ces qualités d'assemblage variables n'impactent que peu la recherche de nouveaux regroupements de métabolites secondaires, cependant, l'utilisation de ces génomes pour des applications différentes peut avoir une grande importance sur la validité des résultats. Des méthodes ont donc été développées pour estimer la qualité d'un assemblage *de novo*. Une première stratégie s'appuie sur la "complétude" d'un assemblage : si l'assemblage d'un génome est complet, il doit contenir l'ensemble de ses gènes (intacts). L'outil BUSCO (Benchmarking Universal Single-Copy Orthologs (Simão et al., 2015)) estime la qualité d'un assemblage en cherchant des gènes conservés au sein d'un niveau taxonomique (par exemple, chez les Actinobactéries) et toujours présents en un seul exemplaire dans les génomes. Identifier chacun de ces gènes complets et en un seul exemplaire dans les génomes est un argument

pour la complétude de ces génomes. À l'inverse, l'absence, la duplication ou la troncature de ces gènes conservés suggèrent des erreurs dans l'assemblage. En 2015, l'analyse de 653 génomes de *Streptomyces* avec BUSCO selon un groupe de gènes communs à toutes les bactéries a révélé que seuls 63,4% des génomes déposés obtenaient un score BUSCO parfait (Studholme, 2016). Une autre catégorie d'estimation de la qualité d'un assemblage est "l'exactitude" qui correspond à une estimation base par base qu'une position soit exempte d'erreur. Ce genre d'approche se base sur les données brutes d'assemblage et recherche des anomalies dans le profil de couverture des lectures comme des déviations par rapport à une couverture lissée qui correspond à l'approche de REAPR (Recognition of Errors in Assemblies Using Reads (Hunt et al., 2013)). Malheureusement les données brutes d'assemblage ne sont que rarement accessibles et ne peuvent donc pas être utilisées pour estimer la qualité d'un grand nombre d'assemblages et l'exactitude est souvent réduite à la valeur de la profondeur de séquençage.

En plus de la qualité des séquences génomiques, tous les génomes ne sont pas annotés de la même façon d'où la nécessité d'une ré-annotation uniforme de tous les génomes.

D- Le genre *Streptomyces*, modèle d'étude de la diversité génétique

1- Définition du genre et caractéristiques générales

Le genre *Streptomyces* fut proposé au début des années 1940, sur la base des caractéristiques physico-chimiques de la membrane cellulaire (Waksman and Henrici, 1943). Les *Streptomyces* sont des bactéries Gram positives à haut pourcentage en bases G et C et appartiennent à l'ordre des *Actinobactéries*, longtemps considérés comme intermédiaire entre les bactéries et les champignons. Ces bactéries sont abondantes dans le sol et l'apparition du genre remonterait à environ 400 millions d'années (Chater, 2016). Leurs fortes capacités à solubiliser des composants de la paroi cellulaire ou de la surface des plantes, des champignons et des insectes (Chater et al., 2010) suggère qu'ils ont joué un rôle dans la dégradation des plantes terrestres précoces et donc dans la formation du sol primitif.

La classification du genre *Streptomyces* fut historiquement très discutée et améliorée au fil du temps en particulier grâce à des avancées technologiques. A cause des antibiotiques synthétisés par les *Streptomyces*, un criblage important à la recherche de nouvelles molécules bioactives des années 1940 à 1970 a mené à une sur-définition du genre *Streptomyces* faisant croître le nombre d'espèces décrites d'environ 40 à plus de 1,000 entre 1940 et 1957 et allant jusqu'à environs 3,000 dans les années 1970 (Williams et al., 1983). Cette période est marquée en particulier par la découverte de la Streptomycine chez *Streptomyces griseus* en 1943 (récompensé du prix Nobel de physiologie ou médecine en 1952). Cette sur-représentation a entraîné le basculement de certaines espèces à l'intérieur ou à l'extérieur du genre au gré des différentes études et présente surtout un très grands nombres d'espèces synonymes. À partir de 1964 et de l'apparition de critères de caractérisation des *Streptomyces* mis en place dans le cadre d'un projet international (International *Streptomyces* Project, ISP), les espèces furent progressivement re-décrites. Cependant, les "frontières" du genre furent très discutées. Par exemple, les *Kitasatospora* furent unifiés au genre *Streptomyces* sur la base de la similarité de leur ARN 16S (Wellington et al., 1992). Dans un second temps, il fut montré que les *Kitasatospora* forment un groupe monophylétique stable à l'extérieur du genre *Streptomyces* (Zhang et al., 1997) re-définissant le genre *kitasatospora* comme initialement décrit (Omura et al., 1982).

Les *Streptomyces* présente un cycle de développement particulier, générant plusieurs "tissus" différents (**figure 13**) : sur un milieu de culture solide, la germination d'une spore produit des filaments mycéliens à croissance apicale (Flärdh, 2003) capables de se ramifier induisant une augmentation très importante de la biomasse de la colonie. En réponse à une diminution des éléments nutritifs dans l'environnement, le mycélium végétatif produit des branches grandissant vers l'espace aérien. Enfin, les régions distales de ces hyphes aériens forment des chaînes de spores par cloisonnement (*ie* septation).

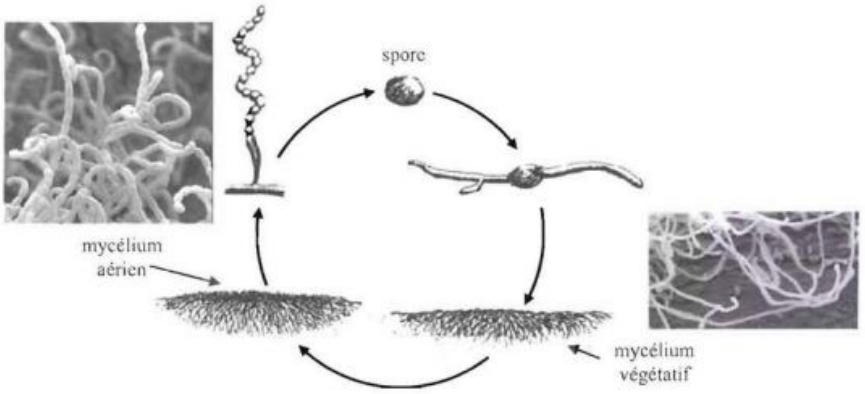

Figure 13 - Schéma du cycle de différenciation chez les *Streptomyces*. (Figure tirée de Choulet (2016))

La majorité des antibiotiques utilisés aujourd'hui sont issus des *Streptomyces* (Chater, 2016). Les analyses génomiques montrent que chaque souche de *Streptomyces* possède le matériel génétique nécessaire à la synthèse de plusieurs métabolites secondaires différents (Charlop-Powers et al., 2015) confortant la recherche de nouvelles molécules bioactives d'intérêts chez ces organismes. Cependant, la capacité des *Streptomyces* à générer des molécules d'intérêt pour l'homme n'est que le premier aspect d'intérêt de ces bactéries. Depuis 2016, plus de 3.200 publications (recherche du mot clef "Streptomyces" dans PUBMED[5]) ont porté sur les *Streptomyces.*

2- Des caractéristiques génomiques originales parmi les bactéries

a- Organisation du chromosome

Les *Streptomyces* possèdent un unique chromosome linéaire de grande taille mesurant en moyenne 8,4 Mb (pouvant aller de 6 à 12 Mb selon les espèces) avec un fort taux en bases G et C (72%). En plus de ce chromosome linéaire, les *Streptomyces* peuvent présenter des plasmides (jusqu'à 7 chez *S. autolyticus*) sous forme circulaire et/ou linéaire.

L'origine de réplication (*oriC*) est localisée dans la région centrale des réplicons linéaires des *Streptomyces* (Musialowski et al., 1994), (Chang and Cohen, 1994). La réplication s'effectue en deux étapes : (i) la réplication bidirectionnelle de l'origine de la réplication vers les télomères, générant des extrémités (3') simple brin au niveau des télomères (Huang et al., 1998); (ii) ces extrémités sont complétées dans un second temps selon un mécanisme amorcé par les protéines terminales (TPs, détaillées plus bas) liées covalemment à l'extrémité 5' (Chen et al., 2002). Le mécanisme d'initiation

[5] https://www.ncbi.nlm.nih.gov/pubmed/

de la réplication chez *Streptomyces* est donc commun à ceux retrouvés chez les bactéries. À l'inverse, le mécanisme de terminaison ("end-patching") est unique parmi les bactéries (Yang et al., 2017).

Les télomères correspondant généralement à des séquences conservées d'environ 170 pb qui contiennent plusieurs séquences palindromiques. Par exemple, la séquence télomérique de *S. coelicolor* A3(2) contient 7 motifs palindromiques, capable de se replier et de générer des structures secondaires (**figure 14**). Deux structures différentes, dites en "oreilles de lapin" et en "feuille de trèfle", ont été proposées pour l'extrémité 3' générée en fin de réplication.

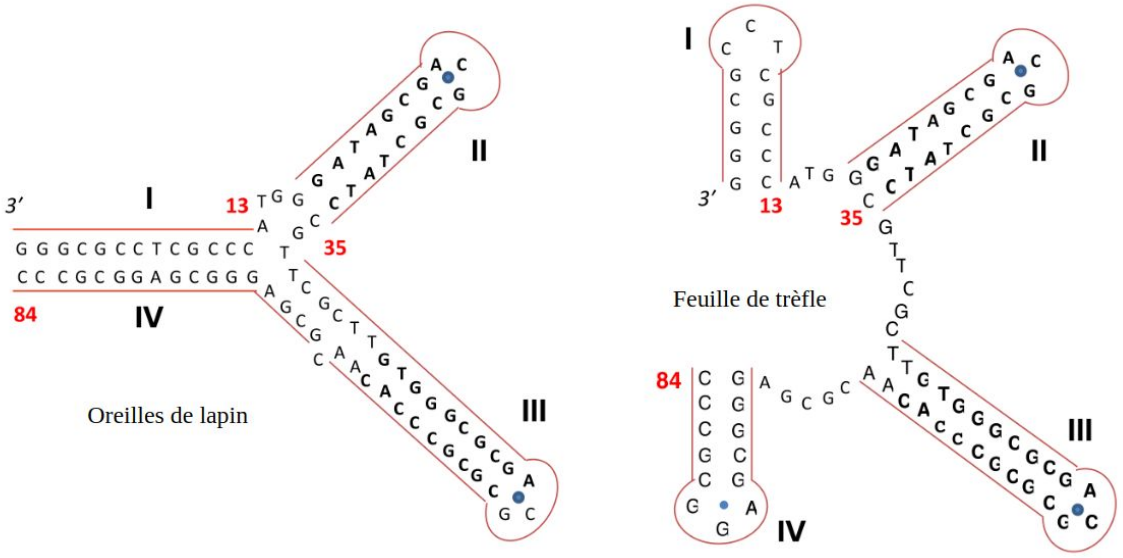

Figure 14 - Structures de l'extrémité 3' des télomères de *S. coelicolor* A3(2). Structures les plus stables (ΔG = −140.1 kJ/mol pour la structure en oreilles de lapin, ΔG = −140.9 kJ/mol pour la structure en feuille de trèfle). Les chiffres romains identifient les différents domaines des structures. (D'après Yang et al., 2017).

Les protéines terminales (TP pour Terminal proteins) liées covalemment au niveau des régions télomériques interagissent entre elles, générant une topologie circulaire de l'ADN (Wang et al., 1999; Yang and Losick, 2001). Les TP sont codées par un gène généralement localisé aux extrémités chromosomiques, le gène *tpg* qui colocalise avec le gène *tap* (Bao and Cohen, 2003). Les protéines Tap contiennent un domaine de liaison capable d'interagir à la fois avec la protéine Tpg et la séquence télomérique créant un complexe protéique terminal essentiel à la terminaison de la réplication et par conséquent au maintien de la forme linéaire du chromosome. Les télomères de chromosomes et des plasmides sont variables, et différents jeu de protéines terminales ont été caractérisées (Yang et al., 2017; Huang et al., 2007).

En plus des séquences télomériques, les extrémités chromosomiques des *Streptomyces* contiennent des régions inversées répétées (ou TIR pour Terminal Inverted Repeat, (Lin et al., 1993; Weaver et al., 2004) de composition et de taille variables.

b- La compartimentation du chromosome

Avec le séquençage et l'assemblage du premier génome de *Streptomyces*, *S. coelicolor* A3(2) (Bentley et al., 2002) puis de *S. avermitilis* (Ikeda et al., 2003) une disposition génétique particulière fut suggérée : l'ensemble des gènes supposés essentiels sont concentrés dans la région centrale du chromosome (dite "core") tandis que les gènes accessoires se situent préférentiellement dans les bras chromosomiques (**figure 15**).

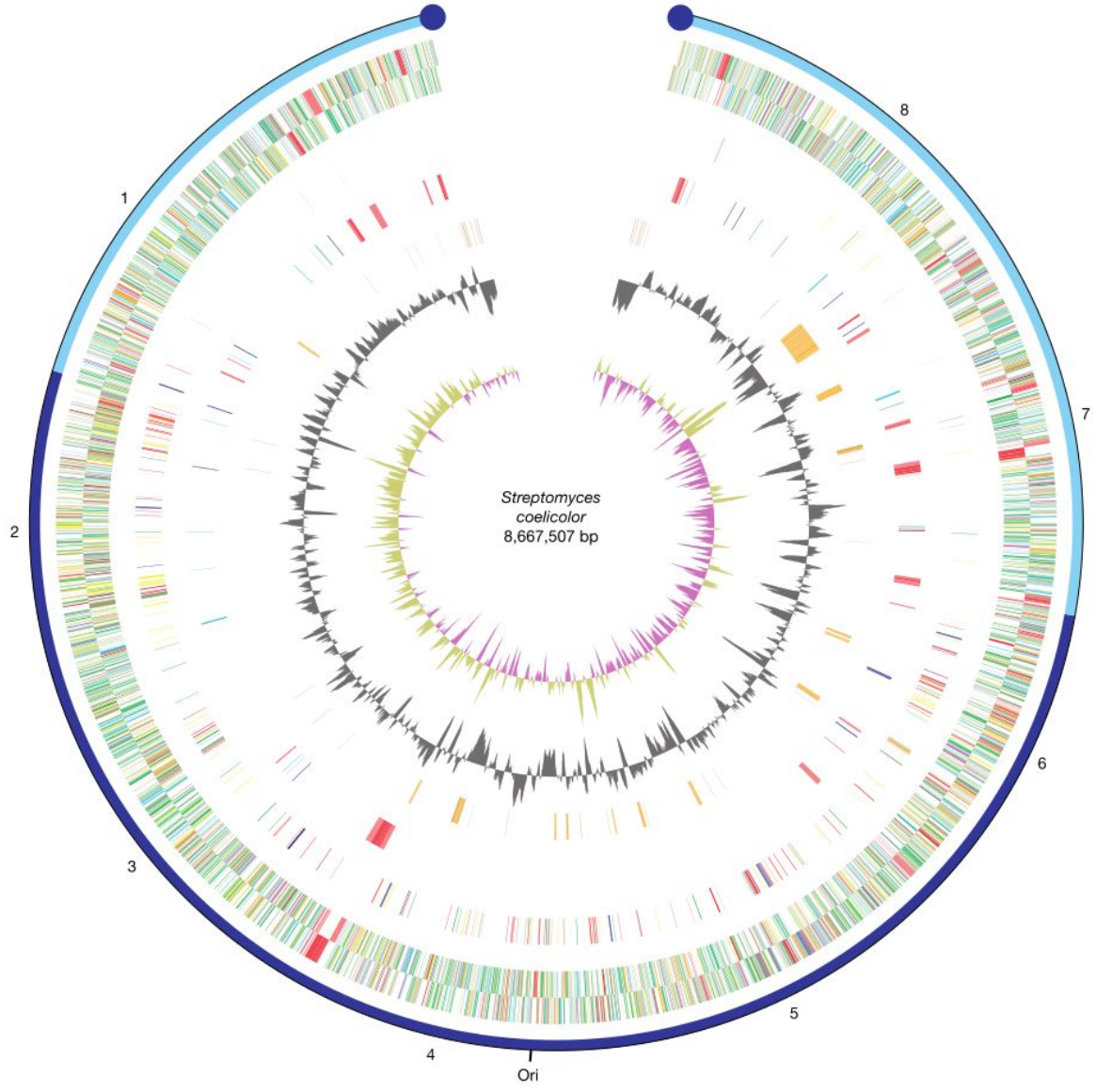

Figure 15 - Représentation schématique du chromosome de *S. coelicolor* A3(2). L'échelle extérieure représente les positions nucléotidiques du chromosome (en mégabases). La région en bleu foncé correspond à la région "core", les régions bleues claires identifient les bras. Les cercles 1 et 2 (de l'extérieur vers l'intérieur) identifient respectivement les gènes localisés sur les brin continu et discontinu. Le cercle 3, les gènes supposés essentiels. Le cercle 4 les gènes accessoires. Les 2 suivants correspondent aux taux en G+C puis au biais en GC. (D'après Bentley et al., 2002).

La conservation de l'ordre des gènes a aussi été estimée au travers d'une analyse comparative du chromosome de *S. ambofaciens* avec ceux de *S. coelicolor* et *S. avermitilis*. L'ordre des gènes a été estimé selon un indice, le GOC (Gene Order Conservation, (Choulet et al., 2006a)). Au sein d'une fenêtre glissante, le GOC est défini comme le ratio du nombre de paires d'orthologues contigus entre les chromosomes comparées sur le nombre de CDS dans la fenêtre. Cette définition locale de la synténie est dérivé du GOC global (Rocha, 2006) décrit dans la section **A1d - Impact de l'organisation en opérons et en regroupements sur la synténie**. En déplaçant la fenêtre le long du chromosome, une diminution progressive de la synténie vers les extrémités chromosomiques a été mis en évidence (**figure 16**). Cette approche a permis de distinguer une région centrale où la synténie est fortement conservée, contrastant avec les extrémités chromosomiques variables. De plus, en ajoutant une espèce plus distante phylogénétiquement dans l'analyse (*S. scabies*) il est apparu que la taille de la région centrale diminue au fur et à mesure que la distance phylogénétique grandit entre les espèces comparées. Une explication envisagée suite à cette observation est l'existence d'un gradient de recombinaison croissant vers les extrémités chromosomiques, excluant les gènes essentiels de ces régions et entraînant un renouvellement rapide des gènes, moteur de la forte capacité d'adaptation des *Streptomyces*.

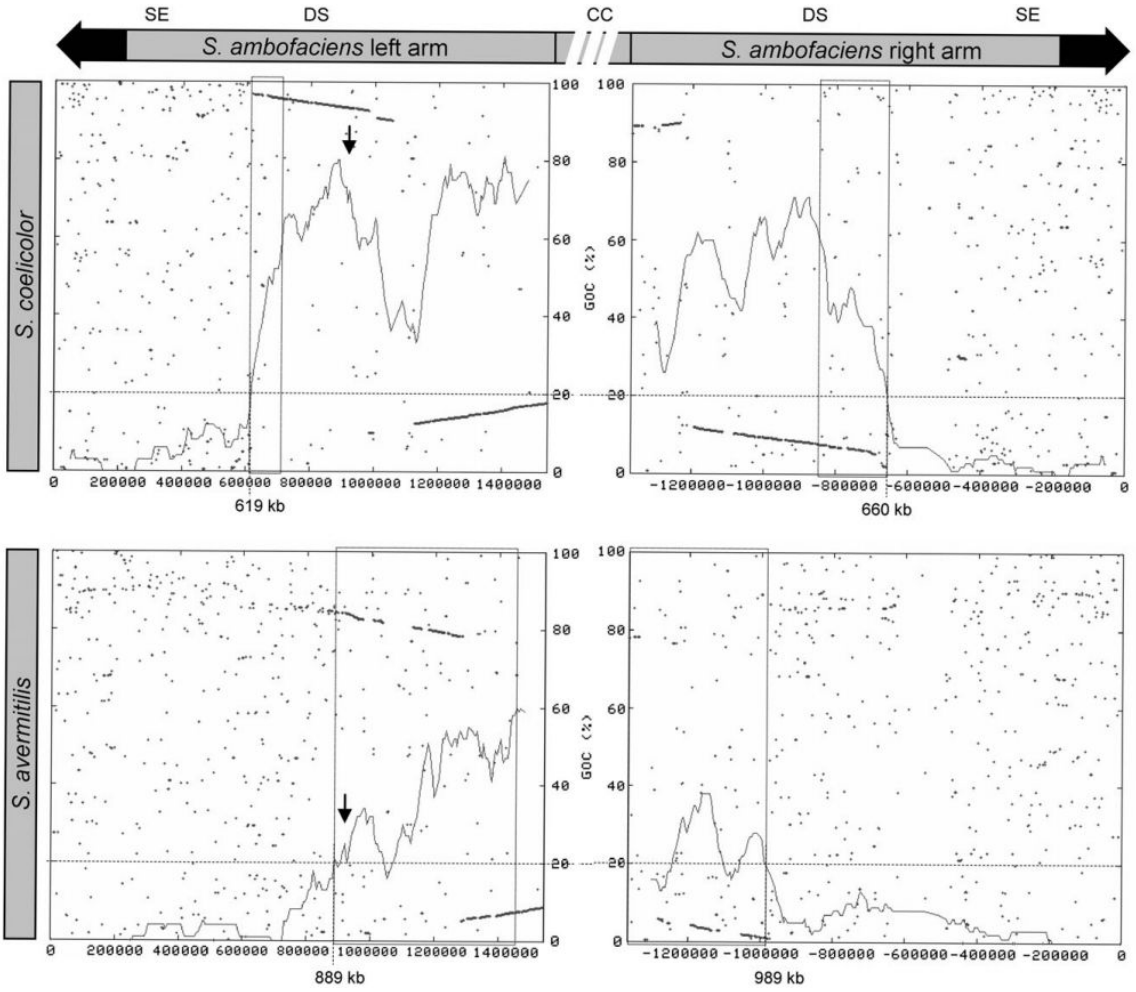

Figure 16 - Profils de GOC le long des bras chromosomiques de *S. ambofaciens* comparées aux génomes de *S. coelicolor* et *S. avermitilis*. Sur chaque graphe, la courbe de GOC est superposé au "dot-plot" de la position des paires d'orthologues. Les rectangles indiquent les région de synténie dégénérée (passage de 60 à 20 % de GOC). Calculs réalisés avec une fenêtre glissante de 100 gènes et un pas de 5 gènes. SE : extrémités spécifiques; DS : régions de synténie dégénérée; CC : région centrale conservée. (D'après Choulet et al., (2006)).

3- La plasticité génomique des *Streptomyces*

Les *Streptomyces* sont des organismes très étudiés, à la fois en raison de leur capacité à synthétiser une très large gamme de molécules bioactives et aussi en raison de l'originalité de l'organisation de leur génome. En plus de ces caractéristiques, les *Streptomyces* sont trouvés dans des environnements changeants illustrant une capacité d'adaptation importante et rapide rendant ces organismes d'autant plus intéressants pour la recherche de mécanismes évolutifs. Aujourd'hui, plus de 200 génomes différents et complets de *Streptomyces* sont disponibles dans la base de données du NCBI[6] illustrant parfaitement l'accumulation de génomes permise par les avancées dans les technologies de

[6] https://www.ncbi.nlm.nih.gov/assembly

séquençage, depuis le séquençage du premier génome complet de *Streptomyces*, *S. coelicolor* A3(2) (Bentley et al., 2002). Ce grand nombre de génomes complets a permis l'émergence de nombreux travaux de génomique comparée et par conséquent des avancées dans la compréhension de l'évolution des génomes de *Streptomyces*.

a- L'instabilité génétique chez *Streptomyces*

Les *Streptomyces* présentent une forte instabilité génétique qui fut d'abord observée lors de culture de colonies en laboratoire où des mutants dépigmentés apparaissent à haute fréquence (environ 0,1%) dans une population sauvage (Leblond and Decaris, 1994; Volff and Altenbuchner, 1998). Cette instabilité est corrélée avec la formation de grandes délétions ou d'amplification dans les bras chromosomiques (Altenbuchner and Cullum, 1985). Chez *S. ambofaciens* une délétion de 2,3 Mb, soit environ un quart du chromosome, a été observée (Fischer et al., 1997). Ces événements de délétions peuvent être accompagnés de différents types d'événements de réarrangements comme la recircularisation du chromosome (Inoue et al., 2003), le remplacement de bras chromosomiques (Uchida et al., 2003) et l'amplification d'ADN (Widenbrant and Kao, 2007).

Le chromosome linéaire des *Streptomyces* peut spontanément se circulariser à la suite de délétions dans les régions terminales (Inoue et al., 2003; Leblond and Decaris, 1994; Volff and Altenbuchner, 1998) qui peuvent atteindre jusqu'à 2 Mb (Fischer et al., 1997). L'analyse de la jonction de circularisation chez 2 mutants obtenu à partir du chromosome de *S. griseus* suite à des événements de délétions (Inoue et al., 2003) a révélé une absence d'homologie de séquence au niveau de la jonction : l'événement de circularisation résulte d'une fusion nucléotidique illégitime entre les deux extrémités du chromosome.

Chez plusieurs espèces de *Streptomyces*, des phénomènes d'amplification d'ADN chromosomique ont également pu être observés (Volff and Altenbuchner, 1998). Il s'agit d'un mécanisme qui engendre la répétition en tandem d'un locus donné en plusieurs dizaines de copies. Ce phénomène ne semble affecter que certains loci : les unités d'ADN amplifiables (ou AUD pour Amplifiable Units of DNA, (Demuyter et al., 1988)). Les amplifications sont généralement retrouvées aux bornes de région délétée, suggérant des mécanismes liés (Aigle et al., 1996; Altenbuchner and Cullum, 1984; Catakli et al., 2003). Des AUD ont été mis en évidence dans plusieurs espèces de *Streptomyces* (Cao et al., 2017; Redenbach et al., 2000; Schmid et al., 1999), elles correspondent souvent à regroupements de métabolites secondaires comme le regroupement des caroténoïdes chez *S. alvus* J1074 (Myronovskyi et al., 2014) ou celui de l'actinorhodine chez *S. coelicolor* (Widenbrant et al., 2008).

b- La variabilité des TIRs

Les séquences inversées répétées (ou TIRs) qui caractérisent les extrémités chromosomiques sont des régions très variables d'une espèce à l'autre. Elles peuvent couvrir plusieurs centaines de kilobases (1.06 Mb chez *S. coelicolor* M600 (Weaver et al., 2004)) ou être restreintes aux séquences palindromiques qui forment les télomères (167 pb chez *S. avermitilis* (Ikeda et al., 2003)).

La taille des TIR est également variable au sein de l'espèce. Chez *S. ambofaciens*, alors qu'elles présentent une taille de 210 kb chez la souche sauvage (Leblond et al., 1996), des TIR de 480 et 850 kb ont été identifiées chez deux souches mutantes (Fischer et al., 1998). Ce polymorphisme résulte de la délétion d'un des deux bras entraînant son remplacement par le second. Cet échange a pu se produire grâce à un événement de recombinaison homologue (crossing-over ou BIR) entre deux gènes (HasR et HasL) situés sur chacun des bras chromosomiques et partageant 99 % d'identité nucléotidique (**figure 17**).

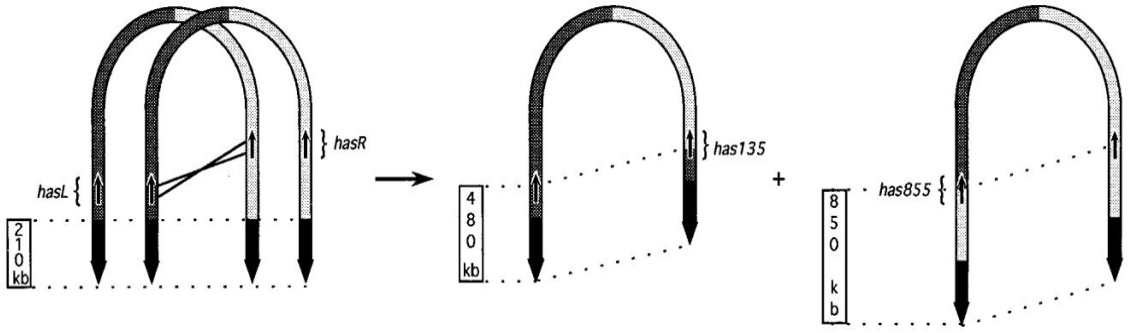

Figure 17 - Représentation schématique de la recombinaison entre les loci *hasL* et *hasT* de copies du chromosome sauvage de *S. ambofaciens* DSM 40697 générant 2 versions mutantes du chromosome : l'une présentant des TIRs de 480 kb et la seconde avec des TIRs de 850 kb. (Figure tirée de Fischer et al., 1998)

Des événements de recombinaison comparables ont aussi été observés chez *S. coelicolor* A3(2) (Widenbrant and Kao, 2007) et *S. griseus* (Uchida et al., 2003) générant des TIR de 405 kb contre 24 kb pour la souche sauvage.

La grande variabilité des TIR impacte logiquement le contenu en gènes. Les deux souches *S. ambofaciens* ATCC23877 et DSM40697, présentent respectivement des TIR de 198 kb et 213 kb (Choulet et al., 2006b). Parmi les 215 CDS prédites dans la souche DSM40697, 65 (~30 %) ne sont pas retrouvées dans les TIRs de la souche ATCC 23877. Réciproquement, 45 (~24 %) des 194 CDSs de ATCC23877 ne sont pas retrouvées dans les TIR de la souche DSM40697 (Choulet et al., 2006b). Ainsi, malgré la proximité phylogénétique de ces deux souches, une fraction significative de gènes distingue leurs TIR (4%).

c- Une forte variabilité génétique

Des travaux de génomique comparée ont été préalablement menés sur le genre *Streptomyces* à partir des génomes complets de *Streptomyces*. Ces études ont été menées en 2012 (Zhou et al., 2012) et 2015 (Kim et al., 2015) et ont utilisées respectivement 5 et 17 génomes, pour aborder la question de l'évolution du contenu en gènes chez les *Streptomyces*. Malgré la différence d'échelle entre ces études, les 2 approches ont révélé que les *Streptomyces* présente un pan-génome illustrant une grande diversité génétique : 17.362 avec 5 espèces (Zhou et al., 2012) et 34.592 gènes avec 17 espèces (Kim et al., 2015). Le pan-génome se définit comme l'ensemble des gènes présents au sein du niveau taxonomique étudié. Il est constitué à la fois du core génome, c'est-à-dire la partie commune à l'ensemble des espèces étudiées et du génome accessoire qui correspond aux gènes présents uniquement dans quelques voire une seule espèce. La taille du pan-génome s'accroît avec le nombre d'espèces considérées (**figure 18**) révélant le caractère ouvert du pan-génome du genre *Streptomyces* où les gènes accessoires sont majoritaires. En effet, le core-génome, pour 17 espèces, s'élève à 2.018 gènes (**figure 18**) ce qui représente 24 à 38 % des génomes de cette étude (Kim et al., 2015).

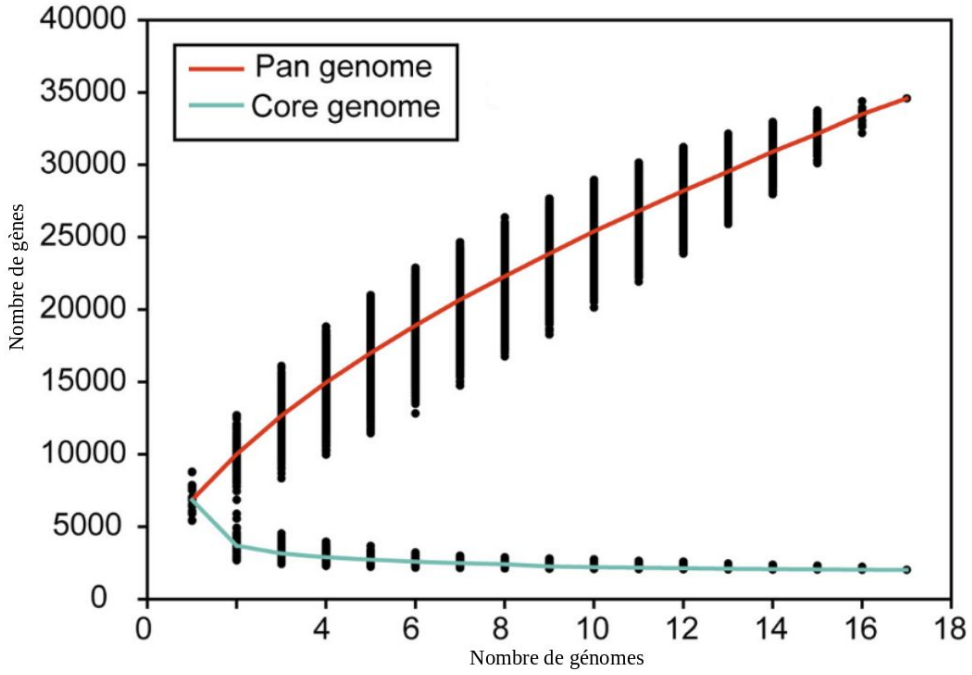

Figure 18 - Evolution du nombre de gènes composant le core et pan-génome selon le nombre de génomes considérés. (Figure tirée de Kim et al., 2015).

E - Objectifs

De nombreux travaux ont exploré la dynamique du génome entre différentes espèces de *Streptomyces* mettant en évidence la plasticité du génome. Cependant, l'ampleur et l'évolution de cette plasticité au sein du genre sont peu décrites. En tirant parti de nombreux génomes de *Streptomyces* complets accessibles sur les bases de données publiques en plus de nouveaux génomes séquencés durant ces travaux, les objectifs de ma thèse étaient d'explorer la dynamique des génomes en fonction de la distance phylogénétique et de modéliser cette dynamique selon les mécanismes moléculaires identifiés.

Pour la suite de ce manuscrit, les résultats seront détaillés en 2 parties distinctes et seront présentés sous forme d'articles scientifiques. En complément des articles, des résultats complémentaires seront détaillés dans chaque partie. La première partie vise à explorer la dynamique des génomes du genre *Streptomyces* et la seconde fera un focus sur l'évolution des génomes des *Streptomyces,* mais à micro-échelle: à partir d'une population de *Streptomyces* isolée depuis un même fragment de sol.
Par la suite, ces résultats seront discutés dans une discussion générale et les perspectives à apporter seront détaillées.

RÉSULTATS

L'ensemble des résultats a été organisé en deux parties et intègre la dynamique du génome des *Streptomyces* à différentes profondeurs phylogénétiques selon la collection d'espèces utilisée. La première partie s'appuie sur un échantillon de génomes représentatif de l'ensemble du genre bactérien. La deuxième partie se concentre sur l'évolution des génomes des *Streptomyces* sur des temps évolutifs beaucoup plus restreints, au sein des souches appartenant à la même espèce et issues d'un même micro-habitat (c'est-à-dire des isolats conspécifiques).

I- Exploration de la dynamique des génomes au niveau du genre *Streptomyces*

A- Constitution de la collection de génomes du genre *Streptomyces*

Afin de comparer les génomes de *Streptomyces*, il a d'abord été nécessaire de sélectionner les génomes d'espèces représentatives du genre parmi l'ensemble des génomes de *Streptomyces* disponibles dans les bases de données. Pour cela, une collecte exhaustive des génomes complets (c'est-à-dire entièrement assemblés) disponibles a été réalisée avec la difficulté que ce nombre a augmenté considérablement tout au long de ce travail de thèse. Le nombre de génomes est passé de 50 en 2016 à 234 en Janvier 2020 selon la base de données RefSeq (O'Leary et al., 2016) . Cela nous a confronté à la question de l'agrandissement constant de notre jeu de données et de son impact sur la validité des résultats. Afin de limiter le risque d'obsolescence de notre jeu de données, nous avons cherché à sélectionner un jeu de génomes représentatifs de la diversité du genre.

1- Un jeu de données représentatif du genre *Streptomyces*

Il s'agit de sélectionner le plus petit nombre d'espèces ou de génomes permettant de décrire l'histoire évolutive de ce genre. Cette démarche permet, d'une part, de s'affranchir de l'accroissement constant de la masse de données génomiques disponibles résultant de l'exploitation de nouvelles technologies de séquençage de l'ADN, et d'autre part, de limiter la quantité de données à traiter dans les approches de bio-informatique.

Le nombre de génomes complets de *Streptomyces* ne cesse de croître (**tableau 1**), donnant de plus en plus de matière à des analyses à grande échelle.

Une question s'est donc posée de façon récurrente tout au long de ce travail : comment estimer le caractère représentatif de la diversité génétique d'un échantillon d'espèces ? Le corollaire, plus pratique, était de définir la taille minimale du panel de souches à considérer comme représentatif.

2016	2017	2018 (15 Octobre)	2019 (6 Février)	2019 (18 Juillet)	2020 (16 Janvier)
~50	~80	124	135	147	234

Tableau 1: Evolution du nombre de génomes de *Streptomyces* complets disponibles dans la base de données RefSeq. Nombre obtenu avec l'utilisation du moteur de recherche du NCBI[7] avec les paramètres suivants : [Organism] = "Streptomyces"; [Assembly level] = "Complete genome"; [Status] = "Latest RefSeq".

Compte tenu de l'évolution du nombre de génomes disponibles au cours de ces dernières années, les résultats acquis et présentés ici n'ont pas été obtenus avec le même jeu de données. Trois collections successives ont été utilisées et seront désignées selon le nombre de génomes qu'elles contiennent : 42, 81, 135 (**tableau 2**) et 234. La collection 234 a été utilisée pour mettre à jour l'article présenté dans la section suivante **(IB- Evolution des régions subtélomériques du chromosome linéaire des *Streptomyces***), les différents résultats présentés ici sont tous obtenus avec les collections présentées dans le **tableau 2**.

[7] https://www.ncbi.nlm.nih.gov/assembly/advanced/

Souche	Collection 42	81	135	GENOME ASSEMBLY
Streptomyces sp. 11 1 2			x	ASM359554v1
Streptomyces sp. 3211		x	x	ASM202838v1
Streptomyces sp. 4F	x	x	x	ASM148470v1
Streptomyces sp. 769	x	x	x	ASM81602v1
S. actuosus ATCC 25421			x	ASM320803v1
Streptomyces sp. ADI95 16			x	ASM372149v1
S. albidoflavus J1074	x	x	x	ASM35952v1
S. albidoflavus SM254		x	x	ASM157738v1
S. albireticuli MDJK11		x	x	ASM219245v1
S. alboflavus MDJK44		x	x	ASM218967v2
S. albulus CK 15			x	ASM93518v3
S. albulus NK660	x	x	x	ASM69523v1
S. albulus ZPM	x	x	x	ASM96351v1
S. albus BK3 25		x	x	ASM175342v1
S. albus DSM 41398	x	x	x	ASM82700v1
S. albus ZD11			x	ASM367534v1
S. alfalfae ACCC40021		x	x	ASM197502v1
S. ambofaciens ATCC 23877	x	x	x	ASM126788v1
S. ambofaciens DSM 40697	x	x	x	ASM163286v1
S. atratus SCSIO ZH16			x	ASM333086v1
S. autolyticus CGMCC0516		x	x	ASM198397v1
S. avermitilis MA 4680	x	x	x	ASM976v2
S. bacillaris ATCC 15855			x	ASM326867v1
S. bingchenggensis BCW 1	x	x	x	ASM9238v1
S. brunneus CR22			x	ASM395571v1
S. cattleya DSM 46488	x	x	x	ASM24016v1
S. cattleya NRRL 8057	x	x	x	ASM23730v1
S. cavourensis 1AS2a			x	ASM280416v1
S. cavourensis TJ430			x	ASM335244v1
Streptomyces sp. CB09001			x	ASM336979v1
Streptomyces sp. CC0208			x	ASM344373v1
Streptomyces sp. CCM MD2014		x	x	ASM77204v1
Streptomyces sp. CdTB01	x	x	x	ASM148456v1
Streptomyces sp. CFMR 7	x	x	x	ASM127809v1
S. chartreusis NRRL 3882			x	NRRL3882
S. clavuligerus F1D 5			x	Scla_1.0
S. clavuligerus F613 1		x	x	ASM169367v1
Streptomyces sp. CLI2509		x	x	ASM228807v1
Streptomyces sp. CMB StM0423			x	ASM284728v1
Streptomyces sp. CNQ 509	x	x	x	ASM101103v1
S. coelicolor A3(2)	x	x	x	ASM20383v1
S. collinus Tu 365	x	x	x	ASM44487v1
S. cyaneogriseus noncyanogenus NMWT 1	x	x	x	ASM93144v1
S. davawensis	x			ASM34932v1
Streptomyces sp. DUT11			x	ASM284852v1
Streptomyces sp. endophyte N2			x	ASM410448v1
Streptomyces sp. fd1 xmd		x	x	ASM200768v1
S. formicae KY5		x	x	ASM255654v1
Streptomyces sp. FR 008		x	x	ASM143176v1
S. fradiae NKZ 259			x	ASM357348v1
S. fulvissimus DSM 40593	x	x	x	ASM38594v1
S. fungicidicus TXX3120			x	ASM366543v1
S. gilvosporeus F607		x	x	ASM208219v1
S. glaucescens GLA O	x	x	x	ASM76121v1
S. globisporus C 1027	x	x	x	ASM26134v2
S. globisporus TFH56			x	ASM314754v1
S. globosus LZH 48			x	ASM332537v1
Streptomyces sp. Go 475			x	ASM333084v1
S. griseochromogenes ATCC 14511		x	x	ASM154262v2
S. griseorubiginosus 3E 1			x	ASM359523v1
S. griseoviridis F1 27			x	ASM399439v1
S. griseus NBRC 13350	x	x	x	ASM1060v1
Streptomyces sp. GSSD 12			x	ASM334496v1
Streptomyces sp. HNM0039			x	ASM309751v1
S. hundungensis BH38			x	ASM362781v1
S. hygroscopicus jinggangensis 5008	x	x	x	ASM24535v1
S. hygroscopicus jinggangensis TL01	x	x	x	ASM34084v1
S. hygroscopicus limoneus KCTC 1717 I	x	x	x	ASM144707v1

Souche	Collection 42	81	135	GENOME ASSEMBLY
S. hygroscopicus XM201		x	x	ASM202187v1
S. koyangensis VK A60T			x	ASM342892v1
Streptomyces sp. KPB2			x	ASM395005v1
S. lavendulae CCM 3239			x	ASM280384v1
S. leeuwenhoekii sleC34	x	x	x	sleC34
S. lincolnensis LC G			x	ASM334444v1
S. lincolnensis NRRL 2936		x	x	ASM168535v1
S. lividans TK24	x	x	x	ASM73910v1
S. lunaelactis MM109			x	ASM305455v1
S. lateoverticillatus CGMCC 15060			x	ASM397071v1
S. lydicus 103		x	x	ASM172948v1
S. lydicus A02	x	x	x	ASM95203v2
S. lydicus GS93 23		x	x	ASM198444v1
S. lydicus WYEC 108			x	ASM399437v1
Streptomyces sp. M2			x	ASM410450v1
Streptomyces sp. M56			x	ASM281240v1
S. malaysiensis DSM4137		x	x	ASM259133v1
Streptomyces sp. Mg1	x	x	x	ASM41226v2
Streptomyces sp. MK45			x	ASM396353v1
Streptomyces sp. MOE7		x	x	ASM209033v1
Streptomyces sp. NEAU S7GS2			x	ASM317327v1
S. nigra 452			x	ASM307405v1
S. niveus SCSIO 3406		x	x	ASM200917v1
S. noursei ATCC 11455		x	x	ASM170427v1
S. olivaceus KLBMP 5084		x	x	ASM176737v1
S. olivoreticuli ATCC 31159			x	ASM339113v1
S. ongäcola HNM0071			x	ASM312236v1
Streptomyces sp. P3			x	ASM303247v1
S. pactum ACT12		x	x	ASM200522v1
Streptomyces sp. PAMC26508	x	x	x	ASM36480v1
S. parvulus 2297		x	x	ASM166004v1
S. peucetius caesius ATCC 27952			x	ASM277753v1
S. pluripotens MUSC 135		x	x	ASM80224v2
S. pluripotens MUSC 137		x	x	ASM81646v4
S. pratensis ATCC 33331	x	x	x	ASM17611v2
S. pristinaespiralis HCCB 10218	x	x	x	ASM127807v1
S. puniciscabiei TW1S1		x	x	ASM173580v1
S. reticuli TUE45	x	x	x	TUE45
Streptomyces sp. RTd22		x	x	ASM165021v1
S. rubrolavendulae MJM4426		x	x	ASM175078v1
Streptomyces sp. S063			x	ASM283267v1
Streptomyces sp. S10 2016		x	x	ASM161179v1
Streptomyces sp. S8		x	x	ASM209499v1
S. sampsonii KJ40		x	x	ASM170419v1
Streptomyces sp. SAT1		x	x	ASM165449v1
S. scabiei 87 22	x	x	x	ASM9130v1
Streptomyces sp. SCSIO 03032		x	x	ASM212830v1
Streptomyces sp. Sge12		x	x	ASM208045v1
Streptomyces sp. SirexAA E	x	x	x	ASM17719v2
Streptomyces sp. SM17			x	ASM291072v2
Streptomyces sp. SM18			x	ASM291077v2
Streptomyces sp. TN58		x	x	ASM194184v1
Streptomyces sp. Tue6075	x	x	x	ASM193163v1
S. venezuelae ATCC 15439	x	x	x	ASM144362v1
S. venezuelae NRRL B 65442	x	x	x	ASM188659v1
S. vietnamensis GIM4 0001	x	x	x	ASM83000v1
S. violaceoruber S21		x	x	ASM208217v1
S. violaceusniger Tu 4113	x	x	x	ASM14781v3
Streptomyces sp. W1SF4			x	ASM395003v1
Streptomyces sp. WAC 01438			x	ASM394552v1
Streptomyces sp. WAC 01529			x	ASM394554v1
Streptomyces sp. WAC 06738			x	ASM394550v1
Streptomyces sp. WAC00288			x	ASM294389v1
S. xiamenensis 318	x	x	x	ASM99378v2
S. xinghaiensis S187		x	x	ASM22070v2
Streptomyces sp. XZHG99			x	ASM294683v1
Streptomyces sp. Z022			x	ASM367532v1
Streptomyces sp. ZFG47			x	ASM326105v1

Tableau 2 - Répartition des différentes souches de *Streptomyces* dans les collections 42, 81 et 135. La colonne GENOME ASSEMBLY présente l'identifiant de chaque souche dans la base de données ASSEMBLY[8].

[8] https://www.ncbi.nlm.nih.gov/assembly/

À l'exception d'une souche, *S. davawensis* qui est spécifique de la collection 42, chaque collection est un sous-ensemble de la collection 135. En effet, lorsque la première collection a été constituée, le traitement des séquences était manuel, et l'état d'assemblage incomplet du génome de *S. davawensis* est passé inaperçu. Par ailleurs, entre les différentes versions de la sélection, certaines souches ont changé de nom (ex. *S. albus* J1074 est devenue *Streptomyces albidoflavus* en octobre 2018). Le **tableau 2** présente les derniers noms de ces génomes, les noms précédents de ceux-ci sont récupérables grâce à l'identifiant NCBI.

a- Réduction du nombre d'espèces/génomes considérés

Le premier critère de sélection des génomes est le caractère complet, c'est-à-dire séquencé avec un haut standard de qualité, ce qui garantit un assemblage et une annotation optimaux. L'ajout constant de nouveaux génomes nous a confronté à la pérennité de nos résultats et à "l'intérêt" d'enrichir la collection de *Streptomyces*.

Les méthodes basées sur les similarités phénotypiques et chimiques tout comme les approches classiques d'analyses de séquences comme le pourcentage de similarité de 16S ou de séquence multilocus (MultiLocus Sequence Typinq, MLST) sont très efficaces pour classifier des bactéries au sein d'un genre, mais manquent de précision pour distinguer les souches (Moore et al., 2010; Tindall et al., 2010). Par exemple, dans le cas du pourcentage de similarité de l'ARN 16S, la séquence est trop conservée pour distinguer des espèces proches (Rosselló-Mora and Amann, 2001). En 2005, la méthode ANI (Average Nucleotide Identity, (Konstantinidis and Tiedje, 2005)) a été proposée comme une nouvelle mesure de proximité entre deux génomes bactériens. Cette méthode calcule un score d'identité globale (voir **MATERIELS ET METHODES C1- La distance ANI**) entre deux génomes qui reflète la relation phylogénétique entre eux. Cette approche présente l'avantage de considérer l'ensemble de la séquence génomique, ce qui n'est évidemment pas le cas des méthodes basées sur la séquence du gène codant l'ARN 16S ou les approches multilocus. De plus, la valeur d'ANI est directement interprétable pour définir la proximité phylogénétique entre deux génomes: deux souches partageant une valeur d'ANI d'au moins 96 % étant considérées comme appartenant à une même espèce, et celles partageant un minimum de 70 % comme appartenant à un même genre (Goris et al., 2007; Richter and Rosselló-Móra, 2009).

L' indice $\overline{ANI}$ (moyenne des deux orientations de l'ANI, voir Matériels et Méthodes) a été calculé pour chaque paire de génomes de la collection 135 (**tableau 2**), générant une matrice symétrique de 135x135. A partir de cette matrice, un regroupement a été effectué par la méthode du lien complet qui consiste en un regroupement hiérarchique agglomératif pour mettre en évidence les groupes de génomes proches selon leur $\overline{ANI}$ et aussi pour avoir une première idée de la diversité génétique représentée dans la collection. Le résultat obtenu a été représenté sous forme de heatmap (**figure 19**).

La **figure 19** présente un clustering de tous les génomes analysés de la collection 135 selon leur $\overline{ANI}$, et permet de mettre en évidence les différentes espèces qui composent ce jeu de données. Chaque génome présente entre 99.99% et 75.34% d'$\overline{ANI}$ avec n'importe qu'elle autre espèce de la collection. Ainsi, selon les critères définis par Richter and Rosselló-Móra (2009), la collection 135 n'est constituée que de *Streptomyces*.

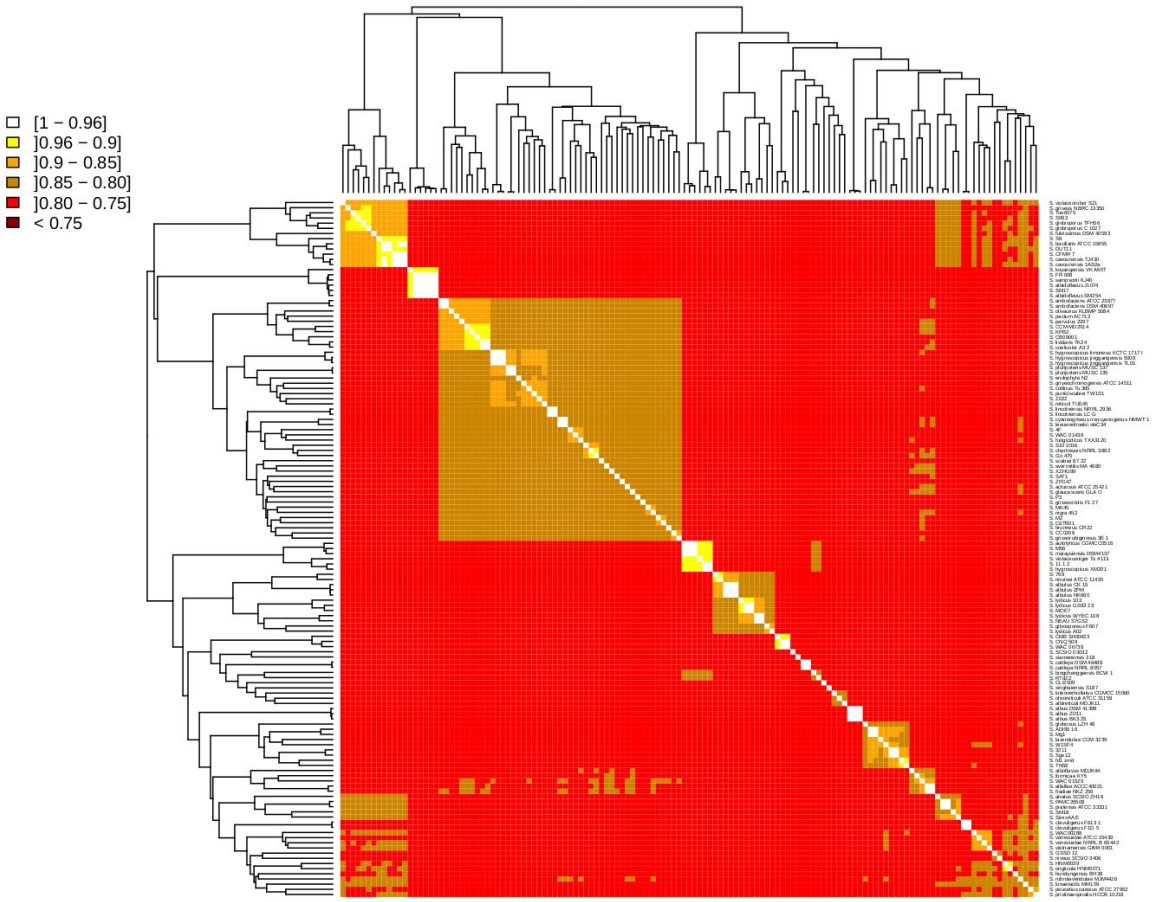

Figure 19 - Carte de chaleur et clustering des différentes espèces de la collection 135 selon leur ANI.

En considérant le seuil de 96% d'ANI comme limite de l'appartenance à une même espèce défini par ces mêmes auteurs, il apparaît que la collection de 135 *Streptomyces* contient 109 espèces dont un représentant de chaque est choisi. Nous avons dérogé à cette règle en incluant ou excluant certaines espèces lors de la sélection des représentants. Ainsi, *S. coelicolor* A3(2) est obligatoirement conservée (référence historique et classiquement utilisée dans de nombreux travaux sur les *Streptomyces*) ainsi que *S. ambofaciens* ATCC 23877 (organisme de référence du laboratoire DynAMic). À l'inverse *Streptomyces* sp. DUT11 est obligatoirement exclue (selon un filtre de qualité décrit dans le point suivant, voir section **IA1b- Estimation de la complétude des génomes**) si cette souche apparaît dans

un regroupement. Pour garder un exemple de relation intra espèce dans la collection, 2 souches de *S. ambofaciens* ont été conservées, les souches ATCC 23877 et DSM 40697.

Notre analyse de proximité phylogénétique montre que des souches appartenant à la même espèce sont nommées différemment dans les bases de données (**tableau 3**) :

Souches	Espèce
Streptomyces sp. FR 008 *S. sampsonii* KJ40 *S. albidoflavus* J1074 *Streptomyces* sp. SM17 *S. albidoflavus* SM254	*S. albidoflavus*
Streptomyces sp. CFMR 7 *S. cavourensis* TJ430 *S. cavourensis* 1AS2a	*S. cavourensis*
S. alfalfae ACCC40021 *S. fradiae* NKZ 259	*S. alfalfae*
S. lividans TK24 *S. coelicolor* A3(2)	*S. coelicolor*
S. pratensis ATCC 33331 *Streptomyces* sp. PAMC 26508	*S. pratensis*
S. bacillaris ATCC 15855 *Streptomyces* sp. DUT11	*S. bacillaris*
S. hygroscopicus XM201 *Streptomyces* sp. 11-1-2	*S. hygroscopicus*
S. autolyticus CGMCC0516 *Streptomyces* sp. M56 *S. malaysiensis* DSM4137	*S. autolyticus*
Streptomyces sp. MOE7 *S. lydicus* GS93 23	*S. lydicus*

Tableau 3 - Souches identifiées comme appartenant à une même espèce selon la méthode ANI. L'espèce est appelée selon le nom déposé le plus anciennement dans les bases de données.

Procédant de la sorte, la collection 81 a été réduite à 59 génomes, puis la collection 135 a été réduite à 109 génomes. Concernant cette dernière, avec l'ajout de *S. ambofaciens* DSM 40697, 110 génomes ont été finalement considérés. Ainsi, la sélection a presque doublé (de 59 à 110) avec l'accroissement de la masse des données. On peut donc supposer, qu'en terme de diversité d'espèces, cette collection est encore loin d'approximer l'ensemble de la diversité du genre *Streptomyces* et que l'ajout de nouveaux génomes dans la collection entraînera l'augmentation du nombre d'espèces.

b- Estimation de la complétude des génomes

L'estimation de la qualité des séquences génomiques a été systématiquement réalisée, grâce à l'outil BUSCO (Benchmarking Universal Single-Copy Orthologs, (Waterhouse et al., 2018)). Cet outil se

base sur des groupes de gènes orthologues retrouvés en copie unique dans les génomes extraits de OrthoDB (Kriventseva et al., 2019) pour fournir une estimation de la "complétude" de l'assemblage et/ou de l'annotation d'un génome. La classification des orthologues au sein d'un clade selon leur universalité (de 'répandu' à 'spécifique') et leur degré de duplication (de copie unique à multi-copies) a permis, par exemple, de construire un paysage des groupes d'orthologues selon ces critères sur un ensemble d'espèces d'*Arthropodes* (**figure 20**). La stratégie de BUSCO est de retrouver (voir MATERIELS ET METHODES) les orthologues en copie unique mis en évidence par le cercle dans la **figure 20** dans toutes les espèces nouvellement séquencées.

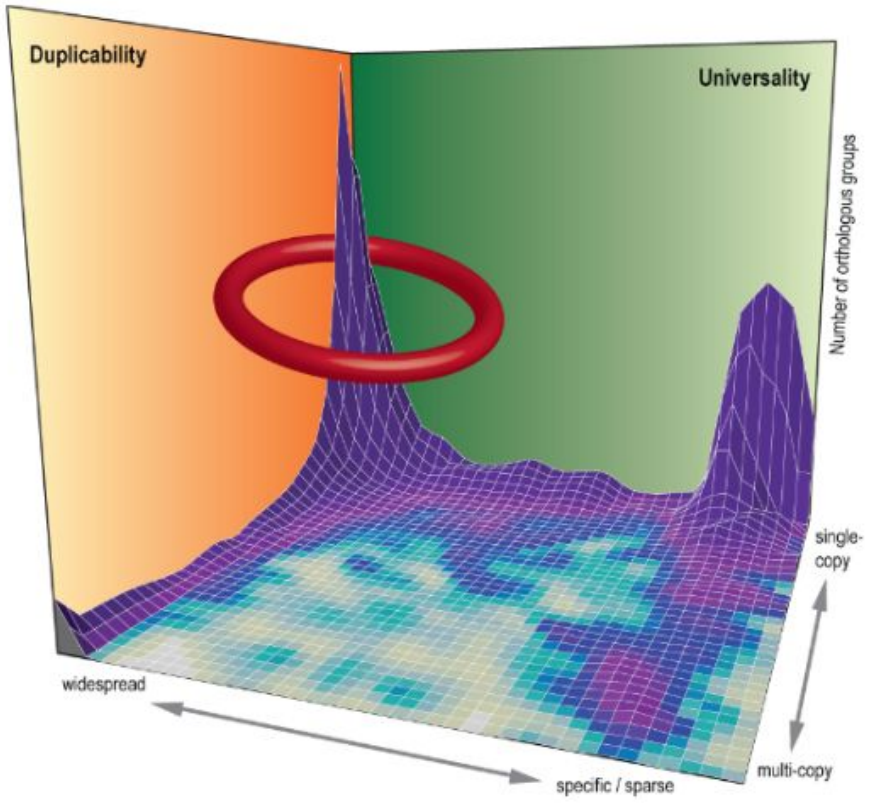

Figure 20 - Paysage du degré de duplication et de l'universalité des groupes d'orthologues d'après Waterhouse (2015). Paysage construit à partir de 80 génomes d'*Arthropodes* extraits de OrthoDB.

BUSCO fournit un score quantitatif sur la complétude d'un génome en terme de contenu en gènes. Les résultats sont divisés en trois catégories: (i) "Complete" correspondant aux gènes complets parfaitement retrouvés (en terme de score d'identité et de la taille de l'alignement), (ii) "Fragmented" correspondant aux gènes partiellement retrouvés (tronqués) et (iii) "Missing" où aucun gène correspondant n'a été identifié. Dans notre cas, l'outil BUSCO a été utilisé pour rechercher les différents groupes d'orthologues de la base de données dédiée aux actinobactéries ***actinbacteria_odb9***

[9] contenant 352 groupes d'orthologues constitués à partir de 412 espèces. La distribution des scores BUSCO est présentée **figure 21**. Les scores correspondent au pourcentage de gènes retrouvés pour chaque catégorie, avec comme valeur médiane, environ 98 % des orthologues retrouvés en copie unique dans la collection de 135 *Streptomyces* révélant une bonne complétude globale de ces génomes. Par conséquent, les autres catégories de scores BUSCO ne représentent qu'une très faible proportion des assignations (valeurs médianes inférieures à 2 %).

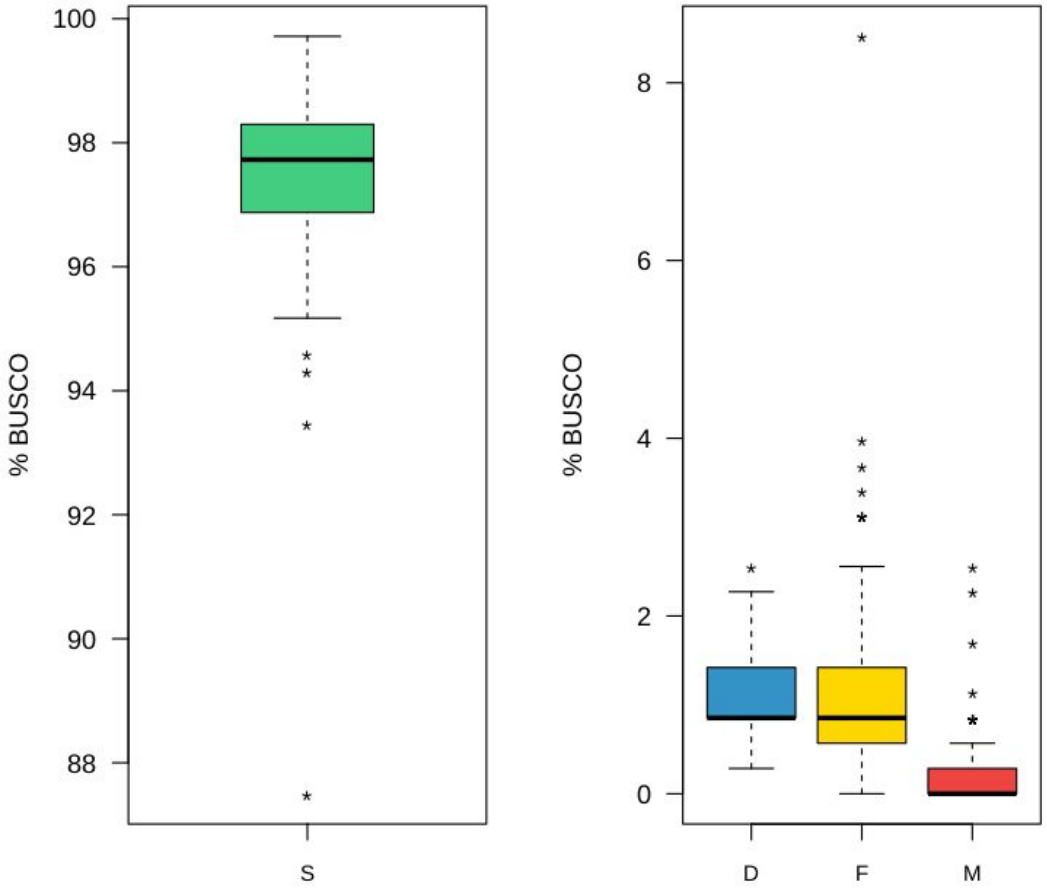

Figure 21 - Représentation sous forme de boîtes à moustaches de la répartition des scores BUSCO selon leur catégorie pour la collection 135. "Complete and Single-copy" = vert (S), "Complete and Duplicated" = bleu (D), "Fragmented" = jaune (F) et "Missing" = rouge (M).

Ces résultats montrent que l'ensemble des génomes de *Streptomyces* utilisés ont une "complétude" satisfaisante selon BUSCO. Un minimum de 93% des gènes recherchés sont parfaitement retrouvés et en un seul exemplaire pour presque tous les génomes. Seule l'espèce *Streptomyces* sp. DUT possède un score "correct" de 87.5%, ce qui est nettement inférieur aux autres espèces. Une grande proportion de ces gènes manquants est par ailleurs retrouvée sous forme fragmentée par BUSCO. Cette espèce est exclue du jeu de données suite à la réduction selon la proximité phylogénétique (voir section **IA1a- Réduction du nombre d'espèces/génomes considérés**).

[9] https://busco.ezlab.org/datasets/actinobacteria_odb9.tar.gz

2- Uniformisation des données génomiques

L'ensemble des travaux effectués durant cette thèse repose sur les séquences complètes des différents génomes récupérés sur les bases de données publiques et donc issues de différentes équipes et technologies de séquençage et annotées par divers programmes d'annotation. Afin de ne pas interpréter ces différences de traitements des données comme des différences intrinsèques entre les génomes comparés, plusieurs étapes préliminaires ont été effectuées pour estimer la qualité globale des séquences génomiques et uniformiser leurs méthodes d'analyse.

a- Annotation

Pour la plupart des génomes disponibles dans les bases de données, une annotation fonctionnelle est disponible (notamment dans NCBI). Cependant, de manière à considérer des annotations homogènes, chacun des génomes a été ré-annotés. Pour cela, la capacité de prédiction de 2 logiciels d'annotation a été comparée sur la collection 42 : RAST (Rapid Annotation using Subsystem Technology, (Aziz et al., 2008) et GLIMMER-3 (Gene Locator and Interpolated Markov ModelER, (Delcher et al., 2007)). RAST est un outil d'annotation des génomes de bactéries et d'archées développé en 2008, il fonctionne en projetant des annotations vérifiées manuellement (la base de données SEED (Overbeek et al., 2005)), sur de nouveaux génomes (Davis et al., 2014). GLIMMER-3 quant à lui utilise le modèle de Markov interpolé pour identifier les régions codantes.

Les annotations obtenues par le logiciel RAST et GLIMMER-3 ont été comparées. Le **tableau 4** présente le bilan de la comparaison des deux méthodes. Il apparaît que GLIMMER-3 détecte en moyenne plus de régions codantes que RAST (respectivement 8166 et 7785). Cependant, en comparant ces annotations avec celles du NCBI, il apparaît que celles fournies par GLIMMER-3 ne correspondent (100% d'identité) qu'à 65% en moyenne à celles du NCBI contre 68% pour RAST.

Les performances des 2 annotateurs sont comparables, cependant l'outil proposé par RAST affine ses annotations grâce au contexte phylogénétique des génomes soumis. Les annotations sélectionnées pour la suite des travaux ont donc été celles générées via le logiciel RAST.

	# Prot. NCBI	# Prot. RAST	# Prot. GLIMMER-3	RAST préd. dans NCBI (%)	GLIMMER-3 préd. dans NCBI (%)
Moyenne	7248	7785	8166	68	65

Tableau 4 - Tableau comparatif des résultats des logiciels d'annotation.
Les 3 premières colonnes correspondent respectivement aux nombres de protéines prédites par le NCBI, par le logiciel RAST et par le logiciel GLIMMER-3. Les deux dernières correspondent au pourcentage de protéines prédites par RAST et GLIMMER-3 communes avec celles prédites par le NCBI.

b- Identification des TIR

La plupart des réplicons linéaires, chromosomes et plasmides, des *Streptomyces* présente des répétitions terminales inversées (TIR, Terminal Inverted Repeats). Les séquences qui composent ces TIR varient d'une espèce à l'autre, et d'un réplicon à l'autre. La taille des TIR varie fortement, de quelques centaines de paires de bases à plusieurs centaines de kilobases, et ces séquences sont connues pour être le siège de nombreux événements de remaniement. Dans l'analyse de la variabilité génomique du chromosome de *Streptomyces*, ces régions sont donc d'un intérêt majeur. Pourtant, leur caractère répété peut empêcher leur identification et leur positionnement lors de l'assemblage des génomes. Tous les génomes ont donc pu ne pas être traités de la même façon par rapport à ces régions. Pour remédier à cette hétérogénéité de traitement une recherche systématique de répétitions terminales a été entreprise sur l'ensemble des chromosomes de notre jeu de données (**MATERIELS ET METHODES II2- Détection des TIR**).

Trois niveaux de complétude ont été identifiés : (i) TIR 'complètes' où des séquences répétées à l'identique et inversées sont identifiées aux deux extrémités du chromosome (**figure 22 a**); (ii) "tronquées" où une section inversée répétée dans la région terminale d'une extrémité forme la seconde extrémité (**figure 22 b**); et (iii) 'absentes', dans ce cas, aucune répétition n'est identifiée aux extrémités chromosomiques (**figure 22 c**).

Figure 22 - Schéma des 3 configurations d'organisation des extrémités chromosomiques observées. **a** : TIR 'complètes'; **b** : TIR 'tronquées'; **c** : Absence de TIR.

Cette approche a permis de mettre en évidence des TIR dans 86 des 135 chromosomes de la collection. Les TIR identifiées dans la collection de *Streptomyces* peuvent représenter jusqu'à 630 kb (chez *S. collinus* Tu 365).

c- Les gènes dupliqués

Plusieurs des approches que nous avons développées dans ce travail se basent sur les relations d'orthologie entre les gènes, et la présence de séquences géniques dupliquées ayant peu divergé (voir **MATERIELS ET METHODES IIB2- Les gènes dupliqués**), peut fausser l'établissement des relations d'orthologie. Le nombre de gènes dupliqués dans les génomes de *Streptomyces* a donc été estimé ; les plasmides ont été exclus de l'analyse, en effet, ils ne sont pas considérés lors des travaux

touchant à l'analyse de la dynamique du chromosome des *Streptomyces*. Au sein de la collection 135, 170 gènes dupliqués en moyenne ont été dénombrés par chromosome, soit en moyenne 2% des gènes totaux d'un chromosome. Pour des travaux à l'échelle du chromosome complet, ce biais apparaît négligeable. Cependant, d'une espèce à l'autre, ce nombre varie fortement. Ceci est dû à la localisation majoritairement terminale de ces gènes; ils composent les TIR (**figure 23 a**). Naturellement, plus les TIR sont grandes plus le nombre de CDS dupliqués dans un génome est important (**figure 23 b**) et sont la source majoritaire des gènes dupliqués. A l'exception des gènes localisés dans les TIR, les gènes dupliqués dans les chromosomes de *Streptomyces* sont marginaux et ne présentent pas d'organisation particulière le long du chromosome.

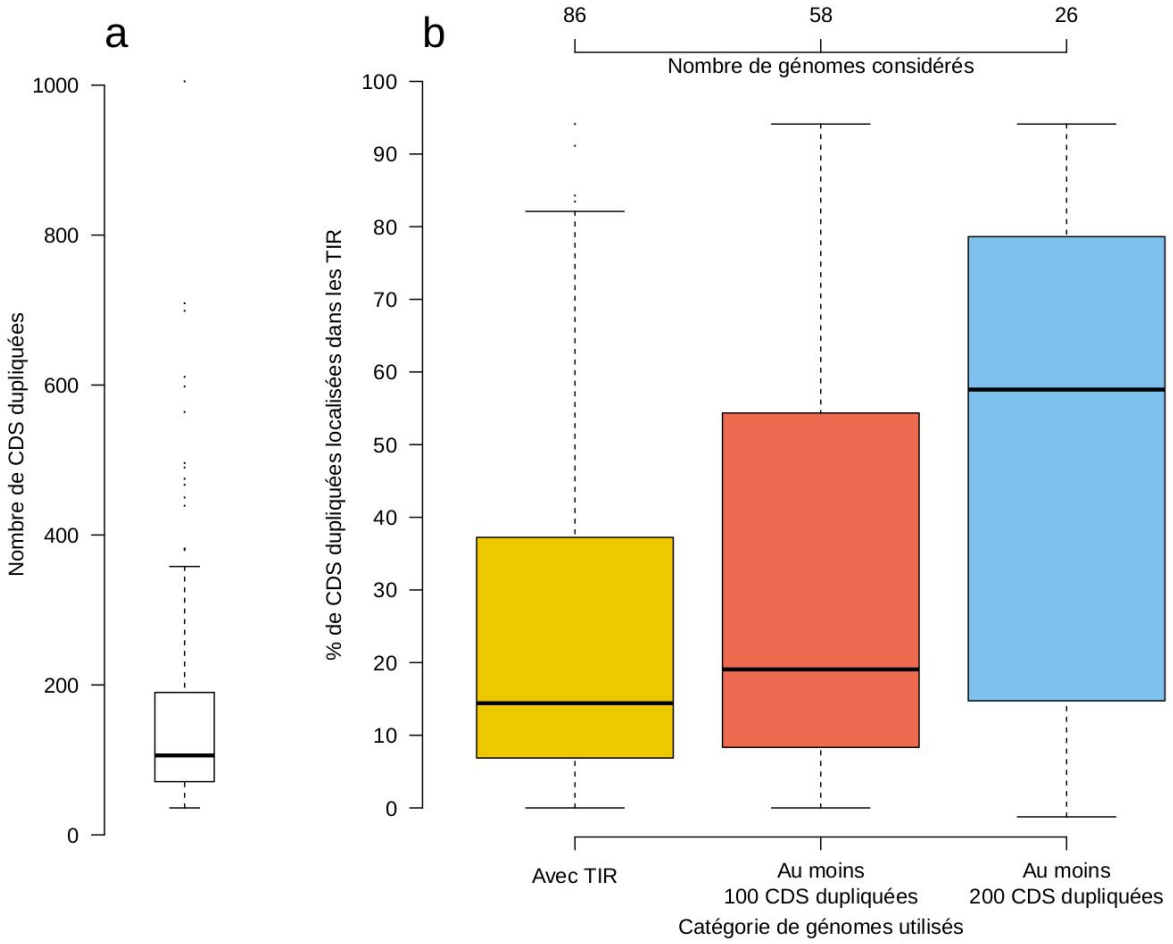

Figure 23 - Distribution du nombre de gènes dupliqués dans les chromosomes de la collection 135 (**a**). Pourcentage de gènes dupliqués localisés dans les TIR par "catégorie" de génomes (**b**). Les génomes sont organisés selon le nombre de CDS identifié dans les TIR et du nombre de gènes dupliqués mis en évidence. La catégorie jaune représente les chromosomes où des TIR ont été identifiées (soit 86 chromosomes). La catégorie rouge restreint l'analyse à tous les chromosomes présentant des TIR et au moins 100 gènes dupliqués (58 chromosomes) et la catégorie bleue aux chromosomes présentant des TIR et au moins 200 gènes dupliqués (26 chromosomes).

d- Uniformisation des données de séquences chromosomiques

Nos travaux sont centrés sur la conservation des gènes et de leur organisation entre espèces, d'où l'importance d'uniformiser les données de séquences chromosomiques considérées dans chaque analyse. Tout d'abord, il était important de distinguer l'information portée par l'ensemble du génome et par les différents réplicons le composant. En effet, l'ensemble de l'information est considérée pour les questions touchant la définition du pan/core génome, mais seule l'information chromosomique est considérée lors de l'analyse de la dynamique du chromosome. Dans ce cas, l'information plasmidique doit être exclue de l'analyse. Les TIR peuvent représenter une proportion importante du chromosome et donc fortement impacter les comparaisons basées sur la synténie dans les régions subtélomériques. Pour éviter que les duplications terminales induisent des biais dans la détection des relations d'orthologie, un seul exemplaire des TIR est conservé par chromosome; si le chromosome présente 2 TIR complètes, une copie est supprimée du chromosome, si l'un des deux exemplaires est tronqué, l'exemplaire incomplet est supprimé. Dans le cas où aucune TIR n'a été détectée, le chromosome est laissé en l'état, il est en effet alors impossible de trancher sur une présence totale ou partielle d'un exemplaire des TIR. Cette étape a aussi pour conséquence de réduire fortement le nombre de gènes dupliqués dans les chromosomes.

Grâce aux résultats précédents, nous avons une collection de génomes complets de *Streptomyces* avec peu d'informations génétiques redondantes (suppression des souches d'une même espèce), avec une qualité d'assemblage (complétude) minimum qui a été vérifiée. Les annotations ont été uniformisées (générées de la même manière) entre tous les génomes et un exemplaire des régions inversées terminales (TIR) a été supprimé de chaque chromosome lorsqu'ils ont été identifiés.
Toutes ces étapes ont été codées de manière à ce que, à partir d'une collection de génomes complets toutes les opérations puissent être effectuées automatiquement. Cela permet une mise à jour aisée de la collection de génomes en amont des étapes d'analyses du contenu en gènes.

B- Evolution du contenu en gènes

Ce second point présentes d'autres résultats servant aussi d'étapes préliminaires à l'analyse de la dynamique des génomes des *Streptomyces* comme la définition des relations d'orthologie et des différents répertoire de gènes. Des résultats directement extrait de l'évolution du contenu en gènes entre les différents génomes de la collection seront aussi présentés ici.

1- Définition des relations d'orthologie

Deux gènes orthologues sont issus par divergence d'une séquence unique présente chez le dernier ancêtre commun aux deux espèces possédant un copie du gène. Cette définition est donc une définition phylogénétique. Cependant dans ces travaux, la définition de l'orthologie est réduite à une qualité d'alignement minimale (voir **MATERIELS ET METHODES IIB1- Les orthologues**) entre les séquences selon la méthode du Bidirectional Best Hit (BBH). La prédiction d'orthologie repose sur un compromis entre sensibilité (obtenir le plus grand nombre de vrais positifs) et sélectivité (obtenir le moindre nombre de faux positifs). La méthode BBH possède une sélectivité importante, mais une sensibilité faible par rapport à d'autres méthodes (Hulsen et al., 2006), en particulier par rapport à celles s'appuyant sur la phylogénie des espèces comme OrthoFinder (Emms and Kelly, 2015). L'écart entre sélectivité et sensibilité de la méthode BBH est d'autant plus important que la distance phylogénétique entre les espèces analysées est grande, cependant cet effet est limité dans notre jeu de données, constitué uniquement d'espèces appartenant à un même genre. Autrement dit, le nombre de relations d'orthologie déterminées ici a été sous-estimé, mais elles possèdent un score de "confiance" élevé.

Par définition, la méthode BBH contraint des relations d'orthologie de type 1:1. Cette contrainte empêche la mise en évidence de relation d'orthologie d'un gène avec plusieurs autres gènes (duplication par exemple). Cependant à l'exception des gènes localisés dans les TIR, peu de familles multigéniques sont présentes dans les génomes de *Streptomyces*. Forcer des relations d'orthologie de type 1:1 a donc un impact faible.

Une base de données des relations d'orthologie a été construite selon la table de relations **figure 24**.

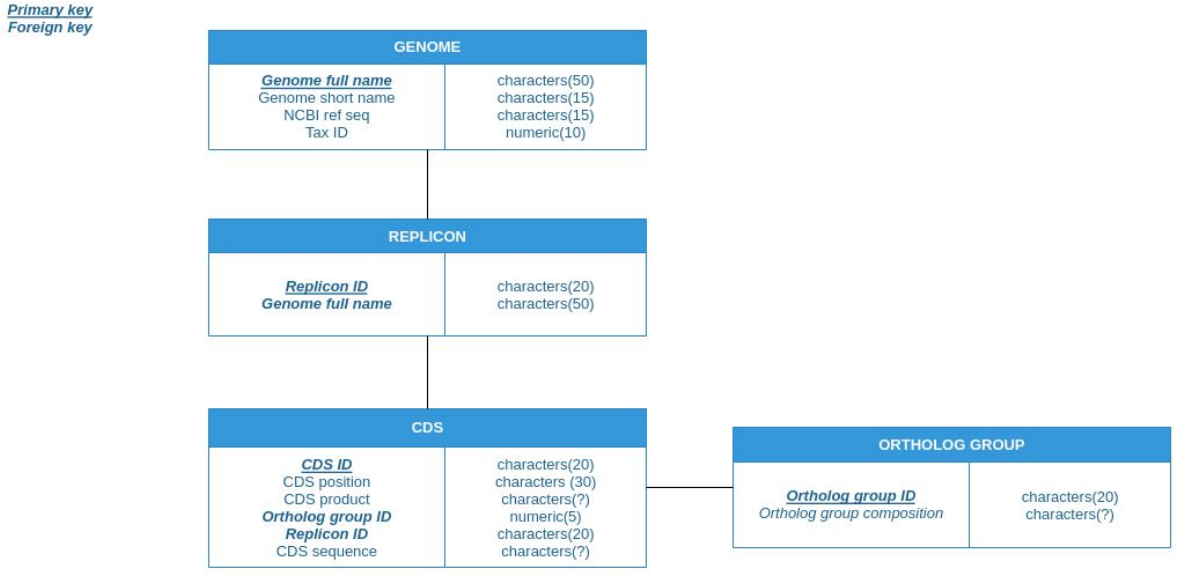

Figure 24 - Table de la base de données relationnelle construite pour enregistrer les relations d'orthologie identifiées.

Sur la collection 81, 85,698 groupes d'orthologues ont été construits (en considérant les plasmides) à partir de 617,124 CDS au total.

2- Le pan-génome et le core-génome du genre *Streptomyces*

110 espèces (**annexe 1**) de *Streptomyces* (obtenues suite à la réduction de la collection 135) ont été utilisées pour effectuer une analyse du pan-génome. À partir des 838,432 gènes contenus dans tous ces génomes le pan-génome, a été construit en regroupant tous les gènes partageant des relations d'orthologie directes ou transitives. Cela a permis d'obtenir un pan-génome de 99,430. En représentant la dimension du pan-génome en fonction du nombre de génomes considérés (**figure 25**), le pan-génome des *Streptomyces* peut être considéré comme ouvert. En effet, la tendance de l'augmentation de la dimension du pan-génome à chaque nouvelle espèce ajoutée est loin d'atteindre une asymptote, reflétant le caractère ouvert du pan-génome des *Streptomyces*.

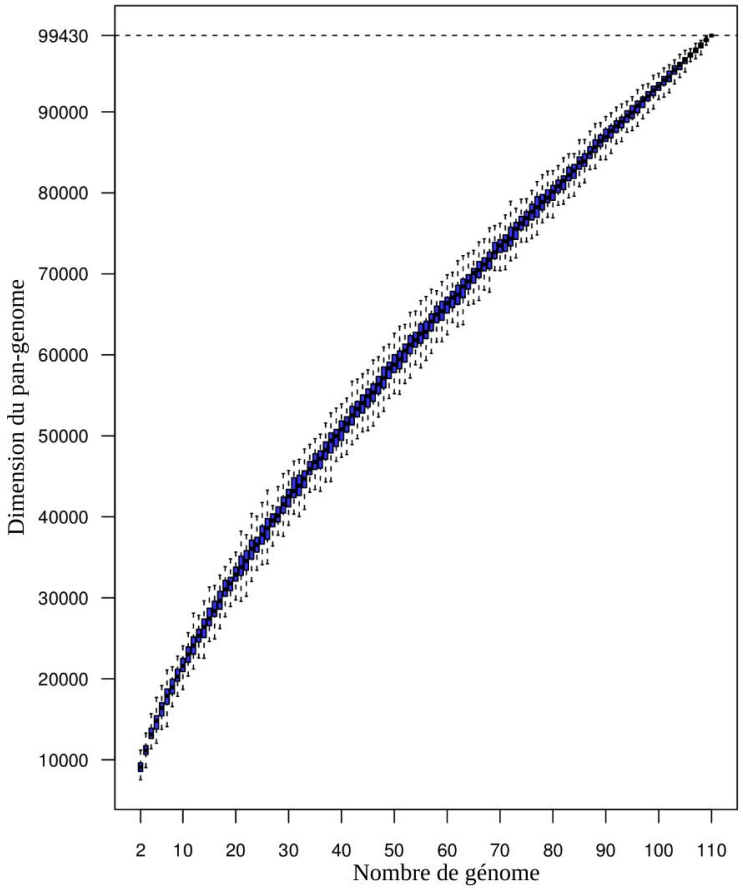

Figure 25- Evolution de la dimension du pan-génome en fonction du nombre de génomes considérés.
Pour un nombre de génomes variant de 2 à 110, la taille du pan-génome a été calculée pour 100 itérations d'une sélection aléatoire de génomes choisis dans notre ensemble de données. Chaque point est une boîte à moustaches[10]. La ligne horizontale donne la valeur médiane.

Cette tendance suggère que le genre *Streptomyces* a un contenu génomique flexible, reflétant la diversité du métabolisme secondaire et la différenciation morphologique.

Le core-génome, c'est-à-dire l'ensemble des gènes partagés par l'ensemble des 110 espèces est constitué de 803 gènes. Le ratio des gènes du core-génome par rapport aux gènes totaux de chacune des 110 espèces est compris entre 14.8% (pour *S. xiamenensis* 318, 5,411 gènes) et 7.8% (pour *S. hygroscopicus* XM201, 10,288 gènes). Cette dimension du core-génome du genre *Streptomyces* est beaucoup plus faible que celle proposée par Kim et al. (2015) obtenue avec une collection de 17 *Streptomyces* (core-génome de 2,018 gènes). Cependant, cette valeur est cohérente avec l'évolution de la dimension du core-génome présentée **figure 26**. Bien que la dimension du core-génome puisse encore réduire lors de l'ajout de nouvelles espèces, la tendance semble tendre vers une asymptote,

[10] La valeur centrale d'une boîte à moustaches est la médiane. Les bords du rectangle sont les quartiles (pour le bord inférieur, un quart des observations ont des valeurs plus petites et trois quarts ont des valeurs plus grandes, le bord supérieur suit le même raisonnement). Les extrémités des moustaches sont calculées en utilisant 1.5 fois l'espace interquartile.

suggérant que cette dimension de core-génome ne devrait plus beaucoup varier avec l'enrichissement en génomes du genre *Streptomyces.*

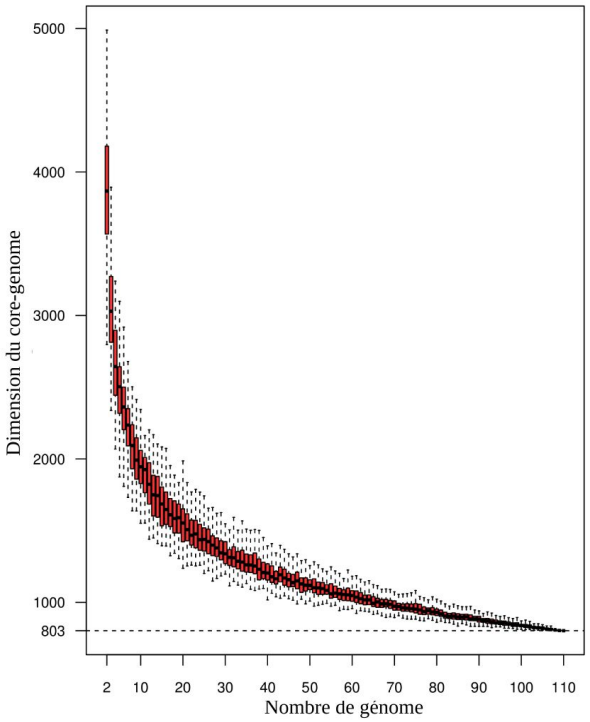

Figure 26- Evolution de la dimension du core-génome en fonction du nombre de génomes considérés.
Pour un nombre de génomes variant de 2 à 110, la taille du core-génome a été calculée pour 100 itérations d'une sélection aléatoire de génomes choisis dans notre ensemble de données. Chaque point est une boîte à moustaches[11]. La ligne horizontale donne la valeur médiane.

3- La phylogénie du genre *Streptomyces*

Le genre *Streptomyces* est un groupe taxonomique situé dans l'ordre des *Actinomycetales*. Les organismes inclus dans cet ordre présentent une importante diversité en terme de morphologie, écologie et caractéristiques génomiques (Ventura et al., 2007). Au sein même du genre *Streptomyces* un niveau élevé de diversité a été mis en évidence : bien que majoritairement trouvées dans le sol, des espèces ont aussi été décrites dans des tissus de plantes (Castillo et al., 2002; Tokala et al., 2002), sur la cuticule de fourmis (Kost et al., 2007) et dans des milieux marins (Zhang et al., 2008). Certaines espèces possèdent une activité pathogène comme *S. scabies,* responsable de la gale commune de la

[11] La valeur centrale d'une boîte à moustaches est la médiane. Les bords du rectangle sont les quartiles (pour le bord inférieur, un quart des observations ont des valeurs plus petites et trois quart ont des valeurs plus grandes, le bord supérieur suit le même raisonnement). Les extrémités des moustaches sont calculées en utilisant 1.5 fois l'espace interquartile.

pomme de terre (Takeuchi et al., 1996). Il existe même des pathogènes humains comme *S. somaliensis* responsable de l'actinomycose (Baril et al., 1999).

Obtenir une phylogénie détaillée du genre *Streptomyces* est de grande importance lors de travaux de génomique comparée en particulier pour explorer la diversité et reconstruire un modèle évolutif. Plusieurs approches de phylogénie ont donc été testées afin de proposer la phylogénie la plus robuste possible du genre *Streptomyces*.

Toutes les espèces qui ont été utilisées dans ces travaux appartiennent à un même genre. Obtenir une phylogénie précise du genre requiert donc une résolution importante dans la phylogénie afin de distinguer différentes souches d'une même espèce. La matrice de distance obtenue à partir de la méthode ANI a permis d'identifier les différentes souches d'une même espèce, mais ne se prête pas à la construction d'arbre phylogénétique robuste. Pour cela, une phylogénie basée sur une approche MLST (MultiLocus Sequence Typing) a été effectuée. De manière à utiliser le maximum d'information génétique pour la construction de la phylogénie, toutes les séquences protéiques du core-génome (803 séquences pour la phylogénie représentée **figure 27**) ont été utilisées (voir **MATERIELS ET METHODES**). Cette approche a permis de générer l'arbre présenté **figure 27** et met en évidence 3 regroupements au sein du genre. Le clade I (rouge) regroupant 36 espèces, le clade II (vert) avec 50 espèces et un dernier groupe regroupant 24 nommé Other (bleu). Ce regroupement en clades est en accord avec la phylogénie proposée par McDonald and Currie (2017), obtenue à partir d'une analyse MLST de 94 gènes de ménage (la dénomination des clades est empruntée à ces auteurs).

La répartition des *Streptomyces* en trois clades a été croisée avec différents critères environnementaux comme le milieu de vie (sol, marin...) ou le lieu de prélèvement (localisation géographique) des échantillons biologiques. Aucune corrélation significative n'a été détectée entre ces critères et l'appartenance à un des clades de cette phylogénie.

À partir de cette phylogénie, la distance cophénétique (*ie* longueur totale des branches entre 2 espèces) a été extraite pour chaque paire d'espèces. Cette nouvelle matrice de distance présente, par rapport à la matrice $\overline{ANI}$, une meilleure distribution des distances pour les espèces les plus éloignées au sein du genre. Cette distance est utilisée par la suite pour les représentations en fonction de la distance phylogénétique.

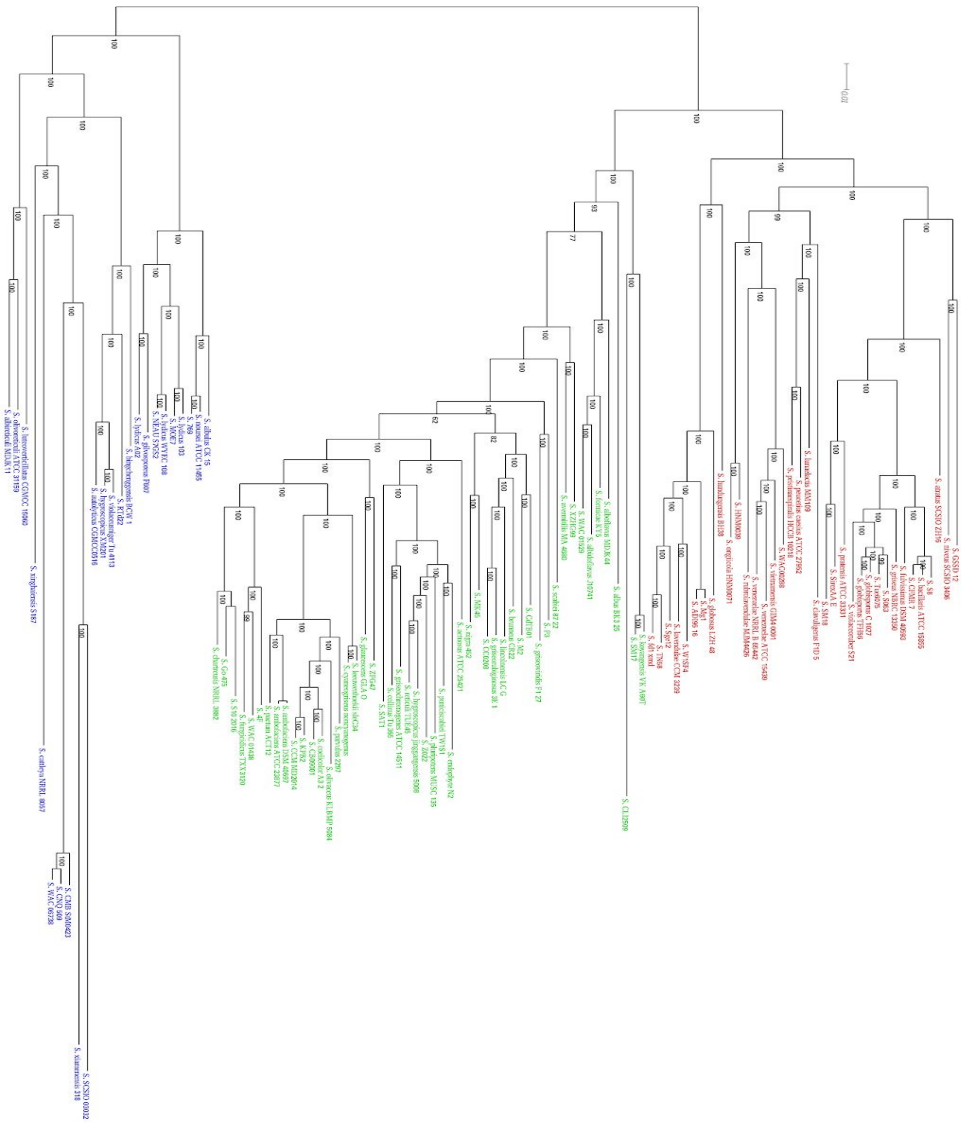

Figure 27 - Arbre phylogénétique du genre *Streptomyces* construit à partir de 110 génomes. Le nom des espèces est coloré selon leur appartenance aux clades I (rouge), II (vert) et Other (bleu). Les valeurs de bootstraps (exprimées en pourcentage) sont indiquées au niveau des noeuds internes.

C- Evolution des régions subtélomériques du chromosome linéaire des *Streptomyces*

A partir des 2 points précédents, nous avons à notre disposition une collection de génomes de *Streptomyces* annotés de la même manière. Ces données servent de base à l'analyse de la dynamique du chromosome de *Streptomyces* et en particulier des régions subtélomériques.
Ce chapitre présente des résultats valorisés par une publication en cours de soumission. Le papier n'étant pas encore finalisé, la version présentée ici est de type "preprint" et se base sur la collection 135. Les différents résultats du papier sont en train d'être mis à jour avec la collection 234.

Publication 1 : Subtelomeres are fast-evolving regions of the linear chromosome in *Streptomyces*. *Jean-Noël Lorenzi, Annabelle Thibessard, Olivier Lespinet, Pierre Leblond*

L'approche présentée ici exploite la profondeur phylogénétique fournie par les génomes complets disponibles du genre *Streptomyces*. La construction du pan-génome a révélé une très grande variabilité génétique, le pan-génome étant constitué de 99,430 gènes ce qui représente plus de 13 fois la valeur médiane du nombre de gènes portés par un génome de *Streptomyces* (7380 gènes). À l'inverse, la taille du core-génome, constitué de 803 gènes ne représente environ que 10% du nombre de gènes médians. La phylogénie obtenue à partir d'une analyse multilocus des 803 gènes du core génome suggère que le genre *Streptomyces* peut être divisé en au moins 3 groupes, dont 2 sont monophylétiques.
La localisation des gènes du core-génome sur le chromosome linéaire des *Streptomyces* a montré que les gènes core sont généralement concentrés dans la partie centrale du chromosome, la région "core" définissant en négatif les bras chromosomiques. Quelques chromosomes présentent une région "core" mais non centrée, présentant par conséquent des bras chromosomiques de tailles très différentes. La dernière configuration mise en évidence est l'absence de région dépourvue de gènes core. La différence de taille entre les chromosomes de *Streptomyces* dépend principalement de la longueur des régions des bras chromosomiques variables qui bordent la région centrale.

Des analyses de conservation de la synténie ont été effectuées pour toutes les paires de génomes possibles dans la collection. En considérant les chromosomes "typiques" (chromosome présentant une région core centrale bordée de 2 bras chromosomiques), il est apparu que les bras chromosomiques sont beaucoup plus remaniés que la région core. En outre, le nombre de cassures de la synténie augmente avec la distance phylogénétique, ce qui démontre la nature cumulative de ces événements.

Plus les chromosomes comparés appartiennent à des souches éloignées, plus les loci remaniés concernent des régions internes du bras chromosomique.

En évaluant la conservation de l'organisation des chromosomes et en estimant le nombre de réarrangements entre chaque paire de chromosomes, il est apparu que les orthologues localisés dans les bras chromosomiques présentent le schéma de réarrangement le plus intense, ce qui est tout à fait cohérent avec les observations précédentes. Cependant, en se concentrant sur la région “core”, l'organisation des gènes du core genome semble être beaucoup plus stables au cours de l'évolution que les gènes non essentiels de la région centrale.

En considérant tous ces éléments ensemble, il apparaît, en outre, que la variation du taux d’orthologues et de la synténie est d’autant plus importante lorsque l'on compare des espèces proches phylogénétiquement. Avec des espèces de plus en plus éloignées, les différences entre les bras chromosomiques et le centre du chromosome semblent s'atténuer.

Subtelomeres are fast-evolving regions of *Streptomyces* linear chromosome

Jean-Noël Lorenzi[1], Annabelle Thibessard[2], Olivier Lespinet[1], Pierre Leblond[2]

[1]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University Paris-Sud, University Paris-Saclay, Gif-sur-Yvette CEDEX
[2]Université de Lorraine, INRA, DynAMic, F-54000 Nancy, France

Key words: *Streptomyces*, linear chromosome, core genome, pangenome, recombination, evolution, horizontal gene transfer

Abstract

Streptomyces possess a large linear chromosome (8-12 Mb) consisting of a conserved central region flanked by variable arms covering several megabases. In order to study the evolution of the chromosome across evolutionary times, a representative panel of *Streptomyces* strain and species (110) whose chromosome are completely sequenced was identified. The pangenome of the genus was modelized and shown to be open with a core-genome reaching 803 genes.

The evolution of the *Streptomyces* chromosome was analyzed by carrying out pairwise comparisons, and by monitoring indexes measuring the conservation of gene order and the rate of orthology. This method, applied at the global level, was also applied at the local level, making it possible to measure recombination intensity along the chromosome. Using the phylogenetic depth offered by the chosen panel, it was possible to infer that the chromosomal arms evolved faster than the central region under the combined effect of rearrangements and the addition of new information from the horizontal transfer.

INTRODUCTION

Horizontal gene transfer (HGT) is the most efficient way to diversify the gene pool of a bacterial group, species or population (Nakamura et al. 2004). It allows the acquisition in a single event of 'ready to use' genes and contributes to extend the accessory genome that is the genes present only in a fraction of the strains. It provides a mean of rapid adaptation to environmental changes by keeping a strong diversity within a population from which adapted individuals can emerge. It can also provide specific advantage to the whole population in nature; some individuals being able to produce compounds in the micro-environment (*e.g.*, antibiotics) for the benefit of the whole group (Baltz 2017). Therefore, the accessory gene pool can be considered as public goods (Cordero et Polz 2014). When the number of genomes considered in the study increases, the accessory genome increases while the proportion of the gene pool shared by all the individuals (*i.e.*, the core genome) decreases. The pan-genome defined as the entire and non-redundant gene set of the group (core + accessory genes) is being said open as it continuously increases with the number of genomes analyzed. The extension of a pan-genome requires to have access to exogenous information, which is favored by life in complex biotic environments, and the ability to perform HGT.

HGT can be decomposed in two steps: firstly, internalization of the exogenous genetic material, which is dependent of three main mechanisms known as conjugation, transduction and natural transformation, and secondly, maintenance of this incoming DNA in the new host genome. Maintenance relies either on the recipient recombination abilities (homologous and illegitimate recombination pathways) or on mechanisms encoded by the mobile genetic elements themselves involved in HGT promotion (*e.g.*, replication of plasmids, integration of Integrative and Conjugative Elements or phages…). Endogenous recombination mechanisms can indeed lead to allele replacement or even involving housekeeping and essential genes to addition of new sequences. The former results from homologous recombination between highly similar gene sequences (it remains undetected if recombination occurs between identical sequences leading to underestimation of homologous recombination rates (Didelot et Maiden 2010), the latter can result from integration of new information by either homologous or illegitimate recombination (Dupuy et al. 2018; Hoff et al. 2018). Although homologous recombination is favored by the genetic

relatedness between the partners, illegitimate recombination would not be subordinate to sequence homology between the incoming DNA and the recipient chromosome. At last, the long-term maintenance of a DNA sequence depends of its impact on bacterial fitness. Hence, most of the incoming DNA is considered as neutral or deleterious while only a few events are assumed to improve the recipient fitness. The more deleterious a DNA sequence is, the more frequent is its loss or replacement within the gene pool (Roth 2010).

Streptomyces are prominent bacteria found in soil and marine environments (Anderson et Wellington 2001). They are committed in complex biotic interactions with bacteria, fungi, plants and insects (Kurth et al. 2014) and involved in biogeochemical recycling thanks to their prolific specialized metabolism (chemical compounds and enzymes). Diversity of the *Streptomyces* genus was empirically exploited for decades for the production of antibiotics, antifungal and other metabolites of medical or biotechnological interest; each species was famous for the production of a few compounds (Aigle et al. 2014). More recently, with the development of new generation sequencing technologies, the facilitated access to genomes revealed that they constitute reservoirs for specialized metabolite gene clusters, but even more intriguing, that closely related strains encode very different arsenals of specialized metabolites (Antony-Babu et al. 2017; Vicente et al. 2018). This high diversity is the revealing sign of the wide-open pan-genome (Kim et al. 2015) which could reflect an intense gene flux that could occur in *Streptomyces*. The ability to diversify the genome content could participate to their adaptability to the soil, which is a naturally challenging ecosystem (Vos et al. 2013).

Genome comparison within a growing set of *Streptomyces* sequenced genomes showed that their linear chromosome (6-12 Mb (Kim et al. 2015)) is typified by a very peculiar organization, with the core genes mostly localized in the central part of the chromosome and accessory genes confined in the chromosomal arms (Bentley et al. 2002; Kim et al. 2015; Kirby 2011). This compartmentalization may find its origin in the linearization process assumed to have accompanied the emergence of the *Streptomyces* genus (Kirby 2011): in a parsimonious hypothesis, acquisition of telomeres and of chromosomal arms would have resulted from a single integration event of a linear replicon within the ancestral circular chromosome. This hypothesis is supported by traces of synteny between the circular *Mycobacterium*

genome (mycobacteria belong to a genus closely related to *Streptomyces*) and the central region of *Streptomyces* chromosome. Consequently, chromosomal arms would have included contingent genes since this early linearization event. This compartmentalization corroborates early studies on genetic instability of *Streptomyces* species (Fischer, Decaris, et Leblond 1997; Thibessard et Leblond 2014) showing that subtelomere regions are, in lab conditions, prone to huge DNA rearrangements (deletion, amplification,…) of over several hundreds of kilobases, presumably because of the lowering of selective pressure exerted on the contingent genes present in these regions.

In this work, we will exploit the phylogenetic depth provided by the availability of complete genome sequences of *Streptomyces* species representing the broader diversity to acquire a dynamic picture of the evolution of the chromosome structure. We will question the basis of the genome diversification in *Streptomyces*. How the different regions of the chromosome evolve (central part *versus* arms)? Is the chromosome compartmentalization a common trait of the *Streptomyces* genus? Are the chromosomal arms more tolerant to DNA rearrangements? Do they constitute fast-evolving regions?

RESULTS

The pan and core genome reflect a great diversity of gene content in the *Streptomyces* genus

Based on a representative collection of 110 *Streptomyces* genomes (see Material and Methods and table I), we addressed the question of the global chromosome dynamics inside the *Streptomyces* genus.

Among our set of 110 species, the genome size can vary in the ratio of one to two, with respectively 5.96 Mb for *Streptomyces xiamenensis* 318 and 12.01 Mb for *Streptomyces hygroscopicus* XM201 (**table 1**). At the same time the estimated gene content, based on the results of the RAST gene predictor (Aziz et al. 2008), can vary from 5,411 genes for *S. xiamenensis* 318 to 10,288 genes for *S. hygroscopicus* XM201. Between these two extreme values, the distribution of the gene content of the 110 genomes studied here is quite variable with a median of 7,380 genes (**figure 1**).

This great variability in gene content seems extrapolable to the whole genus as attested by the size of the pan-genome built from this species set (**figure 2a**). Thus the pan-genome of the 110 species appears to be extremely open with 99,430 genes which represent more than 13 times the gene-content median value (**table 1**). The curve obtained for the pan-genome (**figure 2a**) is far from reaching an asymptote and it is expected that this value will further increase by adding new genomic data.

On the other hand, we observed that the size of the core-genome (**figure 2b**) reached 803 genes which represent about a tenth of the gene-content median value (**table 1**). This core-genome size is lower than those obtained in previous studies (Kim et al. 2015; Zhou et al. 2012), but remains consistent with them, considering the larger number of genomes included in our study.

The amino acid sequence of the 803 genes of the core-genome was used to perform a multilocus phylogenetic analysis (see Material and Methods). The reconstructed tree (**figure 3**) was in accordance with that previously proposed by McDonald and Currie (McDonald et Currie 2017), who suggested that the *Streptomyces* genus can be divided into two major monophyletic clades (clade I and clade II) and a third group of other *Streptomyces* lineages (referred to as "others") (**figure 3**).

Distribution of the core genes along the chromosome

For all the species, all the 803 genes of the core genome were located on the chromosome. Their distribution along the linear chromosome generally showed a localization limited to the central part

which defined the 'core-region'. In **figure 4.A** is illustrated the chromosomal location of the core-genome for three species with different chromosome sizes; *S. xiamenensis* 3185.96 Mb, *S. ambofaciens* ATCC 23877, 8.3 Mb, *S. hygroscopicus* XM201, 12.01 Mb. The core-region (i.e. region containing the core-genes, see MM) spread over 5.2, 5 and 7.5 Mb respectively; the flanking regions were called chromosomal arms and were defined negatively with respect to the core-region. This survey was extended to the whole genome set and revealed that most of the chromosomes (85 of 110) displayed a core-region roughly centered in the chromosome; these chromosomes will later be referred to as the "typical architecture chromosomes". The remaining genomes harboured either a core-region that is not central, resulting in unbalanced arms (arbitrarily defined as chromosomes that display one arm twice longer than the other; 9/110) or a core-region spreading all over the chromosome (arbitrarily defined as chromosomes whose arms represent less than 10% of the total chromosome size; 16/110), leaving almost no arms. Further, by plotting the size of the core-region and that of the chromosomal arms as a function of the chromosome size (**figure 4.B**), it appeared that the size of the arms increased faster than that of the core-region according to the chromosome size. Thus, while the core-region size ranged from 4.6 to 7.7 Mb, the chromosomal arm size (sum of the two arm sizes) ranged from 0.6 to 5.2 Mb. The chromosomal size difference mainly depends upon the length of the variable regions flanking the core-region. In other words, the expansion of the *Streptomyces* genome would depend on the acquisition of accessory genes in the chromosomal arms. Reciprocally, chromosomal size reduction would proceed preferentially in the chromosome arms. Meanwhile the core-region remains rather stable in size across the species.

Consistently with these observations, it is well known that, under laboratory conditions, terminal regions rearrange and that mutants with large deletions, amplification or circularization appear at high frequencies from the wild strain (Inoue et al. 2003; Leblond et Decaris 1994; Townsend et al. 2003). These rearrangements are notably the result of the repair of double-stranded breaks either by homologous or illegitimate recombination, Non Homologous End Joining (NHEJ), a mechanism recently demonstrated in *Streptomyces* (Hoff et al. 2016, 2018). Indeed, the induction of double-strand breaks, particularly in the terminal regions, trigger all genomic rearrangements that appear spontaneously (Hoff et al. 2018). For instance, DSBs repair by homologous recombination events

caused TIR size alteration during growth in lab conditions and could explain the large variability of the TIR size at the genus level observed in our study (**table 1**). Moreover, these terminal recombinations also occur between replicons as evidenced by the formation of hybrid chromosomes and plasmids in *S. coelicolor* (Vivian et Hopwood 1973) or *Streptomyces rimosus* (Pandza et al. 1998). This inter-replicon recombination mechanism is likely to provoke, through a single mutational event, the erasing of the ancestral synteny, and generates evolutionary jumps in the genetic diversification of *Streptomyces*. Altogether, these terminal rearrangements could account not only for the TIR size variability but also its sequence diversity as well as the diversity of the chromosomal structure (*i.e.*, typical, unbalanced, armless).

Evolution of the genome plasticity along the *Streptomyces* chromosome

In a will to identify the origin of the variability of *Streptomyces* chromosomes, we assessed the relative impact, on the one hand, of the shuffling of the ancestral skeleton, and on the other hand, of gene flow by loss, acquisition or replacement. For that purpose, several indicators were monitored in pairwise comparisons of genomes of more or less closely related strains (**figure 5**). All these indicators take into account the number of ortholog genes and/or their relative positions in the two compared genomes.

Evolution of the gene content in *Streptomyces* chromosome (OR)

The Orthology Rate (OR), defined as the rate of genes conserved in both compared genomes whatever their location **(figure 6.a),** was used to estimate the impact of gene fluxes. High OR values indicate a strong gene content conservation while low values reveal a high proportion of specific genes (either lost by one strain or acquired through horizontal transfer by the other). Depending on the relationship between the two compared genomes this number varied from 0.3 to 0.93. At the intra-clade level, the OR values reached a minimum of 0.4 (clade I) and 0.36 (clade II) underlining the intensity of gene fluxes even within a monophyletic clade.

In order to visualize a possible difference depending on the chromosomal location, this analysis was conducted locally in a window of various sizes, sliding along the chromosome. The result obtained for *S. ambofaciens* as a reference in 3 pairwise comparisons with species belonging to clade II and sharing

various phylogenetic distances (*S. coelicolor* A3(2), *S. reticuli* TUE45, *S. albus* BK3 25) is shown **figure 6.a**. *S. ambofaciens* ATCC 23877 was chosen as the reference since it is our lab model and belongs to a species cluster including the model *Streptomyces coelicolor* A3(2). As expected, the level of orthology follows the phylogenetic distance: the more two strains were related, the higher the OR value was. Further, there is a strong and progressive decline in the chromosomal arms. This showed that the chromosomal arms accumulate specific genes more intensely than the central part of the chromosome.

Evolution of the architecture of *Streptomyces* chromosome (GOC and NOC)

Gene Order Conservation (GOC) as previously defined by Choulet et al. (2006) and a GOC-derived parameter called NOC for Neighbour Orthologue Conservation were monitored to assess genome shuffling; GOC has been defined as the relative frequency with which two contiguous genes in a genome have their respective orthologs also contiguous in another genome (**figure 5**).

This analysis was performed locally in a window of various sizes, sliding along the chromosome. When considering the GOC parameter at the local scale, it should be noticed that the particular organisation of *Streptomyces* chromosome could rise two clues: first, since orthologues are scarce in some chromosomal areas (notably in the arms), some windows could be devoid of orthologs, making the calculation inapplicable (**figure 7A**); second, because when the orthologs were scattered, the probability to have a neighbour that is also an ortholog strongly drops, therefore the number of contiguous orthologous genes in a given window may not be significant. Consequently, these results must be interpreted with caution, especially as the phylogenetic distance increases.

In order to overcome these limitations, the NOC (Neighbour Orthologue Conservation) was defined: it consists of the ratio of the number of genes involved in contiguous pairs in the reference strain that are still contiguous in the compared strain (**figure 5**).

In contrast with GOC, the NOC index can be applied and will be valid whatever the orthology level between the compared genomes. However, NOC windows, defined as 5% of the total number of orthologous genes organized as contiguous pairs in the considered pairwise comparison, will cover regions of highly different physical sizes. For example, a window in a poorly conserved region will

extend over a larger chromosomal area than a window in a more conserved region. Further, depending on the phylogenetic distance between the compared strains, a common chromosomal locus will be included in windows of different sizes. As a consequence, the trend of the NOC cannot be compared to those of OR and GOC.

Finally, both GOC and NOC indexes turn out to be complementary and relevant to assess genome plasticity across evolutionary times.

GOC and NOC analyses were performed for all of the 11.990 pairwise comparisons of species of our dataset; each genome chosen as a reference has been compared to the other 109 genomes. Because of the limitation mentioned above, we chose to restrict the presentation to intra-clade pairwise comparisons. As an example, the results obtained for *S. ambofaciens* ATCC 23877 (clade II) and *S. griseus* NBRC 13350 (clade I) used as a reference are presented as heatmaps, respectively in **figure 7** and **Figure S1**. When considering "typical" chromosomes, it appeared that the chromosomal arms are much more shuffled than the central part. Further, the synteny breaks occurrences increase together with the phylogenetic distance, demonstrating the cumulative nature of these events. The more the chromosomes compared belong to distant strains, the more the shuffled loci concern more internal regions within the chromosomal arm. This observation could be interpreted as a frequency of rearrangement increasing towards the ends of the chromosome, confirming the notion of a gradient of loss of synteny (called degenerate synteny), proposed by Choulet et al. (2006). The NOC index (assessing the conservation of adjacent orthologous genes) showed a similar trend as the GOC and confirmed that the arms are recombinogenic regions. The GOC and NOC results obtained for *S. ambofaciens* as a reference in the pairwise comparisons with the 3 species *S. coelicolor* A3(2), *S. reticuli* TUE45, *S. albus* BK325 is shown **figure 6.b and 6.c**. This focus, plotted as curves, allowed to visualize the progressive decline of both these indexes within the arms when getting closer to the chromosomal ends.

Rearrangement within the core-region

We further assessed conservation of chromosome organisation by applying the Double-Cut-and-Join (DCJ) method (Yancopoulos et al. 2005), which infers a distance between two chromosomes based on

the number of DNA rearrangement events. This distance can be estimated by the minimal number of shuffling events (inversion, transposition) needed to transform a given chromosome to the reference chromosome. We distinguished 3 gene categories: the core genes occupying by definition the core-region (i), the "non-core" orthologues (*i.e.* orthologous genes excluding core genes) either localized in the core-region (ii) or in the chromosomal arms (iii) as schematized in **figure 8A**.

By plotting the DCJ distance against the phylogenetic distance between *S. ambofaciens* as a reference and all the other chromosomes **figure 8B**), it appeared that these 3 gene sets behaved differently. The orthologues from the chromosomal arms displayed the most intense rearrangement pattern, which is fully consistent with the previous observations (*i.e.*, GOC and NOC analyses). Focusing on the core-region, the trend of the core and the "non-core" orthologues surprisingly showed distinct behaviors. The organization of the core genes appears to be much more stable across the evolutionary times than the non-core genes of the core-region. This unexpected observation raises the hypothesis that the genes vertically inherited from the distant common ancestor (*i.e.*, core-genes are most likely inherited from the ancestor of the genus) occupy crucial coordinates along the chromosome. In contrast, the genes which have been acquired (or lost) more recently, at different evolutionary times since the appearance of the genus, seem to occupy positions which are flexible revealing an intense recombination activity.

In other words, ancestral genes appeared to be maintained at stable positions of the chromosome, and this could reflect a conserved chromosome skeleton. These data question the compartmentalization of the *Streptomyces* chromosome with unstable arms framing a more constrained central region; the underlying architectural constraints may be related to main cellular processes such as chromosomal replication or *in vivo* 3-dimensional chromosome organization which is correlated to gene expression and cell division. Replication is a major structuring force of the bacterial chromosome. The two replichores, *i.e.*, the two halves of the chromosome defined by the position of the origin of replication and the opposite terminus, have a symmetrical organisation, both in terms of nucleotide composition (skews) and gene organisation. In fast-growing bacteria, the essential genes are most frequently oriented in the direction of continuous replication (Rocha et Danchin 2003). In addition, genes located close to the origin have a higher copy number per cell due to replication reinitiation (the so-called gene-dosage

effect), and therefore show a stronger expression. Chromosomal partition and its coupling with cell division is also a structuring force. Thus, the macrodomains *ori* and *ter* (Valens et al. 2004) which represent about 1 Mb each in *E. coli* (Niki et al. 2000) are positioned in the cells in such a way that they promote an equitable and efficient segregation of the son chromosomes in the daughter cells. The choreography of the chromosomes is regulated by the condensation of these domains and their interaction with the actors of cell division (Gitai et al. 2005). The perturbation of these domains, notably by chromosomal rearrangements such as translocations and inversions (Lesterlin et al. 2005), is strongly counter-selected, revealing the constraint of chromosome organization. The three-dimensional structuring of the genome is also linked to global gene expression. Thus, in prokaryotes as in eukaryotes where this phenomenon was first identified, DNA folding is strongly influenced by the position of highly expressed genes, the latter forming the boundaries of folding domains (CID, TAD (Pombo et Dillon 2015)). Hence, the positioning of highly expressed genes could determine the structuring of the nucleotide *in vivo* and could contribute to the course of cell replication/segregation/division processes. Consequently, the organization of these genes or loci could constitute a strong evolutionary constraint. In a global transcriptome analysis of *S. ambofaciens* ATCC 23877 it was shown that the majority of the 803 core-genes (61%) belong to the most expressed genes in at least one of the growth conditions tested (S. Bury-Moné, unpublished). Moreover, about 75% show a transcriptional orientation in the direction of the continuous replication *versus* only 55% for the rest of the genes (not shown). Therefore, the 803 core genes whose organization appeared particularly conserved could constitute such a chromosomal skeleton.

Chromosomal arms are fast evolving regions

In order to better characterize the decline of synteny along the chromosomal arms, the chromosome used as reference was splitted in portions equivalent by the number of genes and OR trends were plotted against the phylogenetic distance through as many pairwise comparisons as possible in our dataset (109). Different tests, by dividing the genome from 6 to 100 pieces of equivalent size, were carried out in order to visualize the behaviors which could typify the different chromosomal regions regarding their evolutionary speed. Dividing the genome into small fragments (100 pieces) had the advantage of precision but increased the amount of data to be displayed. In fact, this attempt gave too many

overlapping curves and finally illegible results (not shown). In contrast, cutting the genome into 10 or 20 limited the amount of curves to be displayed but result in observing the trend of the index over large portions of the chromosome and therefore averaging the trends, and thus could ultimately prevent distinguishing the peculiar behaviour of some limited chromosomal areas (*i.e.*, significantly smaller than the portion under consideration). It appeared that 6-piece cutout is sufficient to reveal the differential behaviour of the chromosomal parts (**figure 9B**).

As shown on **figure 9**, dividing the chromosome into 6 pieces gave clear results and revealed compelling different evolutionary patterns. Considering *S. ambofaciens* which possesses a 'typical' chromosome (i.e., with two arms flanking the central region**, figure 4**), the two terminal portions (portions I and VI) showed drastically lower values for both GOC and OR than the rest of the chromosome. This reflects that these chromosomal regions are prone to a higher intensity of gene fluxes and recombination than the central part of the chromosome. Further, the decline of the two indexes was more rapid, *i.e.* declined immediately from the closest pairs of compared species, for the two terminal regions than in the rest of the genome. Hence, in the distance range 0-0.1, the portions 1 and 6 showed a more significant OR decline (curve slope α of -4.5) than the portions 2-5 which showed a decline of -2 (**figure 9**). Hence, the greater relative decline of OR (2.25 fold stronger) in comparison of closely related strains *versus* more loosely related pairs could reflect the strong intensity of gene fluxes within the *Streptomyces* genome. Frequent acquisition (horizontal gene transfer) and loss (recombination) of genetic material may rapidly renew and diversify the accessory genes mostly concentrated in the chromosomal arms. This rapid turnover may lead to saturate these regions, phenomenon which may be observable when the phylogenetic distances are increasing. Hence, when the synteny is erased by the intense gene fluxes (reflected by very low OR indexes), it is no longer possible to distinguish the evolution of the terminal parts *versus* the central parts of the chromosome (*i.e.*, the slopes are not dissimilar; slopes $\alpha_{2,3,4,5}$= -0.4, $\alpha_{1,6}$= -0.34 within the phylogenetic distance range of 0.1-0.6).

In contrast to *S. ambofaciens* where the chromosome has a 'typical' organization, *S. avermitilis* MA-4680 possesses an unbalanced chromosome, with a left arm at least 5 times the size of the right arm as revealed by the distribution of the core genes (**figures 4 and 9.A**). When the chromosome is cut into 10 portions, 2 of them (numbered 1 and 2) showed significantly lower values of OR and span the left

terminal region of the chromosome. Similarly, when the chromosome is cut into 20 portions, 5 of them (numbered 1, 2, 3, 4 and 20) showed low OR values. Portions 1-4 corresponded to the left arm while the portion 20 correspond to the right one. the latters spanning the left terminal region of the chromosome. The 6-piece cutout revealed the same trend: only portion 1 corresponding to the left arm showed a rapid and strong decline for both GOC and OR indexes (**figure 9.B**). Therefore, it seems that the unbalanced chromosome of *S. avermitilis* MA-4680 also displayed an asymmetrical evolution pattern, consistent with the faster evolution of the arms compared to the core-region.

The case of *S. xiamenensis* 318 illustrates a third type of trend with all the 6 chromosomal portions showing roughly the same behaviour. This is consistent with the almost total absence of chromosomal arms in that species. This species possesses a rather short chromosome (5.96 Mb) characterized by the predominant space occupied by the core genes.

***Streptomyces* chromosome organisation and genome dynamics**

Altogether, these results allow us to propose a global hypothesis about the mechanisms shaping the *Streptomyces* chromosome. Using *S. ambofaciens* ATCC 23877 as a reference, **figure 10** captures the dynamics of the *Streptomyces* chromosome considering the phylogenetic depth (pairwise comparisons with *S. coelicolor* A3(2), *S. reticuli* TUE45, and *S. albus* BK325). Both OR and GOC indexes revealing gene fluxes and synteny shuffling, respectively, showed similar trends along the genome. *S. ambofaciens* ATCC 23877 which belongs to the "typical" chromosome category (the arms corresponding to 1.33 and 2 Mb on the left and right extremities respectively, **figure 4**). This overview emphasizes that both gene flux and shuffling are operating more intensively in the arms than in the central part. What appears clearly is that this phenomenon is all the more intense as the phylogenetic distance is short. Hence the variation of OR and GOC is higher when *S. ambofaciens* ATCC 23877 is compared to *S. coelicolor* A3(2) (**figure 10**, black dotted curves) than with the more distant *S. reticuli* TUE45 (**figure 10**, green curves), and *S. albus* BK325 (**figure 10**, blue curves) where the differences between chromosomal arms and center of the chromosome appeared mitigated. This phenomenon may result from the strong intensity of gene fluxes and shuffling, both revealing recombinogenic activity in the chromosomal arms. The rapid accumulation of both types of event may saturate the chromosomal

arms rapidly and prevent to observe these phenomena at long phylogenetic distances. A previous study limited to the few complete genome sequences available at that time, showed that chromosomal regions separating the core region from chromosomal arms were riddled with short indels (1 to 10 genes) forming a gradient toward the terminal ends (Choulet, Gallois, et al. 2006). This discovery allowed us to define the concept of 'degenerated synteny' and suggested that diversification occurred mainly by the cumulation of indels through evolutionary times. The results reported here obtained at the genus level can be put into perspective with those obtained between closely related strains (*i.e.*, belonging to the same species) isolated from the same micro-niche at the millimetre scale (Tidjani et al. 2019). Indeed, it has been shown that, although separated by a short evolutionary time, strains isolated from the same sample showed significant genomic differences consisting of insertions and deletions (indels) preferentially located in the chromosomal arms. A close analysis revealed that the indels distribution formed a gradient increasing in intensity towards the chromosomal ends.

Altogether, by combining data collected at short and long genetic distances, it appears that the image of the contemporary genomes, i.e., with a central conserved region and variable chromosomal arms, could result from an insertion and deletion cumulative process. This process could be established preferentially towards the chromosomal ends by increasing recombination frequencies towards the ends. Alternatively, these more frequent indels could reflect the lower selection pressure on the coded functions in the chromosomal arms. This hypothesis seems less plausible to us because it would imply that the genes would be arranged in these regions with decreasing importance towards the chromosomal ends.

As shown in the **figure 11**, we propose that the accumulation of indels according to this recombination gradient accompany the divergence of strains from their common ancestor. In addition to this progressive process, interreplicon followed by intrachromosomal recombination events may trigger sporadic and massive divergence leaps.

MATERIALS AND METHODS

Average Nucleotide Identity (ANIb) computation

The Average Nucleotide Identity (ANI) was calculated by using the BLASTn algorithm, ANIb (Goris et al. 2007). Firstly, the query genome is fragmented into consecutive 1000 parts, and each of them is then were aligned with the sequence of the reference genome using BLASTn (v2.2.31+) (Altschul et al. 1997). The ANIb score corresponds to the average value of the nucleotide identity percentages of the query fragments with a positive match with the reference (alignment greater than 70% with at least 30% of nucleotide identity, (Goris et al. 2007). Since the ANIb score is not reciprocal (*i.e.* the ANIb score of genome A versus genome B may be slightly different than that of genome B versus genome A), we used the average of the two reciprocal $\underline{ANIb}$ values as final score.

Constitution of a set of genomes representative of the *Streptomyces genus*

The 135 complete genomes of *Streptomyces* available in the NCBI database (O'Leary et al. 2016) on the date of 02/06/2019 constituted our basic set of genomic data. In order to reduce the size of the dataset and avoid redundancies, the genomes have been clustered using their $\underline{ANIb}$ value. Genomes sharing an $\underline{ANIb}$ values greater than or equal to 96% were considered to represent strains belonging to the same species (Richter et Rosselló-Móra 2009) and in consequence only one was kept in the final dataset. *Streptomyces ambofaciens* DSM 40697 and ATCC 23877 were both kept in our set since they represent the model species of our lab, and provide a comparison between closely related strains. In total, the genomes of 110 strains or species were collected (**table 1**). The assembly completeness of the genomes retained in the dataset was assessed using BUSCO (Simão et al. 2015) with the actinobacteria dataset (*odb9*[1]).

Terminal Inverted Repeats detection and identification

Terminal Inverted Repeats (TIR) are defined as perfect DNA repeats at the ends of the chromosome or linear plasmids. However, these repeats may be present or not in the assembly published in the database depending of their size which can largely vary (from some tenths of nucleotides to several hundreds of kilobases) and the sequencing technologies used for genome sequencing (i.e., sequencing from

[1] https://busco.ezlab.org/v2/datasets/actinobacteria_odb9.tar.gz

recombinants BACs or cosmids, whole genome sequencing using long or short reads). The longer is the repeat the weaker is the chance to identify it in a single piece in a sequence scaffold. To detect the presence of TIRs and/or to identify their size, we searched for DNA repeats at the ends of all the genomes considered. For that purpose, we search for an alignment between the first magabase of the chromosome and the end of the sequence (BLASTn v2.2.31+ (Altschul et al. 1997)). When a terminal repeat was detected (99% of nucleotide identity), we deduced the TIR length from the position of the beginning of the hit in the query (first megabase) (**table 1**). If identified, only one single copy of TIRs was kept in the final sequence of the chromosome.

Genome annotation, orthology assignment

All the genome sequences were automatically annotated on the RAST Server (Aziz et al. 2008) with the RASTClassic pipeline in order to have an homogeneous annotation process, and fueled the orthology assignment process with a valid and consistent annotation dataset. For each pair of genomes, orthologs were defined by identifying reciprocal BLASTp best hits (BBH) (Fang et al. 2010; Tatusov 1997) with at least 40% of identity, 70% of coverage and an E-value lower than $1e^{-10}$.

Core and pan-genome computation

The core-genome was determined as the set of orthologs in all the genomes of a given dataset, and the pan-genome as the sum of core-genomes plus the specific genes found in the genomes of the same set. In order to estimate the impact of the number of genomes available on the size of the core- and pan-genome, we computed both for growing sets of genome varying from 2 to 110 genomes. For any set of size n lower than 110, core- and pan-genomes were computed by performing 100 iterations of a random selection of n different genomes among the 110 genomes available in our dataset. Core- and pan-genomes dimension for each set size can be represented as a whisker box with the lower and upper whiskers corresponding to the first and the last quartile of the distribution. Considering the whole genome set, the core-genome contained 803 genes. Based on their position, the core-region has been defined as the area including 95% of them.

Maximum-likelihood based phylogenetic inference

For each genome, the protein sequence of the 803 genes of the core-genome were retrieved. After alignment using MUSCLE v3.8.31 (Edgar 2004) and trimming by Gblocks to eliminate poorly aligned

positions (Castresana 2000), the 803 multiple sequence alignments were concatenated into a unique multiple sequence alignment including 255,632 amino acid positions. The multiple alignment was then submitted to RAxML (Stamatakis 2014) with the LG substitution model for maximum likelihood based tree inference. One hundred bootstrap replications were performed. The phylogenetic tree has been represented with the Dendroscope software (Huson et al. 2007).

Gene Order Conservation (GOC), Ortholog Rate (OR) and NOC (Neighbour Orthologue Conservation) indexes

The Gene Order Conservation (GOC) was calculated according to Rocha (2006) and reflects the conservation of the vertically inherited set of genes, that is to say the ancestral chromosome architecture. A strong GOC value means a strong conservation of the gene organization, a low value reveals the shuffling of the orthologs along the genome, meaning that recombination breaks the ancestral synteny. This value is calculated by dividing the number of orthologs involved in contiguous pairs by the number of orthologs between two compared genomes. A sliding window (1% or 5% of the total number of genes of the reference genome) is moved along the reference genome to get local values which can give a trend along the chromosome (**figure 5**). The Ortholog Rate for its part considers the conservation of the orthologs among other genes in a sliding window moved along the reference genome in a pairwise comparison. It allows to measure the influence of the acquired/lost genes in the conservation of genome architecture (**figure 5**). The Neighbour Orthologue Conservation consists of the ratio of the number of genes involved in contiguous pairs in the reference strain that are still contiguous in the compared strain (**figure 5**).

The Double-Cut and Join (DCJ)

The DCJ model is a genomic rearrangement model used to define an editing distance between genomes based on the order and orientation of genes (Yancopoulos et al. 2005). Rearrangement events that happen in genomes, such as inversions, translocations, fusions, and fissions can be represented. In our case, a variant of the DCJ model was used, the restricted DCJ model (Kováč et al. 2011). The application of the DCJ algorithm requires that the two genomes being compared have the same gene content. Therefore, this method could only be used on the core-genome genes or on the orthologous genes for a given pair of species.

GOC and NOC Heat map representation

All the GOC or NOC profiles for a given reference genome have been computed as previously described. Each GOC/NOC profile, *i.e.* all GOC/NOC values ordered as the order of the windows for a pair of chromosomes, can be seen as a vector of GOC/NOC index values. For all the GOC/NOC profiles obtained for a reference genome, the corresponding vector has been converted to a color code according to a color scale. All the color code have been ordered according to the cophenetic distance of the species represented by the color code to the reference genome.

TABLES AND FIGURES

Name	Genome Assembly ID	Length (pb)	# plasmid	# CDS	GC rate (%)	TIR length (pb)
Streptomyces sp. 4F	ASM148470v1	8047771	0	7250	72.28	334071
Streptomyces sp. 769	ASM81602v1	10100774 [237512]	1	8956 [224]	71.6	155968
S. actuosus ATCC 25421	ASM320803v1	8145579	0	7399	72.53	17181
Streptomyces sp. ADI95 16	ASM372149v1	8184000 [902421]	4	7362 [861]	72.04	-
S. albireticuli MDJK11	ASM219245v1	8144417	0	7024	72.82	133288
S. alboflavus MDJK44	ASM218967v2	9622415 [345134]	2	8796 [335]	72.0	-
S. albulus CK 15	ASM93518v3	9336218	0	8241	72.35	-
S. albus BK3 25	ASM175342v1	8308430	0	6871	72.64	-
S. alfalfae ACCC40021	ASM197502v1	8625867	0	7525	72.09	149241
S. ambofaciens ATCC 23877	ASM126788v1	8303940 [89658]	1	7413 [128]	72.19	202695
S. ambofaciens DSM 40697	ASM163286v1	8137876	0	7250	72.33	212696
S. atratus SCSIO ZH16	ASM333086v1	9641288	0	8700	69.59	288246
S. autolyticus CGMCC0516	ASM198397v1	10029028 [155632]	7	8243 [36]	71.21	-
S. avermitilis MA 4680	ASM976v2	9025608 [94287]	1	8038 [94]	70.7	50
S. bacillaris ATCC 15855	ASM326867v1	7888441	0	6953	71.95	-
S. bingchenggensis BCW 1	ASM9238v1	11936683	0	10216	70.75	139
S. brunneus CR22	ASM395571v1	10393987	0	9540	70.85	96011
S. cattleya NRRL 8057	ASM23730v1	6283062 [1809491]	1	5538 [1651]	73.01	-
Streptomyces sp. CB09001	ASM336979v1	7787608	0	7085	71.95	-
Streptomyces sp. CC0208	ASM344373v1	9320089	0	8531	70.59	-
Streptomyces sp. CCM MD2014	ASM77204v1	8274043	0	7501	72.13	14382
Streptomyces sp. CdTB01	ASM148456v1	9902731 [288836]	1	9157 [287]	71.53	400291
Streptomyces sp. CFMR 7	ASM127809v1	8207742 [99537]	1	7095 [95]	72.03	11739
S. chartreusis NRRL 3882	NRRL3882	8983317	0	8083	71.23	39831
S. clavuligerus F1D 5	Scla_1.0	6900908 [1158217]	2	5831 [1079]	72.53	34
Streptomyces sp. CLI2509	ASM228807v1	7088673 [144028]	1	6198 [149]	73.3	25387
Streptomyces sp. CMB StM0423	ASM284728v1	8029398	0	6930	73.14	9667
Streptomyces sp. CNQ 509	ASM101103v1	8039333	0	7028	73.07	32534
S. coelicolor A3(2)	ASM20383v1	8667507 [387340]	2	7920 [400]	72.0	21653
S. collinus Tu 365	ASM44487v1	8272925 [104361]	2	7182 [101]	72.55	631364
S. cyaneogriseus noncyanogenus NMWT 1	ASM93144v1	7762396	0	6681	72.86	-
Streptomyces sp. endophyte N2	ASM410448v1	8428700	0	7398	71.83	53557
Streptomyces sp. fd1 xmd	ASM200768v1	7929999	0	7236	72.51	-
S. formicae KY5	ASM255654v1	9611874	0	8317	71.38	35482
S. fulvissimus DSM 40593	ASM38594v1	7905758	0	7072	71.48	-
S. fungicidicus TXX3120	ASM366543v1	6740768 [926728]	1	5998 [841]	72.22	-
S. gilvosporeus F607	ASM208219v1	8482298	0	7624	70.95	-
S. glaucescens GLA O	ASM76121v1	7453200 [170574]	1	6575 [140]	72.91	14128
S. globisporus C 1027	ASM26134v2	7608611 [174988]	2	6925 [154]	71.54	18
S. globisporus TFH56	ASM314754v1	7488586 [177935]	2	6847 [155]	71.54	-
S. globosus LZH 48	ASM332537v1	6863360 [672390]	2	6196 [539]	73.65	-
Streptomyces sp. Go 475	ASM333084v1	8570609	0	7729	71.96	6902
S. griseochromogenes ATCC 14511	ASM154262v2	10764674	0	9839	70.76	-
S. griseorubiginosus 3E 1	ASM359523v1	9512378	0	8551	70.94	113251
S. griseoviridis F1 27	ASM399439v1	8963414	0	7785	72.38	58707
S. griseus NBRC 13350	ASM1060v1	8545929	0	7306	72.23	132910
Streptomyces sp. GSSD 12	ASM334496v1	8454852	0	7201	71.19	-
Streptomyces sp. HNM0039	ASM309751v1	7289495	0	6363	72.46	-
S. hundungensis BH38	ASM362781v1	8393044	0	7541	70.94	65122
S. hygroscopicus jinggangensis 5008	ASM24535v1	10145833 [237851]	2	9158 [243]	71.84	15
S. hygroscopicus XM201	ASM202187v1	12012215	0	10288	70.75	-
S. koyangensis VK A60T	ASM342892v1	7220839	0	6365	73.03	20855
Streptomyces sp. KPB2	ASM395005v1	8082236	0	7256	72.21	40616
S. lavendulae CCM 3239	ASM280384v1	8691711 [241081]	1	7854 [231]	72.62	237734
S. leeuwenhoekii sleC34	sleC34	7903895 [218596]	2	6857 [263]	72.68	388272
S. lincolnensis LC G	ASM334444v1	9513637	0	8622	71.06	-
S. lincolnensis NRRL 2936	ASM168535v1	8396100 [174091]	2	7724 [149]	69.75	-
S. lunaelactis MM109	ASM305455v1	7367863	0	6574	72.05	-
S. luteoverticillatus CGMCC 15060	ASM397071v1	8201357	0	7329	72.22	-
S. lydicus 103	ASM172948v1	9307519	0	8534	70.68	52505
S. lydicus A02	ASM95203v2	9125666	0	7970	70.8	8910
S. lydicus WYEC 108	ASM399437v1	8718751	0	8025	70.69	-
Streptomyces sp. M2	ASM410450v1	7868178 [848015]	3	7052 [812]	72.13	-
Streptomyces sp. Mg1	ASM41226v2	10677137	0	9604	70.63	131471
Streptomyces sp. MK45	ASM396353v1	8399509	0	7522	71.99	107206
Streptomyces sp. MOE7	ASM209033v1	9641634 [45805]	1	8456 [52]	70.78	-
Streptomyces sp. NEAU S7GS2	ASM317327v1	7641029	0	6992	71.91	107944
S. nigra 452	ASM307405v1	7990492	0	7105	70.46	-
S. niveus SCSIO 3406	ASM200917v1	9815884	0	8811	71.45	170209
S. noursei ATCC 11455	ASM170427v1	8180260	0	7384	72.41	141244
S. olivaceus KLBMP 5084	ASM176737v1	8809793	0	7593	71.11	170471
S. olivoreticuli ATCC 31159	ASM339113v1	7180417	0	6166	72.45	-
S. ongiicola HNM0071	ASM312236v1	9851971	0	8820	71.37	-
Streptomyces sp. P3	ASM303247v1	8550793	0	7375	72.06	239610

S. pactum ACT12	ASM200522v1	7149446 [617085]	1	6520 [443]	72.73	14
S. parvulus 2297	ASM166004v1	8023114	0	7391	70.59	143511
S. peucetius caesius ATCC 27952	ASM277753v1	7346075	0	6561	69.94	-
S. pluripotens MUSC 135	ASM80224v2	7337497 [318607]	2	6583 [312]	70.99	800
S. pratensis ATCC 33331	ASM17611v2	8532592	0	7498	71.54	419828
S. pristinaespiralis HCCB 10218	ASM127807v1	9698948	0	8823	71.09	112391
S. puniciscabiei TW1S1	ASM173580v1	8353915 [982309]	3	7437 [832]	72.72	-
S. reticuli TUE45	TUE45	11142275	0	9796	71.28	19211
Streptomyces sp. RTd22	ASM165021v1	6543262	0	5602	74.78	39368
S. rubrolavendulae MJM4426	ASM175078v1	7614683	0	7110	71.49	39495
Streptomyces sp. S063	ASM283267v1	9083372	0	8241	71.26	134209
Streptomyces sp. S10 2016	ASM161179v1	7529075 [72789]	1	6546 [78]	72.29	12184
Streptomyces sp. S8	ASM209499v1	7472530	0	6503	73.15	-
S. scabiei 87 22	ASM9130v1	10148695	0	9009	71.45	18488
Streptomyces sp. SCSIO 03032	ASM212830v1	6287975	0	5506	73.52	-
Streptomyces sp. Sge12	ASM208045v1	7983613 [127085]	1	7257 [119]	72.17	162199
Streptomyces sp. SirexAA E	ASM17719v2	7414440	0	6663	71.75	-
Streptomyces sp. SM17	ASM291072v2	6975788 [204126]	3	6107 [186]	73.35	-
Streptomyces sp. SM18	ASM291077v2	7703166	0	6783	71.84	14611
Streptomyces sp. TN58	ASM194184v1	7585034	0	6936	72.3	193939
Streptomyces sp. Tue6075	ASM193163v1	7931832	0	6994	71.57	12088
S. venezuelae ATCC 15439	ASM144362v1	9054831	0	8130	71.74	48229
S. venezuelae NRRL B 65442	ASM188659v1	8222198 [158122]	1	7422 [158]	72.42	-
S. vietnamensis GIM4 0001	ASM83000v1	8867142 [286635]	1	8016 [271]	71.99	49821
S. violaceoruber S21	ASM208217v1	7916045	0	6979	72.65	4122
S. violaceusniger Tu 4113	ASM14781v3	10657107 [481206]	2	9087 [523]	70.88	-
Streptomyces sp. W1SF4	ASM395003v1	7272878 [795265]	2	6764 [656]	73.14	-
Streptomyces sp.WAC 01438	ASM394552v1	8138328 [62839]	1	7278 [72]	71.96	49769
Streptomyces sp. WAC 01529	ASM394554v1	8270461	0	7269	71.32	58922
Streptomyces sp. WAC 06738	ASM394550v1	8324535	0	7207	73.18	11391
Streptomyces sp. WAC00288	ASM294389v1	7467777 [385082]	4	6808 [379]	72.73	-
S. xiamenensis 318	ASM99378v2	5961401	0	5411	72.02	-
S. xinghaiensis S187	ASM22070v2	7137891	0	6223	73.14	-
Streptomyces sp. XZHG99	ASM294683v1	8541354 [168817]	2	8026 [210]	69.92	21003
Streptomyces sp. Z022	ASM367532v1	8085191 [72694]	1	7187 [55]	72.12	26248
Streptomyces sp. ZFG47	ASM326105v1	9269371 [969901]	1	8198 [772]	70.76	191546

Table 1: Principal informations of the 110 *Streptomyces* species.
Principals informations of the 110 *Streptomyces* species. **Name**: scientific name; **Genome Assembly ID**: unique identifier used in the NCBI Genome Assembly database; **Length (pb)**: chromosome length [cumulative length of plasmids] expressed in pb; **# plasmid**: number of plasmids ; **# CDS**: CDS in chromosome [CDS in plasmids]; **GC rate (%)**: percentage of cytosine and guanine in the sequence; **TIR length (pb)**: Size of one copy of a TIR expressed in pb.

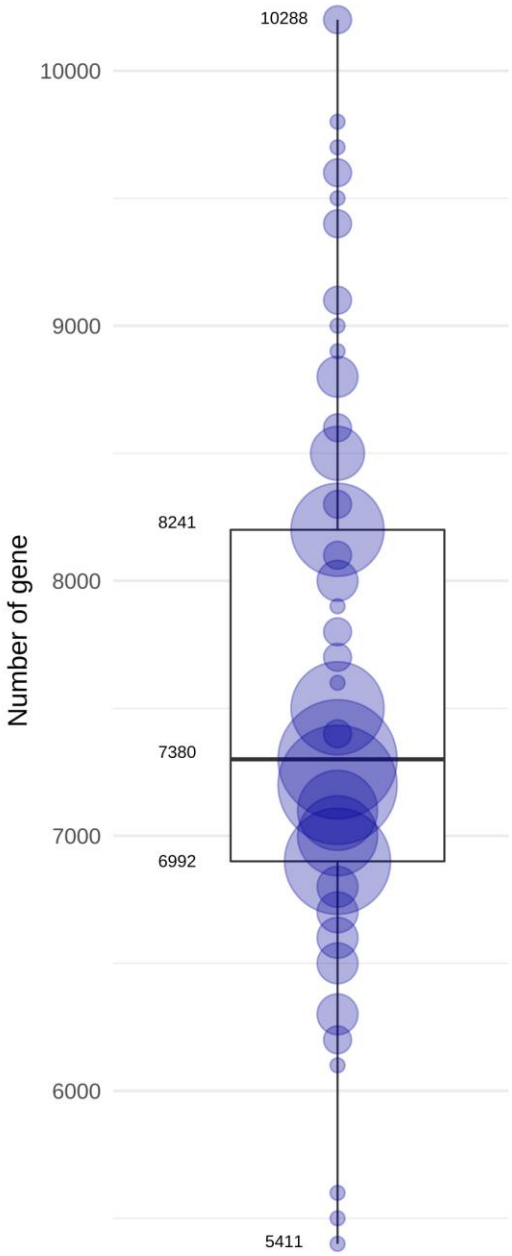

Figure 1: Genomes size distribution of the 110 *Streptomyces* species.
The lower and upper whiskers of the boxplot show the first quartile and the last quartile, the horizontal central line indicates the median value. The blue circles indicate the number of genomes in our dataset along the distribution where the radius of each circle is proportional to the number of genome in this range of size.

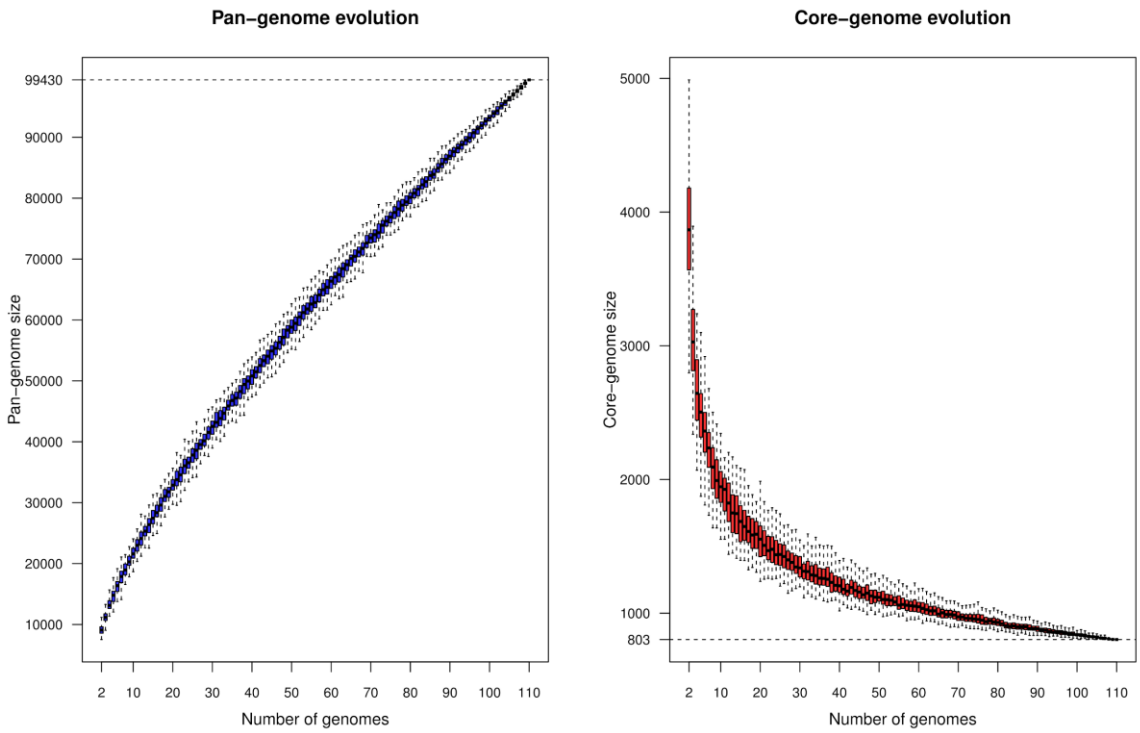

Figure 2: Pan-genome and core-genome evolution according to the number of genomes.
For a number of genomes varying from 2 to 110, the size of the pan-genome and core-genome was approaches after 100 iterations of a random selection of genomes picked in our dataset. Each point is a whisker box with the lower and upper whiskers corresponding to the first and the last quartile of the distribution. The horizontal line gives the median value. The pan- and core genomes were estimated to be 99,430 and 803 genes, respectively, for the whole set of species.

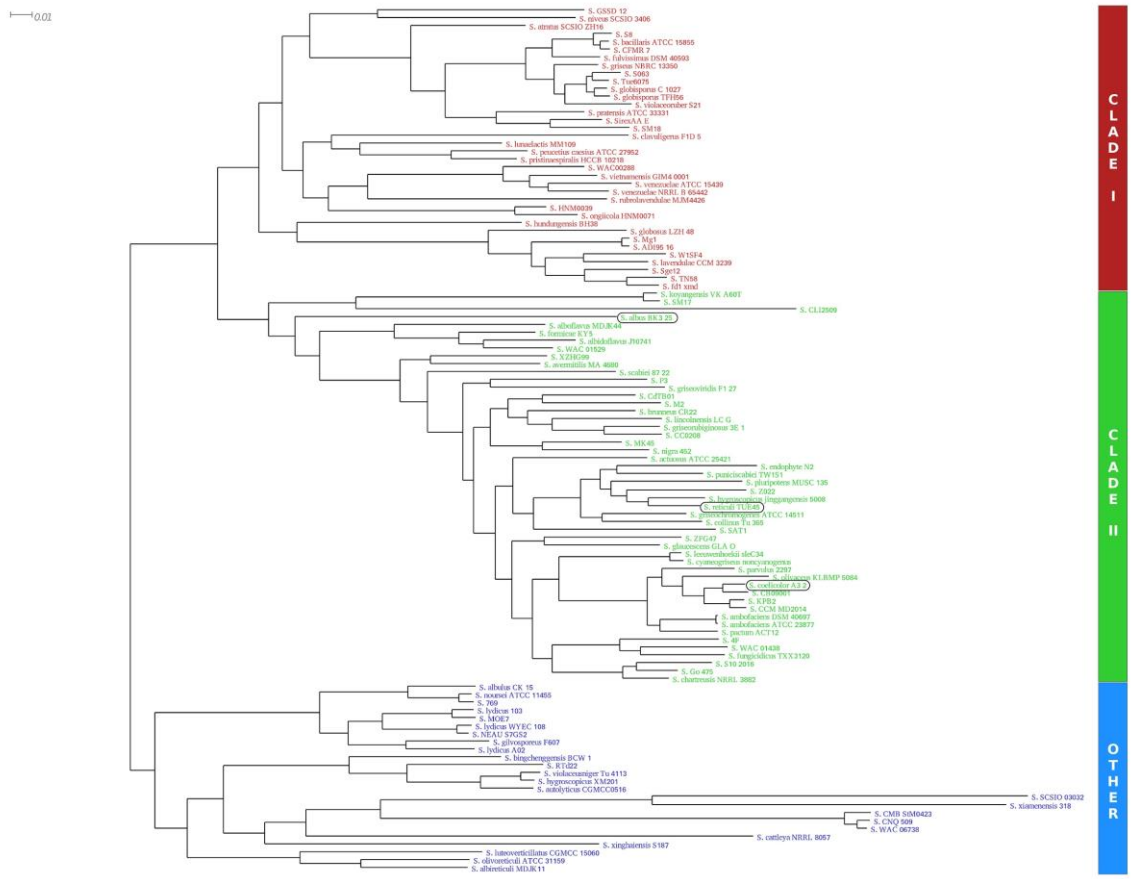

Figure 3: Phylogenetic tree of the 110 Streptomyces species.
The Maximum likelihood approach was used to reconstruct the phylogenetic relationships between our species/strains from the 803 genes of the Streptomyces core genome. The different monophyletic clades identified and consistent with McDonald and Currie (McDonald et Currie 2017) are indicated. The numbers on the nodes correspond to the bootstrap values obtained after 100 bootstrap replications. Species used in figure 5 are identified by a tick.

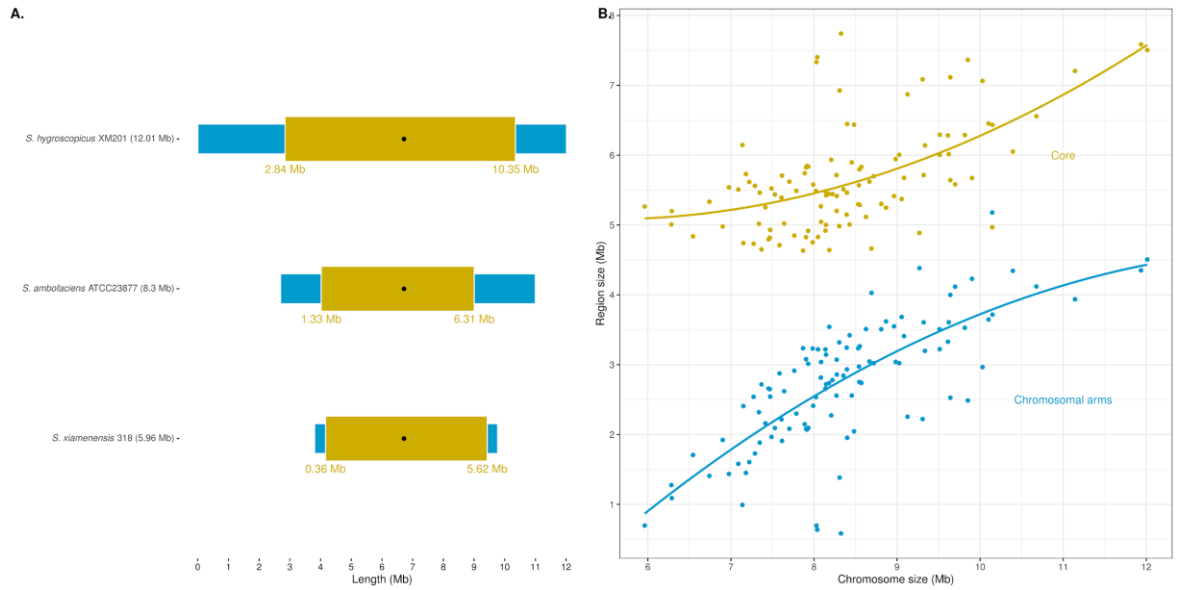

Figure 4: Core-genome and chromosomal arms across *Streptomyces* species.
A. Density of core genes along some remarkable *Streptomyces* chromosomes (e.g. S. *ambofaciens* ATCC 23877, …).
B. Chromosomal arms versus core-genome across *Streptomyces* species. The chromosomal arms are the regions devoid of core-genes (the sizes right and left arms are cumulated). The core-region is defined as the region occupied by 95% of the core-genome.

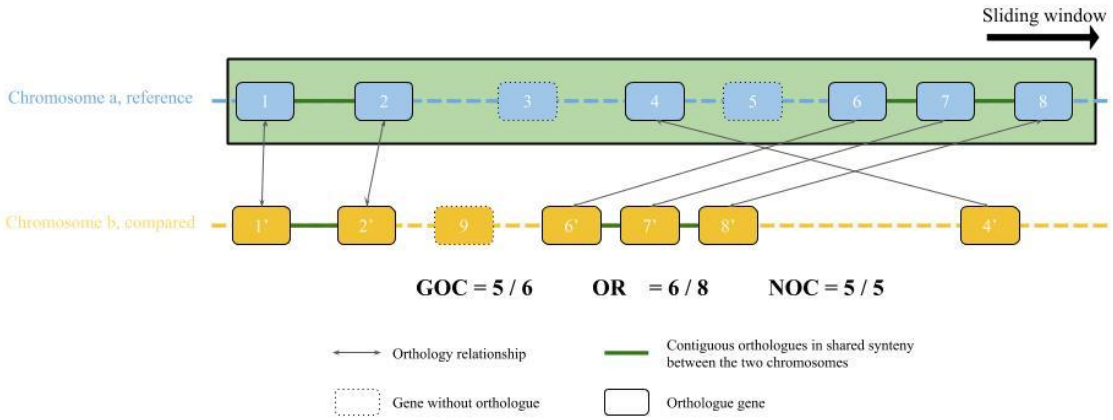

Figure 5 : Definition of indexes for genome conservation analyses.
Pairwise comparison are achieved using a sliding window and comparing a reference strain (species a) to another (species b). Here is considered a window containing 8 genes noted from 1 to 8 in the reference species. For the GOC index, the pair of contiguous genes present in the window in the reference species are searched and counted as contiguous pairs of orthologues in the whole genome of species b related to the number of orthologues defined by the window. For the OR index, genes in the window in the reference are searched and counted in the compared genome and normalized by the number of genes in the window.

$$GOC = \frac{Number\ of\ contiguous\ orthologues\ in\ shared\ synteny\ between\ the\ two\ chromosomes}{Orthologues\ number\ between\ the\ two\ chromosomes}$$

$$OR = \frac{Orthologues\ number\ between\ the\ two\ chromosomes}{Genes\ number\ in\ the\ reference\ chromosome}$$

$$NOC = \frac{Number\ of\ contiguous\ orthologues\ in\ shared\ synteny\ between\ the\ two\ chron}{Number\ of\ genes\ involved\ in\ a\ group\ of\ contiguous\ orthologues\ in\ the\ reference}$$

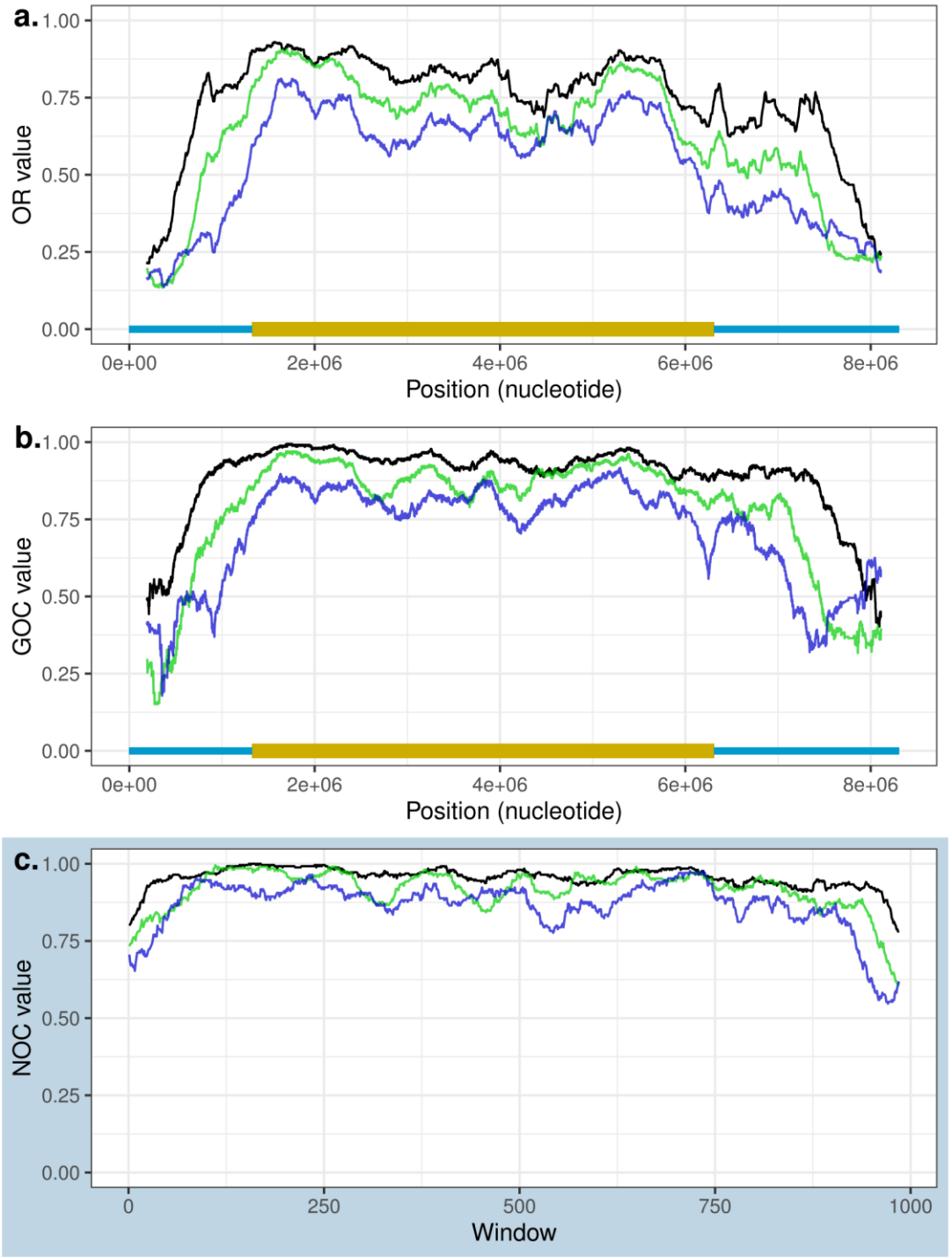

Figure 6: GOC, OR and NOC profiles along the Streptomyces chromosome.
Profiles are obtained by plotting GOC, OR or NOC values obtained by pairwise comparison of S. *ambofaciens* ATCC 23877 as a reference with three species showing phylogenetic distance from tight (vs. *S. coelicolor* A3(2), black curve), to average (*S. reticuli* TUE45, green curve), to weak *(S. albus* BK3 25, blue curve). GOC and OR values were calculated using a sliding window of size equal to 5% of the chromosomal gene content of *S. ambofaciens* ATCC 23877. NOC values were calculated using a sliding window of size equal to 5% of the number of genes involved in a group of contiguous orthologues of *S. ambofaciens* ATCC 23877 and the given other species. For GOC and OR, each window is located on the chromosome by the position of its central gene. For NOC, each window is not located on the chromosome, only their relative order is presented.

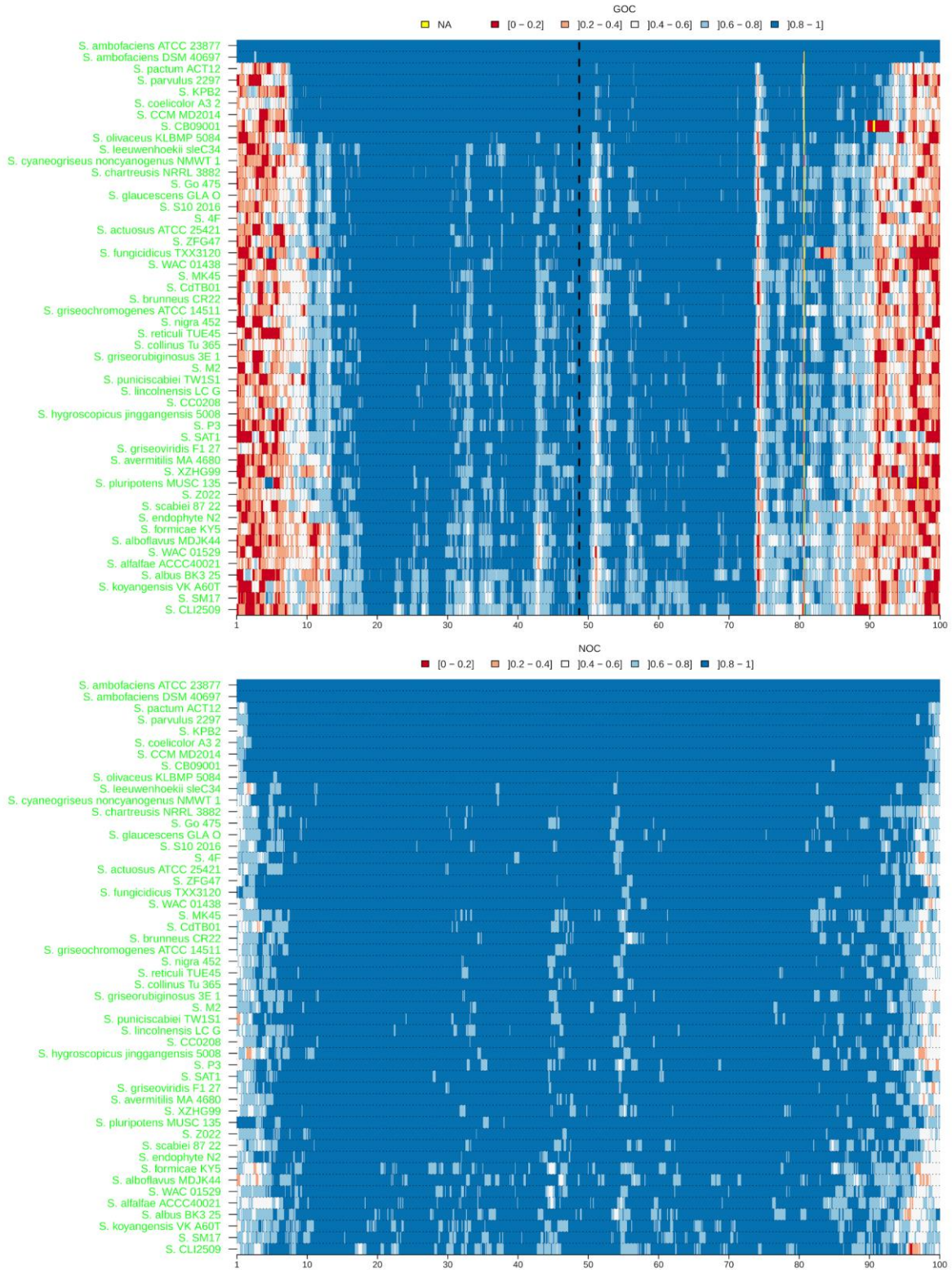

Figure 7: *S. ambofaciens* GOC and NOC heatmaps.
Each line corresponds to the index profile of the reference (*S. ambofaciens ATCC 23877*) against another *Streptomyces* species belonging to clade II. Each chromosomal region was colored according to its GOC value (top) or NOC value (bottom), from red to blue, each range of 0.2 is differently colored (blue > 0.8 - light blue > 0.6 - white > 0.4 - light red > 0.2 - red > 0). On GOC heatmap, yellow indicates a window for which the orthology rate is 0: GOC cannot be calculated. Dotted vertical line represents the *oriC* location on GOC heatmap. Species are arranged top down from the closest to farthest species. GOC was calculated using a window of 1% of the total number of *S. ambofaciens* ATCC 23877 genes. NOC using

a window 1% based of the total number of contiguous pairs of othologs in the reference strain that are still contiguous in the compared strain.

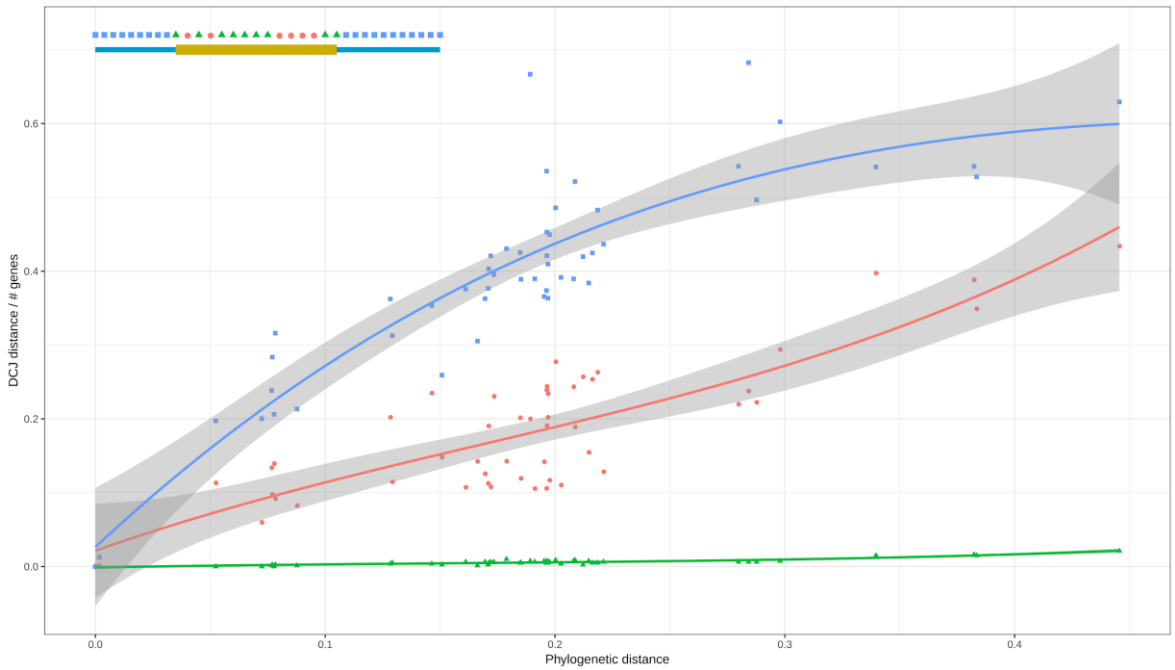

Figure 8: Propensity to rearrangements of *S. ambofaciens* ATCC 23877 genes.
Three categories of genes have been defined for *S. ambofaciens* ATCC 23877 according to their location and level of conservation with other species: core genes (green triangle), orthologues genes outside the core region (blue square) and orthologues genes inside the core region (red dot). For each category, a weighted version of the DCJ (Kováč et al. 2011) was calculated in pairwise comparisons and plotted against the phylogenetic distance between *S. ambofaciens* ATCC 23877 and the compared strains. A regression (polynomial degree 3) was applied.

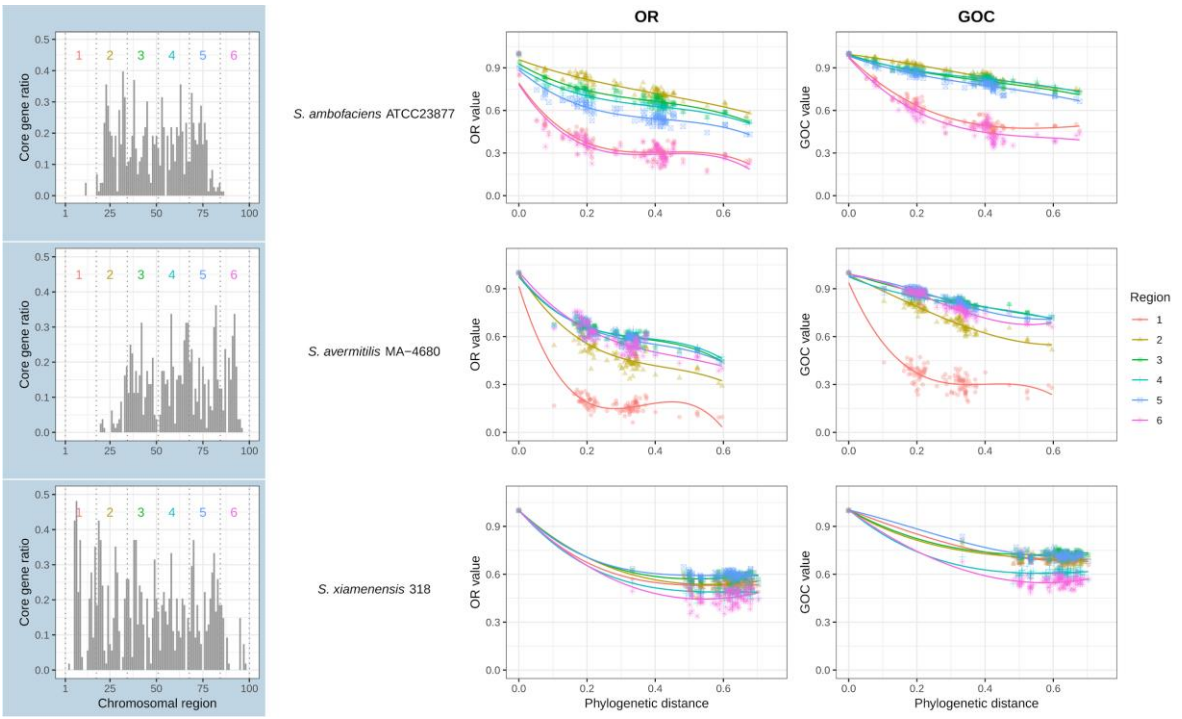

Figure 9: *Streptomyces* subtelomeres are fast-evolving regions.
Three reference species, *S. ambofaciens* ATCC 23877, *S. avermitilis* MA-4680 and *S. xiamenensis* 318 were compared using OR and GOC indexes to the other species of our sample. The chromosome of each representative species was split into 6 parts containing the same number of genes. The OR and GOC values for these 6 chromosomal parts were calculated in pairwise comparisons and plotted against the phylogenetic distance between the reference and the compared strains. A regression (polynomial degree 3) was applied.

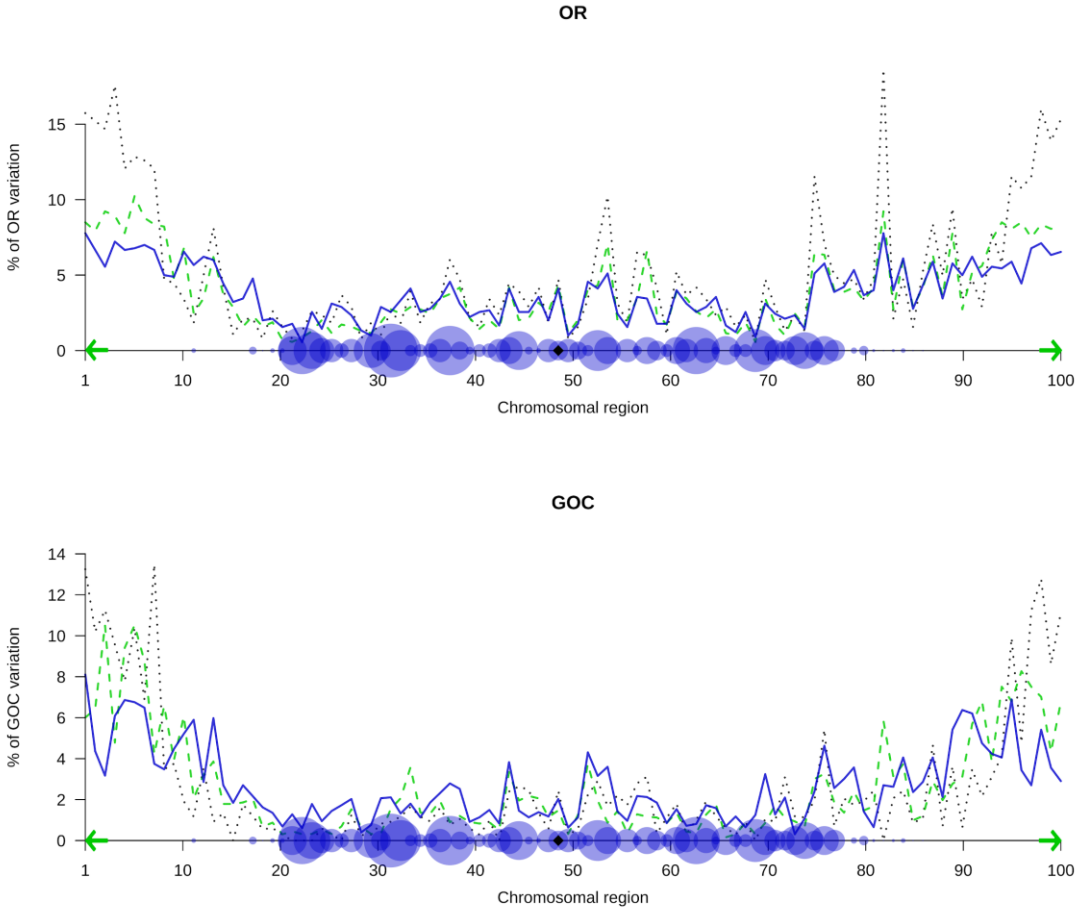

Figure 10: *S. ambofaciens* ATCC 23877 chromosome organisation and gene content dynamic.
Ortholog Rate (OR) and GOC values of S. *ambofaciens* ATCC 23877 versus three species showing different evolutionary distance (*S. coelicolor* A3(2) (black), *S. reticuli* TUE45 (green), *S. albus* BK 25 (blue)). The chromosome was divided into 100 regions containing the same number of genes. Green arrows identified the TIRs, black diamond corresponds to the replication origin and the blue circles to the core gene density (the radius of the circles are proportional to the number of core genes in the region).

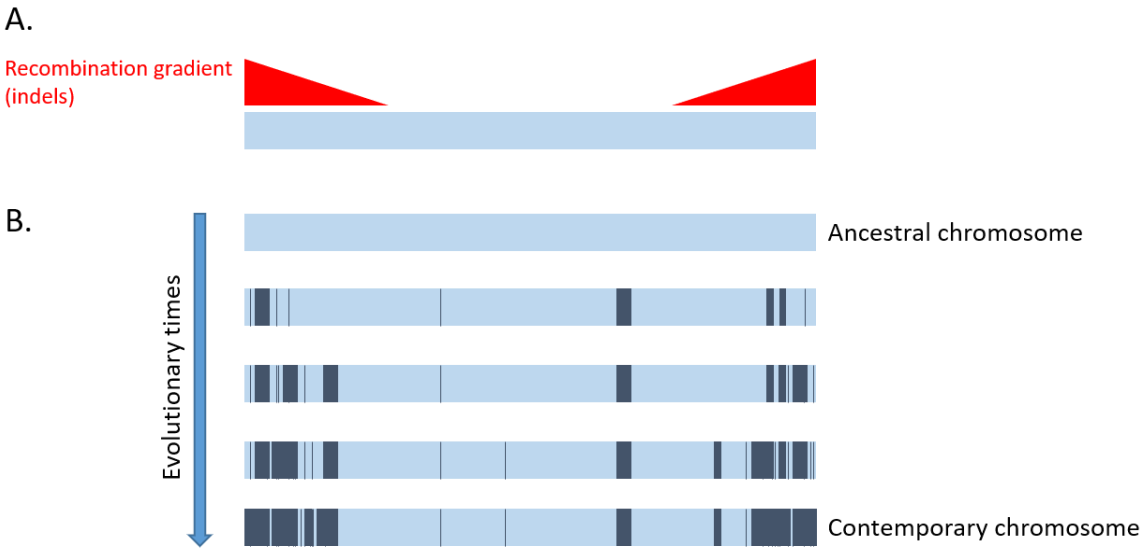

Figure 11: Accumulation of recombinaison events from an ancestral *Streptomyces* chromosome.
A. Schematic representation of the recombination gradient (red triangles) on a *Streptomyces* chromosome (blue rectangle). **B**. Rectangles represent the linear chromosome of hypothetical *Streptomyces* chromosomes. Chromosomes are ordered from top (ancestral chromosome) to bottom (contemporary chromosome) relative to evolutionary time. The darkblue rectangles represent the accumulation of recombination events.

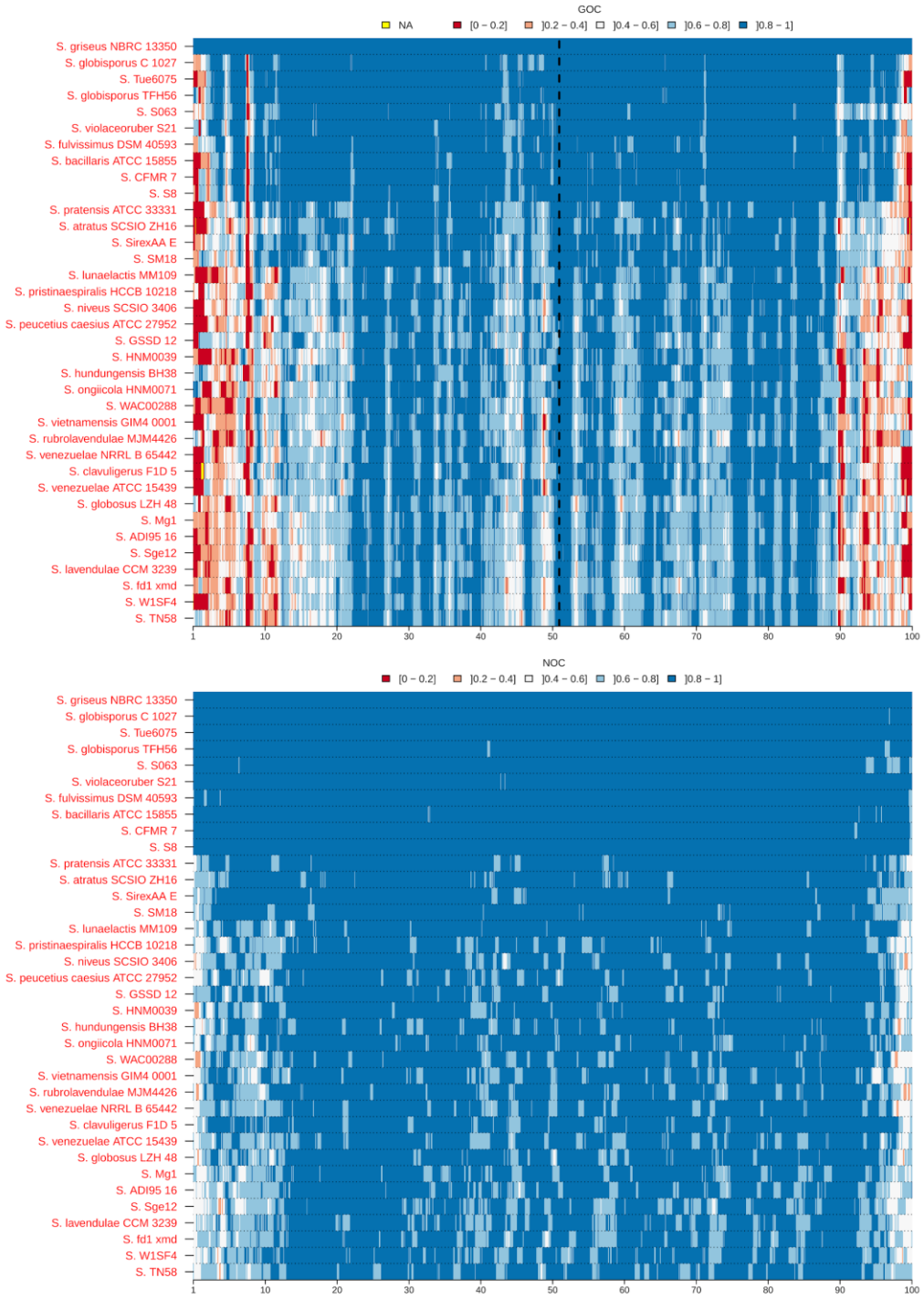

Figure S1: ***S. griseus*** **NBRC 13350 GOC and NOC heatmaps.**
Each line corresponds to the index profile of the reference (*S. ambofaciens ATCC 23877*) against another *Streptomyces* species belonging to clade II. Each chromosomal region was colored according to its GOC value (top) or NOC value (bottom), from red to blue, each range of 0.2 is differently colored (blue > 0.8 - light blue > 0.6 - white > 0.4 - light red > 0.2 - red > 0). On GOC heatmap, yellow indicates a window for which the orthology rate is 0: GOC cannot be calculated. Dotted vertical line represents the *oriC* location on GOC heatmap. Species are arranged top down from the closest to farthest species. GOC was calculated using a window of 1% of the total number of *S. ambofaciens* ATCC 23877 genes. NOC using a window 1% based of the total number of contiguous pairs of othologs in the reference strain that are still contiguous in the compared strain.

BIBLIOGRAPHY

Aigle, Bertrand et al. 2014. « Genome mining of Streptomyces ambofaciens ». *Journal Of Industrial Microbiology & Biotechnology* 41(2): 251-63.

Altschul, Stephen F. et al. 1997. « Gapped BLAST and PSI-BLAST: a new generation of protein database search programs ». *Nucleic acids research* 25(17): 3389–3402.

Anderson, A S, et E M Wellington. 2001. « The taxonomy of Streptomyces and related genera. » *International Journal of Systematic and Evolutionary Microbiology* 51(3): 797-814.

Antony-Babu, Sanjay et al. 2017. « Multiple Streptomyces Species with Distinct Secondary Metabolomes Have Identical 16S RRNA Gene Sequences ». *Scientific Reports* 7(1): 1-8.

Aziz, Ramy K. et al. 2008. « The RAST Server: rapid annotations using subsystems technology ». *BMC genomics* 9(1): 75.

Baltz, Richard H. 2017. « Gifted Microbes for Genome Mining and Natural Product Discovery ». *Journal of Industrial Microbiology & Biotechnology* 44(4): 573-88.

Bentley, Stephen D. et al. 2002. « Complete genome sequence of the model actinomycete Streptomyces coelicolor A3 (2) ». *Nature* 417(6885): 141–147.

Castresana, J. 2000. « Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis ». *Molecular Biology and Evolution* 17(4): 540-52.

Choulet, Frédéric, Bertrand Aigle, et al. 2006. « Evolution of the terminal regions of the Streptomyces linear chromosome ». *Molecular biology and evolution* 23(12): 2361–2369.

Choulet, Frédéric, Alexandre Gallois, et al. 2006. « Intraspecific variability of the terminal inverted repeats of the linear chromosome of Streptomyces ambofaciens ». *Journal of bacteriology* 188(18): 6599–6610.

Cordero, Otto X., et Martin F. Polz. 2014. « Explaining Microbial Genomic Diversity in Light of Evolutionary Ecology ». *Nature Reviews Microbiology* 12(4): 263-73.

Didelot, Xavier, et Martin C. J. Maiden. 2010. « Impact of recombination on bacterial evolution ». *Trends in Microbiology* 18(7): 315-22.

Dupuy, Pierre, Laurent Sauviac, et Claude Bruand. 2018. « Stress-Inducible NHEJ in Bacteria: Function in DNA Repair and Acquisition of Heterologous DNA ». *Nucleic Acids Research* 47(3): 1335-49.

Edgar, Robert C. 2004. « MUSCLE: multiple sequence alignment with high accuracy and high throughput ». *Nucleic acids research* 32(5): 1792–1797.

Fang, Gang, Nitin Bhardwaj, Rebecca Robilotto, et Mark B. Gerstein. 2010. « Getting Started in Gene Orthology and Functional Analysis ». *PLoS Computational Biology* 6(3). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2845645/ (11 décembre 2019).

Fischer, G., B. Decaris, et P. Leblond. 1997. « Occurrence of Deletions, Associated with Genetic Instability in Streptomyces Ambofaciens, Is Independent of the Linearity of the Chromosomal DNA. » *Journal of Bacteriology* 179(14): 4553-58.

Gitai, Zemer, Martin Thanbichler, et Lucy Shapiro. 2005. « The Choreographed Dynamics of Bacterial Chromosomes ». *Trends in Microbiology* 13(5): 221-28.

Goris, Johan et al. 2007. « DNA–DNA hybridization values and their relationship to whole-genome sequence similarities ». *International journal of systematic and evolutionary microbiology* 57(1): 81–91.

Hoff, Grégory et al. 2016. « Multiple and variable NHEJ-like genes are involved in resistance to DNA damage in Streptomyces ambofaciens ». *Frontiers in microbiology* 7.

Hoff, Grégory et al.. 2018. « Genome plasticity is governed by double strand break DNA repair in Streptomyces ». *Scientific Reports* 8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5869714/ (27 novembre 2018).

Huson, Daniel H. et al. 2007. « Dendroscope: An interactive viewer for large phylogenetic trees ». *BMC bioinformatics* 8(1): 460.

Inoue, Shoichi et al. 2003. « Chromosomal Circularization in Streptomyces griseus by Nonhomologous Recombination of Deletion Ends ». *Bioscience, Biotechnology, and Biochemistry* 67(5): 1101-8.

Kim, Ji-Nu et al. 2015. « Comparative Genomics Reveals the Core and Accessory Genomes of Streptomyces Species ». *Journal of Microbiology and Biotechnology* 25(10): 1599-1605.

Kirby, Ralph. 2011. « Chromosome diversity and similarity within the Actinomycetales ». *FEMS microbiology letters* 319(1): 1–10.
Kováč, Jakub, Robert Warren, Marília D.V. Braga, et Jens Stoye. 2011. « Restricted DCJ Model: Rearrangement Problems with Chromosome Reincorporation ». *Journal of Computational Biology* 18(9): 1231-41.
Kurth, Florence et al. 2014. « Streptomyces-Induced Resistance Against Oak Powdery Mildew Involves Host Plant Responses in Defense, Photosynthesis, and Secondary Metabolism Pathways ». *Molecular Plant-Microbe Interactions®* 27(9): 891-900.
Leblond, Pierre, et Bernard Decaris. 1994. « New Insights into the Genetic Instability of Streptomyces ». *FEMS Microbiology Letters* 123(3): 225-32.
Lesterlin, Christian et al. 2005. « Roles for Replichores and Macrodomains in Segregation of the Escherichia Coli Chromosome ». *EMBO reports* 6(6): 557-62.
McDonald, Bradon R., et Cameron R. Currie. 2017. « Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces ». *mBio* 8(3): e00644–17.
Nakamura, Yoji, Takeshi Itoh, Hideo Matsuda, et Takashi Gojobori. 2004. « Biased Biological Functions of Horizontally Transferred Genes in Prokaryotic Genomes ». *Nature Genetics* 36(7): 760.
Niki, Hironori, Yoshiharu Yamaichi, et Sota Hiraga. 2000. « Dynamic Organization of Chromosomal DNA in Escherichia Coli ». *Genes & Development* 14(2): 212-23.
O'Leary, Nuala A. et al. 2016. « Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation ». *Nucleic Acids Research* 44(D1): D733-45.
Pandza, Suada et al. 1998. « Recombination between the Linear Plasmid PPZG101 and the Linear Chromosome of Streptomyces Rimosus Can Lead to Exchange of Ends ». *Molecular Microbiology* 28(6): 1165-76.
Pombo, Ana, et Niall Dillon. 2015. « Three-Dimensional Genome Architecture: Players and Mechanisms ». *Nature Reviews Molecular Cell Biology* 16(4): 245-57.
Richter, Michael, et Ramon Rosselló-Móra. 2009. « Shifting the genomic gold standard for the prokaryotic species definition ». *Proceedings of the National Academy of Sciences* 106(45): 19126–19131.
Rocha, Eduardo P. C. 2006. « Inference and Analysis of the Relative Stability of Bacterial Chromosomes ». *Molecular Biology and Evolution* 23(3): 513-22.
Rocha, Eduardo P. C., et Antoine Danchin. 2003. « Essentiality, Not Expressiveness, Drives Gene-Strand Bias in Bacteria ». *Nature Genetics* 34(4): 377-78.
Roth, John R. 2010. « Genetic Adaptation: A New Piece for a Very Old Puzzle ». *Current Biology* 20(1): R15-17.
Simão, Felipe A. et al. 2015. « BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs ». *Bioinformatics* 31(19): 3210-12.
Stamatakis, Alexandros. 2014. « RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies ». *Bioinformatics* 30(9): 1312-13.
Tatusov, R. L. 1997. « A Genomic Perspective on Protein Families ». *Science* 278(5338): 631-37.
Thibessard, Annabelle, et Pierre Leblond. 2014. « Subtelomere Plasticity in the Bacterium Streptomyces ». In *Subtelomeres*, éd. Edward J Louis et Marion M Becker. Berlin, Heidelberg: Springer, 243-58. https://doi.org/10.1007/978-3-642-41566-1_14 (27 novembre 2019).
Tidjani, Abdoul-Razak et al. 2019. « Massive Gene Flux Drives Genome Diversity between Sympatric Streptomyces Conspecifics ». *mBio* 10(5): e01533-19.
Townsend, Jeffrey P., Kaare M. Nielsen, Daniel S. Fisher, et Daniel L. Hartl. 2003. « Horizontal Acquisition of Divergent Chromosomal DNA in Bacteria: Effects of Mutator Phenotypes ». *Genetics* 164(1): 13-21.
Valens, Michèle et al. 2004. « Macrodomain Organization of the Escherichia Coli Chromosome ». *The EMBO Journal* 23(21): 4330-41.
Vicente, Cláudia et al. 2018. « Comparative Genomics among Closely Related Streptomyces Strains Revealed Specialized Metabolite Biosynthetic Gene Cluster Diversity ». *Antibiotics* 7(4): 86.
Vivian, A., et D. A. Hopwood. 1973. « Genetic Control of Fertility in Streptomyces coelicolor A3(2) :

New Kinds of Donor Strains ». *Microbiology*, 76(1): 147-62.
Vos, Michiel, Alexandra B. Wolf, Sarah J. Jennings, et George A. Kowalchuk. 2013. « Micro-Scale Determinants of Bacterial Diversity in Soil ». *FEMS Microbiology Reviews* 37(6): 936-54.
Yancopoulos, Sophia, Oliver Attie, et Richard Friedberg. 2005. « Efficient sorting of genomic permutations by translocation, inversion and block interchange ». *Bioinformatics* 21(16): 3340-46.
Zhou, Zhan, Jianying Gu, Yong-Quan Li, et Yufeng Wang. 2012. « Genome Plasticity and Systems Evolution in Streptomyces ». *BMC Bioinformatics* 13(S10): S8.

II- Évolution du génome des *Streptomyces* isolés à partir d'un micro-habitat

Cette partie utilise les données collectées par Maxime Toussaint et Abdoul Razak Tidjani, doctorants dans l'unité DynAMic (Université de Lorraine). Ces ressources ont permis d'aborder les mécanismes évolutifs au sein d'une population de *Streptomyces* sympatriques, c'est-à-dire vivant dans le même micro-habitat, à l'échelle micro- et centimétrique. Ces organismes sont susceptibles d'interagir autant d'un point de vue physiologique que génétique.

A- Obtention d'une collection sympatrique de *Streptomyces*

L'équipe DynAMic a prélevé 9 micro-agrégats (de l'ordre du mm^3) de sol forestier (sol nu et rhizosphère de hêtre) dans la forêt de Montiers-sur-Saulx. De ces prélèvements, 129 souches de *Streptomyces* ont été isolées et 46 ont été sélectionnées sur leur capacité à sporuler efficacement en laboratoire.

Une première caractérisation taxonomique a été effectuée par amplification PCR, séquençage du gène codant l'ARNr 16S et calcul du pourcentage d'identité entre ces séquences nucléotidiques (993 positions considérées). Cinq clades distincts ont été distingués ; les souches présentent 100% d'identité au sein d'un même groupe (**tableau 5**).

	Groupe I	Groupe II	Groupe III	Groupe IV	Groupe V
Groupe I		97.6 %	97.1 %	97.4 %	97.3 %
Groupe II			97.8 %	98.7 %	99.1 %
Groupe III				98.4 %	98.7%
Groupe IV					99.5 %
Groupe V					
Effectif	3	7	2	2	32

Tableau 5 - Matrice d'identité de la séquence nucléotidique codant pour l'ARNr 16S entre les 5 groupes et effectif de chaque groupe.

Dans un second temps, pour estimer le degré de clonalité des souches et en particulier au sein du groupe 5 (qui regroupe 32 des 46 souches isolées), les profils alléliques et la phylogénie basée sur une analyse MLST (MultiLocus Sequence Typing) ont été effectués à partir de 5 gènes de ménage: *recA*, *trpB*, *gyrB*, *rpoB* et *atpD*. La phylogénie réalisée avec la concaténation de ces séquences (**figure 28**) a

confirmé les regroupements observés en ARNr 16S et forment des regroupements monophylétiques à l'exception du groupe II.

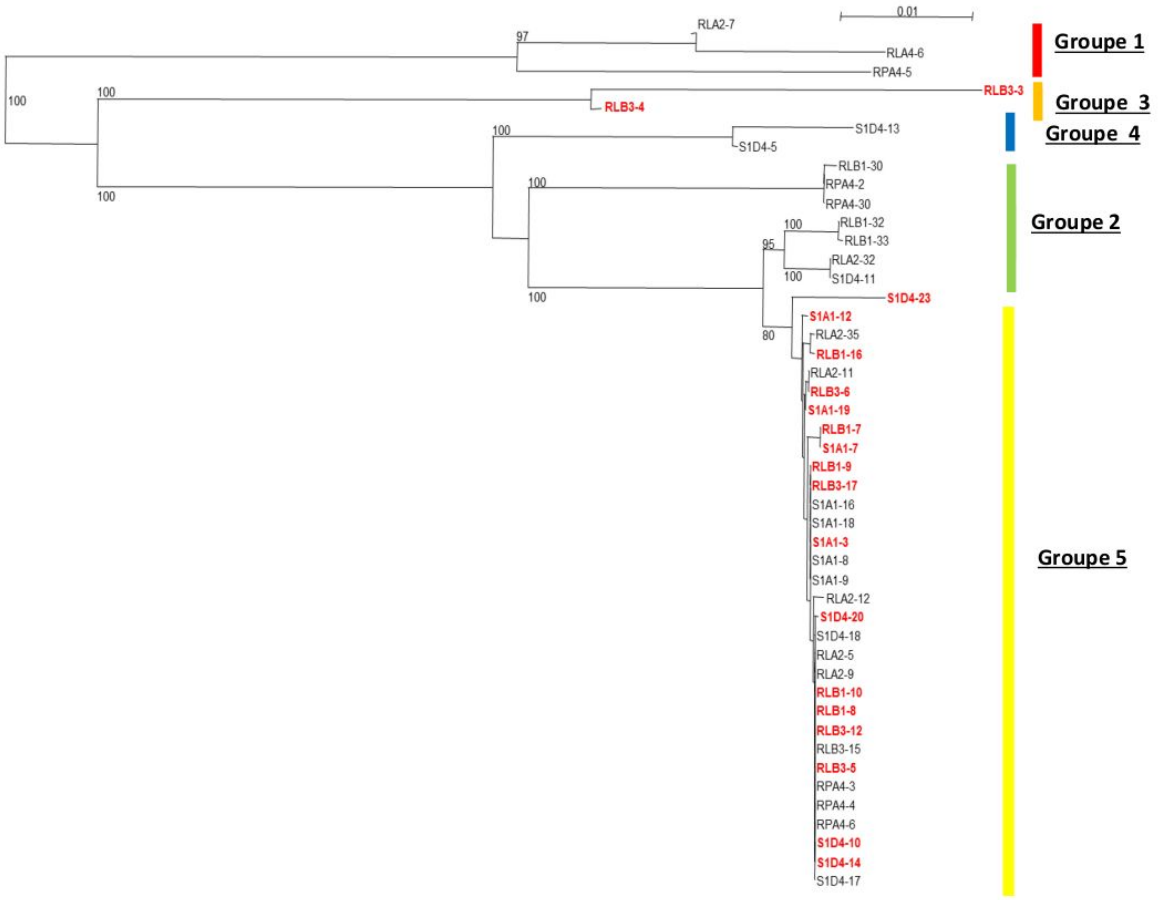

Figure 28 - Analyse MLST pour les 46 souches de la collection extraite de (Toussaint, 2018). L'arbre a été construit selon la méthode de distance Neighbor-Joining et corrigé avec la méthode Kimura 2-parameter. Les séquences nucléotidiques partielles des gènes *recA* (552 pb), *rpoB* (727 pb), *trpB* (473 pb), *atpD* (570 pb) et (*gyrB* 865 pb) ont été concaténées (3184 positions) et utilisées pour générer l'arbre. La robustesse des branches a été statistiquement éprouvée par 100 répliques de bootstrap. Seules les valeurs de bootstrap supérieures à 70 % sont indiquées. Toutes les positions contenant des gaps ont été éliminées. L'échelle représente le nombre de substitution par nucléotide. Les souches en rouge correspondent aux souches sélectionnées pour la campagne de séquençage selon la technologie paired-end.

Les groupes ont donc été renommés de manière à ne former que des groupes monophylétiques : groupe I = 1, groupe II = 4 + 5, groupe III = 2, groupe IV = 3 et groupe V = 6. Les profils alléliques identifiés dans le groupe 6 partagent 99,88% d'identité. Cela montre que les 32 souches du groupe 6 appartiennent à la même espèce, mais présentent cependant une certaine diversité, malgré un temps évolutif court écoulé depuis leur dernier ancêtre commun. C'est à partir de ces observations qu'une approche de génomique a été entreprise afin de comprendre sur quelles bases génétiques reposent

cette diversité intra-espèce. Pour cela, 18 souches (en rouge dans la **figure 28**) ont été sélectionnées et leur ADN génomique séquencé par la technologie Illumina.

1- Première campagne de séquençage génomique

a- Séquençage selon la technologie Illumina paired-end

Dix-neuf souches ont été séquencées (**tableau 6**) selon la technologie paired-end (voir MATERIELS ET METHODES). En complément, la souche RLB1-9 a aussi été séquencée selon la technologie Nanopore (MinIon) et assemblée en 3 contigs correspondant au chromosome de la souche ainsi que 2 réplicons extrachromosomiques (plasmide linéaire et bactériophage présumé).

	RLB1-10	RLB1-16	RLB1-7_S5	RLB1-9	RLB3-12	RLB3-17	RLB3-3	RLB3-4	RLB3-6
# reads	190,582	200,955	118,611	307,195	214,387	202,612	355,384	158,652	228,116
Taille moyenne	248	246	248	243	247	248	245	248	247

	S1A1-12	S1A1-19	S1A1-3	S1A1-7	S1D4-10	S1D4-14	S1D4-20	S1D4-23
# reads	240,990	115,264	75,795	65,944	178,963	42,509	256,939	23,149
Taille moyenne	247	244	248	244	247	248	246	247

Tableau 6 - Tableau récapitulatif des données de séquençage obtenues sur 19 souches de *Streptomyces*

b- Mise en évidence des variations génomiques

Les lectures (reads) de séquençage des isolats ont été alignées à l'outil Bowtie2 sur la référence que constitue le génome de la souche RLB1-9 (Langmead and Salzberg, 2012) (voir MATERIEL ET METHODES). Une région dépourvue de lectures correspond à une région spécifique de la souche RLB1-9. Seules les régions de plus de 10 kb sans lectures ont été visualisées *via* un script R *ad hoc* présenté **figure 29**. Considérer des régions de 10 kb permet de s'assurer que ces régions spécifiques ne correspondent pas à des régions mal séquencées en particulier avec les génomes présentant une faible profondeur de séquençage.

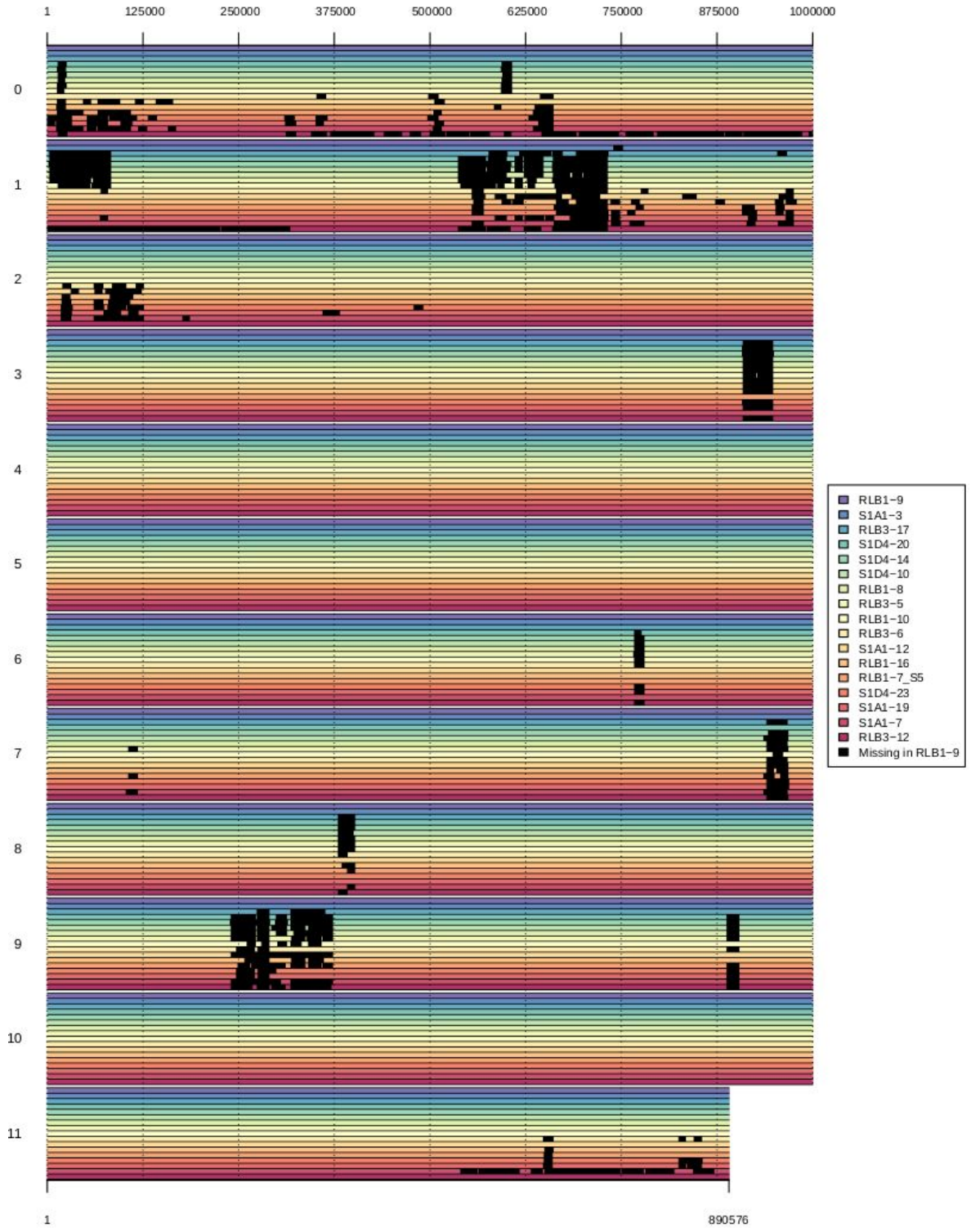

Figure 29 - Alignements des données paired-end contre le chromosome de RLB1-9. Les valeurs en ordonnée représentent les positions du génome en mégabases, celles en abscisse en kilobases. Chaque souche est représentée par un code et une couleur (voir légende) et les blocs noirs correspondent aux régions spécifiques de RLB1-9 d'au moins 10 kb. Les souches sont organisées selon leur proximité phylogénétique avec RLB1-9, de la plus proche à la plus éloignée.

Cette approche a révélé que les souches bien qu'appartenant à la même espèce possèdent des quantités d'ADN spécifique importantes. Ainsi, RLB1-9 possède des régions spécifiques allant de 13 kb avec la souche S1A1-3 jusqu'à environ 1.3 Mb avec la souche RLB3-12.

2- Deuxième campagne de séquençage : séquençage et assemblage de 11 souches conspécifiques

Une deuxième campagne de séquençage a été menée sur ce projet, en particulier pour obtenir un niveau d'assemblage comparable à la souche RLB1-9. Pour cela 11 souches de *Streptomyces* provenant de la même niche écologique ont été séquencés et assemblés.

Publication 2 : **Genome sequence of 11 *Streptomyces* sp. conspecific strains.** *Abdoul-Razak Tidjani, Jean-Noël Lorenzi, Maxime Toussaint, Erwin van Dijk, Delphine Naquin, Olivier Lespinet, Cyril Bontemps, Pierre Leblond.*

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Genome Sequences of 11 Conspecific *Streptomyces* sp. Strains

Abdoul-Razak Tidjani,[a] Jean-Noël Lorenzi,[a,b] Maxime Toussaint,[a] Erwin van Dijk,[b] Delphine Naquin,[b] Olivier Lespinet,[b] Cyril Bontemps,[a] Pierre Leblond[a]

[a]Université de Lorraine, INRA, DynAMic, Nancy, France
[b]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University Paris-Sud, University Paris-Saclay, Gif-sur-Yvette, France

ABSTRACT The genomes of 11 conspecific *Streptomyces* strains, i.e., from the same species and inhabiting the same ecological niche, were sequenced and assembled. This data set offers an ideal framework to assess the genome evolution of *Streptomyces* species in their ecological context.

Streptomyces species are soil-dwelling bacteria that harbor large linear chromosomes (1). We report here the genome sequences of 11 sympatric *Streptomyces* strains belonging to the same species. In order to select conspecifics, we sampled soil grains at the cubic milliliter scale from a French forest (maximal distance of 8 cm from each other). After dissolution in sterile water and spreading of serial dilutions on *Streptomyces* isolation medium (SIM) (2), the 16S rRNA sequences of the strains were determined and analyzed using BLAST (3), and their phylogenetic relationships were characterized by multilocus sequence analysis (MLSA) using the Molecular Evolutionary Genetics Analysis version 7 (MEGA7) software (4). After growth in liquid Hickey-Tresner medium at 30°C for 30 h, DNA purification was performed using the salting-out method (5), followed by chloroform extraction. The targeted genes (rRNA gene and the MLSA genes) were amplified using universal (16S rRNA gene [6]) and specific (MLSA [7]) primers. The 11 strains selected showed identical 16S rRNA gene sequences and minimal MLSA divergence. These strains are related to *Streptomyces olivochromogenes* (strain DSM40451), with an average identity of 99.93% for the 16S rRNA gene sequences (8). A hybrid assembly using Oxford Nanopore technology for scaffolding and Illumina technology for sequence improvement was performed (Table 1). Base calling of these sequences was performed using the Oxford Nanopore base callers Albacore (v0.8.4 or v2.0.2) or Guppy (v0.3.0). Nanopore reads (minimum quality mean, 7) were generated on minION or gridION systems. When strains were multiplexed, Porechop (v0.2.4, using default settings) was used for demultiplexing (and adaptor trimming). The coverage ranged from 41× to 344×. The Illumina paired-end libraries were created using the Illumina Nextera kit, except for RLB1-8 and RLB1-9, for which sonication (Covaris) and adaptor ligation (Illumina TruSeq) were used instead. Paired reads were generated using a MiSeq reagent kit v3 (150 cycles) and the Genome Analyzer system (Illumina). The minimum read size was set to 10 bp, and adaptor trimming was performed using Cutadapt (v1.15, using default settings). The coverages of the paired-end reads (length, 75 to 300 bp) ranged from 58× to 320×. The hybrid assembly was performed using Unicycler (9) v0.4.2 or v0.4.3 (using default settings) to assemble 1 to 19 large contigs covering the whole genome of each strain, enabling the acquisition of each linear chromosome in one scaffold and the identification of extrachromosomal elements when present. One or two extrachromosomal linear or circular replicons were identified in 5 of the 11 strains by *in silico* prediction or pulsed-field gel electrophoresis experiments (10). The total genome sizes ranged from 11.76 to 12.45 Mb, positioning these strains among the largest bacterial genomes (Table 1).

Citation Tidjani A-R, Lorenzi J-N, Toussaint M, van Dijk E, Naquin D, Lespinet O, Bontemps C, Leblond P. 2019. Genome sequences of 11 conspecific *Streptomyces* sp. strains. Microbiol Resour Announc 8:e00863-19. https://doi.org/10.1128/MRA.00863-19.

Editor Christina A. Cuomo, Broad Institute

Address correspondence to Pierre Leblond, pierre.leblond@univ-lorraine.fr.

Received 24 July 2019
Accepted 28 August 2019
Published 19 September 2019

TABLE 1 Genome features, sequencing statistics, and accession numbers of the 11 conspecific *Streptomyces* strains

Strain or replicon	Illumina sequencing information		Oxford Nanopore sequencing information						Replicon size (bp)[c]	Genome size (bp)	Total no. of CDS[a]	G+C content (%)	TIR[b] (kb)	GenBank accession no.
	No. of reads (approx coverage [×])	SRA accession no.	No. of reads (approx coverage [×])	N_{50} of raw reads (kb)	Flow cell type(s)	Sequencing kit(s)	Base caller	SRA accession no.						
RLB1-8	15,381,622 (320)	SRR9661592	655,482 (150)	4.3	FAH18893 (9.5), FAH18988 (9.5), FAH24488 (9.5)	sqk-lsk308	albacore_2.0.2	SRR9710048	11,765,340	11,765,340	10,635	70.2	357	CP041650
RLB1-9	18,329,970 (115)	SRR9661591	88,694 (41)	7.6	FAF19789 (9.4)	sqk-lsk308	albacore_0.8.4	SRR9710047	11,940,408	12,200,709	10,838	70.2	311	CP041654
pRLB1-9.1									154,158[C]		175	69,0		CP041653
pRLB1-9.2									106,143[L]		111	68.7	24	CP041652
RLB3-5	3,196,108 (67)	SRR9661590	144,521 (56)	7.1	FAH24352 (9.5), FAH29240 (9.4)	sqk-lsk308, sqk-lsk108	albacore_2.0.2	SRR9710050	11,898,970	11,898,970	10,731	70.2	365	CP041651
RLB3-6	3,274,272 (68)	SRR9661589	299,155 (52)	4.6	FAF19789 (9.4)	sqk-lsk308	guppy_0.3.0	SRR9710049	12,338,263	12,448,281	11,255	70.1	587	CP041602
pRLB3-6.1									110,314[C]		101	70.6		CP041601
RLB3-17	3,976,622 (83)	SRR9661596	202,455 (50)	5.6	FAF19789 (9.4)	sqk-lsk308	guppy_0.3.0	SRR9710052	12,023,175	12,023,175	10,934	70.2	451	CP041610
S1A1-3	3,243,184 (68)	SRR9661595	198,567 51)	5.2	FAF19789 (9.4)	sqk-lsk308	guppy_0.3.0	SRR9710051	12,042,091	12,042,091	10,920	70.2	393	CP041611
S1A1-7	3,504,210 (73)	SRR9661594	533,299 (73)	2.8	FAF19789 (9.4)	sqk-lsk308	guppy_0.3.0	SRR9710054	11,713,151	12,005,504	10,580	70.3	513	CP041604
pS1A1-7.1									292,353[C]		252	69.7		CP041603
S1A1-8	3,191,318 (66)	SRR9661593	780,962 (53)	1.1	FAF19789 (9.4)	sqk-lsk308	guppy_0.3.0	SRR9710053	12,036,971	12,036,971	10,918	70.2	394	CP041612
S1D4-14	2,794,454 (58)	SRR9661598	2,066,754 (344)	3.2	FAF19789 (9.4)	sqk-lsk308	guppy_0.3.0	SRR9710056	11,723,487	11,934,498	10,591	70.2	369	CP041607
pS1D4-14.1									112,196[L]		118	68.7	0	CP041605
pS1D4-14.2									98,815[C]		138	69.1		CP041606
S1D4-20	3,327,172 (69)	SRR9661597	1,424,402 (255)	4.1	FAF19789 (9.4)	sqk-lsk308	guppy_0.3.0	SRR9710055	11,851,257	12,245,276	10,742	70.2	373	CP041609
pS1D4-20.1									394,019[L]		329	69.1	68	CP041608
S1D4-23	3,543,760 (74)	SRR9661599	400,348 (61)	3.5	FAF19789 (9.4)	sqk-lsk308	guppy_0.3.0	SRR9710057	12,057,712	12,057,712	10,971	70.2	421	CP041613

[a] As determined through automatic annotation by the NCBI Prokaryotic Genome Annotation Pipeline. CDS, coding sequences.

[b] TIR, terminal inverted repeat.

[c] L/C, linear (L) or circular (C) replicon configuration, as predicted by the assembler and tested by pulsed-field gel electrophoresis (not shown).

Data availability. Genome sequences and raw sequence reads are available from GenBank and the NCBI Sequence Read Archive under the accession numbers shown in Table 1.

ACKNOWLEDGMENTS

This work was funded by the French National Research Agency (grants ANR LABEX ARBRE and ANR-11-LABX-0002-01), by the French National Institute for Agricultural Research (INRA), and by Région Lorraine (now called Région Grand Est).

REFERENCES

1. Thibessard A, Leblond P. 2014. Subtelomere plasticity in the bacterium *Streptomyces*, p 243–258. *In* Louis EJ, Becker MM (ed), Subtelomeres. Springer, Berlin, Germany.
2. D'Costa VM, McGrann KM, Hughes DW, Wright GD. 2006. Sampling the antibiotic resistome. Science 311:374–377. https://doi.org/10.1126/science.1120800.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
4. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874. https://doi.org/10.1093/molbev/msw054.
5. Kieser T, Bibb M, Buttner M, Chater K, Hopwood D. 2000. Practical *Streptomyces* genetics. John Innes Foundation, Norwich, United Kingdom.
6. Rintala H, Nevalainen A, Rönkä E, Suutari M. 2001. PCR primers targeting the 16S rRNA gene for the specific detection of streptomycetes. Mol Cell Probes 15:337–347. https://doi.org/10.1006/mcpr.2001.0379.
7. Guo Y, Zheng W, Rong X, Huang Y. 2008. A multilocus phylogeny of the *Streptomyces griseus* 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. Int J Syst Evol Microbiol 58:149–159. https://doi.org/10.1099/ijs.0.65224-0.
8. Tidjani AR, Lorenzi JN, Toussaint M, van Dijk E, Naquin D, Lespinet O, Bontemps C, Leblond P. 2019. Massive gene flux drives genome diversity between sympatric *Streptomyces* conspecifics. mBio 10:e01533-19. https://doi.org/10.1128/mBio.01533-19.
9. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
10. Leblond P, Francou FX, Simonet JM, Decaris B. 1990. Pulsed-field gel electrophoresis analysis of the genome of *Streptomyces ambofaciens* strains. FEMS Microbiol Lett 60:79–88. https://doi.org/10.1111/j.1574-6968.1990.tb03866.x.

B- Diversité génétiques des souches conspécifiques

À partir des génomes précédemment séquencés, une analyse de la diversité génétique de ces souches de *Streptomyces* conspécifiques a été effectuée. Elle a permis d'explorer l'impact du transfert horizontal dans la diversification génétique d'une population naturelle de *Streptomyces* tout en observant une dynamique dans l'acquisition de cette diversité, et ce, même à micro-échelle, aussi bien temporelle que spatiale.

Ces travaux ont été valorisés par une publication:

Publication 3 : **Massive gene flux drives genome diversity between sympatric *Streptomyces* conspecifics.** *Abdoul Razak Tidjani, Jean-Noël Lorenzi, Maxime Toussaint, Erwin van Dijk, Delphine Naquin, Olivier Lespinet, Cyril Bontemps, and Pierre Leblond.*

À partir des 11 génomes précédemment séquencés, l'appartenance à une même espèce de *Streptomyces* a été vérifiée grâce au critère ANI qui varie entre 98.68% et 99.99% selon les paires de génomes considérés. Le pan-génome de ces 11 souches consiste en 13,814 gènes, les 2 souches les plus proches diffèrent seulement de 12 gènes, les plus distantes de 1,393 gènes. Au sein de ce pan-génome, environ un tiers des gènes (5 036 gènes) ne sont retrouvés que dans certaines souhes, constituant le génome accessoire. Le core-génome est lui composé de 8,778 gènes.

Les comparaisons par paire des 11 génomes ont permis d'identifier des perturbations dans la synténie chromosomique avec des événements d'insertion ou de délétion (indels). Le nombre de rupture dans la synténie entre deux souches varie de 1 à 124 et augmente de façon linéaire avec la distance phylogénétique, ce qui indique qu'elles s'accumulent au fil du temps. De plus, la distribution des indels est hétérogène le long du chromosome ; une densité accrue des indels dans les régions terminales a été mise en évidence. Cette accumulation d'indels dans les régions terminales s'accentue à mesure que la distance phylogénétique entre les souches comparées augmente.

Au sein de cette population de 11 souches de *Streptomyces*, 51 regroupements biosynthétiques de gènes (BGC pour Biosynthetic Gene clusters) uniques dans la population ont été identifiés, chaque souche possédant environ 35 regroupements. Par ailleurs, quelques individus seulement de la population produisent un anti-microbien contre une souche de *Bacillus* (isolé depuis le même échantillon de sol). De plus, comme aucune des souches n'inhibe ses congénères, nous pouvons supposer que cet ensemble de *Streptomyces* forme une population coopérative, où une souche produisant l'activité antimicrobienne protège ses congénères non producteurs.

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Ecological and Evolutionary Science

Massive Gene Flux Drives Genome Diversity between Sympatric *Streptomyces* Conspecifics

Abdoul-Razak Tidjani,[a] Jean-Noël Lorenzi,[a,b] Maxime Toussaint,[a] Erwin van Dijk,[b] Delphine Naquin,[b] Olivier Lespinet,[b] Cyril Bontemps,[a] Pierre Leblond[a]

[a]Université de Lorraine, INRA, DynAMic, Nancy, France
[b]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University Paris-Sud, University Paris-Saclay, Gif-sur-Yvette, France

ABSTRACT In this work, by comparing genomes of closely related individuals of *Streptomyces* isolated at a spatial microscale (millimeters or centimeters), we investigated the extent and impact of horizontal gene transfer in the diversification of a natural *Streptomyces* population. We show that despite these conspecific strains sharing a recent common ancestor, all harbored significantly different gene contents, implying massive and rapid gene flux. The accessory genome of the strains was distributed across insertion/deletion events (indels) ranging from one to several hundreds of genes. Indels were preferentially located in the arms of the linear chromosomes (ca. 12 Mb) and appeared to form recombination hot spots. Some of them harbored biosynthetic gene clusters (BGCs) whose products confer an inhibitory capacity and may constitute public goods that can favor the cohesiveness of the bacterial population. Moreover, a significant proportion of these variable genes were either plasmid borne or harbored signatures of actinomycete integrative and conjugative elements (AICEs). We propose that conjugation is the main driver for the indel flux and diversity in *Streptomyces* populations.

IMPORTANCE Horizontal gene transfer is a rapid and efficient way to diversify bacterial gene pools. Currently, little is known about this gene flux within natural soil populations. Using comparative genomics of *Streptomyces* strains belonging to the same species and isolated at microscale, we reveal frequent transfer of a significant fraction of the pangenome. We show that it occurs at a time scale enabling the population to diversify and to cope with its changing environment, notably, through the production of public goods.

KEYWORDS *Streptomyces*, conjugation, gene transfer, plasticity, population

Streptomyces organisms are prominent soil-dwelling bacteria found in all terrestrial ecosystems, typified by their complex differentiation cycle and their central roles in biogeochemical cycles and the homeostasis of the soil. They are known as the most prolific bacterial genus for the production of specialized metabolites (e.g., antibiotics and antifungals) and enzymes of high biotechnological and medical interest (1). In soil, these compounds are involved in interactions with the surrounding microbial communities and contribute to plant health and growth (2). Each species or strain was reported to produce one or only a few compounds, but access to the whole-genome sequences revealed a much more important biosynthetic gene reservoir (3, 4).

The genome of *Streptomyces* is remarkably large (6 to 12 Mb) and possesses one of the few linear bacterial chromosomes ever described. The genome is highly compartmentalized, with genes conserved across the species distributed at the center of the chromosome and variable genes in the terminal parts (5, 6). The biosynthetic gene clusters (BGCs) are consistently enriched in the variable part of the chromosome. The proportion of the flexible genome is impressive and is reflected at the genus level by

Citation Tidjani A-R, Lorenzi J-N, Toussaint M, van Dijk E, Naquin D, Lespinet O, Bontemps C, Leblond P. 2019. Massive gene flux drives genome diversity between sympatric *Streptomyces* conspecifics. mBio 10:e01533-19. https://doi.org/10.1128/mBio.01533-19.

Editor Julian E. Davies, University of British Columbia

Address correspondence to Cyril Bontemps, cyril.bontemps@univ-lorraine.fr, or Pierre Leblond, pierre.leblond@univ-lorraine.fr.

Received 13 June 2019
Accepted 25 July 2019
Published 3 September 2019

a wide-open pangenome, i.e., the entire gene set of all species of the genus (7–9). The *Streptomyces* genomes can be completed by large linear and circular plasmids. Linear chromosomes and plasmids share the same invertron structure that consists of long terminal inverted repeats (TIRs) ending with telomeric sequences (10–12). In addition, the *Streptomyces* genomes are recombinogenic, as revealed either by the high frequency of spontaneous rearrangements of large DNA fragments (13) or by multilocus sequence analysis-based studies at the intra- and interspecific levels showing that recombination rates exceed those observed within many bacterial species (14).

Gene flux, i.e., the gain and loss of genetic material, crucially depends first on the access to exogenous information that can be achieved thanks to horizontal gene transfer (HGT) and second on the recombination capacities of the recipient genome that promotes the integration of genetic material in the genome (15). Regarding HGT, *Streptomyces* exhibits little talent for natural competence, and transducing phages have remained elusive (16, 17) despite extensive research. Thus, conjugation remains the only gene exchange mechanism described in *Streptomyces*, and the genomes of these bacteria are known to be rich in conjugative elements (integrated in the chromosome or of plasmid origin). Two types of conjugative elements coexist: integrated and conjugative element (ICE) depending on a type IV secretion system, and actinomycete ICE (AICE) whose transfer depends on a DNA translocase (TraB) (18, 19). The latter is the most widespread in *Streptomyces* and is capable of mobilizing chromosomal DNA in *trans*, i.e., not physically linked to the element (20).

These gene fluxes contribute to increase the flexibility of the genome and explain the diversity observed at the genus level. McDonald and Currie (21) recently showed that the *Streptomyces* genus is ancient (380 million years old) and that acquisition and retention of genes through HGT seem rare among nonclosely related lineages. Yet, nothing is known regarding the intensity and short-term impact of HGT at the population scale. Only some recent examples have shown that *Streptomyces* sharing an identical 16S rRNA gene sequences can exhibit differential phenotypes, including antimicrobial activities (22–24), suggesting rapid genome diversification.

Genome sequencing of closely related strains locally cooccurring has already revealed the existence of a large diversity in terms of gene content within different bacterial species (25–29). However, little is known regarding bacteria living in soil. Only a few examples are documented, and similarly, they seemed to indicate that variability in gene or allelic content can impact the ecology of soil bacteria. For instance, the polymorphism in quorum-sensing genes in *Bacillus subtilis* living in the same cubic centimeter impacted the ability to communicate and led to kin differentiation (30). In the same way, genomic comparison of two closely related groups of *Myxococcus* strains sampled at a centimeter scale enabled the identification of a 150-kb region involved in their sexual isolation (31). It seems thus relevant to assess genome diversification among soil inhabiting congenics, i.e., closely related strains, to unravel the genetic bases underlying the phenotypic diversity that supports niche differentiation and adaptation. The deeper the phylogenetic relationships (infraspecific level) are, the more recent the revealed events will be. Beyond analysis of 16S rRNA gene sequences, multilocus sequence analysis (MLSA), average nucleotide identity (ANI) across the genome, or phylogenomic tree reconstruction enable us to reach close phylogenetic relationships and to infer recent molecular events.

In soil ecosystems, the heterogeneity of the matrix constitutes a powerful driver of microbial growth and dynamics (32, 33). Biotic interactions in soil are most likely to occur between genetically related cells, produced by cell division, living close together on the same soil particle (34). Thus, bacteria in soil do not live as individual cells but are more likely structured in populations composed of closely related strains deriving from a recent common ancestor. The physical and phylogenetic proximities within these populations maximize the potentiality of gene exchanges. These latter reduce genetic isolation, disrupting the confinement of bacteria resulting from their reproduction scheme. However, the true extent of the gene flux that breaks clonality, and the impact of this on the function of natural bacterial populations, remains to be explored.

For that purpose, we undertook the first comparative genomics study of *Streptomyces* conspecifics. We favored the isolation of such closely related strains by sampling *Streptomyces* at the microscale and sequenced and compared their genomes. Next, we identified a high level of divergence in terms of the presence/absence of genes between them and mapped the events along the linear chromosome. This allowed us to identify recombination hot spots, some of which include BGCs. Next, we revealed the high prevalence of actinomycete integrative and conjugative elements (AICEs). We then consider the role of AICEs as motor of genome diversification and how this can later contribute to population functioning.

RESULTS

High levels of genetic diversity within a *Streptomyces* soil population. To uncover microevolutionary processes, we isolated and sequenced the genomes of strains from a *Streptomyces* population inhabiting soil aggregates. For that purpose, four grains of soil, on the order of cubic millimeters in size and distant from each other from 2 cm to a maximum of 8 cm, were sampled from a clod of soil from which we isolated 129 strains. Based on identical 16S rDNA-coding gene sequences and a high degree of identity (>99.8%, nucleotide [nt]) measured with a five-gene multilocus sequence typing (MLST) scheme (see Fig. S1A in the supplemental material), we identified 32 highly related strains at the infraspecific level. The closest species to our group of strains was *Streptomyces olivochromogenes*, with in average 99.93% identity in the 16S rRNA gene and 94.5% ANIblast with the strain DSM40451. The genomes of 11 strains, representative of the different MLSTs identified in the cluster 6 of the population (Fig. S1A), were fully sequenced by a combination of Nanopore (Oxford Nanopore Technologies) and Illumina technologies (see Table S1). For each strain, a single scaffold for each replicon (i.e., chromosome and plasmids) was obtained with high-quality standards enabling full genome comparison. Phylogenomic tree construction (Fig. S1B) and calculation of ANIb percentages (ranging from 98.68% to 99.99%) between all isolates confirmed that they belonged to the same species and diverged recently from a common ancestor. The genomes ranged from 11.75 Mb to 12.44 Mb in size, positioning our strains among the biggest bacterial sequenced genomes (35). The chromosomes are linear with terminal inverted repeats (TIRs) of various sizes (from 311 kb to 587 kb) (see Table S2), as previously reported at the intra- and interspecific levels (12, 36). Following genome annotation (RAST), we determined that the pangenome of our population consisted of 13,814 genes (Fig. 1A). By extrapolating, we found that it reaches a maximum of 23,772 genes by 1,000 genomes. In contrast, for 11 strains randomly chosen in the genus, the pangenome was found to be approximately twice that of the population (23,672 versus 13,814) (Fig. 1B). Regarding the population pangenome, approximately one-third (5,036 genes) was not shared across the population and constituted the accessory genome (Fig. 1A). No two strains shared the same gene content; the closest differed by 12 predicted genes, while the two more distant isolates possessed 1,014 and 1,393 specific genes. Approximately one-quarter (27.6%; 1,392 genes) of the accessory genes were harbored on circular or linear extrachromosomal elements (98.8 kb to 394 kb), which were predicted *in silico* or experimentally observed by pulsed-field gel electrophoresis (PFGE) (see Fig. S2 and Table S2). These elements were found in single strains, and their mobile nature highlights the role of HGT as a driver of specificity, enabling the transfer of numerous genes simultaneously.

Genome diversity is unevenly distributed in numerous indels along the linear chromosome. Pairwise genome comparisons ($n = 65$) among the 11 sequenced genomes enabled us to identify disruptions in chromosomal synteny with insertion or deletion events (indels) of a minimum of one gene. The numbers of breaks differing between two strains ranged from 1 to 124 and increased linearly with the phylogenetic distance, indicating that they accumulate over time (see Fig. S3). When a break was shared by at least two strains, we considered it a single genetic event that may have been vertically inherited. This allowed consideration of 452 independent events that occurred during the population diversification (see Table S3). In extreme cases, the

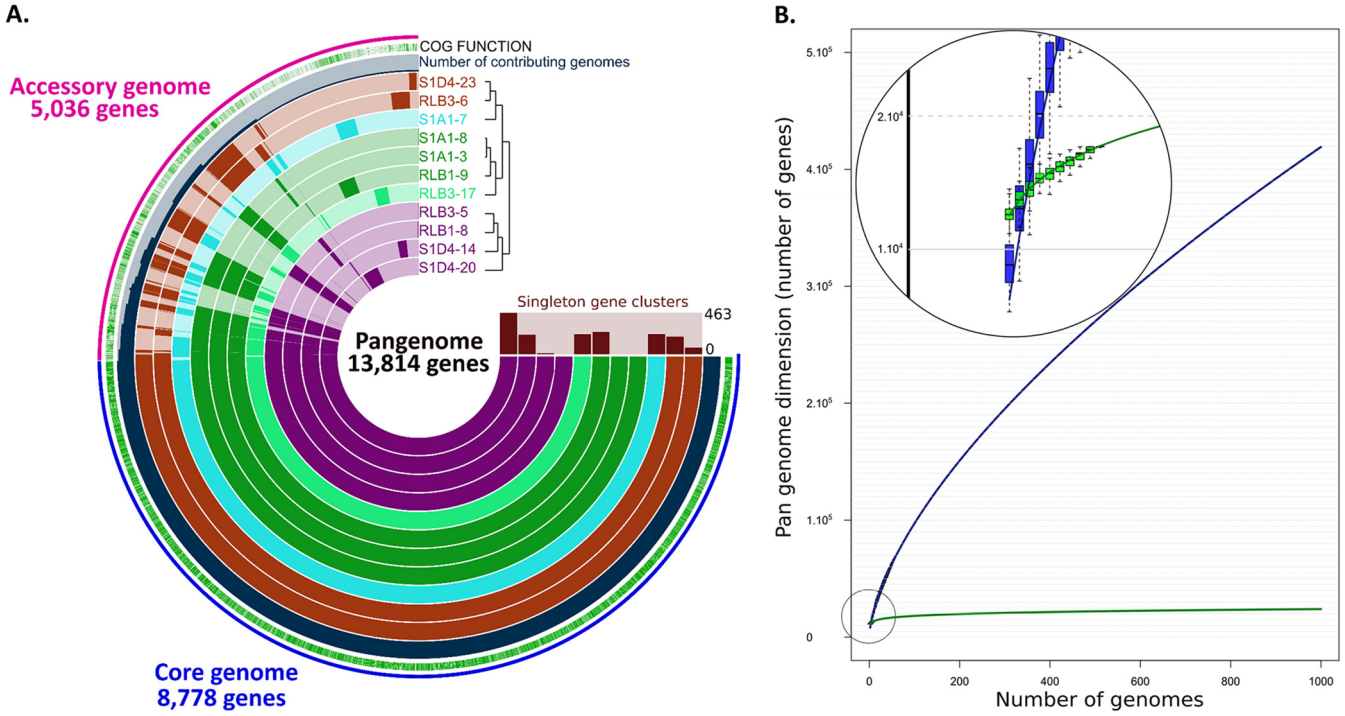

FIG 1 Pangenome analysis of the *Streptomyces* population. (A) Comparative Anvi'o genomic analysis of the 11 conspecific strains isolated in this study. The inner layers represent individual genomes organized regarding their phylogenetic relationships as indicated by the dendrogram. In the layers, dark colors indicate the presence of a gene group and light color its absence. The core (8,778 genes) and the accessory (5,036 genes) genomes are indicated in blue and pink, respectively, in the outmost layer. The blue layer represents the number of genomes among the population contributing to each gene group, and the green layer describes the gene groups in which at least one gene was functionally annotated using cluster of orthologous genes (COGs). (B) Comparison of the *Streptomyces* pangenome evolution at the genus and the population levels. The evolution of the pangenome of 59 complete *Streptomyces* genomes at the genus level is represented in blue. Its extrapolation to 1,000 genomes did not show any shift in the trend of the curve, indicating an open pangenome. The evolution of the population pangenome is represented in green. Its extrapolation rapidly reaches a plateau (see zoomed section).

closest strains differed by only two events, while the most distinct strains differed by 235.

Comparative genomics of the 11 strains shows that indels were scattered along the chromosome but tended to form hot spots (Fig. 2A) with locations where at least three distinct gene assemblies were encountered (Fig. 3 and 4). Remarkably, indel distribution was heterogeneous along the chromosome, with increased density in the terminal regions of the chromosome, and was exaggerated as the phylogenetic distance of strains increased (e.g., RLB1-8 versus RLB3-6). Mapping the core genome of the *Streptomyces* genus (971 genes [J.N. Lorenzi, A. Thibessard, O. Lespinet, P. Leblond, unpublished]) identified the two distal positions, which delineate the core region from chromosomal arms of approximately 2.4 Mb and 1.4 Mb which are devoid of conserved genes on the left and right arms, respectively. The indel density was approximately 5-fold higher in the arms than in the core region (Fig. 2B). Moreover, the compilation of all the pairwise comparisons shows that the density of indels increases from these positions toward the ends of the chromosome, forming a gradient of events. Although the size of the indel in the core region appears larger than in the arms, the total numbers of variable genes follow the same trend, i.e., a strong increase in the arms. The highest variability was noticed at the very ends of the chromosome, as no two strains were identical in terms of the size of their terminal inverted repeats (Table S2). As previously reported, this variability might result from the formation of indels in the TIRs followed by chromosome arm replacement (36–38).

Conjugative elements as the motor of rapid genome diversification. Genome mining for the presence of integrative and conjugative elements in our population allowed the identification of one ICE and 25 AICEs. These 26 elements are likely to be functional (i.e., excisable and self-transmissible), since all the key functions (integrase

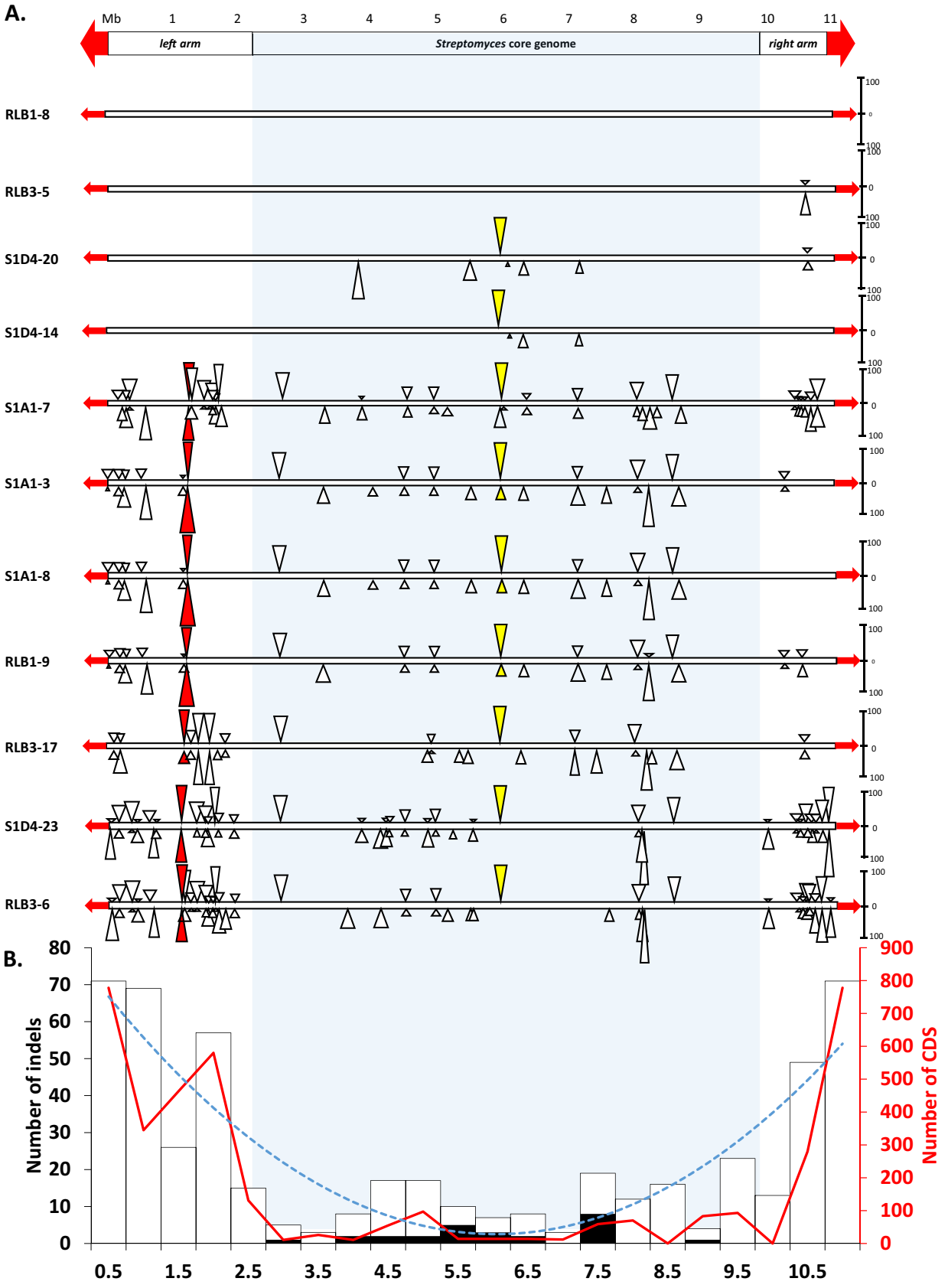

FIG 2 Distribution of insertion and deletion events along the linear chromosome of *Streptomyces*. (A) The scheme at the top represents the *Streptomyces* chromosome with a megabase scale. The position of the core genome of the *Streptomyces* genus is highlighted by a light gray

(Continued on next page)

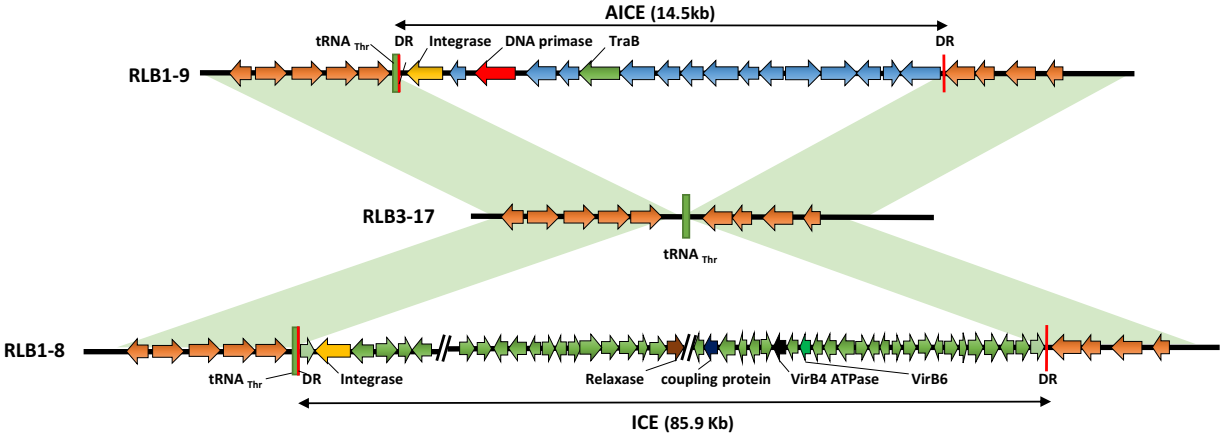

FIG 3 ICE/AICE insertion hot spot. The scheme is illustrating a hot spot for conjugative and integrative element insertion. Two different elements, AICE and ICE, are inserted in the $tRNA_{Thr}$ genes of strains RLB1-9 and RLB1-8, respectively. The $tRNA_{Thr}$ insertion site remains empty in RLB3-17. Direct repeats (DR) flanking the mobile element are represented in red (43 nucleotides for RLB1-9 and 42 nucleotides for RLB1-8). Only the key genes used to identify the elements are shown. Regions in light green correspond to syntenic regions.

and Tra translocase) and *cis* sequences (flanking *att* sites) are present in each of them. Each genome has on average five elements, which make our population particularly rich compared to two elements found on average in a previous study on *Actinobacteria* (18). All the identified elements in our population were present in indels, meaning that they were not part of the core genome, highlighting their remarkable dynamics among the population. They were mainly inserted within the central region of the chromosome (Fig. 2B) and tended to form hots pots with several distinct elements being aggregated in the same target sequence. As illustrated in Fig. 3, the same tRNA target site was occupied by two distinct elements; while an AICE was present in RLB1-9, an ICE was observed in RLB1-8. In the third strain (RLB3-17), the target site was empty, which suggests that at least two of these strains experienced different conjugative events. While conjugative transfer is acknowledged as a powerful driver of diversity in telluric bacteria, their mobility in the soil remains difficult to demonstrate (39, 40). Here, the fact that each strain of the population is unique in terms of AICE/ICE content suggests that the exchanges are intense in the soil.

HGT fosters social cohesiveness of bacterial population. *Streptomyces* organisms are prolific antibiotic producers, and specialized metabolite (SM) production has already been reported to be highly variable among related strains of *Streptomyces* (22–24). We tested our *Streptomyces* isolates in pairwise inhibitory and resistance bioassays and found that none was able to inhibit its conspecifics. On the other hand, some but not all produced antimicrobial activity against a *Bacillus* strain isolated from the same soil, named "killer strains." To correlate the inhibitory phenotype with the presence/absence of biosynthetic gene clusters (BGCs), we performed comparative genomic analyses using antiSMASH (41). We identified 51 unique BGCs within the population, each strain possessing around 35 clusters (not shown). Interestingly, some

FIG 2 Legend (Continued)
frame. The terminal inverted repeats (TIRs) are shown as red arrows. The bottom portion shows illustrations of pairwise genome comparisons (among the 10 possible pairs) within the population using strain RLB1-8 as a reference. Rectangles represent the linear chromosome of each strain. The strains are ordered from top to bottom relative to their phylogenetic distance to the reference. Triangles above the chromosome represent insertions in the reference strain, while triangles below correspond to insertions in the compared strain. For the sake of clarity, only the insertions of at least 10 predicted genes are shown. The height of a triangle reflects the number of genes involved in the insertion (the scale on the right of the chromosome indicates the number of genes). The colored triangles correspond to examples of insertion hot spots (the yellow and red triangles correspond to the hot spots depicted in Fig. 3 and 4, respectively). (B) The scheme represents the distribution of all the indel events identified within the population by genome pairwise comparisons. Each histogram bar corresponds to the number of indels within a 0.5-Mb window. The proportion of ICE/AICE insertions within a window is shown in black. The dotted blue line ("smiley" curve) corresponds to the polynomial trend curve (order 2, R^2 value) of the indel distribution. The red curve shows the number of genes involved in indels within a window.

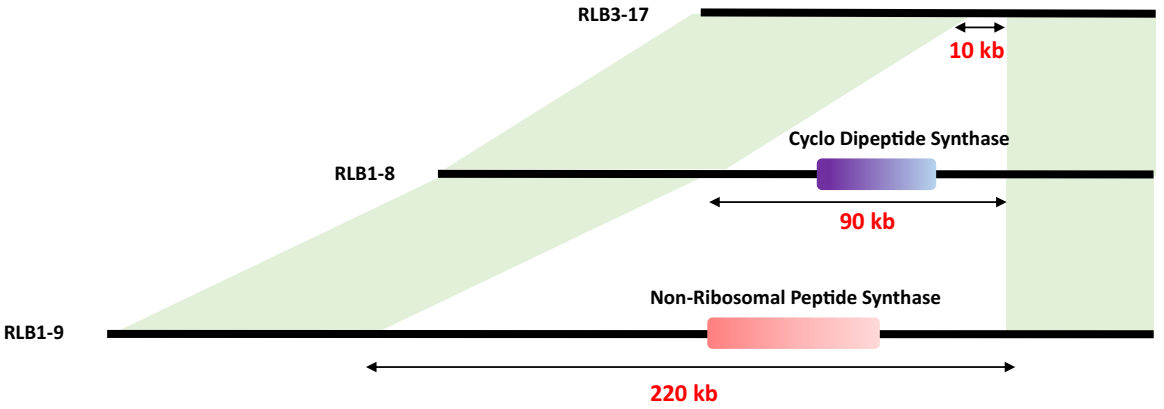

FIG 4 Identification of biosynthetic gene clusters in a variability hot spot. The scheme is illustrating a hot spot of variability observed between three strains (RLB3-17, RLB1-8, and RLB1-9). In two of them, different biosynthetic gene clusters were predicted by antiSMASH: one including a tRNA-dependent cyclodipeptide synthase in RLB1-8 and one including a nonribosomal peptide synthase in RLB1-9. Regions in light green correspond to syntenic regions.

of these BGCs were present in a variability hot spot (Fig. 4), with three strains exhibiting different gene content at these locations. The strain RLB3-17 contained no BGC, while RLB1-8 harbored a cyclodipeptide synthase BGC and RLB1-9 a nonribosomal peptide synthetase (NRPS) BGC. The disruption of a key gene belonging to the NRPS BGC (Fig. 4) only present in the *Streptomyces* killer strains (including RLB1-9) abolished the inhibitory activity against *Bacillus* (see Fig. S4), proving that this BGC is responsible for the antimicrobial activity of the killer strains under these conditions.

DISCUSSION

In this work, we isolated *Streptomyces* conspecifics at a soil microscale, and we revealed by genome comparison the extent of the massive gene flux that occurs in this natural population. This is reflected by a large pangenome, with almost clonal strains harboring strain-specific gene content. The size of the pangenome is influenced by three parameters: (i) the capacity to capture new DNA sequences, (ii) ecological selection of adaptive gene sets, and (iii) the evolutionary time enabling the accumulation of genome divergence (42). The fact that the population pangenome appears reduced in comparison with that of the genus (Fig. 1B) might reflect a constrained flux of genetic information within the microhabitat that is dependent of the diversity of the incoming information mostly arising from sister strains. However, the genetic diversity remained high in our population, implying an intense rate of gain and loss of genome regions over a short evolutionary time.

In *Streptomyces*, we have shown recently that repair of chromosomal double-strand breaks (DSB) experimentally induced in the terminal regions triggers the formation of chromosomal rearrangements (43). The formation of short insertions was also associated with DSB repair depending of nonhomologous end joining (NHEJ). Although the underlying mechanisms are mostly unknown, the plasticity of the subtelomeric regions appears as a hallmark of chromosome linearity in many organisms, from the limited range of bacterial taxa harboring linear chromosomes (44) to eukaryotes (45), including humans (46). In *Streptomyces*, tolerance to DNA rearrangements in the terminal regions could find its origins in the ancestral chromosomal linearization event. It is assumed that the ancestral actinomycete chromosome (i.e., harboring the ancestral core genes) was circular and linearized by recombination with an incoming linear replicon, leading to the addition of chromosomal arms harboring mostly contingency genes (47). However, if the ancestral gain of chromosomal arms can explain the tolerance to DNA rearrangements, it does not shed light on the evolutionary mechanisms that led to the contemporary variability of the *Streptomyces* genome (8, 9, 48). With respect to this, previous work has shown that chromosomal regions separating the core region from chromosomal arms were riddled with short indels (1 to 10 genes) forming a gradient

toward the terminal ends (5). We speculate that the indels, observed at the population level, when cumulated over evolutionary times, gave rise to the genetic compartmentalization observed within the genus (5).

Considering the high prevalence and plasticity of AICEs, we further propose that conjugative mechanisms are responsible for the massive gene flux observed in indels. To mobilize DNA, the translocase TraB binds to specific 8-mer repeats (named *clt* for *cis* acting locus of transfer) present on the AICE and assumed to be also present on the chromosome (*clt* like). Thus, beyond promoting self-transfer of the elements, TraB is able to transfer chromosomal markers in *trans* (20). This property was characterized and exploited in early studies on *Streptomyces* genetics (49). While conjugative transfer of chromosomal DNA was proved to be efficient under laboratory conditions (49, 50), the extent (frequency and size of transferred DNA stretches) remains to be elucidated. The indels ranging from 1 to 241 genes give an indication of the size of the chromosomal DNA stretches that may be transferred through conjugational processes. This is reminiscent of the distributive conjugal transfer in mycobacteria that enables the transfer of chromosomal fragments creating chimeric chromosomes in the recipient cells (51, 52). It is tempting to speculate that several DNA fragments could be transferred at once and generate multiple and diverse recombination events in the recipient *Streptomyces* hyphae.

At the ecological level, genome plasticity increases the functional potential at the strain but also the population level. For instance, it can enable the specific expression of specialized metabolites that can constitute public goods. Bacterial populations (i.e., sympatric closely related strains) can act as social units, meaning that their competitive interactions are more intense toward individuals not belonging to the population rather than between members of the population. Variability of SM production and resistance has already been reported as a driver of ecological cohesive units in *Vibrio* (53) and *Streptomyces* (23). Accordingly, differential SM production and resistance can be used as a proxy to understand the ecological dynamics of the bacterial population (53–55). Here, we confirm that the intense gene flux that *Streptomyces* conspecifics experienced promotes this variability, as previously proposed by Vetsigian et al. (23).

The production of SM within a bacterial population may underpin two principal ecological outcomes, bearing in mind that actual interactions in the soil may be more complex or different than under laboratory conditions. The production of antimicrobial activity and its respective resistance may benefit the sole carrier of the antibiotic production gene and mediate competition with conspecifics as well as with other species. Alternatively, an antibiotic produced by only some individuals of the population where the rest are resistant supports the hypothesis that antibiotics can constitute public goods, benefiting nonproducing but resistant conspecifics. Since none of our isolates was able to inhibit its conspecifics, it supports the hypothesis that they formed a cooperative population. The acquisition of the capacity to produce antimicrobial activity together with the capacity to resist its toxicity can result from a single HGT event thanks to the clustering of biosynthetic and resistance genes in the same genetic region, which is a typical trait of BGCs in actinomycetes. Therefore, the emergence of individuals able to overcome their conspecifics by producing a lethal activity is easily achieved by HGT. However, as we show in this work, our *Streptomyces* population experienced a massive and intense gene flux, which, in addition to favoring the acquisition of BGCs by individuals can also promote a rapid dissemination of all or parts of BGCs to the rest of the population. The replacement of a complete BGC obviously diversifies the antimicrobial arsenal, while its loss accompanied or not by that of the resistant determinant can result in a high intrapopulational SM variability. Further, recombination within or between BGCs may be enhanced thanks to their highly redundant structure (i.e., modular organization of NRPS and polyketide synthase [PKS]) and foster the emergence of new BGCs and the production of new antimicrobial

compounds. The intense turnover of the BGC and resistance gene sets would constantly modulate the interactions within the population and, through this dynamic state, favor the social cohesiveness of the population.

MATERIALS AND METHODS

Strain isolation and maintenance. Four grains of soil on the order of cubic millimeters in size and distant from each other from 2 cm to a maximum of 8 cm were sampled from a clod of soil collected in the Montiers-sur-Saulx forest in France (GPS coordinates: 48°32′37.248′′N, 5°18′21.946′′E) and stored for 48 h at 4°C before processing. For bacterial isolation, each soil aggregate was dissolved 1:100 (wt/vol) in sterile water by vortexing for 15 min. Serial dilutions from 10^{-1} to 10^{-3} were spread on *Streptomyces* isolation medium (SIM) (56), and bacterial colonies showing a typical *Streptomyces* phenotype were randomly picked after 7 days of incubation at 30°C. After three consecutive subcultures on mannitol soy agar plates (57), strains were stored as spore suspensions in 20% glycerol (57).

Genome sequencing, annotation, and DNA manipulation. Total genomic DNA was isolated as previously described by Kieser et al. (57). Each genome was sequenced by one-dimension MinION (Oxford Nanopore Technologies, UK) (mean coverage of 108×) and further corrected with MiSeq sequencing (Illumina, CA, USA) (mean coverage 50×), enabling the acquisition of each chromosome in one scaffold and the identification of extrachromosomal elements when present. Sequencing and assembling were performed via the I2BC NGS platform (France). Genome sequencing data and genome accession numbers are listed in Table S2 in the supplemental material. Coding sequence prediction and annotation were performed using the NCBI Prokaryotic Genome Annotation Pipeline (58). Genes involved in SM biosynthesis were predicted with antiSMASH (41).To disrupt a gene cluster putatively involved in the biosynthesis of a NRPS in strain RLB1-9, a 897-bp sequence internal to the peptide synthetase gene (open reading frame [ORF]-07845) was amplified and cloned into the suicide vector pIJ8668 to give pIJ8668-*intORF-07845* (59). Intergeneric *Escherichia coli*/*Streptomyces* conjugation (60) followed by selection for apramycin resistance (pIJ8668) allowed us to select NRPS-encoding gene disruption by homologous recombination. Independent mutant clones (i.e., from distinct transformation assays) were tested for their antimicrobial activities. Pulsed-field gel electrophoresis (PFGE) was performed on whole genomic DNA in agarose plugs as described previously (61).

Phylogenetic analyses. 16S rRNA genes and MLST genes (*atpD*, *gyrB*, *recA*, *rpoB*, and *trpB*) were amplified using primers and PCR conditions described by Rintala et al. (62) and Guo et al. (63), respectively. Neighbor-joining phylogenetic trees were built using single or concatenated MLST gene (3,178 positions) sequences with a Kimura's two-parameter distance correction and 100 bootstrap replicates (MEGA7 software [64]). Phylogenomic analyses were performed with the nucleotidic sequences of the core genome identified by a Best BLAST Hit analysis (65), subsequently aligned with Muscle (v3.8.31 [66]), and corrected with Gblocks (67). From the 8,778 genes of the population core genome, we selected the 5,213 genes (5,149,602 nucleotide positions) shared with *Streptomyces avermitilis*, the reference strain used as a root in the phylogenetic reconstruction. Maximum likelihood phylogenomic tree (GTR, 100 bootstrap replicates) was built with RaxML (68). Python scripts (unpublished) were used to calculate the average nucleotide identity (ANIb) (69) and to identify and to extrapolate pangenomes. Pangenome visualization was performed using Anvi'o (70), and extrapolation graphs were computed with the UpsetR package in script R. Functional AICEs and ICEs were identified first by searching signature genes (replication, transfer, excision, and integration) using BLASTP (65) as described by Bordeleau et al. (18) and then by identifying direct repeats flanking the element. Breaks of synteny were identified using an in-lab developed software (unpublished and available upon request).

Antibacterial activity assays. Drops of *Streptomyces* spore suspensions (5 μl of spore stock, 10^7 spores per ml) were spotted and incubated for 5 days at 30°C on GA medium plates (54). Plates were overlaid with a Luria-Bertani top (0.4% agar), inoculated with *Bacillus* (optical density at 600 nm [OD_{600}] of 0.1), and incubated at 4°C for 2 h and then at 30°C overnight. Growth inhibition of *Bacillus* was assessed by the observation of a zone of clearance. To test *Streptomyces-Streptomyces* inhibitory capacities, spores of a first strain were spotted on mannitol soy medium plates for a 3-day incubation at 30°C. Then, a similar deposit of spores of the second strain (indicator) was adjacency spotted (5 mm) and incubated for 5 days at 30°C. The absence of phenotypic variation in comparison with control strains inoculated alone was considered a neutral interaction, and growth or sporulation alterations were recorded as antagonism.

Accession number(s). The genome accession numbers for the strains in this study are CP041650 to CP041654 and CP041601 to CP041613.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.01533-19.

FIG S1, PDF file, 0.6 MB.

FIG S2, PDF file, 0.4 MB.

FIG S3, PDF file, 0.4 MB.

FIG S4, PDF file, 0.4 MB.

TABLE S1, PDF file, 0.4 MB.
TABLE S2, PDF file, 0.5 MB.
TABLE S3, PDF file, 0.8 MB.

ACKNOWLEDGMENTS

This work was supported by a grant overseen by the French National Research Agency (ANR) as part of the Investissements d'Avenir program (ANR-11-LABX-0002-01, Lab of Excellence ARBRE).

We thank the research system SOERE-OPE (Système d'Observation et d'Experimentation au long terme pour la Recherche en Environnement, Observatoire Pérenne de l'Environnement) for access to the sampling site.

REFERENCES

1. Barka EA, Vatsa P, Sanchez L, Gaveau-Vaillant N, Jacquard C, Meier-Kolthoff JP, Klenk H-P, Clément C, Ouhdouch Y, van Wezel GP. 2016. Taxonomy, physiology, and natural products of *Actinobacteria*. Microbiol Mol Biol Rev 80:1–43. https://doi.org/10.1128/MMBR.00019-15.
2. Olanrewaju OS, Babalola OO. 2019. *Streptomyces*: implications and interactions in plant growth promotion. Appl Microbiol Biotechnol 103:1179–1188. https://doi.org/10.1007/s00253-018-09577-y.
3. Aigle B, Lautru S, Spiteller D, Dickschat JS, Challis GL, Leblond P, Pernodet J-L. 2014. Genome mining of *Streptomyces ambofaciens*. J Ind Microbiol Biotechnol 41:251–263. https://doi.org/10.1007/s10295-013-1379-y.
4. Baltz RH. 2017. Gifted microbes for genome mining and natural product discovery. J Ind Microbiol Biotechnol 44:573–588. https://doi.org/10.1007/s10295-016-1815-x.
5. Choulet F, Aigle B, Gallois A, Mangenot S, Gerbaud C, Truong C, Francou F-X, Fourrier C, Guérineau M, Decaris B, Barbe V, Pernodet J-L, Leblond P. 2006. Evolution of the terminal regions of the *Streptomyces* linear chromosome. Mol Biol Evol 23:2361–2369. https://doi.org/10.1093/molbev/msl108.
6. Redenbach M, Kieser HM, Denapaite D, Eichner A, Cullum J, Kinashi H, Hopwood DA. 1996. A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb *Streptomyces coelicolor* A3(2) chromosome. Mol Microbiol 21:77–96. https://doi.org/10.1046/j.1365-2958.1996.6191336.x.
7. Tian X, Zhang Z, Yang T, Chen M, Li J, Chen F, Yang J, Li W, Zhang B, Zhang Z, Wu J, Zhang C, Long L, Xiao J. 2016. Comparative genomics analysis of *Streptomyces* species reveals their adaptation to the marine environment and their diversity at the genomic level. Front Microbiol 7:998. https://doi.org/10.3389/fmicb.2016.00998.
8. Kim J-N, Kim Y, Jeong Y, Roe J-H, Kim B-G, Cho B-K. 2015. Comparative genomics reveals the core and accessory genomes of *Streptomyces* species. J Microbiol Biotechnol 25:1599–1605. https://doi.org/10.4014/jmb.1504.04008.
9. Zhou Z, Gu J, Li Y-Q, Wang Y. 2012. Genome plasticity and systems evolution in *Streptomyces*. BMC Bioinformatics 13 Suppl 10:S8. https://doi.org/10.1186/1471-2105-13-S10-S8.
10. Lin YS, Kieser HM, Hopwood DA, Chen CW. 1993. The chromosomal DNA of *Streptomyces lividans* 66 is linear. Mol Microbiol 10:923–933. https://doi.org/10.1111/j.1365-2958.1993.tb00964.x.
11. Huang CH, Lin YS, Yang YL, Huang SW, Chen CW. 1998. The telomeres of *Streptomyces* chromosomes contain conserved palindromic sequences with potential to form complex secondary structures. Mol Microbiol 28:905–916. https://doi.org/10.1046/j.1365-2958.1998.00856.x.
12. Chen CW, Huang C-H, Lee H-H, Tsai H-H, Kirby R. 2002. Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. Trends Genet 18:522–529. https://doi.org/10.1016/S0168-9525(02)02752-X.
13. Thibessard A, Leblond P. 2014. Subtelomere plasticity in the bacterium *Streptomyces*, p 243–258. *In* Louis EJ, Becker MM (ed), Subtelomeres. Springer, Berlin, Germany.
14. Doroghazi JR, Buckley DH. 2010. Widespread homologous recombination within and between *Streptomyces* species. ISME J 4:1136–1143. https://doi.org/10.1038/ismej.2010.45.
15. Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304. https://doi.org/10.1038/35012500.
16. Stuttard C. 1982. Temperate phages of *Streptomyces venezuelae*: lysogeny and host specificity shown by phages SV1 and SV2. Microbiology 128:115–121. https://doi.org/10.1099/00221287-128-1-115.
17. Burke J, Schneider D, Westpheling J. 2001. Generalized transduction in *Streptomyces coelicolor*. Proc Natl Acad Sci U S A 98:6289–6294. https://doi.org/10.1073/pnas.101589398.
18. Bordeleau E, Ghinet MG, Burrus V. 2012. Diversity of integrating conjugative elements in actinobacteria: coexistence of two mechanistically different DNA-translocation systems. Mob Genet Elements 2:119–124. https://doi.org/10.4161/mge.20498.
19. Thoma L, Muth G. 2016. Conjugative DNA-transfer in *Streptomyces*, a mycelial organism. Plasmid 87–88:1–9. https://doi.org/10.1016/j.plasmid.2016.09.004.
20. Pettis GS. 2018. Spreading the news about the novel conjugation mechanism in *Streptomyces* bacteria. Environ Microbiol Rep 10:503–510. https://doi.org/10.1111/1758-2229.12659.
21. McDonald BR, Currie CR. 2017. Lateral gene transfer dynamics in the ancient bacterial genus *Streptomyces*. mBio 8:e00644-17. https://doi.org/10.1128/mBio.00644-17.
22. Antony-Babu S, Stien D, Eparvier V, Parrot D, Tomasi S, Suzuki MT. 2017. Multiple *Streptomyces* species with distinct secondary metabolomes have identical 16S rRNA gene sequences. Sci Rep 7:11089. https://doi.org/10.1038/s41598-017-11363-1.
23. Vetsigian K, Jajoo R, Kishony R. 2011. Structure and evolution of *Streptomyces* interaction networks in soil and *in silico*. PLoS Biol 9:e1001184. https://doi.org/10.1371/journal.pbio.1001184.
24. Vicente CM, Thibessard A, Lorenzi J-N, Benhadj M, Hôtel L, Gacemi-Kirane D, Lespinet O, Leblond P, Aigle B. 2018. Comparative genomics among closely related *Streptomyces* strains revealed specialized metabolite biosynthetic gene cluster diversity. Antibiotics (Basel) 7:e86. https://doi.org/10.3390/antibiotics7040086.
25. Harris HMB, Bourin MJB, Claesson MJ, O'Toole PW. 2017. Phylogenomics and comparative genomics of *Lactobacillus salivarius*, a mammalian gut commensal. Microb Genom 3:e000115. https://doi.org/10.1099/mgen.0.000115.
26. Park CJ, Andam CP. 2019. Within-species genomic variation and variable patterns of recombination in the tetracycline producer *Streptomyces rimosus*. Front Microbiol 10:552. https://doi.org/10.3389/fmicb.2019.00552.
27. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP. 2014. Diversification of bacterial genome content through distinct mechanisms over different timescales. Nat Commun 5:5471. https://doi.org/10.1038/ncomms6471.
28. Levade I, Terrat Y, Leducq J-B, Weil AA, Mayo-Smith LM, Chowdhury F, Khan AI, Boncy J, Buteau J, Ivers LC, Ryan ET, Charles RC, Calderwood SB, Qadri F, Harris JB, LaRocque RC, Shapiro BJ. 2017. *Vibrio cholerae* genomic diversity within and between patients. Microb Genom 3:e000142. https://doi.org/10.1099/mgen.0.000142.
29. Bruns H, Crüsemann M, Letzel A-C, Alanjary M, McInerney JO, Jensen PR, Schulz S, Moore BS, Ziemert N. 2018. Function-related replacement of bacterial siderophore pathways. ISME J 12:320–329. https://doi.org/10.1038/ismej.2017.137.
30. Stefanic P, Kraigher B, Lyons NA, Kolter R, Mandic-Mulec I. 2015. Kin discrimination between sympatric *Bacillus subtilis* isolates. Proc Natl Acad Sci U S A 112:14042–14047. https://doi.org/10.1073/pnas.1512671112.

31. Wielgoss S, Didelot X, Chaudhuri RR, Liu X, Weedall GD, Velicer GJ, Vos M. 2016. A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium *Myxococcus xanthus*. ISME J 10:2468–2477. https://doi.org/10.1038/ismej.2016.34.
32. Vos M, Wolf AB, Jennings SJ, Kowalchuk GA. 2013. Micro-scale determinants of bacterial diversity in soil. FEMS Microbiol Rev 37:936–954. https://doi.org/10.1111/1574-6976.12023.
33. Wolf AB, Vos M, de Boer W, Kowalchuk GA. 2013. Impact of matric potential and pore size distribution on growth dynamics of filamentous and non-filamentous soil bacteria. PLoS One 8:e83661. https://doi.org/10.1371/journal.pone.0083661.
34. Raynaud X, Nunan N. 2014. Spatial ecology of bacteria at the microscale in soil. PLoS One 9:e87217. https://doi.org/10.1371/journal.pone.0087217.
35. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, Connor R, Fiorini N, Funk K, Hefferon T, Holmes JB, Kim S, Kimchi A, Kitts PA, Lathrop S, Lu Z, Madden TL, Marchler-Bauer A, Phan L, Schneider VA, Schoch CL, Pruitt KD, Ostell J. 2019. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 47:D23–D28. https://doi.org/10.1093/nar/gky1069.
36. Wenner T, Roth V, Fischer G, Fourrier C, Aigle B, Decaris B, Leblond P. 2003. End-to-end fusion of linear deleted chromosomes initiates a cycle of genome instability in *Streptomyces ambofaciens*. Mol Microbiol 50:411–425. https://doi.org/10.1046/j.1365-2958.2003.03698.x.
37. Fischer G, Wenner T, Decaris B, Leblond P. 1998. Chromosomal arm replacement generates a high level of intraspecific polymorphism in the terminal inverted repeats of the linear chromosomal DNA of *Streptomyces ambofaciens*. Proc Natl Acad Sci U S A 95:14296–14301. https://doi.org/10.1073/pnas.95.24.14296.
38. Uchida T, Miyawaki M, Kinashi H. 2003. Chromosomal arm replacement in *Streptomyces griseus*. J Bacteriol 185:1120–1124. https://doi.org/10.1128/jb.185.3.1120-1124.2003.
39. Clerc S, Simonet P. 1996. Efficiency of the transfer of a pSAM2-derivative plasmid between two strains of *Streptomyces lividans* in conditions ranging from agar slants to non-sterile soil microcosms. FEMS Microbiol Ecol 21:157–165. https://doi.org/10.1111/j.1574-6941.1996.tb00343.x.
40. Sullivan JT, Ronson CW. 1998. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. Proc Natl Acad Sci U S A 95:5145–5149. https://doi.org/10.1073/pnas.95.9.5145.
41. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, de los Santos ELC, Kim HU, Nave M, Dickschat JS, Mitchell DA, Shelest E, Breitling R, Takano E, Lee SY, Weber T, Medema MH. 2017. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res 45:W36–W41. https://doi.org/10.1093/nar/gkx319.
42. McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes. Nat Microbiol 2:17040. https://doi.org/10.1038/nmicrobiol.2017.40.
43. Hoff G, Bertrand C, Piotrowski E, Thibessard A, Leblond P. 2018. Genome plasticity is governed by double strand break DNA repair in *Streptomyces*. Sci Rep 8:5272. https://doi.org/10.1038/s41598-018-23622-w.
44. Casjens S, Palmer N, van Vugt R, Huang WM, Stevenson B, Rosa P, Lathigra R, Sutton G, Peterson J, Dodson RJ, Haft D, Hickey E, Gwinn M, White O, Fraser CM. 2002. A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. Mol Microbiol 35:490–516. https://doi.org/10.1046/j.1365-2958.2000.01698.x.
45. Ricchetti M, Dujon B, Fairhead C. 2003. Distance from the chromosome end determines the efficiency of double strand break repair in subtelomeres of haploid yeast. J Mol Biol 328:847–862. https://doi.org/10.1016/S0022-2836(03)00315-2.
46. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 437:94–100. https://doi.org/10.1038/nature04029.
47. Bentley SD, Chater KF, Cerdeño-Tárraga A-M, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang C-H, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream M-A, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA. 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417:141–147. https://doi.org/10.1038/417141a.
48. Hsiao N-H, Kirby R. 2008. Comparative genomics of *Streptomyces avermitilis*, *Streptomyces cattleya*, *Streptomyces maritimus* and *Kitasatospora aureofaciens* using a *Streptomyces coelicolor* microarray system. Antonie Van Leeuwenhoek 93:1–25. https://doi.org/10.1007/s10482-007-9175-1.
49. Hopwood DA. 2006. Soil to genomics: the *Streptomyces* chromosome. Annu Rev Genet 40:1–23. https://doi.org/10.1146/annurev.genet.40.110405.090639.
50. Hopwood DA, Kieser T, Wright HM, Bibb MJ. 1983. Plasmids, recombination and chromosome mapping in *Streptomyces lividans* 66. J Gen Microbiol 129:2257–2269. https://doi.org/10.1099/00221287-129-7-2257.
51. Gray TA, Krywy JA, Harold J, Palumbo MJ, Derbyshire KM. 2013. Distributive conjugal transfer in mycobacteria generates progeny with meiotic-like genome-wide mosaicism, allowing mapping of a mating identity locus. PLoS Biol 11:e1001602. https://doi.org/10.1371/journal.pbio.1001602.
52. Derbyshire KM, Gray TA. 2014. Distributive conjugal transfer: new insights into horizontal gene transfer and genetic exchange in mycobacteria. Microbiol Spectr 2:MGM2-0022-2013. https://doi.org/10.1128/microbiolspec.MGM2-0022-2013.
53. Cordero OX, Wildschutte H, Kirkup B, Proehl S, Ngo L, Hussain F, Le Roux F, Mincer T, Polz MF. 2012. Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. Science 337:1228–1231. https://doi.org/10.1126/science.1219385.
54. Abrudan MI, Smakman F, Grimbergen AJ, Westhoff S, Miller EL, van Wezel GP, Rozen DE. 2015. Socially mediated induction and suppression of antibiosis during bacterial coexistence. Proc Natl Acad Sci U S A 112:11054–11059. https://doi.org/10.1073/pnas.1504076112.
55. Butaitė E, Baumgartner M, Wyder S, Kümmerli R. 2017. Siderophore cheating and cheating resistance shape competition for iron in soil and freshwater *Pseudomonas* communities. Nat Commun 8:414. https://doi.org/10.1038/s41467-017-00509-4.
56. D'Costa VM, McGrann KM, Hughes DW, Wright GD. 2006. Sampling the antibiotic resistome. Science 311:374–377. https://doi.org/10.1126/science.1120800.
57. Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA. 2000. Practical Streptomyces genetics. John Innes Foundation, Norwich, United Kingdom.
58. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569.
59. Sun J, Kelemen GH, Fernández-Abalos JM, Bibb MJ. 1999. Green fluorescent protein as a reporter for spatial and temporal gene expression in *Streptomyces coelicolor* A3(2). Microbiology 145:2221–2227. https://doi.org/10.1099/00221287-145-9-2221.
60. Kim M-K, Ha H-S, Choi S-U. 2008. Conjugal transfer using the bacteriophage phiC31 att/int system and properties of the attB site in *Streptomyces ambofaciens*. Biotechnol Lett 30:695–699. https://doi.org/10.1007/s10529-007-9586-0.
61. Leblond P, Francou FX, Simonet J-M, Decaris B. 1990. Pulsed-field gel electrophoresis analysis of the genome of *Streptomyces ambofaciens* strains. FEMS Microbiol Lett 72:79–88. https://doi.org/10.1111/j.1574-6968.1990.tb03866.x.
62. Rintala H, Nevalainen A, Rönkä E, Suutari M. 2001. PCR primers targeting the 16S rRNA gene for the specific detection of streptomycetes. Mol Cell Probes 15:337–347. https://doi.org/10.1006/mcpr.2001.0379.
63. Guo Y, Zheng W, Rong X, Huang Y. 2008. A multilocus phylogeny of the *Streptomyces griseus* 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. Int J Syst Evol Microbiol 58:149–159. https://doi.org/10.1099/ijs.0.65224-0.
64. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874. https://doi.org/10.1093/molbev/msw054.
65. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
66. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. https://doi.org/10.1186/1471-2105-5-113.
67. Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein se-

quence alignments. Syst Biol 56:564–577. https://doi.org/10.1080/10635150701472164.
68. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
69. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91. https://doi.org/10.1099/ijs.0.64483-0.
70. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319. https://doi.org/10.7717/peerj.1319.

III - Diversité des regroupements de gènes codant des métabolites spécialisés entre espèces proches

Ce point présente les résultats d'une étude menée principalement par Cláudia Vicente (Université de Lorraine INRA, Dynamic) qui s'intéresse aux métabolites spécialisés de différentes souches proches de *Streptomyces.*

Ces travaux ne se basent pas sur les mêmes données que décrites jusqu'à présent, mais sur une collection de 5 souches de *Streptomyces* précédemment isolées (Benhadj et al., 2018). Ils fournissent un aperçu de la diversité génétique des *Streptomyces* en considérant les métabolites spécialisés, et ce, même à une échelle phylogénétique restreinte.

Publication 4 : **Comparative Genomics among Closely Related *Streptomyces* Strains Revealed Specialized Metabolite Biosynthetic Gene Cluster Diversity.** *Cláudia M. Vicente, Annabelle Thibessard, Jean-Noël Lorenzi, Mabrouka Benhadj, Laurence Hôtel, Djamila Gacemi-Kirane, Olivier Lespinet, Pierre Leblond, Bertrand Aigle*

Cette analyse se base sur une collection de 5 souches environnementales de *Streptomyces* (toutes prélevées dans une région humide d'Algérie (lac Fetzara) (Benhadj et al., 2018)). Une analyse multilocus utilisant 5 gènes (*atpD*, *gyrB*, *recA*, *rpoB* et *trpB*) a permis de définir les relations phylogénétiques de ces souches avec plusieurs espèces de *Streptomyces* déjà définies. 4 clades ont ainsi été distingués, représentés à chaque fois par au moins une espèce déjà déposée des *Streptomyces*. En complément, en calculant les scores d'ANIb (variant de 81.4% à 99.9%), il est apparu que parmi les 5 souches environnementales, 3 appartiennent à la même espèce. Les 2 autres appartiennent à 2 espèces différentes.

L'identification des SMBGC (pour Specialized Metabolite Biosynthetic Gene Cluster) dans chaque souche a mis en évidence de 28 à 33 SMBGC selon les souches considérées. Les génomes présentant le moins et le plus de regroupements de gènes biosynthétiques correspondent, comme attendus, respectivement au plus petit et plus grand génome. Entre les 4 clades, un core SMBGC a été défini, il consiste en 8 SMBG retrouvés dans tous les clades. Les SMBGC spécifiques à chaque souche ont été identifiés, révélant que chacune des souches présentes entre 1 et 15 SMBGC qui leur sont spécifiques. Parmi ces regroupements spécifiques, plusieurs sont inconnus (de 1 à 6 par souche). Bien qu'une caractérisation de ces regroupements soit encore nécessaire, ils reflètent la richesse chimique qui peut être trouvée à des niveaux phylogénétiques proches, et suggèrent que le séquençage de souches appartenant à la même espèce continuera à produire de l'originalité.

antibiotics

MDPI

Communication

Comparative Genomics among Closely Related *Streptomyces* Strains Revealed Specialized Metabolite Biosynthetic Gene Cluster Diversity

Cláudia M. Vicente [1], Annabelle Thibessard [1], Jean-Noël Lorenzi [2], Mabrouka Benhadj [1,3], Laurence Hôtel [1], Djamila Gacemi-Kirane [4], Olivier Lespinet [2], Pierre Leblond [1,*] and Bertrand Aigle [1,*]

1 Université de Lorraine, INRA, DynAMic, F-54000 Nancy, France; claudia.morgado-vicente@univ-lorraine.fr (C.M.V.); annabelle.thibessard@univ-lorraine.fr (A.T.); mabrouka.benhadj@gmail.com (M.B.); laurence.hotel@univ-lorraine.fr (L.H.)
2 Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University Paris-Sud, University Paris-Saclay, Gif-sur-Yvette CEDEX, France; jean-noel.lorenzi@i2bc.paris-saclay.fr (J.-N.L.); olivier.lespinet@i2bc.paris-saclay.fr (O.L.)
3 Biomolecules and Application Laboratory, Faculty of Exact Sciences and Natural and Life Sciences, University of Tebessa, Tebessa 12002, Algeria
4 Department of Biochemistry, Faculty of Science, University Badji Mokhtar Annaba, Annaba 23000, Algeria; dj_gacemi@yahoo.fr
* Correspondence: pierre.leblond@univ-lorraine.fr (P.L.); bertrand.aigle@univ-lorraine.fr (B.A.); Tel.: +33-372-745-143 (P.L.); +33-372-745-141 (B.A.)

Received: 12 September 2018; Accepted: 28 September 2018; Published: 2 October 2018

Abstract: Specialized metabolites are of great interest due to their possible industrial and clinical applications. The increasing number of antimicrobial resistant infectious agents is a major health threat and therefore, the discovery of chemical diversity and new antimicrobials is crucial. Extensive genomic data from *Streptomyces* spp. confirm their production potential and great importance. Genome sequencing of the same species strains indicates that specialized metabolite biosynthetic gene cluster (SMBGC) diversity is not exhausted, and instead, a pool of novel specialized metabolites still exists. Here, we analyze the genome sequence data from six phylogenetically close *Streptomyces* strains. The results reveal that the closer strains are phylogenetically, the number of shared gene clusters is higher. Eight specialized metabolites comprise the core metabolome, although some strains have only six core gene clusters. The number of conserved gene clusters common between the isolated strains and their closest phylogenetic counterparts varies from nine to 23 SMBGCs. However, the analysis of these phylogenetic relationships is not affected by the acquisition of gene clusters, probably by horizontal gene transfer events, as each strain also harbors strain-specific SMBGCs. Between one and 15 strain-specific gene clusters were identified, of which up to six gene clusters in a single strain are unknown and have no identifiable orthologs in other species, attesting to the existing SMBGC novelty at the strain level.

Keywords: *Streptomyces*; strain; specialized metabolites; biosynthetic gene cluster

1. Introduction

Microorganisms are unmatched in their capacity to produce chemically diverse specialized metabolites [1]. *Streptomyces* in particular, which are ubiquitous filamentous bacteria with a differentiating mycelial life cycle, have a significant role in biotechnology industry. They are employed in the production of more than half of all antibiotics used in human and veterinary medicine,

as well as a large assortment of other high-value molecules such as immune suppressants, anticancer, and anti-parasitic molecules [2,3]. The increasing incidence of antibiotic resistant infections and the lack of new antimicrobial agents have driven the race for natural product discovery. The significant advances made in high-throughput DNA sequencing, along with the relatively low cost of whole genome sequencing, have greatly contributed to this effort [4]. Strategies to address the antimicrobial shortage include both the intensive screening of uncultivable microorganisms through metagenomics and environmental DNA cloning for heterologous expression [5,6], as well as further genome mining of cultivable isolates. Consequently, correlation data between species and specialized metabolomes are accumulating progressively.

Streptomyces spp. are amongst the best-characterized bacteria, and the vastly increasing genomic data that is available has revealed the presence of thousands of specialized metabolite biosynthetic gene clusters (SMBGCs), highlighting their enormous potential for natural product biosynthesis [7,8]. Estimates say that no more than 10% have been identified to date, and furthermore, the majority of these SMBGCs appear to be "silent" [9]. A recent bioinformatic analysis of six *Streptomyces albus* strains led to the identification of strain-specific SMBGCs, speaking to the chemical diversity at the strain level [10]. It has also been shown that *Streptomyces* species with identical 16S rRNA sequences can have distinct specialized metabolomes [11]. Moreover, studies in actinobacteria *Salinispora* demonstrated an important biosynthetic pathway diversity in marine populations, providing insight into new chemical diversity-generating mechanisms [12].

A *Streptomyces* strain collection that included isolates with interesting biological activities against both Gram-positive and Gram-negative indicator strains was recently characterized [13]. Here, we report the genomic analysis of a small group of these environmental *Streptomyces* strains, including those antimicrobial-producing isolates and the associated SMBGC diversity that was encountered. This analysis highlights that the use of closely related strains to identity new SMBGCs in *Streptomyces* constitutes a promising approach for the identification of novel specialized biosynthetic pathways.

2. Results and Discussion

2.1. Analyzed Strains are Phylogenetically Diverse

The genomes of five environmental strains from a previously constructed *Streptomyces* collection [13], isolates E5N91, E2N166, E2N171, E1N211, E5N298, were sequenced and examined. Draft genome sequences have 7.0–8.5 Mb with an average G+C content of 71.8%, which is in good accordance with the genus (Table 1).

Table 1. Strains analyzed in this work.

Strain	G + C Content (%)	Genome Size (Mb)
E5N91	71.9	8.51
E2N166	70.2	7.91
E2N171	72.3	7.00
E1N211	72.1	7.32
E5N298	71.9	7.87
Streptomyces sp. ETH9427	72.1	7.75

A comparative genomics analysis of the nucleotide sequences, including the taxonomically close *Streptomyces* sp. ETH9427 recently sequenced [14], was carried out through a multilocus sequence typing (MLST) using five genes: *atpD* (ATP synthase, β subunit), *gyrB* (DNA gyrase, subunit B), *recA* (recombination protein), *rpoB* (RNA polymerase, β subunit), and *trpB* (tryptophan synthase, β subunit). It highlighted the taxonomic relationships between the strains and with other model *Streptomyces* species (Figure 1). Four clades are distinguishable in the constructed phylogenetic tree: the first clade with strain E5N91, *Streptomyces coelicolor* A3(2) [7], and *S. lividans* TK24 [15], the second containing strain E5N298 and the recently described *Streptomyces* sp. M1013 [16], clade 3, including strain E2N166

and *S. ambofaciens* strains ATCC 23877 [17] and DSM 40697 [18], and finally strains E2N171, E1N211, and ETH9427 grouped with *Streptomyces* sp. 4F (accession number CP013142.1) in clade 4 (Figure 1).

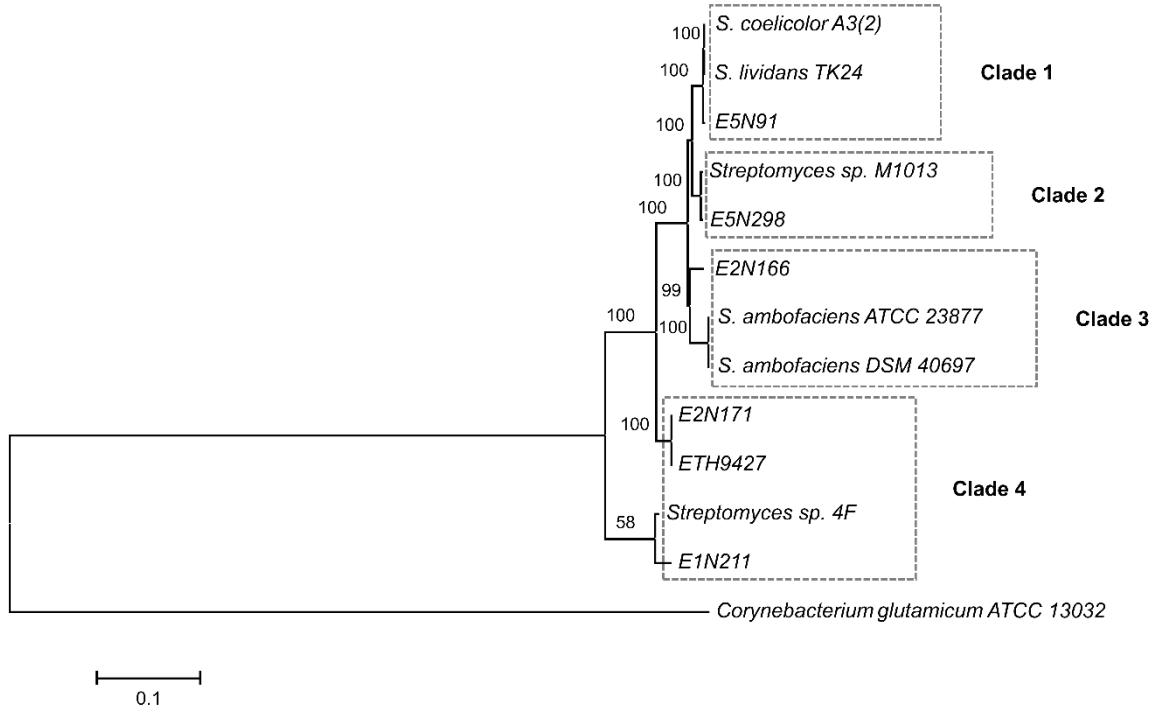

Figure 1. Phylogenetic tree of sequenced isolated and reference strains. Constructed based on the concatenated sequence alignment of five loci (*atpB-gyrB-recA-rpoB-trpB*), using inference by neighbor-joining with 100 bootstrap replicates. Only positions with a minimum 92% site coverage were used and the final dataset has 9141 positions in total. The scale bar indicates 10% estimated sequence divergence. The tree is rooted on *Corynebacterium glutamicum* ATCC 13032.

Taxonomic affiliations were further assessed by calculating the average nucleotide identity (ANI) using JSpeciesWS [19], and are resumed in Table 2. The ANI in clade 1 between strain E5N91 and *S. coelicolor* and with *S. lividans* is 97.32% (with 75.5% of the genome nucleotides aligned) and 97.65% (77.0% of nucleotides aligned), respectively. In clade 2, the ANI value is 98.19% (80.9% nucleotides aligned) between strain E5N298 and *Streptomyces* sp. M1013, and in clade 3 between strain E2N166, *S. ambofaciens* ATCC 23877, and *S. ambofaciens* DSM 40697, the ANI values are 88.86% (65.3% alignment) and 88.75% (with 66.0% nucleotides aligned), respectively. The values in clade 4 are as follows: the E2N171 strain's ANI value is 98.98% with strain E1N211 (83.5% of aligned nucleotides), 99.14% (84.8% alignment) with strain ETH9427, and 94.56% (76.3% of nucleotides aligned) with *Streptomyces* sp. 4F; the ANI value of strain E1N211 is 99.90% (86.4% of aligned nucleotides) with strain ETH9427, and 94.41% (71.1% alignment) with *Streptomyces* sp. 4F; and the ETH9427 strain's ANI value is 94.13% (70.7% of nucleotides aligned) with *Streptomyces* sp. 4F. These results suggest that strain E5N91 belongs to a species that is taxonomically closely related to *S. coelicolor* and *S. lividans* in clade 1, as do strain E5N298 and *Streptomyces* sp. M1013 in clade 2. Moreover, the ANI value among the six studied strains varies between 81.40–99.90%, except in those from clade 4, where the ANI value is 99.00–99.90%, well above the species boundary [20] and indicating that strains E2N171, E1N211, and ETH9427 belong to the same species. The closest relative to these strains is *Streptomyces* sp. 4F, and as such, all are clustered in clade 4. However, these strains and *Streptomyces* sp. 4F belong to distinct but related species, as seen by the ANI values between 94.13–94.56% and a low branch bootstrap in the phylogenetic tree (Figure 1). Still, these strains have highly similar 16S ribosomal DNA sequences (dissimilarities between 0.1–1.1%), which would mistakenly suggest that most of the strains belong to the same species. These results indicate that 16S single-gene analysis is not the most accurate tool for streptomycetes

phylogeny analysis and inferring intraspecific genetic diversity, as already seen previously [21,22]. Instead, taxonomic relationships are better resolved using genome comparison indexes, such as ANI, and multi-gene analysis, such as MLST [23,24], which in some cases can even match whole-genome comparisons [25].

Table 2. Average nucleotide identity values of the compared strains.

	SLI	*E5N91*	*M1013*	*E5N298*	*ATCC*	*DSM*	*E2N166*	*E2N171*	*E1N211*	*ETH9427*	*4F*
SCO	99.0	97.3	91.5	91.3	86.9	86.9	87.5	81.7	81.8	81.9	81.8
SLI		97.7	91.8	91.6	87.0	87.0	87.7	81.6	81.7	81.8	81.7
E5N91			91.2	91.0	86.6	86.6	87.4	81.4	81.4	81.6	81.3
M1013				98.2	87.1	87.1	87.9	81.5	81.5	81.7	81.4
E5N298					87.2	87.2	87.9	81.7	81.7	81.8	81.6
ATCC						99.0	88.9	81.8	81.7	81.8	81.9
DSM							88.8	81.7	81.6	81.7	81.9
E2N166								82.1	82.1	82.2	82.0
E2N171									99.0	99.1	94.6
E1N211										99.9	94.4
ETH9427											94.1

2.2. Specialized Metabolism Diversity

The identification of SMBGCs was performed using antiSMASH 3.0 [26] and then edited to take into account previously published experimental data. A total of 28–33 SMBGCs was predicted in the isolated strains (Figure 2), with strain E2N171 harboring the least (28) and strain E5N91 harboring the most (33) gene clusters, expectantly coinciding with the smallest and largest genome, respectively. All six strains harbor SMBGCs belonging to the principal classes such as polyketide synthases (PKS), non-ribosomal synthetases (NRPS), and PKS–NRPS hybrids. Interestingly, the analysis revealed that a single strain within its clade can contain up to 48% strain-specific SMBGCs, suggesting that strain-level search for new SMBGCs diversity is not exhausted, and genome sequencing of several strains from a single species can lead to the identification of new gene clusters, as already seen previously [10,12].

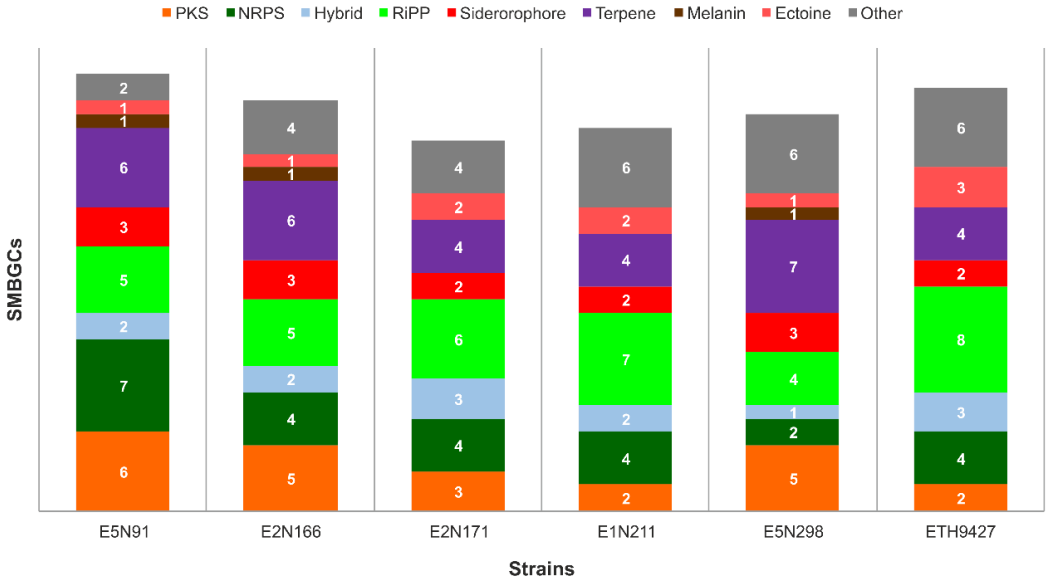

Figure 2. Specialized metabolite biosynthetic gene clusters identified in analyzed strains. Genome sequences were analyzed with antiSMASH 3.0 for gene cluster identification. The "Hybrid" category refers only to PKS-NRPS hybrid gene clusters. Other hybrid clusters that were detected were separated into their constituent parts to better visualize cluster diversity. Clusters encoding bacteriocins, lantipeptides, and lassopeptides are grouped in the category "RiPP", and the category "Other" includes other KS pathways and less common clusters such as phenazines and butyrolactones, among others.

2.3. The Core and Conserved Specialized Metabolites

Some of the identified gene clusters are well known and usually found in *Streptomyces* spp. The core SMBGCs are gene clusters that are present in the majority of streptomycetes and in the examined strains were selected based on conservation, with at least 58% of the strains sharing all of the core SMBGCs that encode for: melanin [27], ectoine osmolytes [28], spore pigment [29], the siderophore desferrioxamine B [30], the terpenes albaflavenone [31], hopene [32], isorenieratene [33], and the volatile geosmin [34]. This set of eight biosynthetic gene clusters is present in all of the strains from clades 1, 2, and 3, although no isorenieratene gene cluster was found in strain E2N166 (Figure 3). Interestingly, none of the strains in clade 4 possess either melanin or isorenieratene encoding gene clusters. Furthermore, strains E2N171, E1N211, and ETH9427 have additional ectoine gene clusters: two in ETH9427, and one in each of the other two strains. These have different gene contents (Figure 3, yellow and green E clusters), and as expected, the phylogenetically closer E2N171 and ETH9427 strains have the same extra ectoine encoding gene cluster, that in the case of ETH9427 is doubled due to its location on the chromosomal terminal inverted repeats. This result is consistent with strains from the same species, which usually share the same core metabolome [35].

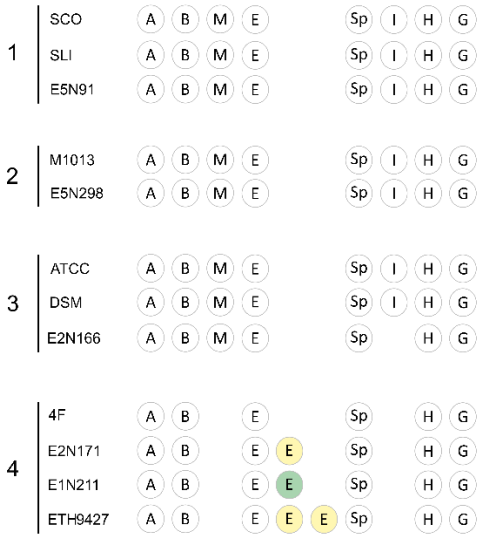

Figure 3. The core specialized metabolites produced by the strains. From the total number of SMBGCs that were identified, the core metabolome constitutes a set of gene clusters frequently conserved across *Streptomyces* species. The letters stand for the specialized metabolites albaflavenone (A), desferrioxamine B (B), melanin (M), ectoine (E), spore pigment (Sp), isorenieratene (I), hopene (H), and geosmin (G). Cluster disposition is not related to chromosomal position. The ectoine clusters in yellow and green represent different gene clusters.

In addition to the core metabolome, other SMBGCs are conserved within each clade (Figure 4). Strain E5N91 contains 18 conserved gene clusters out of 33 in total, which are also present in *S. coelicolor* in clade 1. Surprisingly, biosynthetic gene clusters for the characteristic metabolites coelichelin [36], coelibactin [37], calcium-dependent antibiotic [38], and coelimycin [39] are conserved in strain E5N91, whereas the actinorhodin [40] and undecylprodigiosin [41] gene clusters are not, even if they are present in phylogenetically closely related species. In clade 2, two-thirds of the gene clusters identified in strain E5N298 are conserved in *Streptomyces* sp. M1013 (21 out of 30 in total). Conversely, in clade 3, strain E2N166 has only nine gene clusters out of 31 in total conserved with *S. ambofaciens* strains ATCC 23877 and DSM 40697. These include the species-distinctive Type II PKS gene cluster encoding kinamycin [42], the NRPS biosynthetic gene cluster for coelichelin, two siderophore and two bacteriocin gene clusters, and a terpene-encoding gene cluster [43]. The E2N171, E1N211, and ETH9427 strains in clade 4 have a remarkably high number of conserved SMBGCs, with each strain showing 54 to

72% of the identified gene clusters that are present in one or both of the other two strains. However, only seven gene clusters are conserved in *Streptomyces* sp. 4F, which is consistent with its taxonomic relationship in clade 4. These results indicate that phylogenetic relationships are accompanied by specialized gene cluster maintenance.

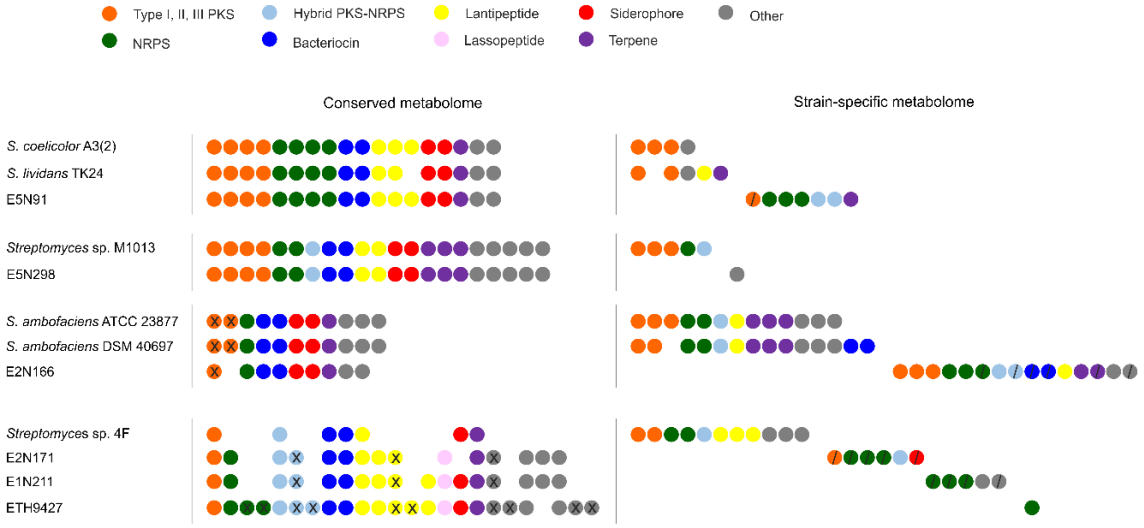

Figure 4. Conserved and strain-specific biosynthetic gene clusters. Within each clade, conserved and strain-specific gene clusters (circles) were identified. Original chromosomal organization is not considered, and clusters with the same position are identical. Colors represent the major classes of specialized metabolites: Type I, II, and III PKS (orange), NRPS (green), hybrid PKS–NRPS (light blue), bacteriocin (dark blue), lantipeptide (yellow), lassopeptide (pink), siderophore (red), terpene (purple), and others that include other KS pathways and less common metabolites such as phenazines, butyrolactones, and indoles (grey). Unknown clusters are indicated with "/" and duplicated clusters with "X".

2.4. Strain-Specific Biosynthetic Gene Clusters

Other gene clusters exist in each of the six examined strains beyond the core and conserved metabolome, the strain-specific SMBGCs. These were determined as gene clusters that are exclusive to a given strain and absent from its phylogenetically closest relatives, based on the nucleotide identity of putative biosynthetic genes. The putative products of these strain-specific biosynthetic pathways comprise all the major classes of specialized metabolites (Figure 4). Each strain harbors at least one strain-specific gene cluster. The fewest strain-specific gene clusters were encountered in strains E5N298 and ETH9427, with only one each. Strains E5N91 and E5N298, which are taxonomically close to their clade's counterparts, dedicate almost opposite amounts of their specialized metabolome to strain-specific products (21% and 3%, respectively), perhaps as a result of horizontal gene transfer events and reflecting different levels of adaptation [44,45]. Conversely, other strains show a significantly higher number of strain-specific gene clusters. Strains E2N171 and E2N166 harbor six and 15 strain-specific gene clusters that represent 21% and 48% of their specialized metabolome, respectively.

The preliminary results of genome alignment and sequence comparison of the relatively close E2N166 and *S. ambofaciens* strains reveals large chromosome reorganizations events, such as a pericentral inversion, and most of the acquired strain-specific gene clusters are unexpectedly located in the central region of the chromosome. Moreover, genome alignment and comparison of the ETH9427 strain with its closest related strain *Streptomyces* sp. 4F and the reference species *S. ambofaciens* ATCC 23477 shows a relatively conserved chromosome organization, particularly between the first two strains (Figure 5). The identified SMBGCs also reflect the shared chromosomal regions, with seven gene clusters shared between ETH9427 and *Streptomyces* sp. 4F, most of which are also maintaining relative

positions, and only three gene clusters between ETH9427 and *S. ambofaciens* ATCC 23477 (Figure 5), supporting the notion that the distribution of SMBGCs is coherent with phylogenetic relationships [46].

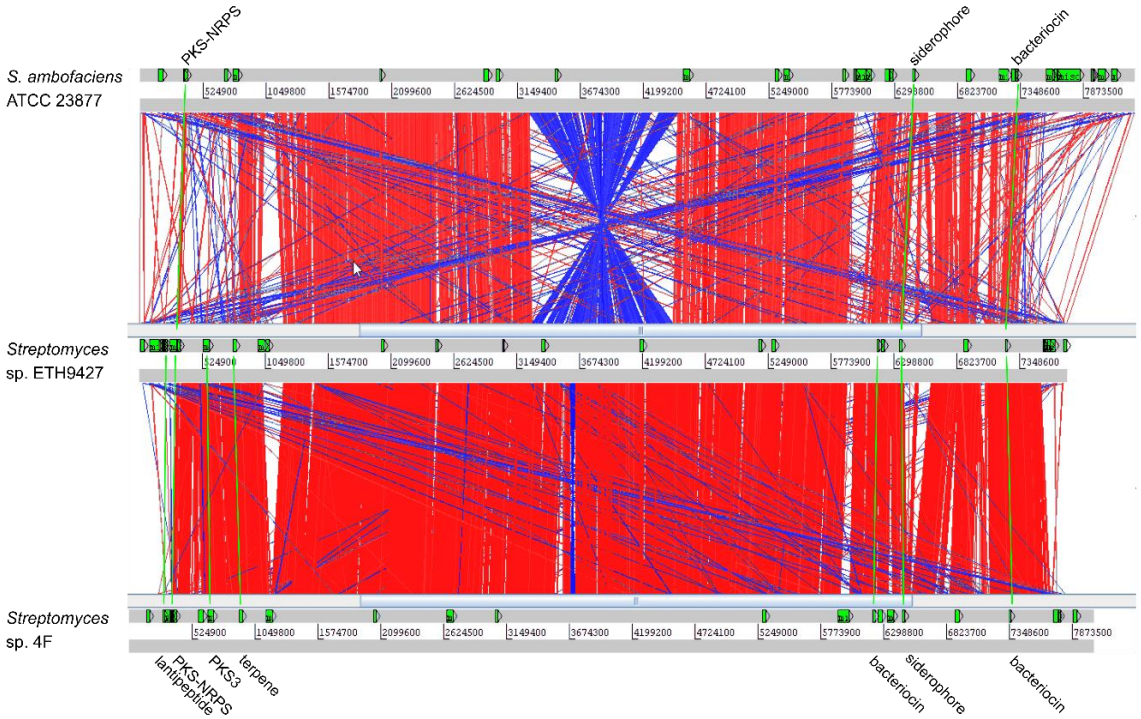

Figure 5. Genome alignment and specialized metabolome comparison of *Streptomyces* sp. ETH9427. Alignment was performed with the Artemis Comparison Tool software (ACT) using an identity threshold of 77% and a score threshold of 500. Synteny regions are represented by red lines, inversion events are represented in blue, and breaks in synteny are seen as blank spaces. SMBGCs are marked in the horizontal panels (green arrows), and conserved gene clusters are linked (green lines).

While some of the identified strain-specific gene clusters have orthologues in other species, four out of the six strains that are used in this work (strains E5N91, E2N166, E2N171, and E1N211) have between one and six unknown biosynthetic gene clusters. These represent 3% to 19% of the metabolome, leading to a total of 16 SMBGCs that could possibly lead to novel chemistry. The strain-specific gene clusters identified in E2N171 and E1N211 (six NRPS, a type I PKS and a siderophore-like and another gene cluster classified as "other") are of particular interest, as they can conceivably be responsible for the antimicrobial activities previously observed in Benhadj et al., namely against nosocomial infections responsible bacterial strains of *Escherichia coli*, *Streptococcus aureus* and *Pseudomonas aeruginosa*. These bioactivities are exclusive to E2N171 and E1N211 amongst the entire collection, and are absent from strains such as E5N298; therefore, the identified strain-specific SMBGCs in these two strains are good candidates for further detailed characterization [13].

3. Materials and Methods

3.1. Strains and DNA Extraction

Five strains from a previously built *Streptomyces* collection were used in this work [13]. In brief, isolates were obtained from water samples of Lake Fetzara (Annaba, Algeria) on different agar media supplemented with antifungals. Additionally, a strain from the ETH collection (Eidgenosische Technische Hochschule Kultursammlungen, Zurich, Switzerland), *Streptomyces* sp. ETH9427 (accession numbers CP029624, CP029625, CP029626), was also analyzed. Genomic DNA was extracted from liquid cultures using a standard "*salting out*" protocol [47].

3.2. Whole-Genome Sequencing and Annotation

Draft genomes were constructed by assembling reads from an Illumina Genome Analyzer (Illumina, San Diego, CA, USA), with pair-ends (c.300 bp) and/or mate-pairs (c.8 kb) DNA fragments libraries, and were processed as described in Thibessard et al. [14]. Sequences were trimmed to eliminate adaptor sequences, and contigs with a size smaller than 1000 pb were eliminated. Draft genomes of the strains E5N91, E2N166, E2N171, E1N211, and E5N298 were submitted into the NCBI GenBank database and given the following accession numbers: RAIE00000000, RAIF00000000, RAIH00000000, RAII00000000, and RAIG00000000, respectively. Whenever possible, sequence assembly and contig ordering were achieved using the CLC Main Workbench (Qiagen, Hilden, Germany), and coding sequence prediction and annotation were automatically performed using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) [48]. Genome alignments were performed with Artemis Comparison Tool (ACT) [49]. Average nucleotide identity (ANI) was determined using the JSpeciesWS server with the BLAST algorithm (http://jspecies.ribohost.com/jspeciesws/) [19].

3.3. Phylogenetic Analysis and Biosynthetic Gene Cluster Identification

The phylogenetic relationships between strains were reconstructed with an MLST analysis [50], using DNA sequences from *atpB*, *gyrB*, *recA*, *rpoB*, and *trpB* genes. Several reference *Streptomyces* species were included in the analysis: *S. ambofaciens* ATCC 23877 (accession number CP012382.1), *S. ambofaciens* DSM 40697 (accession number CP012949.1), *S. coelicolor* A3(2) (accession number NC_003888.3), *S. lividans* TK24 (accession number CP009124.1), *Streptomyces* sp. M1013 (accession number MQUH00000000.1), and *Streptomyces* sp. 4F (accession number CP013142.1). The actinobacteria *Corynebacterium glutamicum* ATCC 13032 (accession number BA000036.3) was used to root the phylogenetic tree. Nucleotide sequence alignments were performed with SeaView software [51] using the muscle algorithm, and a concatenated gene sequence (20592 nt in total) was used to construct a neighbor-joining phylogenetic tree with Kimura's 2-parameter distance correction and 100 bootstrap replicates with MEGA7 software [52]. Overall, there was a good separation and statistical support for most of the branches in the tree. Biosynthetic gene clusters were predicted using the antiSMASH 3.0 server (https://antismash.secondarymetabolites.org/#!/start) without ClusterFinder algorithm [26].

4. Conclusions

The efforts for *Streptomyces* environmental strains genome sequencing continue to show the genus SMBGC's diversity. Strain de-replication of isolates can be achieved using multi-gene sequence comparison instead of the traditional 16S rRNA, and the analysis of isolated strains reveals a pool for novel specialized metabolites in taxonomically closely related strains. Encountered gene cluster diversity supports phylogenetic relationships, where closely related strains share a higher number of SMBGCs. Horizontal gene transfer as a source of genetic diversity does not distort phylogenetic examination, as these are rare events, and the presence or absence of a SMBGC is not enough to blur phylogenetic relationships. Strain-level analysis has already been demonstrated to lead to the identification of new biosynthetic pathways. Though the strain-specific gene clusters identified here require further characterization, they nonetheless reflect that chemical richness can be found at this level, and imply that the deep sequencing of strains belonging to same species will continue to produce originality.

Author Contributions: Conceptualization, C.M.V., P.L. and B.A.; methodology, C.M.V.; validation, P.L. and B.A.; formal analysis, C.M.V.; investigation, C.M.V., L.H. and J.-N.L.; resources, M.H., D.K.-G.; data curation, C.M.V., P.L., A.T. and O.L.; writing—original draft preparation, C.M.V.; writing—review and editing, B.A. and P.L.; visualization, all authors; supervision, P.L. and B.A.; project administration, B.A.; funding acquisition, B.A. and P.L.

Funding: This work was founded by the French National Research Agency through two programs (ANR Streptoflux ANR-07-BLAN-0096, ANR MiGenIs ANR-13-BSV6-0006), the French National Institute for Agricultural

Research (INRA), Région Lorraine and the Laboratory Excellence ARBRE (ANR-11-LABX-0002-01). CMV was also supported by the AgreenSkillsPlus Program (FP7-609398.0000) and the Région Grand Est.

Acknowledgments: The authors acknowledge the support of the DynAMic laboratory by the Impact Biomolecules project of the "Lorraine Université d'Excellence (Investissements d'avenir—ANR)".

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bérdy, J. Bioactive microbial metabolites. *J. Antibiot.* **2005**, *58*, 1–26. [CrossRef] [PubMed]
2. Worrall, J.A.R.; Vijgenboom, E. Copper mining in *Streptomyces*: Enzymes, natural products and development. *Nat. Prod. Rep.* **2010**, *27*, 742–756. [CrossRef] [PubMed]
3. Newman, D.J.; Cragg, G.M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, *79*, 629–661. [CrossRef] [PubMed]
4. Corre, C.; Challis, G.L. New natural product biosynthetic chemistry discovered by genome mining. *Nat. Prod. Rep.* **2009**, *26*, 977–986. [CrossRef] [PubMed]
5. Lewis, K. Platforms for antibiotic discovery. *Nat. Rev. Drug. Discov.* **2013**, *12*, 371–387. [CrossRef] [PubMed]
6. Rebets, Y.; Kormanec, J.; Luzhetskyy, A.; Bernaerts, K.; Anné, J. Cloning and Expression of Metagenomic DNA in *Streptomyces* lividans and Subsequent Fermentation for Optimized Production. *Methods Mol. Biol.* **2017**, *1539*, 99–144. [CrossRef] [PubMed]
7. Bentley, S.D.; Chater, K.F.; Cerdeño-Tárraga, A.-M.; Challis, G.L.; Thomson, N.R.; James, K.D.; Harris, D.E.; Quail, M.A.; Kieser, H.; Harper, D.; et al. Complete genome sequence of the model actinomycete *Streptomyces* coelicolor A3(2). *Nature* **2002**, *417*, 141–147. [CrossRef] [PubMed]
8. Medema, M.H.; Fischbach, M.A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **2015**, *11*, 639–648. [CrossRef] [PubMed]
9. Nett, M.; Ikeda, H.; Moore, B.S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat. Prod. Rep.* **2009**, *26*, 1362–1384. [CrossRef] [PubMed]
10. Seipke, R.F. Strain-level diversity of secondary metabolism in *Streptomyces* albus. *PLoS ONE* **2015**, *10*, e0116457. [CrossRef] [PubMed]
11. Antony-Babu, S.; Stien, D.; Eparvier, V.; Parrot, D.; Tomasi, S.; Suzuki, M.T. Multiple *Streptomyces* species with distinct secondary metabolomes have identical 16S rRNA gene sequences. *Sci. Rep.* **2017**, *7*, 11089. [CrossRef] [PubMed]
12. Ziemert, N.; Lechner, A.; Wietz, M.; Millán-Aguiñaga, N.; Chavarria, K.L.; Jensen, P.R. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E1130–E1139. [CrossRef] [PubMed]
13. Benhadj, M.; Gacemi-Kirane, D.; Toussaint, M.; Hotel, L.; Bontemps, C.; Duval, R.E.; Aigle, B.; Leblond, P. Diversity and antimicrobial activities of *Streptomyces* isolates from Fetzara Lake, north eastern Algeria. *Ann. Biol. Clin.* **2018**, *76*, 81–95. [CrossRef] [PubMed]
14. Thibessard, A.; Vicente, C.M.; Bertrand, C.; Aigle, B.; Leblond, P. Draft genome of *Streptomyces* sp. ETH9427, whole genome shotgun sequence. *Genome Announc.* **2018**, in press.
15. Rückert, C.; Albersmeier, A.; Busche, T.; Jaenicke, S.; Winkler, A.; Friðjónsson, Ó.H.; Hreggviðsson, G.Ó.; Lambert, C.; Badcock, D.; Bernaerts, K.; et al. Complete genome sequence of *Streptomyces* lividans TK24. *J. Biotechnol.* **2015**, *199*, 21–22. [CrossRef] [PubMed]
16. Haas, D.; Gerbaud, C.; Sahin, N.; Pernodet, J.-L.; Lautru, S. Draft Genome Sequence of *Streptomyces* sp. M1013, a Close Relative of Streptomyces ambofaciens and *Streptomyces* coelicolor. *Genome Announc.* **2017**, *5*. [CrossRef] [PubMed]
17. Thibessard, A.; Haas, D.; Gerbaud, C.; Aigle, B.; Lautru, S.; Pernodet, J.-L.; Leblond, P. Complete genome sequence of *Streptomyces* ambofaciens ATCC 23877, the spiramycin producer. *J. Biotechnol.* **2015**, *214*, 117–118. [CrossRef] [PubMed]
18. Thibessard, A.; Leblond, P. Complete Genome Sequence of *Streptomyces* ambofaciens DSM 40697, a Paradigm for Genome Plasticity Studies. *Genome Announc.* **2016**, *4*. [CrossRef] [PubMed]

19. Richter, M.; Rosselló-Móra, R.; Oliver Glöckner, F.; Peplies, J. JSpeciesWS: A web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* **2016**, *32*, 929–931. [CrossRef] [PubMed]
20. Rosselló-Móra, R.; Amann, R. Past and future species definitions for Bacteria and Archaea. *Syst. Appl. Microbiol.* **2015**, *38*, 209–216. [CrossRef] [PubMed]
21. Alam, M.T.; Merlo, M.E.; Takano, E.; Breitling, R. Genome-based phylogenetic analysis of *Streptomyces* and its relatives. *Mol. Phylogenet. Evol.* **2010**, *54*, 763–772. [CrossRef] [PubMed]
22. Rong, X.; Liu, N.; Ruan, J.; Huang, Y. Multilocus sequence analysis of *Streptomyces* griseus isolates delineating intraspecific diversity in terms of both taxonomy and biosynthetic potential. *Antonie. Van. Leeuwenhoek* **2010**, *98*, 237–248. [CrossRef] [PubMed]
23. Rong, X.; Huang, Y. Taxonomic evaluation of the *Streptomyces* hygroscopicus clade using multilocus sequence analysis and DNA-DNA hybridization, validating the MLSA scheme for systematics of the whole genus. *Syst. Appl. Microbiol.* **2012**, *35*, 7–18. [CrossRef] [PubMed]
24. Labeda, D.P.; Dunlap, C.A.; Rong, X.; Huang, Y.; Doroghazi, J.R.; Ju, K.-S.; Metcalf, W.W. Phylogenetic relationships in the family Streptomycetaceae using multi-locus sequence analysis. *Antonie Van Leeuwenhoek* **2017**, *110*, 563–583. [CrossRef] [PubMed]
25. Jiménez, G.; Urdiain, M.; Cifuentes, A.; López-López, A.; Blanch, A.R.; Tamames, J.; Kämpfer, P.; Kolstø, A.-B.; Ramón, D.; Martínez, J.F.; et al. Description of Bacillus toyonensis sp. nov., a novel species of the Bacillus cereus group, and pairwise genome comparisons of the species of the group by means of ANI calculations. *Syst. Appl. Microbiol.* **2013**, *36*, 383–391. [CrossRef] [PubMed]
26. Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H.U.; Bruccoleri, R.; Lee, S.Y.; Fischbach, M.A.; Müller, R.; Wohlleben, W.; et al. AntiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **2015**, *43*, W237–W243. [CrossRef] [PubMed]
27. Bernan, V.; Filpula, D.; Herber, W.; Bibb, M.; Katz, E. The nucleotide sequence of the tyrosinase gene from Streptomyces antibioticus and characterization of the gene product. *Gene* **1985**, *37*, 101–110. [CrossRef]
28. Malin, G.; Lapidot, A. Induction of synthesis of tetrahydropyrimidine derivatives in *Streptomyces* strains and their effect on Escherichia coli in response to osmotic and heat stress. *J. Bacteriol.* **1996**, *178*, 385–395. [CrossRef] [PubMed]
29. Davis, N.K.; Chater, K.F. Spore colour in Streptomyces coelicolor A3(2) involves the developmentally regulated synthesis of a compound biosynthetically related to polyketide antibiotics. *Mol. Microbiol.* **1990**, *4*, 1679–1691. [CrossRef] [PubMed]
30. Neilands, J.B. Microbial iron compounds. *Annu. Rev. Biochem.* **1981**, *50*, 715–731. [CrossRef] [PubMed]
31. Gürtler, H.; Pedersen, R.; Anthoni, U.; Christophersen, C.; Nielsen, P.H.; Wellington, E.M.; Pedersen, C.; Bock, K. Albaflavenone, a sesquiterpene ketone with a zizaene skeleton produced by a streptomycete with a new rope morphology. *J. Antibiot.* **1994**, *47*, 434–439. [CrossRef] [PubMed]
32. Poralla, K.; Muth, G.; Härtner, T. Hopanoids are formed during transition from substrate to aerial hyphae in Streptomyces coelicolor A3(2). *FEMS Microbiol. Lett.* **2000**, *189*, 93–95. [CrossRef] [PubMed]
33. Krügel, H.; Krubasik, P.; Weber, K.; Saluz, H.P.; Sandmann, G. Functional analysis of genes from Streptomyces griseus involved in the synthesis of isorenieratene, a carotenoid with aromatic end groups, revealed a novel type of carotenoid desaturase. *Biochim. Biophys. Acta* **1999**, *1439*, 57–64. [CrossRef]
34. Gerber, N.N.; Lechevalier, H.A. Geosmin, an earthly-smelling substance isolated from actinomycetes. *Appl. Microbiol.* **1965**, *13*, 935–938. [PubMed]
35. Choudoir, M.J.; Pepe-Ranney, C.; Buckley, D.H. Diversification of Secondary Metabolite Biosynthetic Gene Clusters Coincides with Lineage Divergence in Streptomyces. *Antibiotics* **2018**, *7*, 12. [CrossRef]
36. Lautru, S.; Deeth, R.J.; Bailey, L.M.; Challis, G.L. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. *Nat. Chem. Biol.* **2005**, *1*, 265–269. [CrossRef] [PubMed]
37. Hesketh, A.; Kock, H.; Mootien, S.; Bibb, M. The role of absC, a novel regulatory gene for secondary metabolism, in zinc-dependent antibiotic production in Streptomyces coelicolor A3(2). *Mol. Microbiol.* **2009**, *74*, 1427–1444. [CrossRef] [PubMed]
38. Chong, P.P.; Podmore, S.M.; Kieser, H.M.; Redenbach, M.; Turgay, K.; Marahiel, M.; Hopwood, D.A.; Smith, C.P. Physical identification of a chromosomal locus encoding biosynthetic genes for the lipopeptide calcium-dependent antibiotic (CDA) of *Streptomyces* coelicolor A3(2). *Microbiology* **1998**, *144 Pt 1*, 193–199. [CrossRef] [PubMed]

39. Takano, E.; Kinoshita, H.; Mersinias, V.; Bucca, G.; Hotchkiss, G.; Nihira, T.; Smith, C.P.; Bibb, M.; Wohlleben, W.; Chater, K. A bacterial hormone (the SCB1) directly controls the expression of a pathway-specific regulatory gene in the cryptic type I polyketide biosynthetic gene cluster of *Streptomyces* coelicolor. *Mol. Microbiol.* **2005**, *56*, 465–479. [CrossRef] [PubMed]
40. Wright, L.F.; Hopwood, D.A. Actinorhodin is a chromosomally-determined antibiotic in *Streptomyces* coelicolar A3(2). *J. Gen. Microbiol.* **1976**, *96*, 289–297. [CrossRef] [PubMed]
41. Rudd, B.A.; Hopwood, D.A. A pigmented mycelial antibiotic in *Streptomyces* coelicolor: Control by a chromosomal gene cluster. *J. Gen. Microbiol.* **1980**, *119*, 333–340. [CrossRef] [PubMed]
42. Aigle, B.; Pang, X.; Decaris, B.; Leblond, P. Involvement of AlpV, a new member of the *Streptomyces* antibiotic regulatory protein family, in regulation of the duplicated type II polyketide synthase alp gene cluster in *Streptomyces* ambofaciens. *J. Bacteriol.* **2005**, *187*, 2491–2500. [CrossRef] [PubMed]
43. Aigle, B.; Lautru, S.; Spiteller, D.; Dickschat, J.S.; Challis, G.L.; Leblond, P.; Pernodet, J.-L. Genome mining of *Streptomyces* ambofaciens. *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 251–263. [CrossRef] [PubMed]
44. Jensen, P.R.; Williams, P.G.; Oh, D.-C.; Zeigler, L.; Fenical, W. Species-specific secondary metabolite production in marine actinomycetes of the genus Salinispora. *Appl. Environ. Microbiol.* **2007**, *73*, 1146–1152. [CrossRef] [PubMed]
45. Penn, K.; Jenkins, C.; Nett, M.; Udwary, D.W.; Gontang, E.A.; McGlinchey, R.P.; Foster, B.; Lapidus, A.; Podell, S.; Allen, E.E.; et al. Genomic islands link secondary metabolism to functional adaptation in marine actinobacteria. *ISME J.* **2009**, *3*, 1193–1203. [CrossRef] [PubMed]
46. Joynt, R.; Seipke, R.F. A phylogenetic and evolutionary analysis of antimycin biosynthesis. *Microbiology* **2018**, *164*, 28–39. [CrossRef] [PubMed]
47. Kieser, T.; Bibb, M.; Buttner, M.; Chater, K.; Hopwood, D.A. *Practical Streptomyces Genetics*; The John Innes Foundation: Norwich, UK, 2000.
48. Tatusova, T.; DiCuccio, M.; Badretdin, A.; Chetvernin, V.; Nawrocki, E.P.; Zaslavsky, L.; Lomsadze, A.; Pruitt, K.D.; Borodovsky, M.; Ostell, J. NCBI prokaryotic genome annotation pipeline. *Nucleic. Acids. Res.* **2016**, *44*, 6614–6624. [CrossRef] [PubMed]
49. Carver, T.; Berriman, M.; Tivey, A.; Patel, C.; Böhme, U.; Barrell, B.G.; Parkhill, J.; Rajandream, M.-A. Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **2008**, *24*, 2672–2676. [CrossRef] [PubMed]
50. Guo, Y.; Zheng, W.; Rong, X.; Huang, Y. A multilocus phylogeny of the *Streptomyces* griseus 16S rRNA gene clade: Use of multilocus sequence analysis for streptomycete systematics. *Int. J. Syst. Evol. Microbiol.* **2008**, *58*, 149–159. [CrossRef] [PubMed]
51. Gouy, M.; Guindon, S.; Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **2010**, *27*, 221–224. [CrossRef] [PubMed]
52. Tamura, K.; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M.; Kumar, S. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **2011**, *28*, 2731–2739. [CrossRef] [PubMed]

DISCUSSION

A- *Streptomyces*, une classification discutée

Le genre *Streptomyces* a été proposé en 1943 (Waksman and Henrici, 1943), et placé dans la famille des *Streptomycetaceae*, sur la base d'observations phénotypiques et des caractéristiques de la paroi cellulaire. L'incorporation de caractéristiques moléculaires a permis des avancées considérables dans la délimitation du genre *Streptomyces* (Stackebrandt et al., 1997) au sein de la classe des Actinobactéries. Cependant, la délimitation du genre reste floue, et plusieurs espèces ont été incluses dans le genre des *Streptomyces* puis exclues selon de nouvelles études et/ou de nouvelles technologies phylogénétiques. Par exemple, *Kitasatospora* a été unifiée avec le genre *Streptomyces* sur la similarité de leur ARN 16S (Wellington et al., 1992). Ensuite, il a été démontré que l'espèce *Kitasatospora* formait un groupe monophylétique stable en dehors du genre *Streptomyces* (Zhang et al., 1997) faisant "revivre" le genre *Kitasatospora* comme il fut premièrement décrit (Omura et al., 1982).

En raison des antibiotiques produits par *Streptomyces*, un important criblage de nouveaux composés bioactifs a été effectué, ce qui a conduit à une sur-classification du genre : entre les années 1940 et 1960, le nombre d'espèces de *Streptomyces* est passé de 40 à plus de 3000 (Waksman and Henrici, 1943). Des critères d'identification ont été introduits suite à un consortium international sur les *Streptomyces* (Shirling and Gottlieb, 1966), afin de réduire le nombre d'espèces synonymes. En outre, des méthodes ont été développées pour améliorer l'estimation de la parenté des espèces à l'intérieur du genre grâce à l'avènement de l'analyse moléculaire et de la génomique. Le premier génome de *Streptomyces* entièrement séquencé, *Streptomyces coelicolor* A3(2) (Bentley et al., 2002), a marqué la possibilité d'utiliser les données du génome entièrement séquencé pour améliorer notre compréhension de l'organisation du genre. McDonald and Currie (2017) ont proposé une phylogénie du genre *Streptomyces* **figure 30** avec une horloge moléculaire qui estime une origine ancienne du genre *Streptomyces*, il y a environ 380 millions d'années. Cette phylogénie met aussi en évidence une subdivision du genre en au moins 2 clades monophylétiques (Clade I et Clade II) et un dernier regroupement (Other Lineages) où le caractère monophylétique n'apparaît pas.

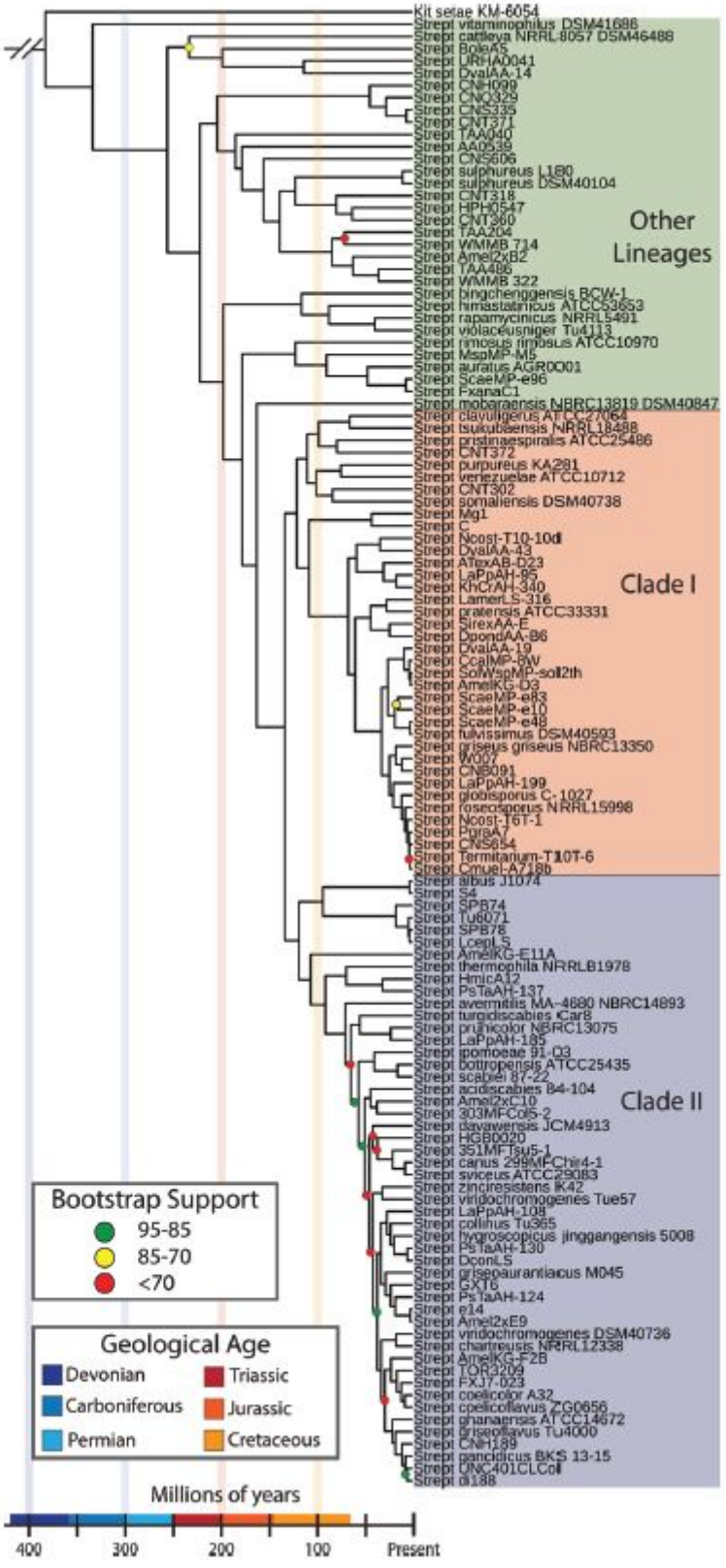

Figure 30- Phylogénie basée sur un alignement multilocus de 94 gènes conservés et une horloge moléculaire du genre *Streptomyces*. La longueur des branches indique les temps de divergence estimés par Reltime (Tamura et al., 2012). Les valeurs de bootstrap sont indiquées par des cercles colorés sur tous les nœuds ayant des valeurs inférieures à 95. *Streptomyces* et *Kitasatospora* sont respectivement abrégés en Strept et Kit.

Figure tirée de (McDonald and Currie, 2017).

Dans ces travaux de thèse, plusieurs phylogénies du genre *Streptomyces* ont été réalisées, et cela en considérant de plus en plus d'espèces différentes. Pour chaque phylogénie, cette subdivision en 3 groupes est retrouvée. De plus, en se basant sur les espèces partagées entre la phylogénie de McDonald and Currie (2017) et nos phylogénies, les mêmes clades (clade I et II) ainsi que le groupe "other" ont pu être mis en évidence. Notre méthode de construction de la phylogénie du genre *Streptomyces* s'appuie sur un core-génome, c'est-à-dire l'ensemble des gènes partagés entre toutes les espèces du genre. Cela empêche l'utilisation d'autres espèces en dehors du genre pour raciner l'arbre. Par conséquent, avec le jeu de données actuelles, (**figure 27**) nous ne pouvons pas être catégorique sur le caractère monophylétique du regroupement "other" qui regroupe de nombreuses espèces distantes des deux autres clades.

Ainsi, avec les données et méthodes d'analyses phylogénétiques actuelles, la définition du genre *Streptomyces* semble claire, en particulier en s'appuyant sur des méthodes de comparaison globale des génomes comme l'ANI pour gagner en précision à courte distance phylogénétique (identification des différentes souches d'une même espèce) et les phylogénies basées sur des alignements multiples de nombreux gènes pour considérer le genre dans sa globalité.

La discussion se porte aujourd'hui sur l'organisation intra-genre où 2 clades sont déjà bien définies à l'opposé du groupe "other" où un enrichissement en espèces semble nécessaire pour gagner en précision. Avec la phylogénie de la collection 234 (non finalisée), il apparaît que le groupe "other" pourrait être divisé en 2 groupes monophylétiques.

B - Évolution des répertoires de gènes chez *Streptomyces*

La variabilité génétique au niveau intra-spécifique a été révélée dès que les méthodologies de séquençage ont permis d'aborder le répertoire de gènes de souches appartenant à la même espèce (Perna et al., 2001). Avec ces travaux, les termes core-génome et génome accessoire ont été inventés pour décrire cette variabilité. Le core-génome fait référence aux familles de gènes conservées que l'on trouve dans tous les membres séquencés jusqu'à présent et dépend du groupe d'isolats considérés. Le génome accessoire fait référence aux gènes dispensables présents dans une fraction des génomes considérés. Le pan-génome est constitué de toutes les familles de gènes qui ont été trouvées dans l'échantillon ou le niveau phylogénétique considéré (espèce, genre). Certaines espèces procaryotes ont des pan-génomes ouverts, tandis que d'autres renferment peu de diversité génomique et présentent un pan-génome fermé. La validité du pan-génome dépend de la représentativité de l'échantillonnage de génomes séquencés par rapport au niveau phylogénétique envisagé.

Ces travaux de thèse ont mis en évidence que nous sommes encore loin d'avoir échantillonné toute la diversité du genre, et ce quel que soit le niveau phylogénétique considéré : aussi bien dans la première partie de ce manuscrit où l'étude porte sur un ensemble de *Streptomyces* couvrant tout le genre que sur la deuxième partie où nous nous sommes concentré sur une population sympatrique de *Streptomyces*.

De plus, en évaluant la diversité génétique par clade, l'évolution de la taille du pan-génome par clade (**figure 31**) montre des tendances différentes. Selon un test de comparaison de médianes, l'évolution de la dimension du pan-génome du groupe "other" est significativement plus importante que celle des clades I et II. À l'inverse, pour le clade I, l'évolution est significativement plus lente que les autres groupes.

Les génomes bactériens sont biaisés en faveur de la suppression de l'ADN (Kuo and Ochman, 2009). Au cours de l'évolution, les génomes ont donc tendance à perdre les gènes neutres suggérant que l'augmentation de la dimension d'un pan-génome rend compte de l'adaptation d'un organisme à son environnement par l'acquisition de nouveaux gènes et ce, principalement par HGT adaptatifs (Sela et al., 2016). La majorité des *Streptomyces* sont retrouvés dans le sol ; c'est un milieu complexe dont les caractéristiques physico-chimiques (pH, température, humidité, type de minéraux...) peuvent être radicalement différentes d'un endroit à l'autre. De plus, dans le sol, le contact avec d'autres espèces bactériennes est fréquent, facilitant la mise en place de transferts horizontaux avec d'autres bactéries.

L'accroissement rapide de la dimension du pan-génome du groupe "other" par rapport aux autres clades peut-être une conséquence de la sous-représentation ou d'une plus grande diversité de ce groupe. En effet, la phylogénie (**figure 27**) montre que le groupe "other" est principalement composé

d'espèces distantes les unes des autres (longues branches dans la phylogénie). Cette diversité phylogénétique plus importante que dans les 2 clades monophylétiques est cohérente avec un apport de gènes spécifiques et accessoires plus important à chaque nouvelle espèce considérée dans le groupe "other" que dans les autres clades.

Dans le cas du clade I, où l'accroissement de la dimension du pan-génome est plus faible que celle du genre (**figure 31**), la tendance ne semble pas être due à une sous-représentation du clade. McDonald and Currie (2017) ont mis en évidence que parmi les espèces composant leur clade I (**figure 30**), nombreuses sont celles isolées depuis des insectes. Cette situation d'association à un hôte diminue la variabilité de l'environnement des bactéries et pourrait justifier cette tendance d'accroissement plus lent du pan-génome pour le clade I. Malheureusement, l'origine précise de l'échantillon biologique d'où sont extraites les séquences, est une information qui n'est pas toujours mentionnée dans les bases de données et même dans les publications associées au dépôt d'une séquence. A cause de cela, il n'a pas été possible d'obtenir suffisamment d'informations sur l'origine des espèces qui composent le clade I (de la collection 110) pour vérifier l'enrichissement de ce clade en *Streptomyces* associé à un hôte.

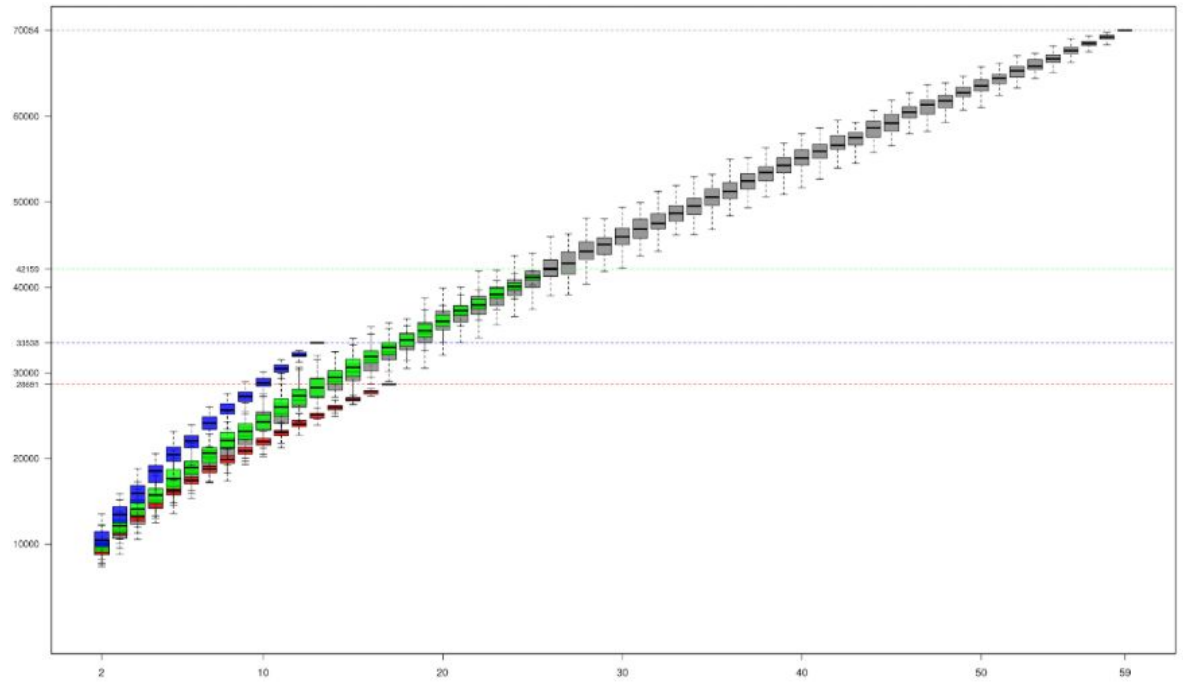

Figure 31- Evolution de la dimension du pan-génome par clade en fonction du nombre de génomes. Pour un nombre de génomes variant de 2 à 59, la taille du pan-génome a été calculée pour 100 itérations d'une sélection aléatoire de génomes choisis dans notre ensemble de données de la collection 59. Chaque point est une boîte à moustaches[12]. La ligne horizontale donne la valeur médiane. Les boîtes à moustaches bleues

[12] La valeur centrale d'une boîte à moustaches est la médiane. Les bords du rectangle sont les quartiles (pour le bord inférieur, un quart des observations a des valeurs plus petites et trois quarts ont des valeurs plus grandes, le bord supérieur suit le même raisonnement). Les extrémités des moustaches sont calculées en utilisant 1.5 fois l'espace interquartile.

correspondent au groupe “other”, les rouges au clade I, les vertes au clade II et les grises à l’ensemble des espèces.

C- Un chromosome compartimenté

1- Extrémités chromosomiques et adaptation

Les extrémités chromosomiques sont des régions hautement variables chez les *Streptomyces*, tant au niveau interspécifique que intraspécifique. Chaque espèce de *Streptomyces* possède donc un grand nombre de fonctions spécifiques, dont les plus étudiées pour des raisons médicales et économiques sont les voies de biosynthèse de métabolites spécialisés (Bérdy, 2005). Outre leur aspect valorisable, ces voies de biosynthèse pourraient être reliées à l'adaptation de chaque espèce à son environnement. Le sol est un milieu complexe et changeant tant au niveau des paramètres physico-chimiques qu'au niveau des interactions avec d'autres organismes. La diversité des molécules produites, dont certaines possèdent une activité antibiotique, peut être interprétée comme la mise en place d'un arsenal de défense contre les organismes en compétition pour les ressources. De plus, la diversité et la multiplicité des antibiotiques produits par une seule souche peut refléter la diversité des organismes compétiteurs qui peuvent être présents dans la même niche écologique (ex: bactéries Gram négative et Gram positive, champignons). Cependant, il n'est pas exclu que ces molécules jouent un tout autre rôle dans le développement des colonies de *Streptomyces*. L'adaptation à l'écosystème du sol peut également être reliée aux capacités de biosynthèse de nombreuses enzymes extracellulaires responsables de la dégradation de polymères retrouvés en abondance dans l'environnement comme la cellulose ou la chitine.

Les régions terminales sont les loci privilégiés d'acquisition de nouvelles fonctions par transfert horizontal. Elles seraient donc très fortement impliquées dans l'adaptation à l'environnement. Chez *S. coelicolor*, il a été mis en évidence que durant la croissance végétative, les gènes correspondant à la région centrale du chromosome sont plus fortement exprimés que les gènes localisés dans les extrémités chromosomiques (Karoonuthaisiri et al., 2005). Cette observation expérimentale va dans le sens d'une organisation du chromosome de *Streptomyces* où les gènes essentiels à la croissance sont localisés dans la région centrale du chromosome. Cela est cohérent avec l'organisation des gènes du core génome mise en évidence sur l'ensemble de notre collection de 110 génomes (**figure 32**).

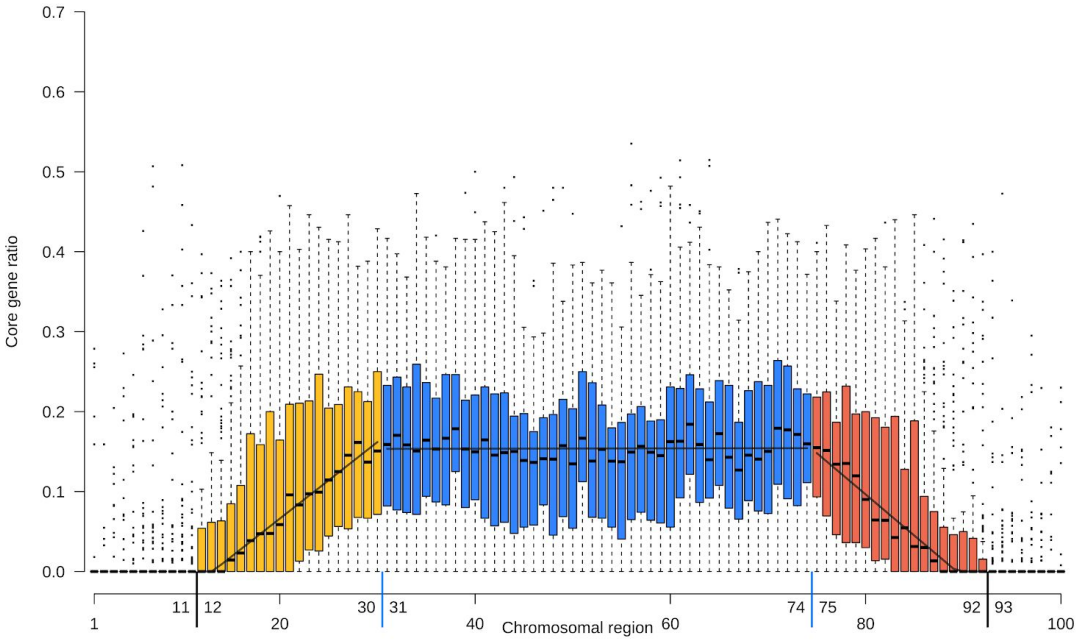

Figure 32- Proportion des gènes core selon leur position dans le chromosome.
Tous les chromosomes de la collection 110 ont été normalisés en subdivisant chaque chromosome en 100 régions successives équivalentes en nombre de gènes. Pour chaque région, le ratio de gènes core a été calculé et présenté sous forme de boîtes à moustaches. 5 régions chromosomiques sont distinguées : les extrémités gauche (régions 1 à 11) et droite (régions 93 à 100), dépourvues de gènes core. La région centrale, en bleue (entre les régions 31 et 74) où le ratio de gène core est le plus élevé et relativement constant. La région jaune (régions 12 à 30) et la rouge (régions 75 à 92) où le ratio de gène core décroît progressivement des bornes de la région bleue vers les extrémités chromosomiques.

Les gènes portés par les bras des extrémités chromosomiques sont, certes, peu exprimés pendant la croissance végétative, mais leur transcription s'accroît pendant la phase stationnaire et au cours de différents stress (Karoonuthaisiri et al., 2005). Les fonctions spécifiques des différentes espèces semblent s'exprimer surtout pendant les phases tardives de croissance. C'est par exemple le cas du métabolisme secondaire qui se met en place au moment de la différenciation du mycélium.

2 - Les TIR ont-elles une fonction ?

Dans la collection 135, des TIR ont été mises en évidence pour 86 chromosomes. Cependant, chez certaines espèces, la duplication terminale est restreinte à la séquence télomérique (ex. *S. avermitilis,* Ikeda et al., 2003). Ainsi, une absence de TIR n'est pas nécessairement une erreur d'assemblage. On peut donc se questionner sur le rôle des TIR chez *Streptomyces*. Les TIR sont extrêmement variables, même à courtes distances phylogénétiques (Choulet et al., 2006b; Weaver et al., 2004) suggérant que ces régions sont sujettes à de nombreux événements de perte et de gain de matériel génétique. Certains auteurs ont mis en avant l'intérêt de posséder des répétitions terminales dans les mécanismes

de restauration des télomères ayant subi des dommages (Uchida et al., 2003). La présence de TIR pourrait n'être qu'une conséquence du processus de réparation des cassures double-brin et des dommages causés aux télomères (Hoff et al., 2018).

La réduction des TIR pourrait, elle, être due à une accumulation de mutations ponctuelles. Chez *S. coelicolor* A3 (2) les TIR mesurent 22 kb mais des souches présentant des TIR de 1,06 Mb ont été isolées en laboratoire. Ces souches correspondraient à la situation ancestrale dans la généalogie des souches de *S. coelicolor* lors de son isolement du sol (Weaver et al., 2004). La réduction de la taille des TIR peut aussi être la conséquence d'un événement de recombinaison intrachromosomique entre deux séquences identiques (Weaver et al., 2004).

3- Compartimentalisation du chromosome de *Streptomyces*

L'analyse du niveau de conservation de la synténie au travers du GOC et du NOC a révélé que la taille de la région centrale conservée entre deux espèces dépend principalement de la distance phylogénétique (**figure 33**). Ainsi, lors de la comparaison entre 2 espèces phylogénétiquement proches, *S. ambofaciens* ATCC 23977 et *S. coelicolor* A3(2), la taille des régions variables pour *S. ambofaciens* représente environ 16% du chromosome (soit environ 1,3 Mb). En considérant une espèce plus distante phylogénétiquement comme *S. reticuli* TUE45, la taille de ces régions représentent environ 25% du chromosome de *S. ambofaciens* (soit environ 2 Mb). Ainsi, pour deux espèces de *Streptomyces* phylogénétiquement éloignées, la taille de leurs régions variables sera plus grandes qu'entre deux espèces étroitement apparentées. Réciproquement, la taille de la région centrale héritée de l'ancêtre commun diminue avec la distance phylogénétique qui sépare les espèces comparées. L'accumulation de la variabilité par les extrémités chromosomiques semble saturer (c'est-à-dire qu'il n'est plus possible de visualiser cette accumulation de la variabilité) avec la distance phylogénétique. En effet, la taille des régions ne varie plus dès lors qu'un score de GOC faible est atteint.

La présence de cette synténie dégénérée, c'est-à-dire la diminution progressive de la conservation de l'ordre des gènes du centre du chromosome vers les extrémités (**figure 33**) montre que les gènes présents dans ces régions ne sont pas issus d'un échange entre extrémités de réplicons linéaires, mais vraisemblablement d'événements de recombinaisons multiples. En effet, le remplacement d'extrémités ferait varier l'ensemble du contenu en gènes depuis le point de recombinaison jusqu'aux télomères. Les régions terminales évolueraient principalement par la fixation d'événements d'insertions/délétions (indels) au cours de l'évolution. La dégénérescence graduelle de la synténie montre que la fréquence de ces événements serait variable suivant la localisation chromosomique et augmenterait à proximité des extrémités. Par saturation en indels, la synténie ancestrale est progressivement effacée. Une hypothèse alternative serait que la fréquence des évènements de

recombinaisons ne change pas le long du chromosome, mais que ces événements sont plus ou moins fixés selon les régions chromosomiques.

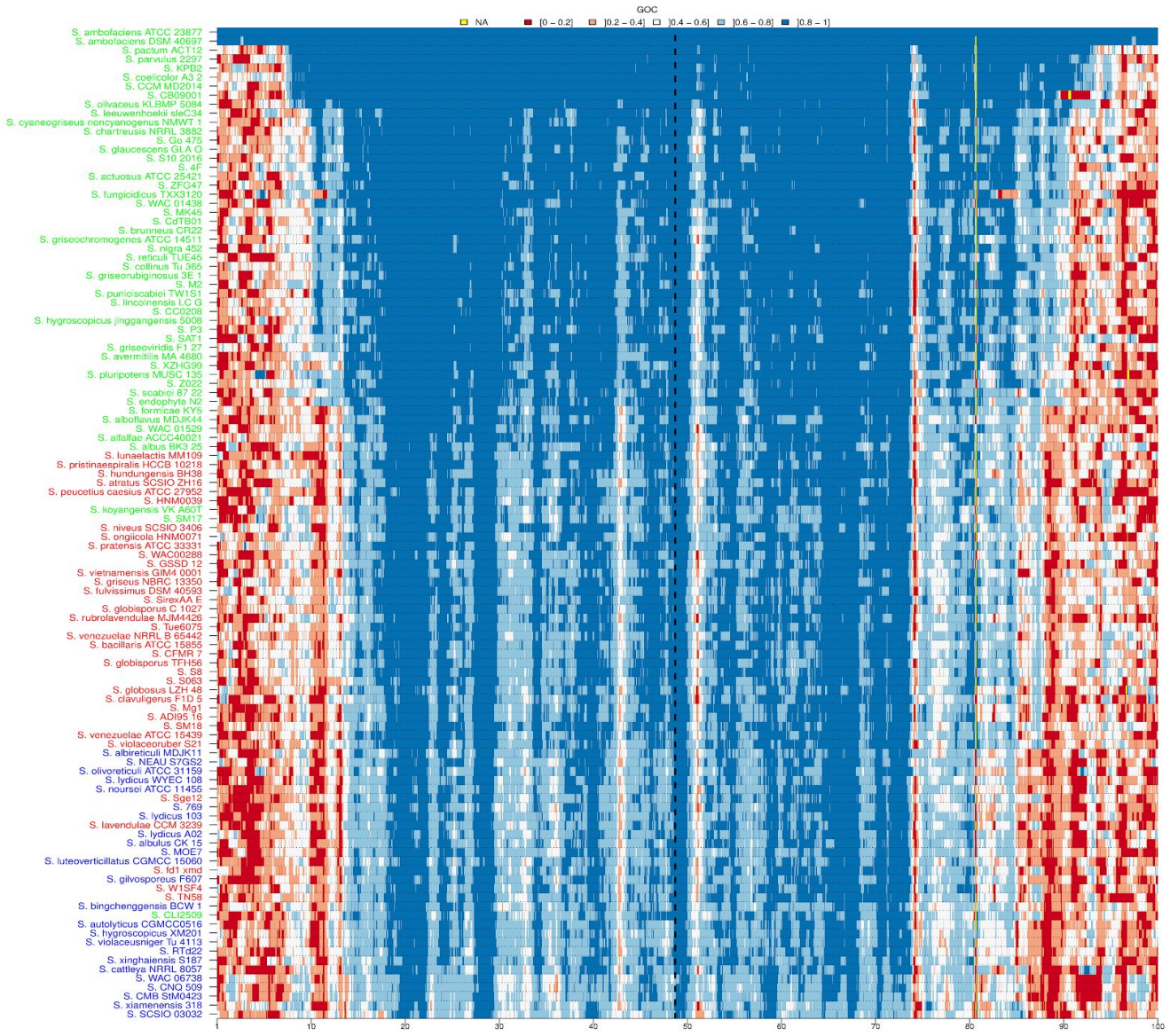

Figure 33 - Carte de chaleur des scores de GOC le long du chromosome de *S. ambofaciens* ATCC 23877 en référence. Chaque ligne correspond au profil de GOC de la référence contre une autre espèce. Les espèces sont organisées de la plus proche à la plus éloignée phylogénétiquement. Le nom de l'espèce est coloré en fonction de son groupe d'espèces (clade I: rouge, clade II: vert, Other: bleu). L'axe des abscisses est exprimé en pourcentage du chromosome de référence. La ligne verticale en pointillées représente l'emplacement du gène *dnaA* délimitant les deux réplichores.

En plus de ces régions variables, il apparaît que la région centrale, classiquement décrite comme conservée, présente de nombreux points où les valeurs de GOC sont faibles, et ce, même à très courte distance phylogénétique comme entre *S. ambofaciens* ATCC 23877 et *S. ambofaciens* DSM 40697 où

une région apparaît en jaune sur la **figure 33** à la position 81%. Une région jaune dans les cartes de chaleur de GOC correspond à une fenêtre de 70 gènes chez *S. ambofaciens* ATCC 23877 qui lui sont spécifiques. Une analyse complémentaire avec l'outil Phaster (Arndt et al., 2016) a mis en évidence que cette région spécifique à *S. ambofaciens* ATCC 23877 correspond à l'insertion d'un phage entre les positions 6,60 et 6,63 Mb du chromosome.

4 - Une région centrale contrainte

Pour estimer le nombre de remaniements de type transposition et inversion entre 2 génomes, la méthode DCJ a été utilisée en considérant les gènes en trois catégories de gènes (**figure 34**) : les gènes appartenant au core-génome du genre, les gènes orthologues n'appartenant pas au core-génome, mais localisés dans la région core et les gènes orthologues localisés en dehors de la région core (c'est-à-dire les bras chromosomiques). Cela a confirmé que les bras chromosomiques présentent une tendance forte à se réarranger. Par ailleurs, dans la région centrale du chromosome, les gènes core et non-core montrent une tendance étonnamment différente : la position relative des gènes core semble être très stable au cours de l'évolution contrairement aux gènes non core de la même région qui semblent, eux, plus fréquemment réarrangés.

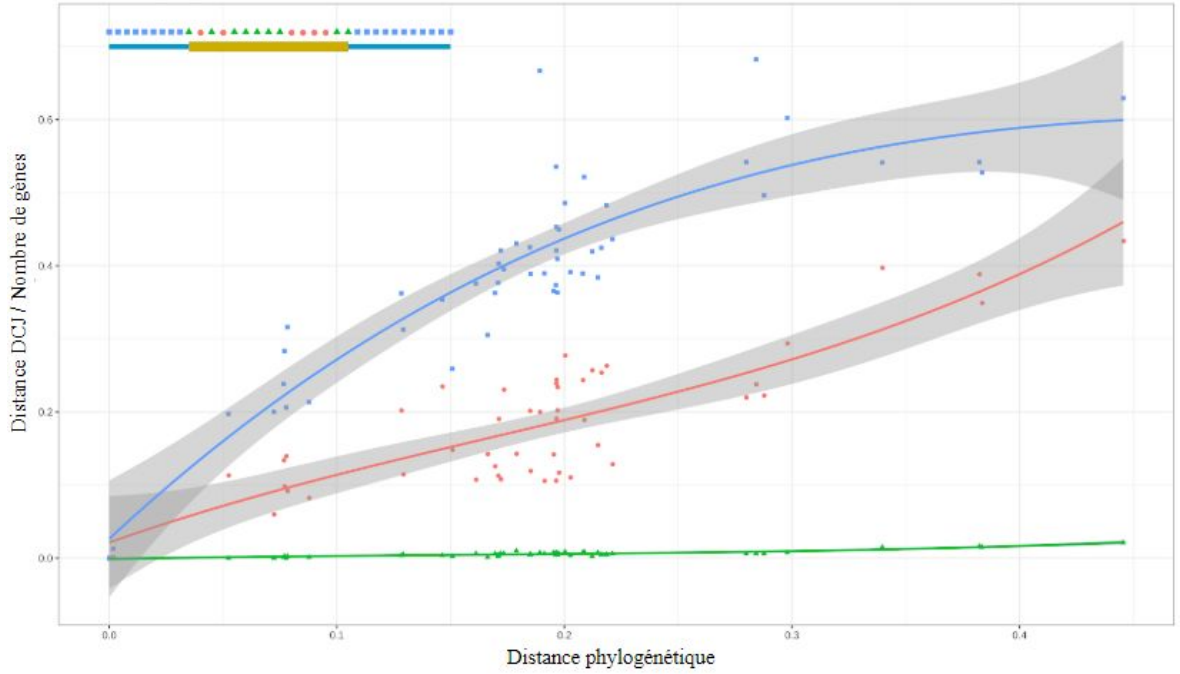

Figure 34 - Représentation de l'évolution de la distance DCJ divisée par le nombre de gènes en fonction de la distance phylogénétique pour trois "catégories" de gènes avec *S. ambofaciens* ATCC 23877 en référence. Le schéma en haut à gauche de la figure illustre ces catégories : la région core du chromosome est en jaune, les bras (régions pauvres en gène du core génome) en bleu. Les carrés bleus représentent les gènes localisés dans les extrémités chromosomiques (en dehors de la région core), les triangles verts représentent les gènes du core-génome et les cercles rouges représentent les gènes non-core dans les régions cores.

Cette observation soulève l'hypothèse que les gènes composant le core génome, probablement hérités verticalement de l'ancêtre commun du genre sont localisés à des endroits clé le long du chromosome. A l'inverse des gènes acquis ou perdus plus récemment, semblent peupler des régions plus sujettes à la recombinaison.

Ces résultats complexifient l'image d'un génome présentant une région stable, le centre du chromosome, et des régions terminales recombinogènes. Ainsi, au coeur même de la région centrale du chromosome, les gènes fortement conservés révéleraient ou formeraient un squelette contraint. La position de ces gènes pourrait avoir été sélectionnée parce qu'elle correspond à des loci stables, moins recombinogènes. Réciproquement, la position de ces gènes, soit parce qu'ils sont essentiels ou fortement exprimés, pourrait définir ces régions stables à travers l'évolution.

Chez *Streptomyces*, lors de la construction d'un support d'expression hétérologue à partir du chromosome de *S. albus* J1074, il a été montré que la position chromosomique du site d'insertion a un impact sur le niveau d'expression des gènes rapporteurs (Bilyk et al., 2017). Étonnement, cet effet n'est pas corrélé à la proximité du site d'insertion avec l'origine de réplication. Chez *E. coli* et *Salmonella* il a été montré que la variation du taux d'expression de gènes rapporteurs est principalement due à l'effet dose (Block et al., 2012; Schmid and Roth, 1987). Par analogie avec les chromosomes bactériens circulaires, dans le chromosome linéaire de *Streptomyces* avec une origine de réplication centrée, si l'effet dose jouait un rôle significatif dans le niveau d'expression hétérologue des gènes rapporteurs, alors plus l'insertion serait localisée près de l'origine de réplication plus le niveau d'expression serait élevé. Or, les mutants obtenus avec des sites d'intégration aux environs de l'origine de réplication présentent des taux d'expression en deçà de la moyenne (Bilyk et al., 2017). Les mutants montrant le plus haut taux d'expression présentent une insertion à 1Mb minimum de l'origine de réplication (**figure 35**), dans une région que les auteurs nomment les "épaules" du chromosome.

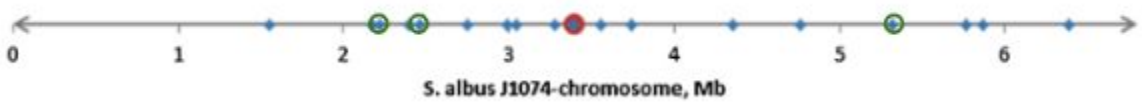

Figure 35- Distribution des loci d'insertion de la cassette de gène dans *S. albus* J1074. Les losanges bleus représentent les positions des sites d'insertion, le cercle rouge représente l'*oriC* et les cercles verts représentent les loci où le taux d'expression hétérologue est maximal.

En observant les courbes de GOC ou du taux d'orthologues chez différents couples d'espèces de *Streptomyces* (**figure 36**), il semble que les régions les plus conservées (c'est-à-dire les endroits avec

les scores de GOC et OR les plus élevés) ne soient pas aux environs de l'origine de réplication, mais en bordure de la région core.

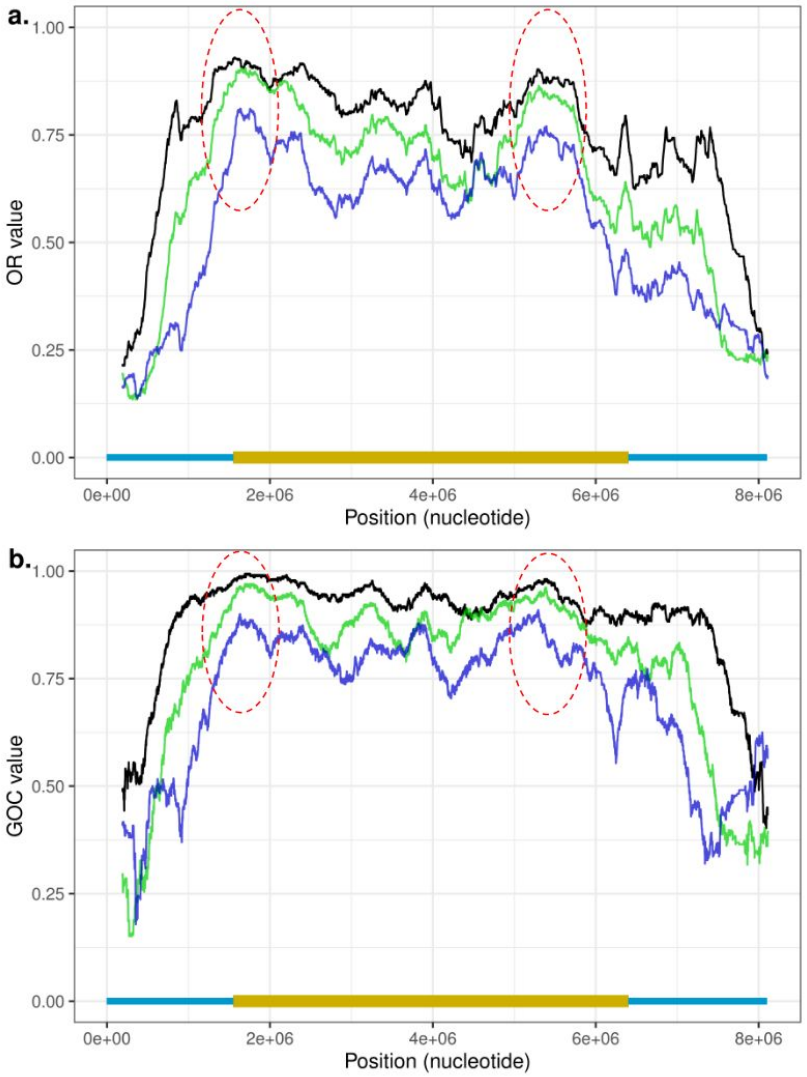

Figure 36 - Profils de OR (**a.**) et de GOC (**b.**) le long du chromosome de *S. ambofaciens* ATCC 23877.
Les profils sont obtenus en traçant les valeurs GOC ou OR obtenues par comparaison du chromosome de *S. ambofaciens* ATCC 23877 avec 3 espèces montrant une distance phylogénétique proche (vs. *S. coelicolor* A3(2), courbe noire), et moyenne (*S. reticuli* TUE45, courbe verte), à distance(*S. albus* BK3 25, courbe bleue). Les valeurs de GOC et OR ont été calculées en utilisant une fenêtre glissante de taille égale à 5 % du contenu en gènes chromosomiques de *S. ambofaciens* ATCC 23877. Les ellipses rouges indiquent la région contenant le maximum de GOC ou OR pour chaque réplichore.

5- Mécanismes de recombinaison et plasticité des régions terminales des *Streptomyces*

Les régions terminales du chromosome des *Streptomyces* subissent de nombreux événements de réarrangements et nos connaissances sur l'impact de différents mécanismes de recombinaison sur la plasticité génomique des *Streptomyces* ont récemment progressé.

La recombinaison homologue (RH) est un mécanisme qui contribue à la stabilité du génome en tant que mécanisme de réparation fidèle, cependant, il a été montré chez *S. ambofaciens* ATCC 23877 que la RH constitue également un des mécanismes principaux de la plasticité de la structure du génome chez les *Streptomyces* (Hoff et al., 2018). La réparation des DSB dans les régions terminales est régulièrement associée à la perte du bras où la DSB a été introduite. Deux régions répétées présentes sur les deux bras en orientation opposée permettent à la recombinaison homologue d'initier soit une recombinaison avec le bras opposé d'un autre chromosome présent dans le compartiment, soit avec le bras opposé du même chromosome. Dans le premier cas, la recombinaison conduit à l'acquisition d'un bras intact, dans le second cas, elle peut initier un événement de “Break Induced Replication” (BIR) (Anand et al., 2013). La survenue de ces événements de BIR est dépendante de la RH et pourrait être facilitée par la configuration spatiale du chromosome. En effet, il existe une interaction entre les protéines terminales fixées au niveau des télomères, permettant de maintenir une proximité physique entre les télomères, ce qui pourrait favoriser des évènements de RH non allélique entre les bras. Ce mécanisme a un impact important sur la plasticité du génome puisque dans certains cas, il peut impliquer des remaniements majeurs du chromosome (perte ou gain de plusieurs centaines de kb).

La voie NHEJ a longtemps été considérée comme propre aux eucaryotes, ce qui suggère que la réparation des DSB chez les procaryotes s'appuie entièrement sur les RH. Cependant, les premières études bioinformatiques ont permis d'identifier des orthologues des acteur majeurs de la voie NHEJ (les protéines Ku) dans plusieurs génomes bactériens phylogénétiquement distants (Aravind and Koonin, 2001; Doherty et al., 2001). Les *Streptomyces* possèdent un grand nombre de gènes de type NHEJ (Hoff et al., 2016) et sont capable d’utiliser cette voie de recombinaison.
Le nombre d’évènements de recombinaison illégitime (type NHEJ) est, a priori, plus faible que ceux de RH. Cependant, les conséquences sur la structure du génome ne sont pas moins importantes. L’existence de mécanismes de recombinaison illégitime est un outil formidable pour l’augmentation de la diversité génomique au sein d’une population. En effet, les mécanismes de recombinaison illégitime n’ont pas la contrainte d’utilisation d’une matrice homologue, et la présence de micro-homologies ou simplement de deux extrémités franches permet la réparation d’une molécule d’ADN rompue. Ainsi, contrairement à la RH où les événements sont contraints par la localisation et l’orientation des séquences répétées, la recombinaison illégitime peut engendrer des structures chromosomiques très diverses.

PERSPECTIVES

Ces travaux ont permis d'améliorer notre compréhension de l'organisation et de la dynamique du chromosome du genre *Streptomyces* au cours de l'évolution.

L'organisation chromosomique en une région "core" centrale regroupant les gènes du core-génome bordée des bras chromosomiques a été identifiée comme l'organisation typique du chromosome des *Streptomyces*. Cette compartimentation du chromosome est associée à une tolérance et/ou une fréquence plus ou moins importante aux événements de réarrangements selon les régions chromosomiques. En effet, en considérant uniquement la position des gènes du core-génome ne suffit pas pour définir l'organisation du chromosome des *Streptomyces*. Les derniers résultats obtenus durant ces travaux ont mis évidence une organisation où les loci des gènes du core génome constituent un squelette ancestral, peu sujet aux évènements de remaniements chromosomiques. Entre ces différents loci, le nombres de remaniements est étonnamment élevé. De plus, en recherchant, pour chaque réplichore du chromosome de *Streptomyces*, la région où la conservation de la synténie est la plus importante, il apparaît que cette région est localisé en fin de la région centrale. Des travaux pour généraliser cette observation sont encore à effectuer. En particulier en intégrant des informations sur l'organisation 3D du chromosome des *Streptomyces*.

MATERIELS ET METHODES

Cette partie vise à présenter l'ensemble des méthodes et outils de génomique comparée développés et utilisés au cours de ma thèse. Seul l'aspect technique est présenté ici, les différentes figures ne sont pas interprétées et servent uniquement à illustrer les méthodes. La discussion du choix de ces approches et des paramètres est réduite au minimum et est effectuée dans la partie RESULTATS auxquelles elles se rapportent.

I- Matériels

Tous les travaux ont été effectués au sein de l'équipe Bio-Informatique Moléculaire au 1er étage du bâtiment 400 du campus de l'Université Paris Sud. Un bureau (101A), un ordinateur (Ubuntu 18.04.02 LTS, 16GiB de mémoire, Intel® Core™ i7-6700K CPU @ 4.00GHz × 8) ont été mis à ma disposition ainsi que 2 serveur de calculs mis à disposition par l'équipe BIM (BIMCAL: Ubuntu 14.04.5 LTS, 250 GiB de mémoire, processeur Intel® Xeon® CPU E5-46200@2.20GHz × 64 et CALBIM: Ubuntu 14.04.5 LTS, 250 GiB de mémoire, processeur Intel® Xeon® CPU E5-46200@2.20GHz × 48).

Les scripts produits ont été développés dans trois langages de programmation différents selon leur fonction :

- Python 3: exploitation des données, comparaisons de séquence.
- R: représentation graphique et analyse statistique.
- bash: automatisation des appels de scripts et appel à des logiciel tiers.

II- Outils bio-informatiques

A- Qualité et mise en forme des génomes

Cette section regroupe toutes les méthodes et outils utilisés pour estimer et corriger la qualité des génomes utilisés durant ces travaux.

1- BUSCO

a- Définition

BUSCO (Simão et al., 2015) est un outil permettant d'estimer la complétude de nouveaux génomes séquencés à partir de leur génome, protéome ou transcriptome. Ici, l'estimation a été faite à partir des protéomes, seul ce protocole est donc détaillé ici.

Le groupe d'orthologues en copie unique (ou groupe BUSCO) pour un niveau phylogénétique donné (Actinobactérie par exemple) est construit à partir de la base de données OrthoDB[13] et est défini comme l'ensemble des gènes présents en copie unique dans au moins 90% des espèces du niveau phylogénétique considéré. L'outil HMMER (Finn et al., 2015) est ensuite utilisé pour rechercher des correspondances entre les groupes BUSCO et chaque séquence du protéome. Toutes les correspondances identifiées sont ensuite classées comme "Complet" si leur longueur correspond aux attentes de la longueur de correspondance du profil BUSCO. Si ceux-ci sont trouvés plus d'une fois, ils sont classés dans la sous-catégorie «Duplicated». S'ils sont uniques, dans la sous-catégorie "Single-copy". Les correspondances qui ne sont que partiellement retrouvées sont classées comme «Fragmented» et les groupes BUSCO pour lesquels aucune correspondance ne réussit les tests d'orthologie sont classés comme "Missing".

b- Représentation graphique

Une représentation est proposée par l'outil BUSCO via l'intermédiaire d'un script R automatiquement généré. Ce graphique (**figure 37**) permet de visualiser pour chaque génome la répartitions des groupes BUSCO dans les catégorie "Missing", "Fragmented", "Complete and Duplicated" et "Complete and Single-copy". Cette représentation permet de comparer visuellement tous les génomes et d'en extraire leurs scores BUSCO, cependant ce type s'avère peu lisible pour de grands jeux de données. Une autre figure (**figure 21**) utilise une représentation sous forme de boxplot où la répartition de score BUSCO est représenté selon le code couleur "Complet and Single-copy (S)" = bleu clair, "Complete and Duplicated (D)" = bleu foncé, "Fragmented (F)" = jaune et "Missing (M)" = rouge.

[13] https://www.orthodb.org/

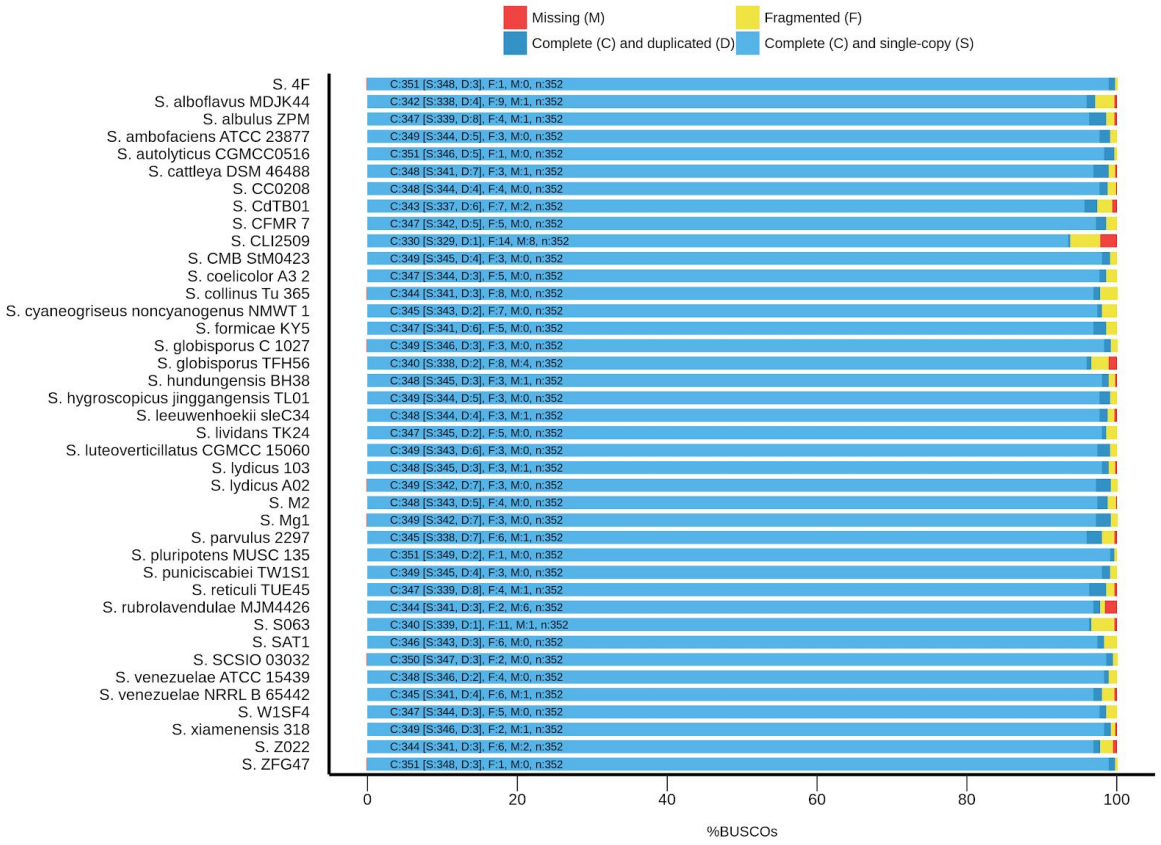

Figure 37 - Représentation graphique des scores BUSCO obtenu pour un sous-ensemble de 20 génomes de *Streptomyces*. Chaque ligne correspond à un génome l'abscisse aux pourcentages des familles BUSCOs où leur catégorie est identifié par un code couleur: "Complet and Single-copy" = bleu clair, "Complete and Duplicated" = bleu foncé, "Fragmented" = jaune et "Missing" = rouge.

2- Détection des TIR

a- Définition

Les TIR correspondent à des régions parfaitement dupliquées et inversées localisées aux extrémités des chromosomes linéaires de *Streptomyces* rendant leur détection aisée par similarité de séquences. Le premier mégabase (dimension du plus grand TIR identifié à ce jour dans une souche mutante de *S. coelicolor* A3(2) (Weaver et al., 2004)) de chaque extrémité a été extrait pour chaque chromosome et aligné avec l'outil BLASTn. Les TIR ont été ensuite mis en évidence en identifiant les régions présentant au minimum 99% d'identité sur l'ensemble de l'alignement et commençant à une extrémité.

Le seuil d'identité n'a pas été fixé à 100%, car il est apparu que le changement d'une simple base interrompant la similarité de 2 extrémités peut fortement impacter sur la dimension finale des TIR. Par exemple, chez *S. collinus* Tu 365 présentant des TIR de 631,370 pb, l'alignement entre les deux

extrémités est interrompu par la délétion d'un unique nucléotide à la position 90,533, aucun autre mésappariement n'a été détecté avant la position 630,196. Dans ce cas, la dimension finale des TIR varie d'un facteur 7 en étant strict sur la définition des TIR. Le choix d'un score d'identité à 99% permet donc de "lisser" de potentielles erreurs de séquençage.

b- Représentation graphique

Aucune représentation graphique spécifique aux TIR n'a été développée, cependant pour les chromosomes où des TIR ont été identifiées, une seule copie est conservée pour effectuer les différents travaux de génomiques comparées (voir section **Mise en forme finale des génomes**). Lors de la représentation des résultats, les régions correspondantes aux TIR sont manuellement dupliquées pour correspondre au mieu à l'organisation réelle du chromosome analysé.

B- Détermination des relations évolutives entre les gènes

Ce point regroupe toutes les méthodes développées touchant à la conservation et perte de gènes. Pour chaque outil développé, la méthode est divisée en deux points : (a) définition et explication de la méthode utilisée et (b) Représentation des méthodes graphiques directement associées.

1- Les orthologues

a- Définition

La définition des orthologues a été effectué selon la méthode BBH (Bidirectional Best Hit) : deux séquences protéiques ont été considérées orthologues lorsqu'elles présentent réciproquement les meilleurs scores d'alignement BLASTp entre les deux génomes comparées (**figure 38**). En plus de la réciprocité des meilleurs résultats BLASTp, les séquences doivent partager au minimum 40% d'identité, s'aligner sur au moins 70% de la séquence la plus petite et présenter une E-value inférieure à 1^{-10}. Ces seuils ont été définis empiriquement de manière à maximiser la sélectivité (obtenir le moins de faux positifs possible).

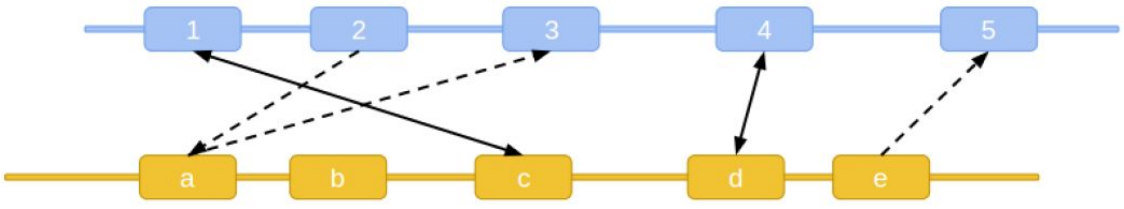

Figure 38 - Schéma de la définition des paires d'orthologues selon la méthode BBH. 2 chromosomes linéaires sont représentés (bleu et jaune), sur chacun d'eux les gènes sont identifiés par des rectangles. Les flèches indiquent les relations misent en évidence entre ces gènes (flèches simples en pointillés : relation non-réciproque de séquences; double flèches : relation réciproque de séquences assimilé à une relation d'orthologie). Dans cet exemple, les gènes 1 et c ainsi que 4 et d sont identifiés comme des paires d'orthologues.

b- Représentation graphique

La visualisation des paires d'orthologues a été abordé de deux manières différentes, sous forme de "dot plot" simple qui permet de représenter les positions relatives des paires de gènes orthologues entre deux chromosomes sur un graphe à deux dimensions (**figure 39**). Cette première représentation permet d'avoir une vue globale de la conservation des gènes entre deux espèces et de mettre en évidence certains grands événements de remaniements chromosomique comme les inversions, mais apporte peu d'informations quantitative.

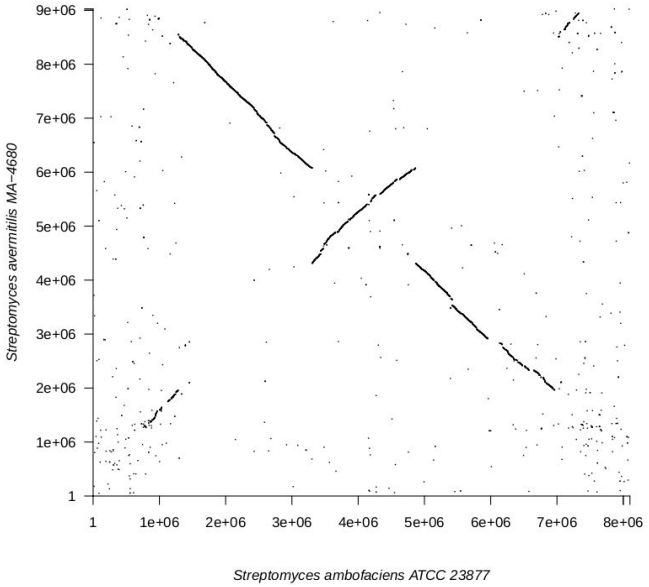

Figure 39 - Représentation sous forme de "dot plot" de la comparaison des paires d'orthologues entre *S. ambofaciens* ATCC 23877 et *S. avermitilis* MA-4680. Les coordonnées de chaque point correspondent aux positions nucléotidiques du couple de gènes orthologues.

La représentation sous forme d'histogramme (**figure 40**) représente le nombre de gènes possédant un orthologue (barre bleue) dans le chromosome cible et le nombre total de gènes (barre grise) dans une région du chromosome de référence. Une région représente 1% de la dimension totale (en nucléotide) du chromosome.

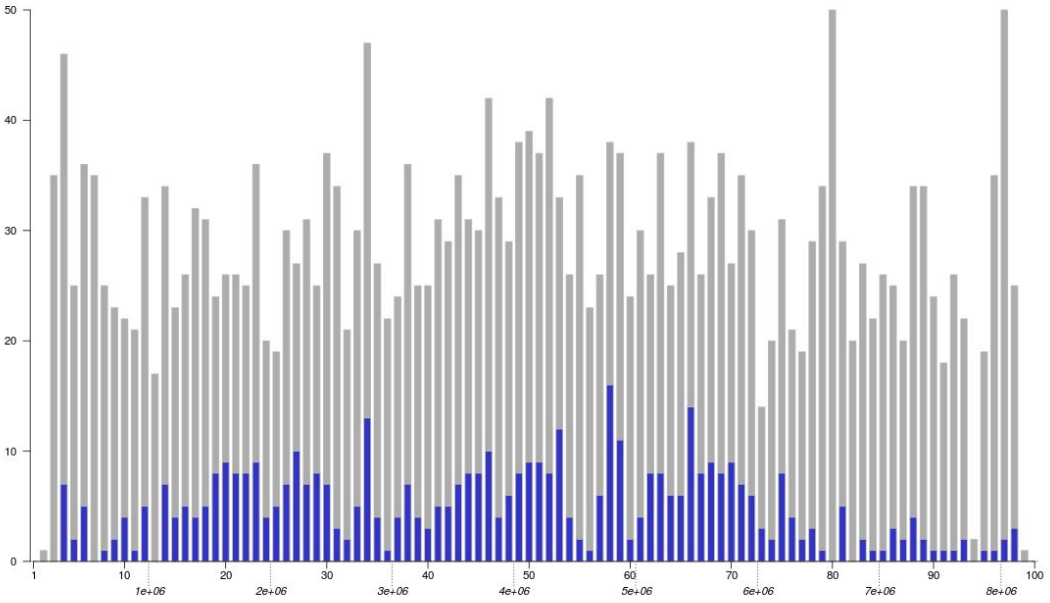

Figure 40 - Représentation sous forme d'histogramme du nombre d'orthologues (barres bleues) et du nombre total de gènes (barres grises) par région de 1% de la dimension totale du chromosome de *S. ambofaciens* ATCC 23877 comparé au chromosome de *S. S. avermitilis* MA-4680. Les régions sont numérotées de 1 à 100 en partant de l'extrémité gauche du chromosome de référence. L'équivalence en position nucléotidique est aussi indiquée sur l'axe des abscisses.

2- Les gènes dupliqués

a- Définition

Les duplications au sein d'un génome ont été identifiées à partir des résultats d'alignement des séquences protéiques d'un génome contre lui-même (**figure 41**). Pour une séquence donnée, toutes les séquences divergeant au maximum de 25% de la taille de la séquence de référence, présentant au moins 80% d'identité, sur 80% de la séquence la plus petite et présentant une E-value inférieure à 1^{-10} sont identifiées comme duplication de la séquence de référence. La réciprocité des alignements au-delà des seuils n'est pas requise pour identifier deux séquences comme dupliquées.

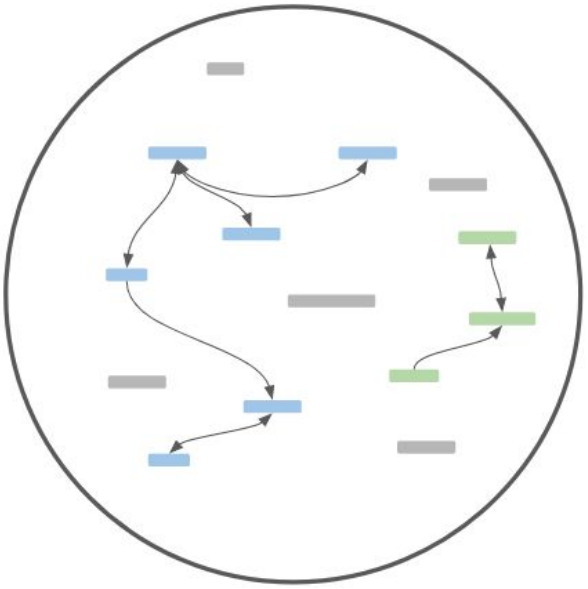

Figure 41 - Schéma du protocole d'identification des familles multigéniques. Chaque rectangle identifie une séquence protéique, les rectangles bleus et verts identifient des gènes appartenant à 2 familles multigéniques différentes, les gris des gènes uniques. Les flèches orientées indiquent les alignements BLASTp satisfaisant les seuils identité >= 80%; alignement >= 80%; E-valuer <= 1^{-10}; variation de taille entre les deux séquences <= 25%.

b- Représentation graphique

La représentation graphique des familles multigéniques a été effectuée sous forme d'un histogramme où chaque barre représente une région de 1% de la dimension totale du chromosome (**figure 42**) et la hauteur indique le nombre de gènes dupliqués dans cette région.

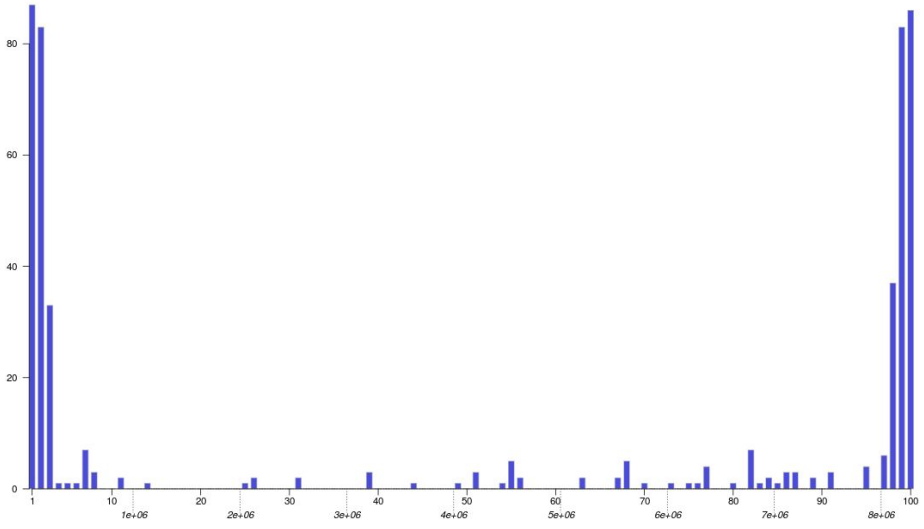

Figure 42 - Représentation sous forme d'histogramme du nombre de gènes dupliqués le long du chromosome de *S. ambofaciens* ATCC 23877. Les régions sont numérotées de 1 à 100 en partant de l'extrémité gauche du chromosome. L'équivalence en position nucléotidique est aussi indiquée sur l'axe des

abscisses. La hauteur des barres indique le nombre de gènes dupliqués par région.

3- Le core et pan-génome

a- Définition

L'ensemble des gènes conservés par toutes les espèces de la collection (core-génome) a été déterminé à partir des paires d'orthologues identifiées par BBH. De la même manière que pour l'identification des relations d'orthologie, cette approche vise à maximiser la sélectivité dans la détection des gènes appartenant au core-génome. Pour cela, un gène est considéré comme appartenant au core-génome s'il possède un orthologue dans toutes les autres espèces de la collection et que toutes les paires d'orthologues sont reliées par une relation d'orthologie, formant ainsi un graphe complet (**figure 43**).

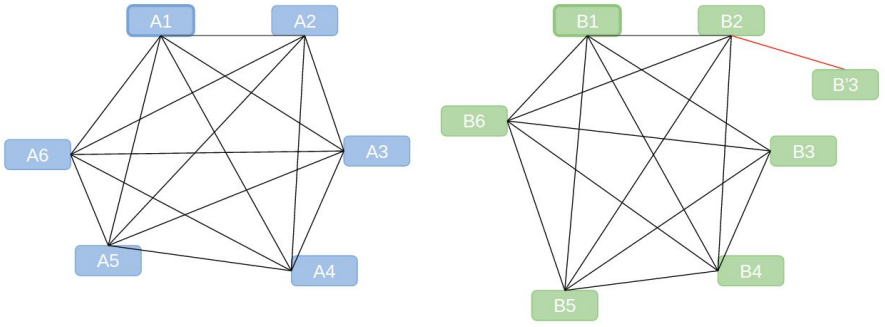

Figure 43 - Schéma de la méthode de construction du core génome. Chaque sommet correspond à un gène identifié par une lettre et un chiffre. Le chiffre indique à quelle espèce appartient le gène (1 à 6). Une arête indique une relation d'orthologie entre 2 sommets. Dans le graphique de gauche, tous les couples de sommets sont reliés par une arête : A1 (et ses orthologues) appartient au core-génome. Dans le graphique de droite, le gène B2 ne possède pas le même orthologue que B1 dans l'espèce 3: le graphe n'est pas complet, B1 (et ses orthologue) n'appartient pas au core-génome

Le pan-génome, c'est-à-dire l'ensemble des gènes de la collection a été obtenu en regroupant tous les gènes partageant des relations d'orthologie directes ou transitives.

Avec ces définitions, l'évolution du nombre de gènes appartenant respectivement au core et pan-génome en fonction du nombre de génomes a été déterminé en considérant des sous-ensembles aléatoires de la collection totale de génomes. 100 tirages aléatoires par sous-ensemble de dimension 2 à N-1 génomes (où N est le nombre de total de génomes de la collection) ont été effectués.

b- Représentation graphique

Plusieurs représentations graphiques ont été développées pour le core et pan-génome afin de répondre à différentes questions posées et à la nature de la collection d'espèces analysées.

Les gènes du core-génome ont été représentés sous forme d'histogramme, de manière similaire à la représentation en histogramme de la localisation des gènes orthologues pour une paire de chromosomes données.

L'évolution du nombre de gènes appartenant respectivement au core et pan-génome en fonction du nombre de génomes a été représenté sous forme de "box plot" où pour chaque dimension utilisé, la médiane et la répartition en quartiles ont été présentées.

Mettre en évidence le nombre de gènes partagés par des sous-ensembles de génomes se visualise facilement sous forme de diagramme de Venn. Cependant, la représentation sous forme de diagramme de Venn est impossible pour un ensemble contenant de nombreux éléments. En effet, le nombre de sous-ensembles possible d'un ensemble de dimension n est de 2^n-1 (le sous-ensemble vide n'est pas considéré ici : un gène appartient au minimum à un génome). Pour contourner cette contrainte, le nombre de gènes composants le pan-génome a été représenté sous forme d'un diagramme de Venn alternatif, selon la méthode "UpSet" (Conway et al., 2017; Lex et al., 2014). Les gènes composants le pan-génome sont au préalable organisés en profil phylogénétique, c'est à dire en une matrice où chaque ligne correspond à un gène et les colonnes aux génomes. Chaque case indique si je le gène *i* est présent dans le génome *j* (Pellegrini et al., 1999). Le package R, "UpSetR" (Conway et al., 2017) permet de visualiser les intersections entre les lignes et les colonnes, c'est-à-dire dans notre cas au nombre de gènes partagés par tous les sous-ensembles possibles de génomes (**figure 44**). Une matrice est générée où chaque génome correspond à une ligne et les sous-ensembles de génomes aux colonnes. Dans une colonne donnée, les génomes appartenant à ce sous-ensemble sont identifiés par des cercles noirs et reliés par un trait noir vertical, les génomes en dehors de ce sous-ensemble sont identifiés par un cercle gris. Le nombre de gènes gènes partagés pour un sous-ensemble donné est représenté sous forme de diagramme en barres au-dessus de la matrice de sous-ensembles. La hauteur des barres correspond au nombre de gènes partagés spécifiquement par le sous-ensemble de génomes alignés.

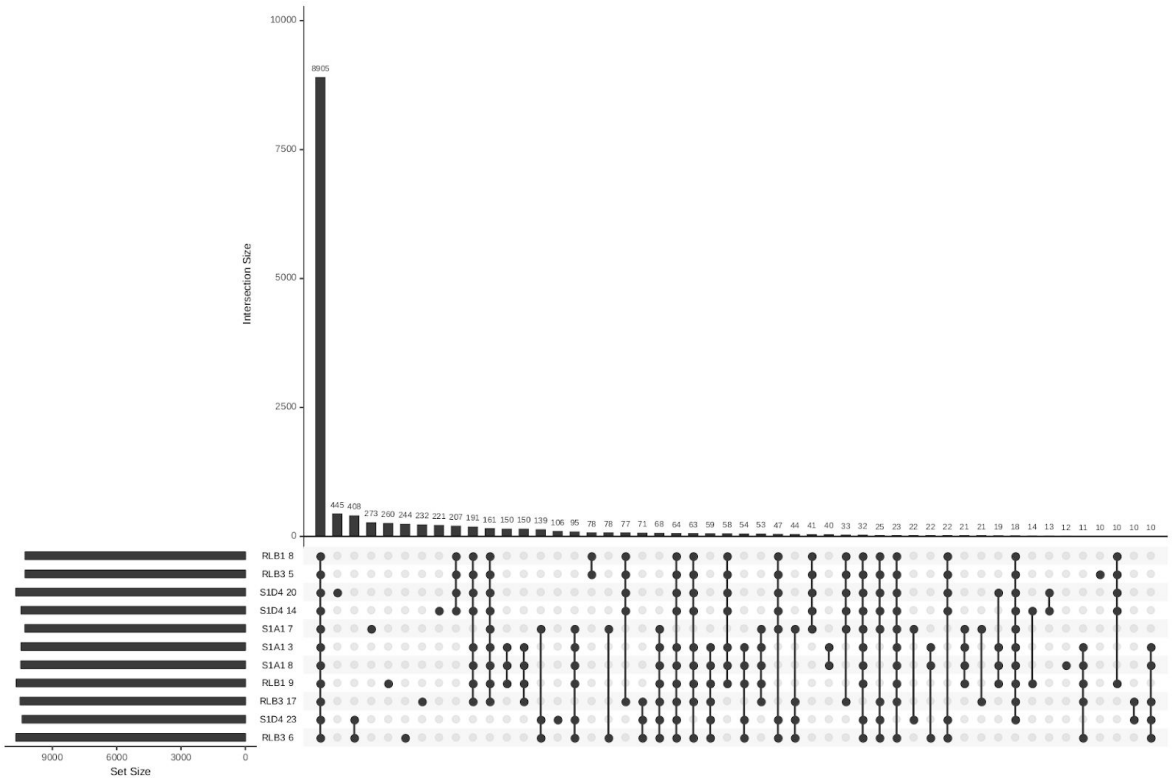

Figure 44 - Représentation selon la méthode "UpSet" des 50 sous-ensembles de génomes présentant le plus de gènes en commun (organisé dans l'ordre décroissant). Les barres horizontales à gauche permettent d'estimer le nombre de gènes contenu dans chaque génome.

B- Evolution de la synténie

1- La conservation de l'ordre des gènes (GOC et NOC)

a- Définition

L'estimation de la synténie c'est-à-dire la conservation de l'ordre des gènes entre deux génomes a été estimé le long d'un chromosome de référence contre un chromosome cible. L'idée de cette approche est d'indiquer, localement, si la synténie est conservée. Cette approche a été développée à partir de l'indice GOC (Gene Order Conservation) proposé par E. Rocha (Rocha, 2003, 2006) comme un estimateur de synténie globale entre deux génomes. Cet indice a été repris F. Choulet (Choulet et al., 2006a), il correspond au GOC "global" défini par E. Rocha, mais est calculé sur une portion du chromosome. Cette portion correspond à une fenêtre définie en nombre de gènes et été a utilisé ici pour visualiser localement l'évolution de la synténie dans les extrémités chromosomiques de *S. ambofaciens* par rapport à des espèces proches. Ici, cette stratégie a été reprise de manière à a estimer

la conservation de la synténie le long d'un chromosome de référence qui est comparé à un second chromosome.

Une formule d'estimation de la synténie a été utilisée: le GOC selon une fenêtre glissante (**figure 45**).

$$GOC = \frac{\#\, ort_{contigus}}{\#\, ort}$$

- $\#\, ort_{adjacents}$ correspond au nombre de paires de gènes qui sont orthologues et contigus dans les 2 génomes comparés dans la fenêtre glissante.
- $\#\, ort$ correspond au nombre de gènes du génome de référence ayant un orthologue dans le génome cible.

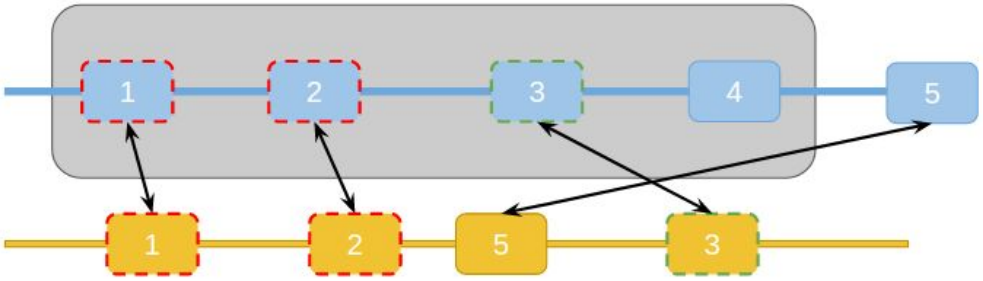

Figure 45 - Schéma de la méthode de calcul du *GOC*. Deux chromosomes linéaires sont représentés (bleu et jaune), sur chacun d'eux, les gènes sont identifiés par des rectangles. Les double flèches indiquent les paires d'orthologues entre les deux chromosomes. Le rectangle gris correspond à la fenêtre glissante (4 gènes). Dans cet exemple, seul les gènes 1 et 2 du chromosome bleu sont adjacents et respectivement orthologues aux gènes adjacents 1' et 2' du chromosome jaune. Sur cet exemple le GOC est donc égal à 2/3 (il n'y a que 3 gènes possédant un orthologue dans la fenêtre).

La fenêtre d'analyse a été définie en nombre de gènes (dans le chromosome de référence), ce nombre est défini en fonction du nombre total de gènes contenu dans le chromosome de référence. Les valeurs de 1% et 5% du nombre de total de gènes du chromosome de référence ont été utilisées, générant respectivement une fenêtre d'analyse de 75 et 375 gènes pour un génome de 7,500 gènes (nombre moyen de gènes dans un chromosome de *Streptomyces*). Le chromosome est parcouru de l'extrémité à gauche à l'extrémité droite par une fenêtre mobile, glissant de manière continue par pas de un gène. Les paires de gènes orthologues ont été au préalable identifiées selon la méthode du BBH décrite dans la **section MATERIELS ET METHODES II.A.1.** La méthode BBH implique qu'un gène ne peut posséder qu'un seul orthologue dans un génome cible, et ne considère donc pas les gènes dupliqués, facilitant la mise en évidence des paires d'orthologues adjacentes entre les 2 génomes.

L'utilisation des deux dimensions de fenêtres permettent d'avoir deux niveaux d'analyses de la synténie, celle de 1% est très sensible au moindre changement de la conservation de l'ordre des gènes entre les deux chromosomes qui induisent de brusque variations dans le score de *GOC* d'une fenêtre à

l'autre. Avec une fenêtre de 5%, considérant plusieurs centaines de gènes, la sensibilité est plus faible, mais permet de suivre l'évolution globale de ces indices le long du chromosome de la référence.

b- Représentation graphique

La représentation graphique des scores de GOC a été effectuée selon 2 approches différentes. L'une visant à comparer spécifiquement deux chromosomes, avec un maximum de précision, l'autre permettant de visualiser l'évolution de la synténie le long d'un chromosome de référence envers plusieurs chromosomes, mais avec une information de type "gros grains".

(i) Comparaison d'un couple de chromosomes

La comparaison d'un couple de chromosomes a été effectuée sous forme de graphe à 2 dimensions où l'axe des abscisses correspond aux positions nucléotidiques des fenêtre de calcul du GOC sur le chromosome du génome de référence. La position nucléotidique d'une fenêtre est réduite la position du gène localisé au centre de la fenêtre. L'axe des ordonnées aux valeurs de GOC obtenues générant un profil sur toute la longueur du chromosome (**figure 46**).

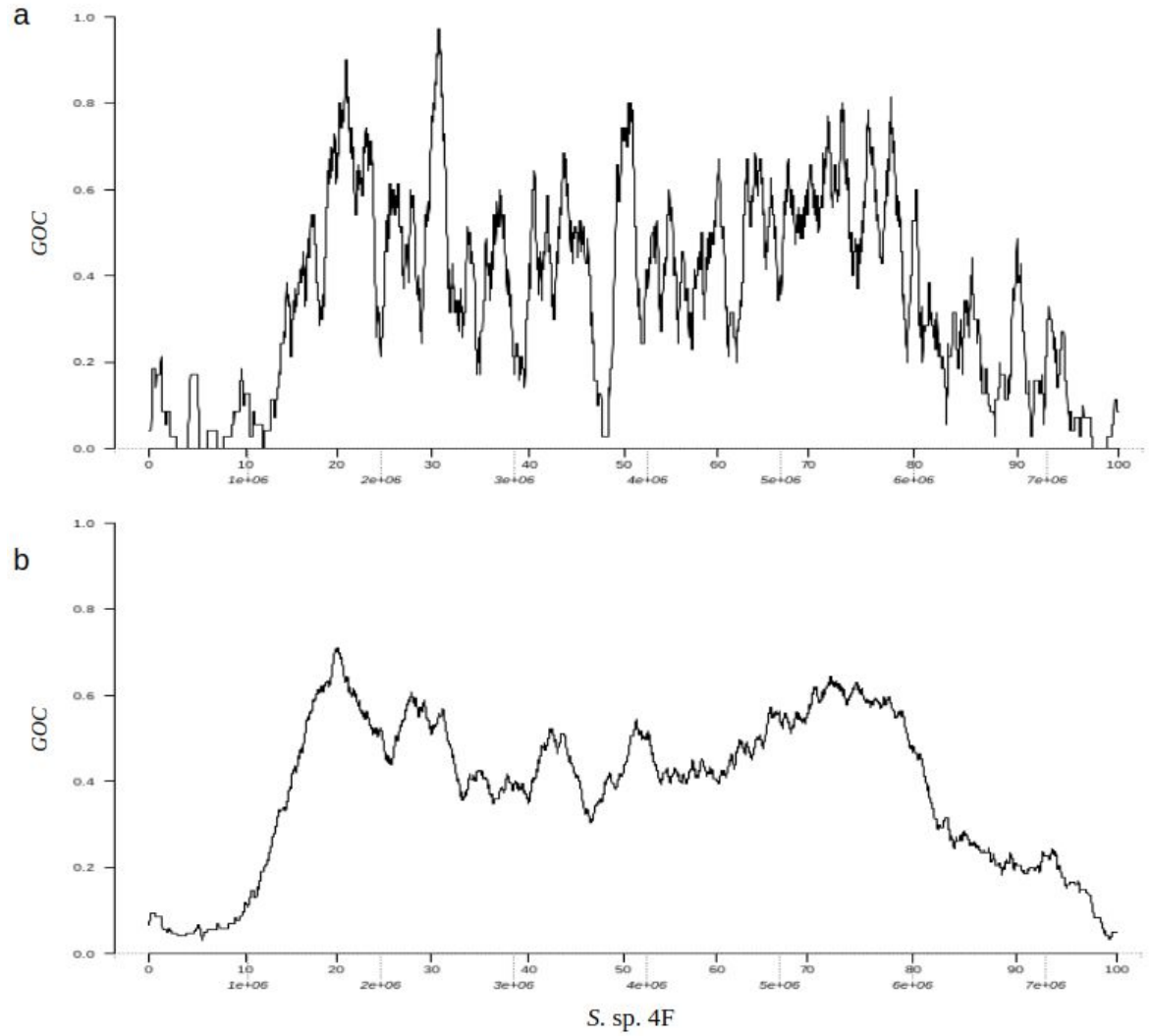

Figure 47 - Profil de *GOC* obtenu en comparant le chromosome de *S.* sp. 4F (référence) et *S. noursei* ATCC 11455. Le graphique **(a)** a été obtenu avec une de 1% du nombre total de gènes de *S.* sp. 4F, le graphique **(b)** avec une fenêtre de 5%. L'axe *y* correspond aux valeurs de *GOC*, l'axe *x* aux positions nucléotidique des fenêtres de calcul du *GOC* (réduite à la position du gène central) et aux positions relatives des fenêtres de *GOC* (en %).

(ii) Comparaison d'un ensemble de chromosomes

Pour représenter le profil de GOC d'un chromosome de référence contre un ensemble de chromosomes en un unique graphique une représentation sous forme de carte de chaleur a été choisie. Chaque profil de GOC a été converti selon une échelle de couleurs. A chaque fenêtre, une couleur est associée selon l'intervalle dans laquelle est contenu la valeur de GOC. Les profils des comparaisons de couple de chromosomes (toujours la même référence) sont organisés selon leur distance phylogénétique en une unique carte de chaleur (**figure 33**), de la plus proche à la plus éloignée. Avec cette méthode, chaque chromosome de la collection peut être comparé à tous les autres en un unique

graphique, permettant d'analyser l'évolution de la synténie d'un chromosome envers tous les autres en tenant compte de la distance phylogénétique entre les espèces.

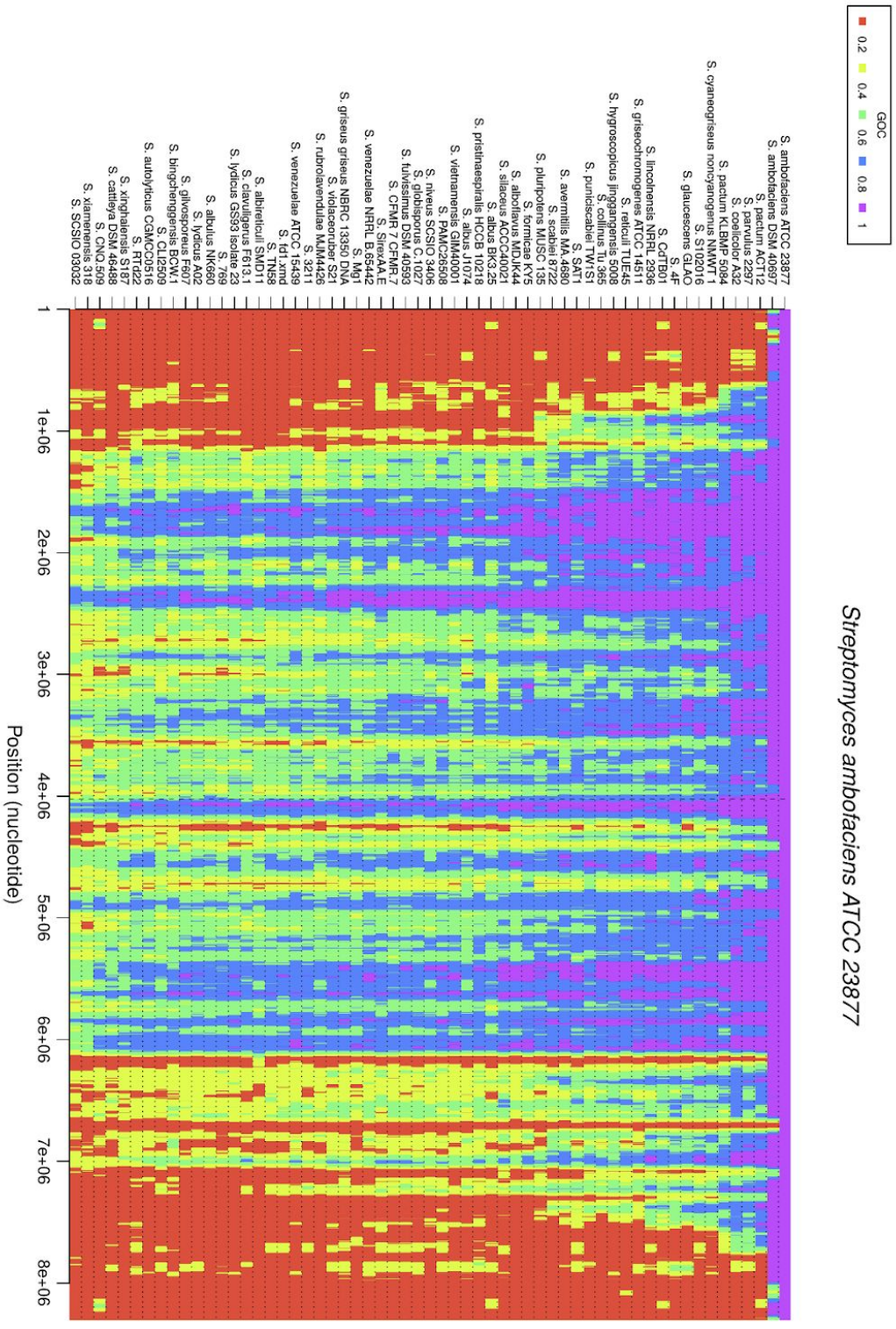

Figure 48 - Carte de chaleur de GOC du chromosome de *S. ambofaciens* ATCC 23877 en référence. Chaque ligne correspond à une espèce cible, son nom est indiqué en légende de l'axe des ordonnées. Ces espèces cibles sont ordonnées en fonction de leur distance phylogénétique avec la référence de la plus proche à la plus éloignée. L'abscisse correspond à la position en nucléotides des fenêtres de gènes. La valeur de GOC associée

à une fenêtre de gènes est convertie selon le code couleur rouge = [0 - 0.2[; jaune = [0.2 - 0.4[; vert = [0.4 - 0.6[; bleu : [0.6 - 0.8[; violet = [0.8 - 1]. Le trait vertical en pointillés correspond à la position de l'origine de réplication.

2- Nombre de réarrangements entre 2 génomes

a- Définition

Une "distance d'édition" a été estimé entre deux génomes, c'est-à-dire le nombre de réarrangements chromosomiques de type inversion et transposition permettant de passer de l'organisation d'un génome A à celle d'un génome B. Pour cela, le modèle DCJ (Double Cut and Join) (Fertin et al., 2017; Yancopoulos et al., 2005; Zanetti et al., 2016) a été utilisé. Le modèle DCJ est un modèle de réarrangement génomique utilisé pour définir une distance d'édition entre les génomes, basée sur l'ordre et l'orientation des gènes et permet l'apparition de nouveaux réplicons stables dans les différents scénarios possible. Dans notre cas, une variante du modèle DCJ a été utilisée, le modèle DCJ restreint (Kováč et al., 2011) : les opérations DCJ décrites par Yancopoulos peuvent imiter des transpositions ou bloquer des échanges en extrayant d'abord une région d'un chromosome, générant un chromosome circulaire provisoire, puis en le ré-insérant à sa place. Dans notre cas, nous nous intéressons à un modèle simplifié où un unique chromosome est utilisé, ainsi chaque excision circulaire doit être immédiatement suivie de sa réincorporation. En pratique, lors de la comparaison de deux génomes, seuls les paires d'orthologues sont considérés. La comparaison de deux génomes avec la méthode DCJ restreinte a été effectué grâce l'outil UNIMOG[14] (Hilker et al., 2012) qui renvoie une distance d'édition entre ces deux génomes, la distance DCJ qui approxime le nombre d'événements minimum de type inversion et transposition pour passer d'une organisation à l'autre.

b- Représentation graphique

Toutes les représentations graphiques de la distance DCJ ont été utilisées (**figure 49**) en fonction de la distance phylogénétique permettant de générer des graphiques à deux dimensions où l'abscisse correspond à la distance phylogénétique et l'ordonnée à la distance DCJ. Chaque point représentant la distance DCJ obtenu en passant de l'organisation du chromosome de référence à un chromosome cible donné (identifié par sa distance phylogénétique à la référence). Ainsi, d'un point à l'autre dans le graphique, le nombre de gènes utilisé pour le calcul de la distance DCJ peut varier, dépendant du nombre de paires d'orthologues entre les espèces comparées. Pour pouvoir comparer des distances de DCJ obtenu avec un nombre différents de gènes, cette valeur a été normalisée par le nombre de gènes utilisés. Cette normalisation a permi de créer un nouvel indice, approximant la proportion de gènes

[14] https://omictools.com/unimog-tool

ayant subi au moins un événement génétique de type inversion ou transposition pour passer d'une organisation à l'autre.

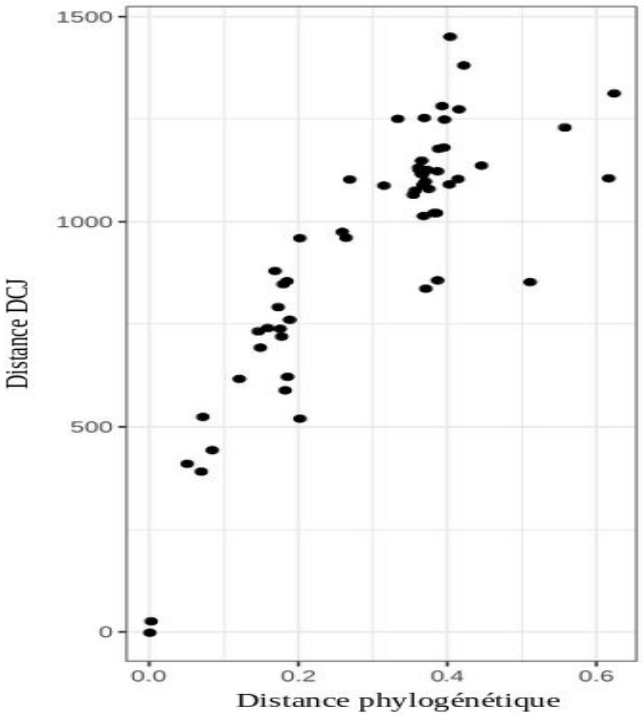

Figure 49 - Evolution de la distance DCJ (**(a)** = valeur brute de DCJ; **(b)** = DCJ normalisé par le nombre d'orthologues entre les génomes comparés) construite à partir des paires d'orthologues entre *S. ambofaciens* ATCC 23877 et 59 autres *Streptomyces*.

C- Distance et arbre phylogénétique

1- La distance ANI

L'ANI a été utilisé ici pour classifier les génomes de *Streptomyces* selon l'algorithme (**figure 50**) décrit par Goris (Goris et al., 2007). Pour chaque paire de génome, un des génomes de la paire (génome de référence) est coupé en fragments successif de 1020 nucléotides. Chacun de ces fragments est ensuite aligné contre la séquence complète de l'autre génome de la paire (génome cible) en utilisant l'algorithme de BLASTN. Le meilleur alignement est conservé s'il présente au moins 30% d'identité et 70% de couverture (par rapport à la taille du fragment de 1020 nucléotides). L'ANI entre le génome de référence et le génome cible correspond à la moyenne des pourcentages d'identités des alignements sélectionnés. Avec ce protocole, la valeur d'ANI est orientée, dans le cas décrit ci-dessus, l'ANI correspond à l'alignement du génome de référence contre la cible ($ANI_{ref \rightarrow cible}$)

et peut être différent de la valeur réciproque ($ANI_{cible \rightarrow ref}$). Pour s'affranchir de cette contrainte, nous avons défini l'ANI moyenné ($\overline{ANI}_{ref\text{-}cible}$) qui correspond à la moyenne de l'$ANI_{ref \rightarrow cible}$ et l'$ANI_{cible \rightarrow ref}$.

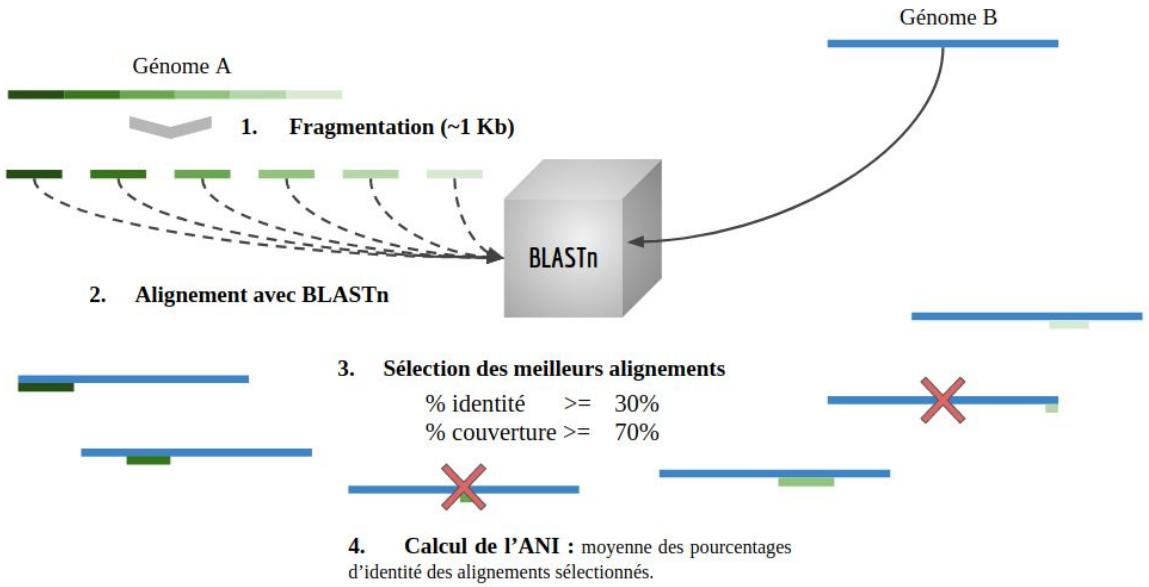

Figure 50 - Schéma du protocole de calcul de l'ANI.
L'$\overline{ANI}_{A-B}$ est obtenu en effectuant l'alignement réciproque et en moyennant les ANI obtenu.

$$\overline{ANI}_{A-B} = \frac{ANI_{A \rightarrow B} + ANI_{B \rightarrow A}}{2}$$

Des outils permettant de calculer des valeurs d'ANI sont disponibles (ANItools (Han et al., 2016), JSpecies (Richter et al., 2015), OrthoANI (Lee et al., 2016)), mais ne permettent pas l'application automatique et à grande échelle de cet algorithme. Un script *ad hoc* a donc été développé pour effectuer cette opération et utilise le programme *NCBI-BLASTn+* (version 2.6.0+).

2- Phylogénie basée sur le génome complet

a- Définition

La construction d'une phylogénie du genre *Streptomyces* a été abordée en maximisant l'information génétique apportée par chaque génome. Pour cela une phylogénie multi-locus à partir du core génome a été effectuée. Les séquences protéiques de chaque gène appartenant au core-génome des Streptomyces ont été extraites pour chaque espèce. Chaque lot de séquences protéiques des gènes du core-génomes ont été alignés indépendamment avec le logiciel MUSCLE (Edgar, 2004). Les alignements ont ensuite été concaténés selon l'ordre d'apparition de ces gènes dans le chromosome de *S. coelicolor* A3(2) puis filtré par Gblocks (Castresana, 2000) pour éliminer les régions pauvres en

informations phylogénétiques. Par exemple, à partir de la collection 110, la concaténation des alignements des séquences du core-génome (803 CDS) a généré un alignement global de 333,377 réduit à 255,632 positions (76% de l'alignement original) avec l'utilisation de Gblocks. L'arbre phylogénétique a été construit par maximum de vraisemblance avec l'outil RAxML (Randomized Axelerated Maximum Likelihood, (Stamatakis, 2014)) avec le modèle de substitution LG (Le and Gascuel, 2008) pour l'inférence d'arbre et 100 étapes de bootstrap.

b- Représentation graphique

La représentation graphique des différents arbres phylogénétiques a été effectuée en utilisant un outil graphique dédié, Dendroscope 3 (Huson et al., 2007; Huson and Scornavacca, 2012).

III- Obtention et exploitation des données DNAseq

A- Séquençage et assemblage des génomes

Le séquençage Illumina de la banque paired-end des 17 souches (**tableau 6**) de la collection a été réalisé en collaboration avec la Plateforme de Génomique Fonctionnelle du laboratoire Nutrition-Génétique et Exposition aux risques environnementaux (NGERE) (INSERM Unité 954 – NGERE, Faculté de Médecine – Bâtiment C 2ème étage, 9 Avenue de la Forêt de Haye– BP 20199 – 54505 Vandoeuvre les Nancy).

En complément, le séquençage et l'assemblage complet de la souche RLB1-9 ont été effectués en utilisant une stratégie combinée de séquençage par les technologies technologies Nanopore (MinIon) et Illumina. Les contigs produits ont ensuite été organisés par alignement avec l'outil Artemis Comparison Tool (Carver et al., 2005), avec le génome de *Streptomyces avermitilis* MA 4680 qui présentait la plus forte parenté avec la souche (99% d'identité du gène ARN 16S) en faisant l'hypothèse d'une forte conservation de l'ordre des gènes entre ces espèces.

B- Détection des régions manquantes

Un protocole de détection des régions manquantes, c'est-à-dire de séquence nucléotidique spécifique à une souche de référence à partir des données brutes de séquençage et d'un génome assemblé a été développé. La stratégie se décompose en deux grandes étapes: (i) un alignement "classique" des lectures sur la référence et (ii) la détection des régions manquantes selon un seuil de couverture.

(i) La qualité des lectures a été testée avec le logiciel FastQC (v0.11.5) et n'a révélé aucun biais. En complément l'impact sur les alignements finaux en prétraitant les données avec Trimmomatic (Bolger et al., 2014) ou non a été testé. Cette étape permet, en particulier, de supprimer des lectures les séquences adaptatrices Illumina, ainsi que d'éditer les premiers et derniers nucléotides ("heading" et "trailing") de chaque read si ces derniers ne passent pas un seuil de qualité donné. Aucune différence entre les alignements obtenus avant ou après pré-traitement des lectures n'a été identifié.

Les alignements ont été effectués avec le logiciel Bowtie2 (Langmead and Salzberg, 2012) en utilisant le génome de RLB1-8 comme référence. Pour qu'un alignement soit validé, les lectures doivent être appariées (*ie* lectures direct et inverse alignées) et l'alignement total ne doit pas dépasser 500 nucléotides. (ii) L'identification des régions manquantes dans le génome de référence correspond à des zones où le nombre de lectures moyennes alignées est inférieur au seuil de significativité. Ce seuil

est défini comme étant le quart du nombre moyen de lectures par positions couvertes dans l'ensemble du génome. Les positions où aucune lecture ne s'aligne ne sont pas prises en compte dans le calcul de la moyenne. Enfin seuls les régions d'au moins 10 kb sont représentées **figure 29**.

BIBLIOGRAPHIE

Achtman, M., Manning, P.A., Kusecek, B., Schwuchow, S., Willetts, N., 1980. A genetic analysis of F sex factor cistrons needed for surface exclusion in Escherichia coli. J. Mol. Biol. 138, 779–795.

Aigle, B., Schneider, D., Morilhat, C., Vandewiele, D., Dary, A., Holl, A.-C., Simonet, J.-M., Decaris, B., 1996. An amplifiable and deletable locus of Streptomyces ambofaciens RP181110 contains a very large gene homologous to polyketide synthase genes. Microbiology, 142, 2815–2824. https://doi.org/10.1099/13500872-142-10-2815

Aldridge, M., Facey, P., Francis, L., Bayliss, S., Del Sol, R., Dyson, P., 2013. A novel bifunctional histone protein in Streptomyces : a candidate for structural coupling between DNA conformation and transcription during development and stress? Nucleic Acids Res. 41, 4813–4824. https://doi.org/10.1093/nar/gkt180

Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S., Ishihama, A., 1999. Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. J. Bacteriol. 181, 6361–6370.

Allardet-Servent, A., Michaux-Charachon, S., Jumas-Bilak, E., Karayan, L., Ramuz, M., 1993. Presence of one linear and one circular chromosome in the Agrobacterium tumefaciens C58 genome. J. Bacteriol. 175, 7869–7874. https://doi.org/10.1128/jb.175.24.7869-7874.1993

Altenbuchner, J., Cullum, J., 1985. Structure of an amplifiable DNA sequence in Streptomyces lividans 66. Mol. Gen. Genet. MGG 201, 192–197. https://doi.org/10.1007/BF00425659

Altenbuchner, J., Cullum, J., 1984. DNA amplification and an unstable arginine gene in Streptomyces lividans 66. Mol. Gen. Genet. MGG 195, 134–138. https://doi.org/10.1007/BF00332735

Anand, R.P., Lovett, S.T., Haber, J.E., 2013. Break-Induced DNA Replication. Cold Spring Harb. Perspect. Biol. 5, a010397. https://doi.org/10.1101/cshperspect.a010397

Aravind, L., Koonin, E.V., 2001. Prokaryotic Homologs of the Eukaryotic DNA-End-Binding Protein Ku, Novel Domains in the Ku Protein and Prediction of a Prokaryotic Double-Strand Break Repair System. Genome Res. 11, 1365–1374. https://doi.org/10.1101/gr.181001

Avery, O.T., Macleod, C.M., McCarty, M., 1944. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. J. Exp. Med. 79, 137–158. https://doi.org/10.1084/jem.79.2.137

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9, 75.

Badrinarayanan, A., Le, T.B., Laub, M.T., 2015. Bacterial chromosome organization and segregation. Annu. Rev. Cell Dev. Biol. 31, 171–199. https://doi.org/10.1146/annurev-cellbio-100814-125211

Bahmed, K., Seth, A., Nitiss, K.C., Nitiss, J.L., 2011. End-processing during non-homologous end-joining: a role for exonuclease 1. Nucleic Acids Res. 39, 970–978. https://doi.org/10.1093/nar/gkq886

Bakshi, S., Siryaporn, A., Goulian, M., Weisshaar, J.C., 2012. Superresolution imaging of ribosomes and RNA polymerase in live Escherichia coli cells. Mol. Microbiol. 85, 21–38. https://doi.org/10.1111/j.1365-2958.2012.08081.x

Bao, K., Cohen, S.N., 2003. Recruitment of terminal protein to the ends of Streptomyces linear plasmids and chromosomes by a novel telomere-binding protein essential for linear DNA replication. Genes Dev. 17, 774–785. https://doi.org/10.1101/gad.1060303

Baril, C., Richaud, C., Baranton, G., Saint Girons, I.S., 1989. Linear chromosome of Borrelia burgdorferi. Res. Microbiol. 140, 507–516.
Baril, L., Boiron, P., Manceron, V., Oumar Ould Ely, S., Jamet, P., Favre, E., Caumes, E., Bricaire, F., 1999. Refractory Craniofacial Actinomycetoma Due to Streptomyces somaliensis That Required Salvage Therapy with Amikacin and Imipenem. Clin. Infect. Dis. 29, 460–461. https://doi.org/10.1086/520246
Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., Horvath, P., 2007. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science 315, 1709–1712. https://doi.org/10.1126/science.1138140
Barrick, J.E., Yu, D.S., Yoon, S.H., Jeong, H., Oh, T.K., Schneider, D., Lenski, R.E., Kim, J.F., 2009. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature 461, 1243. https://doi.org/10.1038/nature08480
Benhadj, M., Gacemi-Kirane, D., Toussaint, M., Hotel, L., Bontemps, C., Duval, R.E., Aigle, B., Leblond, P., 2018. Diversité et activités antimicrobiennes de souches de Streptomyces de la Fetzara (Algérie). Ann. Biol. Clin. (Paris) 76, 81–95. https://doi.org/10.1684/abc.2017.1316
Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.-M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3 (2). Nature 417, 141–147.
Bérdy, J., 2005. Bioactive Microbial Metabolites. J. Antibiot. (Tokyo) 58, 1–26. https://doi.org/10.1038/ja.2005.1
Bertrand, K., Squires, C., Yanofsky, C., 1976. Transcription termination in vivo in the leader region of the tryptophan operon of Escherichia coli. J. Mol. Biol. 103, 319–337. https://doi.org/10.1016/0022-2836(76)90315-6
Bibb, M.J., Ward, J.M., Kieser, T., Cohen, S.N., Hopwood, D.A., 1981. Excision of chromosomal DNA sequences from Streptomyces coelicolor forms a novel family of plasmids detectable in Streptomyces lividans. Mol. Gen. Genet. MGG 184, 230–240. https://doi.org/10.1007/BF00272910
Bilyk, B., Horbal, L., Luzhetskyy, A., 2017. Chromosomal position effect influences the heterologous expression of genes and biosynthetic gene clusters in Streptomyces albus J1074. Microb. Cell Factories 16, 5. https://doi.org/10.1186/s12934-016-0619-z
Bird, A.P., 1995. Gene number, noise reduction and biological complexity. Trends Genet. 11, 94–100. https://doi.org/10.1016/S0168-9525(00)89009-5
Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S.Y., Medema, M.H., Weber, T., 2019. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 47, W81–W87. https://doi.org/10.1093/nar/gkz310
Block, D.H.S., Hussein, R., Liang, L.W., Lim, H.N., 2012. Regulatory consequences of gene translocation in bacteria. Nucleic Acids Res. 40, 8979–8992. https://doi.org/10.1093/nar/gks694
Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170
Bryant, J.A., Sellars, L.E., Busby, S.J.W., Lee, D.J., 2014. Chromosome position effects on gene expression in Escherichia coli K-12. Nucleic Acids Res. 42, 11383–11392. https://doi.org/10.1093/nar/gku828
Burke, J., Schneider, D., Westpheling, J., 2001. Generalized Transduction in Streptomyces coelicolor. Proc. Natl. Acad. Sci. U. S. A. 98, 6289–6294.
Campo, N., Dias, M.J., Daveran‐Mingot, M.-L., Ritzenthaler, P., Bourgeois, P.L., 2004. Chromosomal constraints in Gram-positive bacteria revealed by artificial inversions. Mol. Microbiol. 51, 511–522. https://doi.org/10.1046/j.1365-2958.2003.03847.x
Cao, G., Zhang, P., Gu, Y., Pang, X., 2017. Gene Overexpression in Streptomyces hygroscopicus Associated with DNA Amplification. Curr. Microbiol. 74, 979–986. https://doi.org/10.1007/s00284-017-1278-y

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.-A., Barrell, B.G., Parkhill, J., 2005. ACT: the Artemis comparison tool. Bioinformatics 21, 3422–3423. https://doi.org/10.1093/bioinformatics/bti553

Castillo, U.F., Strobel, G.A., Ford, E.J., Hess, W.M., Porter, H., Jensen, J.B., Albert, H., Robison, R., Condron, M. a. M., Teplow, D.B., Stevens, D., Yaver, D., 2002. Munumbicins, wide-spectrum antibiotics produced by Streptomyces NRRL 30562, endophytic on Kennedia nigriscans. Microbiol.-SGM 148, 2675–2685.

Castresana, J., 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol. Biol. Evol. 17, 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Catakli, S., Andrieux, A., Leblond, P., Decaris, B., Dary, A., 2003. Spontaneous chromosome circularization and amplification of a new amplifiable unit of DNA belonging to the terminal inverted repeats in Streptomyces ambofaciens ATCC 23877. Arch. Microbiol. 179, 387–393. https://doi.org/10.1007/s00203-003-0534-7

Chang, P.C., Cohen, S.N., 1994. Bidirectional replication from an internal origin in a linear streptomyces plasmid. Science 265, 952–954. https://doi.org/10.1126/science.8052852

Charlop-Powers, Z., Owen, J.G., Reddy, B.V.B., Ternei, M.A., Guimarães, D.O., de Frias, U.A., Pupo, M.T., Seepe, P., Feng, Z., Brady, S.F., 2015. Global biogeographic sampling of bacterial secondary metabolism. eLife 4. https://doi.org/10.7554/eLife.05048

Chater, K.F., 2016. Recent advances in understanding Streptomyces. F1000Research 5. https://doi.org/10.12688/f1000research.9534.1

Chater, K.F., Biró, S., Lee, K.J., Palmer, T., Schrempf, H., 2010. The complex extracellular biology of Streptomyces. FEMS Microbiol. Rev. 34, 171–198. https://doi.org/10.1111/j.1574-6976.2009.00206.x

Chayot, R., Montagne, B., Mazel, D., Ricchetti, M., 2010. An end-joining repair mechanism in Escherichia coli. Proc. Natl. Acad. Sci. 107, 2141–2146. https://doi.org/10.1073/pnas.0906355107

Chen, C.W., Huang, C.-H., Lee, H.-H., Tsai, H.-H., Kirby, R., 2002. Once the circle has been broken: dynamics and evolution of Streptomyces chromosomes. Trends Genet. 18, 522–529. https://doi.org/10.1016/S0168-9525(02)02752-X

Chen, W., He, F., Zhang, X., Chen, Z., Wen, Y., Li, J., 2010. Chromosomal instability in Streptomyces avermitilis: major deletion in the central region and stable circularized chromosome. BMC Microbiol. 10, 198.

Choulet, F., 2006. Evolution du génome des Streptomyces: transfert horizontal et variabilité des extrémités chromosomiques.

Choulet, F., Aigle, B., Gallois, A., Mangenot, S., Gerbaud, C., Truong, C., Francou, F.-X., Fourrier, C., Guérineau, M., Decaris, B., 2006a. Evolution of the terminal regions of the Streptomyces linear chromosome. Mol. Biol. Evol. 23, 2361–2369.

Choulet, F., Gallois, A., Aigle, B., Mangenot, S., Gerbaud, C., Truong, C., Francou, F.-X., Borges, F., Fourrier, C., Guérineau, M., 2006b. Intraspecific variability of the terminal inverted repeats of the linear chromosome of Streptomyces ambofaciens. J. Bacteriol. 188, 6599–6610.

Cohen, A., Bar-Nir, D., Goedeke, M.E., Parag, Y., 1985. The integrated and free states of Streptomyces griseus plasmid pSG1. Plasmid 13, 41–50. https://doi.org/10.1016/0147-619X(85)90054-X

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.-A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S., Barrell, B.G., 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393,

537–544. https://doi.org/10.1038/31159
Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honoré, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R.M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J.R., Barrell, B.G., 2001. Massive gene decay in the leprosy bacillus. Nature 409, 1007–1011. https://doi.org/10.1038/35059006
Conway, J.R., Lex, A., Gehlenborg, N., 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics 33, 2938–2940. https://doi.org/10.1093/bioinformatics/btx364
Couturier, E., Rocha, E.P.C., 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol. Microbiol. 59, 1506–1518. https://doi.org/10.1111/j.1365-2958.2006.05046.x
Dandekar, T., Snel, B., Huynen, M., Bork, P., 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. 23, 324–328.
Davis, J.J., Olsen, G.J., Overbeek, R., Vonstein, V., Xia, F., 2014. In search of genome annotation consistency: solid gene clusters and how to use them. 3 Biotech 4, 331–335. https://doi.org/10.1007/s13205-013-0152-2
de Vries, R., 2010. DNA condensation in bacteria: Interplay between macromolecular crowding and nucleoid proteins. Biochimie, Special section “DNA and Chromosomes: Physical and Biological Approaches” 92, 1715–1721. https://doi.org/10.1016/j.biochi.2010.06.024
Delcher, A.L., Bratke, K.A., Powers, E.C., Salzberg, S.L., 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23, 673–679.
Della, M., Palmbos, P.L., Tseng, H.-M., Tonkin, L.M., Daley, J.M., Topper, L.M., Pitcher, R.S., Tomkinson, A.E., Wilson, T.E., Doherty, A.J., 2004. Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine. Science 306, 683–685. https://doi.org/10.1126/science.1099824
Demerec, M., Hartman, P.E., 1959. Complex Loci in Microorganisms. Annu. Rev. Microbiol. 13, 377–406. https://doi.org/10.1146/annurev.mi.13.100159.002113
Demuyter, P., Leblond, P., Decaris, B., Simonet, J.-M., 1988. Characterization of Two Families of Spontaneously Amplifiable Units of DNA in Streptomyces ambofaciens. Microbiology 134, 2001–2007. https://doi.org/10.1099/00221287-134-7-2001
Dillon, S.C., Dorman, C.J., 2010. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat. Rev. Microbiol. 8, 185–195. https://doi.org/10.1038/nrmicro2261
Doherty, A.J., Jackson, S.P., Weller, G.R., 2001. Identification of bacterial homologues of the Ku DNA repair proteins. FEBS Lett. 500, 186–188. https://doi.org/10.1016/S0014-5793(01)02589-3
Duncan, B.K., Miller, J.H., 1980. Mutagenic deamination of cytosine residues in DNA. Nature 287, 560–561.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Emms, D.M., Kelly, S., 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16, 157. https://doi.org/10.1186/s13059-015-0721-2
Espeli, O., Mercier, R., Boccard, F., 2008. DNA dynamics vary according to macrodomain topography in the E. coli chromosome. Mol. Microbiol. 68, 1418–1427. https://doi.org/10.1111/j.1365-2958.2008.06239.x
Ferdows, M.S., Serwer, P., Griess, G.A., Norris, S.J., Barbour, A.G., 1996. Conversion of a linear to a circular plasmid in the relapsing fever agent Borrelia hermsii. J. Bacteriol. 178, 793–800.

https://doi.org/10.1128/jb.178.3.793-800.1996
Fertin, G., Jean, G., Tannier, E., 2017. Algorithms for computing the double cut and join distance on both gene order and intergenic sizes. Algorithms Mol. Biol. 12, 16.
Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A., Eddy, S.R., 2015. HMMER web server: 2015 update. Nucleic Acids Res. 43, W30–W38. https://doi.org/10.1093/nar/gkv397
Fischer, G., Decaris, B., Leblond, P., 1997. Occurrence of deletions, associated with genetic instability in Streptomyces ambofaciens, is independent of the linearity of the chromosomal DNA. J. Bacteriol. 179, 4553–4558. https://doi.org/10.1128/jb.179.14.4553-4558.1997
Fischer, G., Wenner, T., Decaris, B., Leblond, P., 1998. Chromosomal arm replacement generates a high level of intraspecific polymorphism in the terminal inverted repeats of the linear chromosomal DNA of Streptomyces ambofaciens. Proc. Natl. Acad. Sci. 95, 14296–14301. https://doi.org/10.1073/pnas.95.24.14296
Flärdh, K., 2003. Growth polarity and cell division in Streptomyces. Curr. Opin. Microbiol. 6, 564–571. https://doi.org/10.1016/j.mib.2003.10.011
Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., Al, E., 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269, 496–512. https://doi.org/10.1126/science.7542800
Frank, A.C., Lobry, J.R., 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238, 65–77.
Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., 1995. The minimal gene complement of Mycoplasma genitalium. science 270, 397–404.
Frit, P., Barboule, N., Yuan, Y., Gomez, D., Calsou, P., 2014. Alternative end-joining pathway(s): Bricolage at DNA breaks. DNA Repair, Recent Developments in Non-Homologous End Joining 17, 81–97. https://doi.org/10.1016/j.dnarep.2014.02.007
Furuya, E.Y., Lowy, F.D., 2006. Antimicrobial-resistant bacteria in the community setting. Nat. Rev. Microbiol. 4, 36–45. https://doi.org/10.1038/nrmicro1325
Garcia, R., Gemperlein, K., Müller, R., 2014. Minicystis rosea gen. nov., sp. nov., a polyunsaturated fatty acid-rich and steroid-producing soil myxobacterium. Int. J. Syst. Evol. Microbiol. 64, 3733–3742. https://doi.org/10.1099/ijs.0.068270-0
Ghinet, M.G., Bordeleau, E., Beaudin, J., Brzezinski, R., Roy, S., Burrus, V., 2011. Uncovering the Prevalence and Diversity of Integrating Conjugative Elements in Actinobacteria. PLOS ONE 6, e27846. https://doi.org/10.1371/journal.pone.0027846
Gil, R., Silva, F.J., Peretó, J., Moya, A., 2004. Determination of the core of a minimal bacterial gene set. Microbiol. Mol. Biol. Rev. MMBR 68, 518–537, table of contents. https://doi.org/10.1128/MMBR.68.3.518-537.2004
Gogarten, J.P., Townsend, J.P., 2005. Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3, 679–687. https://doi.org/10.1038/nrmicro1204
Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B.S., Cao, Y., Askenazi, M., Halling, C., Mullin, L., Houmiel, K., Gordon, J., Vaudin, M., Iartchouk, O., Epp, A., Liu, F., Wollam, C., Allinger, M., Doughty, D., Scott, C., Lappas, C., Markelz, B., Flanagan, C., Crowell, C., Gurson, J., Lomo, C., Sear, C., Strub, G., Cielo, C., Slater, S., 2001. Genome Sequence of the Plant Pathogen and Biotechnology Agent Agrobacterium tumefaciens C58. Science 294, 2323–2328. https://doi.org/10.1126/science.1066803
Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M., 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 57, 81–91.
Grainger, D.C., Hurd, D., Harrison, M., Holdstock, J., Busby, S.J.W., 2005. Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the E. coli

chromosome. Proc. Natl. Acad. Sci. 102, 17693–17698. https://doi.org/10.1073/pnas.0506687102
Griffith, F., 1928. The Significance of Pneumococcal Types. Epidemiol. Infect. 27, 113–159. https://doi.org/10.1017/S0022172400031879
Guirouilh-Barbat, J., Lambert, S., Bertrand, P., Lopez, B.S., 2014. Is homologous recombination really an error-free process? Front. Genet. 5. https://doi.org/10.3389/fgene.2014.00175
Hilker, R., Sickinger, C., Pedersen, C.N.S., Stoye, J., 2012. UniMoG—a unifying framework for genomic distance calculation and sorting based on DCJ. Bioinformatics 28, 2509–2511. https://doi.org/10.1093/bioinformatics/bts440
Hiller, N.L., Janto, B., Hogg, J.S., Boissy, R., Yu, S., Powell, E., Keefe, R., Ehrlich, N.E., Shen, K., Hayes, J., Barbadora, K., Klimke, W., Dernovoy, D., Tatusova, T., Parkhill, J., Bentley, S.D., Post, J.C., Ehrlich, G.D., Hu, F.Z., 2007. Comparative Genomic Analyses of Seventeen Streptococcus pneumoniae Strains: Insights into the Pneumococcal Supragenome. J. Bacteriol. 189, 8186–8195. https://doi.org/10.1128/JB.00690-07
Hoefler, B.C., Konganti, K., Straight, P.D., 2013. De Novo Assembly of the Streptomyces sp. Strain Mg1 Genome Using PacBio Single-Molecule Sequencing. Genome Announc. 1. https://doi.org/10.1128/genomeA.00535-13
Hoff, G., Bertrand, C., Piotrowski, E., Thibessard, A., Leblond, P., 2018. Genome plasticity is governed by double strand break DNA repair in Streptomyces. Sci. Rep. 8. https://doi.org/10.1038/s41598-018-23622-w
Hoff, G., Bertrand, C., Zhang, L., Piotrowski, E., Chipot, L., Bontemps, C., Confalonieri, F., McGovern, S., Lecointe, F., Thibessard, A., 2016. Multiple and variable NHEJ-like genes are involved in resistance to DNA damage in Streptomyces ambofaciens. Front. Microbiol. 7.
Huang, C.-H., Lin, Y.-S., Yang, Y.-L., Huang, S., Chen, C.W., 1998. The telomeres of Streptomyces chromosomes contain conserved palindromic sequences with potential to form complex secondary structures. Mol. Microbiol. 28, 905–916. https://doi.org/10.1046/j.1365-2958.1998.00856.x
Huang, C.-H., Tsai, H.-H., Tsay, Y.-G., Chien, Y.-N., Wang, S.-L., Cheng, M.-Y., Ke, C.-H., Chen, C.W., 2007. The telomere system of the Streptomyces linear plasmid SCP1 represents a novel class. Mol. Microbiol. 63, 1710–1718. https://doi.org/10.1111/j.1365-2958.2007.05616.x
Hulsen, T., Huynen, M.A., de Vlieg, J., Groenen, P.M., 2006. Benchmarking ortholog identification methods using functional genomics data. Genome Biol. 7, R31. https://doi.org/10.1186/gb-2006-7-4-r31
Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., Otto, T.D., 2013. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 14, R47. https://doi.org/10.1186/gb-2013-14-5-r47
Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., Rupp, R., 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8, 460.
Huson, D.H., Scornavacca, C., 2012. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. Syst. Biol. 61, 1061–1067. https://doi.org/10.1093/sysbio/sys062
Hutchison, C.A., Chuang, R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., Pelletier, J.F., Qi, Z.-Q., Richter, R.A., Strychalski, E.A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K.S., Smith, H.O., Glass, J.I., Merryman, C., Gibson, D.G., Venter, J.C., 2016. Design and synthesis of a minimal bacterial genome. Science 351, aad6253. https://doi.org/10.1126/science.aad6253
Huynen, M., Snel, B., Lathe, W., Bork, P., 2000. Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. Genome Res. 10, 1204–1210. https://doi.org/10.1101/gr.10.8.1204
Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M., Ōmura, S., 2003. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. Nat. Biotechnol. 21, 526–531.

Inoue, S., Higashiyama, K., Uchida, T., Hiratsu, K., Kinashi, H., 2003. Chromosomal Circularization in Streptomyces griseus by Nonhomologous Recombination of Deletion Ends. Biosci. Biotechnol. Biochem. 67, 1101–1108. https://doi.org/10.1271/bbb.67.1101

Jacob, F., Perrin, D., Sánchez, C., Monod, J., 1960. L'opéron : groupe de gènes à expression coordonnée par un opérateur [C. R. Acad. Sci. Paris 250 (1960) 1727–1729]. C. R. Biol., Retour sur l'operon lac 328, 514–520. https://doi.org/10.1016/j.crvi.2005.04.005

Johnston, C., Martin, B., Fichant, G., Polard, P., Claverys, J.-P., 2014. Bacterial transformation: distribution, shared mechanisms and divergent control. Nat. Rev. Microbiol. 12, 181–196. https://doi.org/10.1038/nrmicro3199

Jun, S., Mulder, B., 2006. Entropy-driven spatial organization of highly confined polymers: lessons for the bacterial chromosome. Proc. Natl. Acad. Sci. U. S. A. 103, 12388–12393. https://doi.org/10.1073/pnas.0605305103

Kahramanoglou, C., Seshasayee, A.S.N., Prieto, A.I., Ibberson, D., Schmidt, S., Zimmermann, J., Benes, V., Fraser, G.M., Luscombe, N.M., 2011. Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. Nucleic Acids Res. 39, 2073–2091. https://doi.org/10.1093/nar/gkq934

Karoonuthaisiri, N., Weaver, D., Huang, J., Cohen, S.N., Kao, C.M., 2005. Regional organization of gene expression in Streptomyces coelicolor. Gene 353, 53–66. https://doi.org/10.1016/j.gene.2005.03.042

Kavenoff, R., Ryder, O.A., 1976. Electron microscopy of membrane-associated folded chromosomes of Escherichia coli. Chromosoma 55, 13–25.

Khvorova, A., Chary, V.K., Hilbert, D.W., Piggot, P.J., 2000. The Chromosomal Location of the Bacillus subtilis Sporulation Gene spoIIR Is Important for Its Function. J. Bacteriol. 182, 4425–4429. https://doi.org/10.1128/JB.182.16.4425-4429.2000

Kim, J.-N., Kim, Y., Jeong, Y., Roe, J.-H., Kim, B.-G., Cho, B.-K., 2015. Comparative Genomics Reveals the Core and Accessory Genomes of Streptomyces Species. J. Microbiol. Biotechnol. 25, 1599–1605. https://doi.org/10.4014/jmb.1504.04008

Kogoma, T., 1997. Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. Microbiol. Mol. Biol. Rev. MMBR 61, 212–238.

Konstantinidis, K.T., Tiedje, J.M., 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. 102, 2567–2572. https://doi.org/10.1073/pnas.0409727102

Koonin, E.V., 2000. How Many Genes Can Make a Cell: The Minimal-Gene-Set Concept. Annu. Rev. Genomics Hum. Genet. 1, 99–116. https://doi.org/10.1146/annurev.genom.1.1.99

Kost, C., Lakatos, T., Böttcher, I., Arendholz, W.-R., Redenbach, M., Wirth, R., 2007. Non-specific association between filamentous bacteria and fungus-growing ants. Naturwissenschaften 94, 821–828. https://doi.org/10.1007/s00114-007-0262-y

Kováč, J., Warren, R., Braga, M.D.V., Stoye, J., 2011. Restricted DCJ Model: Rearrangement Problems with Chromosome Reincorporation. J. Comput. Biol. 18, 1231–1241. https://doi.org/10.1089/cmb.2011.0116

Kowalczykowski, S.C., 2000. Initiation of genetic recombination and recombination-dependent replication. Trends Biochem. Sci. 25, 156–165.

Kuo, C.-H., Moran, N.A., Ochman, H., 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res. 19, 1450–1454. https://doi.org/10.1101/gr.091785.109

Kuo, C.-H., Ochman, H., 2009. Deletional Bias across the Three Domains of Life. Genome Biol. Evol. 1, 145–152. https://doi.org/10.1093/gbe/evp016

Kyndt, T., Quispe, D., Zhai, H., Jarret, R., Ghislain, M., Liu, Q., Gheysen, G., Kreuze, J.F., 2015. The genome of cultivated sweet potato contains *Agrobacterium* T-DNAs with expressed genes: An example of a naturally transgenic food crop. Proc. Natl. Acad. Sci. 112, 5844–5849. https://doi.org/10.1073/pnas.1419685112

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S., Ussery, D.W., 2015. Insights from 20 years of bacterial

genome sequencing. Funct. Integr. Genomics 15, 141–161. https://doi.org/10.1007/s10142-015-0433-4
Land, M.L., Hyatt, D., Jun, S.-R., Kora, G.H., Hauser, L.J., Lukjancenko, O., Ussery, D.W., 2014. Quality scores for 32,000 genomes. Stand. Genomic Sci. 9, 20. https://doi.org/10.1186/1944-3277-9-20
Langer-Safer, P.R., Levine, M., Ward, D.C., 1982. Immunological method for mapping genes on Drosophila polytene chromosomes. Proc. Natl. Acad. Sci. U. S. A. 79, 4381–4385.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.
Lapierre, P., Gogarten, J.P., 2009. Estimating the size of the bacterial pan-genome. Trends Genet. 25, 107–110. https://doi.org/10.1016/j.tig.2008.12.004
Lawrence, J.G., Retchless, A.C., 2009. The Interplay of Homologous Recombination and Horizontal Gene Transfer in Bacterial Speciation, in: Gogarten, M.B., Gogarten, J.P., Olendzenski, L.C. (Eds.), Horizontal Gene Transfer. Humana Press, Totowa, NJ, pp. 29–53. https://doi.org/10.1007/978-1-60327-853-9_3
Le, S.Q., Gascuel, O., 2008. An Improved General Amino Acid Replacement Matrix. Mol. Biol. Evol. 25, 1307–1320. https://doi.org/10.1093/molbev/msn067
Le, T.B., Laub, M.T., 2014. New approaches to understanding the spatial organization of bacterial genomes. Curr. Opin. Microbiol., Growth and development: eukaryotes/ prokaryotes 22, 15–21. https://doi.org/10.1016/j.mib.2014.09.014
Le, T.B.K., Imakaev, M.V., Mirny, L.A., Laub, M.T., 2013. High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. Science 342, 731–734. https://doi.org/10.1126/science.1242059
Leblond, P., Decaris, B., 1994. New insights into the genetic instability of streptomyces. FEMS Microbiol. Lett. 123, 225–232. https://doi.org/10.1111/j.1574-6968.1994.tb07229.x
Leblond, P., Fischer, G., Francou, F.-X., Berger, F., Guérineau, M., Decaris, B., 1996. The unstable region of Streptomyces ambofaciens includes 210 kb terminal inverted repeats flanking the extremities of the linear chromosomal DNA. Mol. Microbiol. 19, 261–271. https://doi.org/10.1046/j.1365-2958.1996.366894.x
Lederberg, J., Tatum, E.L., 1946. Gene recombination in Escherichia coli. Nature 158, 558. https://doi.org/10.1038/158558a0
Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H., 2014. UpSet: Visualization of Intersecting Sets. IEEE Trans. Vis. Comput. Graph. 20, 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248
Libby, E.A., Roggiani, M., Goulian, M., 2012. Membrane protein expression triggers chromosomal locus repositioning in bacteria. Proc. Natl. Acad. Sci. U. S. A. 109, 7445–7450. https://doi.org/10.1073/pnas.1109479109
Lin, Y.-S., Chen, C.W., 1997. Instability of artificially circularized chromosomes of Streptomyces lividans. Mol. Microbiol. 26, 709–719. https://doi.org/10.1046/j.1365-2958.1997.5991975.x
Lin, Y.-S., Kieser, H.M., Hopwood, D.A., Chen, C.W., 1993. The chromosomal DNA of Streptomyces lividans 66 is linear. Mol. Microbiol. 10, 923–933. https://doi.org/10.1111/j.1365-2958.1993.tb00964.x
Lind, P.A., Andersson, D.I., 2008. Whole-genome mutational biases in bacteria. Proc. Natl. Acad. Sci. 105, 17878–17883. https://doi.org/10.1073/pnas.0804445105
Lindahl, T., 1993. Instability and decay of the primary structure of DNA. Nature 362, 709–715. https://doi.org/10.1038/362709a0
Liu, G., Chater, K.F., Chandra, G., Niu, G., Tan, H., 2013. Molecular Regulation of Antibiotic Biosynthesis in Streptomyces. Microbiol Mol Biol Rev 77, 112–143. https://doi.org/10.1128/MMBR.00054-12
Łukasik, P., Chong, R.A., Nazario, K., Matsuura, Y., Bublitz, D.A.C., Campbell, M.A., Meyer, M.C., Van Leuven, J.T., Pessacq, P., Veloso, C., Simon, C., McCutcheon, J.P., 2019. One Hundred

Mitochondrial Genomes of Cicadas. J. Hered. 110, 247–256. https://doi.org/10.1093/jhered/esy068
Makarova, K.S., Wolf, Y.I., Koonin, E.V., 2013. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res. 41, 4360–4377. https://doi.org/10.1093/nar/gkt157
Maniloff, J., 1996. The minimal cell genome: "on being the right size." Proc. Natl. Acad. Sci. 93, 10004–10006. https://doi.org/10.1073/pnas.93.19.10004
McCutcheon, J.P., Moran, N.A., 2012. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10, 13–26. https://doi.org/10.1038/nrmicro2670
McDonald, B.R., Currie, C.R., 2017. Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces. mBio 8, e00644–17.
Medini, D., Donati, C., Tettelin, H., Masignani, V., Rappuoli, R., 2005. The microbial pan-genome. Curr. Opin. Genet. Dev., Genomes and evolution 15, 589–594. https://doi.org/10.1016/j.gde.2005.09.006
Mehta, A., Haber, J.E., 2014. Sources of DNA Double-Strand Breaks and Models of Recombinational DNA Repair. Cold Spring Harb. Perspect. Biol. 6, a016428–a016428. https://doi.org/10.1101/cshperspect.a016428
Mercier, R., Petit, M.-A., Schbath, S., Robin, S., El Karoui, M., Boccard, F., Espéli, O., 2008. The MatP/matS Site-Specific System Organizes the Terminus Region of the E. coli Chromosome into a Macrodomain. Cell 135, 475–485. https://doi.org/10.1016/j.cell.2008.08.031
Michaels, M.L., Miller, J.H., 1992. The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-hydroxyguanine (7,8-dihydro-8-oxoguanine). J. Bacteriol. 174, 6321–6325. https://doi.org/10.1128/jb.174.20.6321-6325.1992
Milo, R., 2013. What is the total number of protein molecules per cell volume? A call to rethink some published values. BioEssays 35, 1050–1055. https://doi.org/10.1002/bies.201300066
Mira, A., Ochman, H., Moran, N.A., 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17, 589–596. https://doi.org/10.1016/S0168-9525(01)02447-7
Moore, E.R.B., Mihaylova, S.A., Vandamme, P., Krichevsky, M.I., Dijkshoorn, L., 2010. Microbial systematics and taxonomy: relevance for a microbial commons. Res. Microbiol., Microbial research commons: From strain isolation to practical use 161, 430–438. https://doi.org/10.1016/j.resmic.2010.05.007
Moore, J.K., Haber, J.E., 1996. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. Mol. Cell. Biol. 16, 2164–2173.
Moran, N.A., Bennett, G.M., 2014. The Tiniest Tiny Genomes. Annu. Rev. Microbiol. 68, 195–215. https://doi.org/10.1146/annurev-micro-091213-112901
Mruk, I., Kobayashi, I., 2014. To be or not to be: regulation of restriction–modification systems and other toxin–antitoxin systems. Nucleic Acids Res. 42, 70–86. https://doi.org/10.1093/nar/gkt711
Müller, C.M., Dobrindt, U., Nagy, G., Emödy, L., Uhlin, B.E., Hacker, J., 2006. Role of histone-like proteins H-NS and StpA in expression of virulence determinants of uropathogenic Escherichia coli. J. Bacteriol. 188, 5428–5438. https://doi.org/10.1128/JB.01956-05
Mushegian, A. R., Koonin, E.V., 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc. Natl. Acad. Sci. U. S. A. 93, 10268–10273. https://doi.org/10.1073/pnas.93.19.10268
Mushegian, Arcady R., Koonin, E.V., 1996. Gene order is not conserved in bacterial evolution. Trends Genet. 12, 289–290.
Musialowski, M.S., Flett, F., Scott, G.B., Hobbs, G., Smith, C.P., Oliver, S.G., 1994. Functional evidence that the principal DNA replication origin of the Streptomyces coelicolor chromosome is close to the dnaA-gyrB region. J. Bacteriol. 176, 5123–5125. https://doi.org/10.1128/jb.176.16.5123-5125.1994
Myronovskyi, M., Tokovenko, B., Brötz, E., Rückert, C., Kalinowski, J., Luzhetskyy, A., 2014.

Genome rearrangements of Streptomyces albus J1074 lead to the carotenoid gene cluster activation. Appl. Microbiol. Biotechnol.
Nagarajan, N., Pop, M., 2013. Sequence assembly demystified. Nat. Rev. Genet. 14, 157–167. https://doi.org/10.1038/nrg3367
Nakamura, Y., Itoh, T., Matsuda, H., Gojobori, T., 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat. Genet. 36, 760. https://doi.org/10.1038/ng1381
Nett, M., Ikeda, H., Moore, B.S., 2009. Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat. Prod. Rep. 26, 1362. https://doi.org/10.1039/b817069j
Niki, H., Yamaichi, Y., Hiraga, S., 2000. Dynamic organization of chromosomal DNA in Escherichia coli. Genes Dev. 14, 212–223. https://doi.org/10.1101/gad.14.2.212
Nuñez, P.A., Romero, H., Farber, M.D., Rocha, E.P.C., 2013. Natural selection for operons depends on genome size. Genome Biol. Evol. 5, 2242–2254. https://doi.org/10.1093/gbe/evt174
O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745. https://doi.org/10.1093/nar/gkv1189
Oliveira, P.H., Touchon, M., Rocha, E.P.C., 2014. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Res. 42, 10618–10631. https://doi.org/10.1093/nar/gku734
Omura, S., Takahashi, Y., Iwai, Y., Tanaka, H., 1982. KITASATOSPORIA, A NEW GENUS OF THE ORDER ACTINOMYCETALES. J. Antibiot. (Tokyo) 35, 1013–1019. https://doi.org/10.7164/antibiotics.35.1013
Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., Vonstein, V., 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 33, 5691–5702. https://doi.org/10.1093/nar/gki866
Park, C., Zhang, J., 2012. High Expression Hampers Horizontal Gene Transfer. Genome Biol. Evol. 4, 523–532. https://doi.org/10.1093/gbe/evs030
Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O., 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc. Natl. Acad. Sci. 96, 4285–4288. https://doi.org/10.1073/pnas.96.8.4285
Perna, N.T., Plunkett, G., Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Pósfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A., Blattner, F.R., 2001. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature 409, 529–533. https://doi.org/10.1038/35054089
Pernodet, J.-L., Simonet, J.-M., Guérineau, M., 1984. Plasmids in different strains of Streptomyces ambofaciens: free and integrated form of plasmid pSAM2. Mol. Gen. Genet. MGG 198,

35–41. https://doi.org/10.1007/BF00328697
Pfreundt, U., Kopf, M., Belkin, N., Berman-Frank, I., Hess, W.R., 2014. The primary transcriptome of the marine diazotroph *Trichodesmium erythraeum* IMS101. Sci. Rep. 4, 6187. https://doi.org/10.1038/srep06187
Picardeau, M., Lobry, J.R., Hinnebusch, B.J., 1999. Physical mapping of an origin of bidirectional replication at the centre of the Borrelia burgdorferi linear chromosome. Mol. Microbiol. 32, 437–445. https://doi.org/10.1046/j.1365-2958.1999.01368.x
Postow, L., Hardy, C.D., Arsuaga, J., Cozzarelli, N.R., 2004. Topological domain structure of the Escherichia coli chromosome. Genes Dev. 18, 1766–1779. https://doi.org/10.1101/gad.1207504
Prieto, A.I., Kahramanoglou, C., Ali, R.M., Fraser, G.M., Seshasayee, A.S.N., Luscombe, N.M., 2012. Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. Nucleic Acids Res. 40, 3524–3537. https://doi.org/10.1093/nar/gkr1236
Randall-Hazelbauer, L., Schwartz, M., 1973. Isolation of the Bacteriophage Lambda Receptor from Escherichia coli. J. Bacteriol. 116, 1436–1446.
Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N.R., Chaudhuri, R., Henderson, I.R., Sperandio, V., Ravel, J., 2008. The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates. J. Bacteriol. 190, 6881–6893. https://doi.org/10.1128/JB.00619-08
Redenbach, M., Kleinert, E., Stoll, A., 2000. Identification of DNA amplifications near the center of the Streptomyces coelicolor M145 chromosome. FEMS Microbiol. Lett. 191, 123–129. https://doi.org/10.1111/j.1574-6968.2000.tb09328.x
Reeves, A.R., Post, D.A., Vanden Boom, T.J., 1998. Physical-genetic map of the erythromycin-producing organism Saccharopolyspora erythraea. Microbiology 144, 2151–2159. https://doi.org/10.1099/00221287-144-8-2151
Richardson, C., Jasin, M., 2000. Frequent chromosomal translocations induced by DNA double-strand breaks. Nature 405, 697–700. https://doi.org/10.1038/35015097
Richter, M., Rosselló-Móra, R., 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. 106, 19126–19131.
Rocha, E.P., 2003. DNA repeats lead to the accelerated loss of gene order in bacteria. TRENDS Genet. 19, 600–603.
Rocha, E.P.C., 2006. Inference and Analysis of the Relative Stability of Bacterial Chromosomes. Mol. Biol. Evol. 23, 513–522. https://doi.org/10.1093/molbev/msj052
Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A., Koonin, E.V., 2002. Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res. 30, 2212–2223. https://doi.org/10.1093/nar/30.10.2212
Rosselló-Mora, R., Amann, R., 2001. The species concept for prokaryotes. FEMS Microbiol. Rev. 25, 39–67. https://doi.org/10.1111/j.1574-6976.2001.tb00571.x
Rückert, C., Kalinowski, J., Heide, L., Apel, A.K., 2014. Draft Genome Sequence of Streptomyces roseochromogenes subsp. oscitans DS 12.976, Producer of the Aminocoumarin Antibiotic Clorobiocin. Genome Announc. 2, e01147-13. https://doi.org/10.1128/genomeA.01147-13
RUDD, B.A.M., HOPWOOD, D.A., 1980. A Pigmented Mycelial Antibiotic in Streptomyces coelicolor: Control by a Chromosomal Gene Cluster. Microbiology, 119, 333–340. https://doi.org/10.1099/00221287-119-2-333
RUDD, B.A.M., HOPWOOD, D.A., 1979. Genetics of Actinorhodin Biosynthesis by Streptomyces coelicolor A3(2). Microbiology, 114, 35–43. https://doi.org/10.1099/00221287-114-1-35
Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., Smith, M., 1977a. Nucleotide sequence of bacteriophage φX174 DNA. Nature 265, 687–695. https://doi.org/10.1038/265687a0

Sanger, F., Nicklen, S., Coulson, A.R., 1977b. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 74, 5463–5467. https://doi.org/10.1073/pnas.74.12.5463
Schmid, E., Büchler, C., Altenbuchner, J., 1999. AUD4, a new amplifiable element from Streptomyces lividans. Microbiology, 145, 3331–3341. https://doi.org/10.1099/00221287-145-12-3331
Schmid, M.B., Roth, J.R., 1987. Gene location affects expression level in Salmonella typhimurium. J. Bacteriol. 169, 2872–2875. https://doi.org/10.1128/jb.169.6.2872-2875.1987
Sela, I., Wolf, Y.I., Koonin, E.V., 2016. Theory of prokaryotic genome evolution. Proc. Natl. Acad. Sci. 113, 11399–11407.
Seol, J.-H., Shim, E.Y., Lee, S.E., 2018. Microhomology-mediated end joining: Good, bad and ugly. Mutat. Res. Mol. Mech. Mutagen. 809, 81–87. https://doi.org/10.1016/j.mrfmmm.2017.07.002
Sfeir, A., Symington, L.S., 2015. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? Trends Biochem. Sci. 40, 701–714. https://doi.org/10.1016/j.tibs.2015.08.006
Shirling, E.B., Gottlieb, D., 1966. Methods for characterization of Streptomyces species. Int. J. Syst. Bacteriol. 16, 313–340.
Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351
Singh, S.S., Singh, N., Bonocora, R.P., Fitzgerald, D.M., Wade, J.T., Grainger, D.C., 2014. Widespread suppression of intragenic transcription initiation by H-NS. Genes Dev. 28, 214–219. https://doi.org/10.1101/gad.234336.113
Skippington, E., Ragan, M.A., 2012. Phylogeny rather than ecology or lifestyle biases the construction of *Escherichia coli – Shigella* genetic exchange communities. Open Biol. 2, 120112. https://doi.org/10.1098/rsob.120112
Smillie, C.S., Smith, M.B., Friedman, J., Cordero, O.X., David, L.A., Alm, E.J., 2011. Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480, 241–244. https://doi.org/10.1038/nature10571
Snel, B., 2001. Genomes in Flux: The Evolution of Archaeal and Proteobacterial Gene Content. Genome Res. 12, 17–25. https://doi.org/10.1101/gr.176501
Sobetzko, P., Travers, A., Muskhelishvili, G., 2012. Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. Proc. Natl. Acad. Sci. 109, E42–E50. https://doi.org/10.1073/pnas.1108229109
Stackebrandt, E., Rainey, F.A., Ward-Rainey, N.L., 1997. Proposal for a New Hierarchic Classification System, Actinobacteria classis nov. Int. J. Syst. Bacteriol. 47, 479–491. https://doi.org/10.1099/00207713-47-2-479
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033
Stavans, J., Oppenheim, A., 2006. DNA–protein interactions and bacterial chromosome architecture. Phys. Biol. 3, R1–R10. https://doi.org/10.1088/1478-3975/3/4/R01
Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S.L., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K.-S., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E.W., Lory, S., Olson, M.V., 2000. Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature 406, 959. https://doi.org/10.1038/35023079
Studholme, D.J., 2016. Genome Update. Let the consumer beware: Streptomyces genome sequence quality. Microb. Biotechnol. 9, 3–7. https://doi.org/10.1111/1751-7915.12344
Stuttard, C., 1982. Temperate Phages of Streptomyces venezuelae: Lysogeny and Host Specificity

Shown by Phages SV1 and SV2. Microbiology 128, 115–121. https://doi.org/10.1099/00221287-128-1-115
Svendsen, J.M., Harper, J.W., 2010. GEN1/Yen1 and the SLX4 complex: solutions to the problem of Holliday junction resolution. Genes Dev. 24, 521–536. https://doi.org/10.1101/gad.1903510
Swiercz, J.P., Nanji, T., Gloyd, M., Guarné, A., Elliot, M.A., 2013. A novel nucleoid-associated protein specific to the actinobacteria. Nucleic Acids Res. 41, 4171–4184. https://doi.org/10.1093/nar/gkt095
Swinger, K.K., Lemberg, K.M., Zhang, Y., Rice, P.A., 2003. Flexible DNA bending in HU-DNA cocrystal structures. EMBO J. 22, 3749–3760. https://doi.org/10.1093/emboj/cdg351
Szostak, J.W., Orr-Weaver, T.L., Rothstein, R.J., Stahl, F.W., 1983. The double-strand-break repair model for recombination. Cell 33, 25–35. https://doi.org/10.1016/0092-8674(83)90331-8
TAKEUCHI, T., SAWADA, H., TANAKA, F., MATSUDA, I., 1996. Phylogenetic Analysis of Streptomyces spp. Causing Potato Scab Based on 16S rRNA Sequences. Int. J. Syst. Evol. Microbiol. 46, 476–479. https://doi.org/10.1099/00207713-46-2-476
Tamura, K., Battistuzzi, F.U., Billing-Ross, P., Murillo, O., Filipski, A., Kumar, S., 2012. Estimating divergence times in large molecular phylogenies. Proc. Natl. Acad. Sci. 109, 19333–19338. https://doi.org/10.1073/pnas.1213199109
Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., DeBoy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Ros, I.M. y, Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J.B., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M., 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." Proc. Natl. Acad. Sci. 102, 13950–13955. https://doi.org/10.1073/pnas.0506758102
Tindall, B.J., Rosselló-Móra, R., Busse, H.-J., Ludwig, W., Kämpfer, P., 2010. Notes on the characterization of prokaryote strains for taxonomic purposes. Int. J. Syst. Evol. Microbiol. 60, 249–266. https://doi.org/10.1099/ijs.0.016949-0
Tokala, R.K., Strap, J.L., Jung, C.M., Crawford, D.L., Salove, M.H., Deobald, L.A., Bailey, J.F., Morra, M.J., 2002. Novel plant-microbe rhizosphere interaction involving Streptomyces lydicus WYEC108 and the pea plant (Pisum sativum). Appl. Environ. Microbiol. 68, 2161–2171. https://doi.org/10.1128/aem.68.5.2161-2171.2002
Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., Karoui, M.E., Frapy, E., Garry, L., Ghigo, J.M., Gilles, A.M., Johnson, J., Bouguénec, C.L., Lescat, M., Mangenot, S., Martinez-Jéhanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M.A., Pichon, C., Rouy, Z., Ruf, C.S., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Médigue, C., Rocha, E.P.C., Denamur, E., 2009. Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. PLOS Genet. 5, e1000344. https://doi.org/10.1371/journal.pgen.1000344
Tourand, Y., Kobryn, K., Chaconas, G., 2003. Sequence-specific recognition but position-dependent cleavage of two distinct telomeres by the Borrelia burgdorferi telomere resolvase, ResT. Mol. Microbiol. 48, 901–911. https://doi.org/10.1046/j.1365-2958.2003.03485.x
Toussaint, M., 2018. Exploitation et exploration de la diversité génétique d'une population naturelle de Streptomyces issue d'un micro-habitat sol (thesis). http://www.theses.fr. Université de Lorraine.
Truong, L.N., Li, Y., Shi, L.Z., Hwang, P.Y.-H., He, J., Wang, H., Razavian, N., Berns, M.W., Wu, X., 2013. Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. Proc. Natl.

Acad. Sci. U. S. A. 110, 7720–7725. https://doi.org/10.1073/pnas.1213431110
Tsai, H.-H., Shu, H.-W., Yang, C.-C., Chen, C.W., 2012. Translesion-synthesis DNA polymerases participate in replication of the telomeres in Streptomyces. Nucleic Acids Res. 40, 1118–1130. https://doi.org/10.1093/nar/gkr856
Tsai, Y.-M., Chang, A., Kuo, C.-H., 2018. Horizontal Gene Acquisitions Contributed to Genome Expansion in Insect-Symbiotic Spiroplasma clarkii. Genome Biol. Evol. 10, 1526–1532. https://doi.org/10.1093/gbe/evy113
Uchida, T., Miyawaki, M., Kinashi, H., 2003. Chromosomal Arm Replacement in Streptomyces griseus. J. Bacteriol. 185, 1120–1124. https://doi.org/10.1128/JB.185.3.1120-1124.2003
Valens, M., Penaud, S., Rossignol, M., Cornet, F., Boccard, F., 2004. Macrodomain organization of the Escherichia coli chromosome. EMBO J. 23, 4330–4341. https://doi.org/10.1038/sj.emboj.7600434
Valkenburg, J.A., Woldringh, C.L., 1984. Phase separation between nucleoid and cytoplasm in Escherichia coli as defined by immersive refractometry. J. Bacteriol. 160, 1151–1157.
Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G.F., Chater, K.F., Sinderen, D. van, 2007. Genomics of Actinobacteria: Tracing the Evolutionary History of an Ancient Phylum. Microbiol. Mol. Biol. Rev. 71, 495–548. https://doi.org/10.1128/MMBR.00005-07
Volff, J.-N., Altenbuchner, J., 2000. A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. FEMS Microbiol. Lett. 186, 143–150. https://doi.org/10.1111/j.1574-6968.2000.tb09095.x
Volff, J.-N., Altenbuchner, J., 1998. Genetic instability of the Streptomyces chromosome. Mol. Microbiol. 27, 239–246. https://doi.org/10.1046/j.1365-2958.1998.00652.x
Waksman, S.A., Henrici, A.T., 1943. The Nomenclature and Classification of the Actinomycetes. J. Bacteriol. 46, 337–341.
Wang, S.-J., Chang, H.-M., Lin, Y.-S., Huang, C.-H., Chen, C.W., 1999. Streptomyces genomes: circular genetic maps from the linear chromosomes. Microbiology 145, 2209–2220. https://doi.org/10.1099/00221287-145-9-2209
Wang, Y., Yang, J.K., Lee, O.O., Li, T.G., Al-Suwailem, A., Danchin, A., Qian, P.-Y., 2011. Bacterial Niche-Specific Genome Expansion Is Coupled with Highly Frequent Gene Disruptions in Deep-Sea Sediments. PLOS ONE 6, e29149. https://doi.org/10.1371/journal.pone.0029149
Ward, J.F., 1994. The Complexity of DNA Damage: Relevance to Biological Consequences. Int. J. Radiat. Biol. 66, 427–432. https://doi.org/10.1080/09553009414551401
Watt, V.M., Ingles, C.J., Urdea, M.S., Rutter, W.J., 1985. Homology requirements for recombination in Escherichia coli. Proc. Natl. Acad. Sci. 82, 4768–4772. https://doi.org/10.1073/pnas.82.14.4768
Weaver, D., Karoonuthaisiri, N., Tsai, H.-H., Huang, C.-H., Ho, M.-L., Gai, S., Patel, K.G., Huang, J., Cohen, S.N., Hopwood, D.A., Chen, C.W., Kao, C.M., 2004. Genome plasticity in Streptomyces: identification of 1 Mb TIRs in the S. coelicolor A3(2) chromosome. Mol. Microbiol. 51, 1535–1550. https://doi.org/10.1111/j.1365-2958.2003.03920.x
Weller, G.R., Kysela, B., Roy, R., Tonkin, L.M., Scanlan, E., Della, M., Devine, S.K., Day, J.P., Wilkinson, A., d'Adda di Fagagna, F., Devine, K.M., Bowater, R.P., Jeggo, P.A., Jackson, S.P., Doherty, A.J., 2002. Identification of a DNA nonhomologous end-joining complex in bacteria. Science 297, 1686–1689. https://doi.org/10.1126/science.1074584
Wellington, E.M.H., Stackebrandt, E., Sanders, D., Wolstrup, J., Jorgensen, N.O.G., 1992. Taxonomic Status of Kitasatosporia, and Proposed Unification with Streptomyces on the Basis of Phenotypic and 16S rRNA Analysis and Emendation of Streptomyces Waksman and Henrici 1943, 339AL. Int. J. Syst. Bacteriol. 42, 156–160. https://doi.org/10.1099/00207713-42-1-156
Werren, J.H., 2011. Selfish genetic elements, genetic conflict, and evolutionary innovation. Proc. Natl. Acad. Sci. 108, 10863–10870. https://doi.org/10.1073/pnas.1102343108
Werren, J.H., Nur, U., Wu, C.-I., 1988. Selfish genetic elements. Trends Ecol. Evol. 3, 297–302.

https://doi.org/10.1016/0169-5347(88)90105-X
Widenbrant, E.M., Kao, C.M., 2007. Introduction of the Foreign Transposon Tn4560 in Streptomyces coelicolor Leads to Genetic Instability near the Native Insertion Sequence IS1649. J. Bacteriol. 189, 9108–9116. https://doi.org/10.1128/JB.00983-07
Widenbrant, E.M., Tsai, H.-H., Chen, C.W., Kao, C.M., 2008. Spontaneous Amplification of the Actinorhodin Gene Cluster in Streptomyces coelicolor Involving Native Insertion Sequence IS466. J. Bacteriol. 190, 4754–4758. https://doi.org/10.1128/JB.00131-08
Williams, S.T., Goodfellow, M., Alderson, G., Wellington, E.M.H., Sneath, P.H.A., Sackin, M.J., 1983. Numerical Classification of Streptomyces and Related Genera. Microbiology 129, 1743–1813. https://doi.org/10.1099/00221287-129-6-1743
Wilson, T.E., Topper, L.M., Palmbos, P.L., 2003. Non-homologous end-joining: bacteria join the chromosome breakdance. Trends Biochem. Sci. 28, 62–66. https://doi.org/10.1016/S0968-0004(03)00005-7
Wright, M.A., Kharchenko, P., Church, G.M., Segrè, D., 2007. Chromosomal periodicity of evolutionarily conserved gene pairs. Proc. Natl. Acad. Sci. 104, 10559–10564. https://doi.org/10.1073/pnas.0610776104
Wyrobek, A.J., 2005. Relative Susceptibilities of Male Germ Cells to Genetic Defects Induced by Cancer Chemotherapies. J. Natl. Cancer Inst. Monogr. 2005, 31–35. https://doi.org/10.1093/jncimonographs/lgi001
Yancopoulos, S., Attie, O., Friedberg, R., 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21, 3340–3346. https://doi.org/10.1093/bioinformatics/bti535
Yang, C.-C., Tseng, S.-M., Pan, H.-Y., Huang, C.-H., Chen, C.W., 2017. Telomere associated primase Tap repairs truncated telomeres of Streptomyces. Nucleic Acids Res. 45, 5838–5849. https://doi.org/10.1093/nar/gkx189
Yang, M.C., Losick, R., 2001. Cytological Evidence for Association of the Ends of the Linear Chromosome in Streptomyces coelicolor. J. Bacteriol. 183, 5180–5186. https://doi.org/10.1128/JB.183.17.5180-5186.2001
Zanetti, J.P.P., Biller, P., Meidanis, J., 2016. Median Approximations for Genomes Modeled as Matrices. Bull. Math. Biol. 78, 786–814. https://doi.org/10.1007/s11538-016-0162-4
Zhang, H., Zhang, W., Jin, Y., Jin, M., Yu, X., 2008. A comparative study on the phylogenetic diversity of culturable actinobacteria isolated from five marine sponge species. Antonie Van Leeuwenhoek 93, 241–248. https://doi.org/10.1007/s10482-007-9196-9
Zhang, Z., Wang, Y., Ruan, J., 1997. A Proposal To Revive the Genus Kitasatospora (Omura, Takahashi, Iwai, and Tanaka 1982). Int. J. Syst. Bacteriol. 47, 1048–1054. https://doi.org/10.1099/00207713-47-4-1048
Zheng, Y., Szustakowski, J.D., Fortnow, L., Roberts, R.J., Kasif, S., 2002. Computational Identification of Operons in Microbial Genomes. Genome Res. 12, 1221–1230. https://doi.org/10.1101/gr.200602
Zhou, Z., Gu, J., Li, Y.-Q., Wang, Y., 2012. Genome plasticity and systems evolution in Streptomyces. BMC Bioinformatics 13, S8. https://doi.org/10.1186/1471-2105-13-S10-S8
Zinder, N.D., Lederberg, J., 1952. GENETIC EXCHANGE IN SALMONELLA. J. Bacteriol. 64, 679–699.

ANNEXE

Name	Genome Assembly ID	Length (pb)	# plasmid	# CDS	GC rate (%)	TIR length (pb)
Streptomyces sp. 4F	ASM148470v1	8047771	0	7250	72.28	334071
Streptomyces sp. 769	ASM81602v1	10100774 [237512]	1	8956 [224]	71.6	155968
S. actuosus ATCC 25421	ASM320803v1	8145579	0	7399	72.53	17181
Streptomyces sp. ADI95 16	ASM372149v1	8184000 [902421]	4	7362 [861]	72.04	-
S. albireticuli MDJK11	ASM219245v1	8144417	0	7024	72.82	133288
S. alboflavus MDJK44	ASM218967v2	9622415 [345134]	2	8796 [335]	72.0	-
S. albulus CK 15	ASM93518v3	9336218	0	8241	72.35	-
S. albus BK3 25	ASM175342v1	8308430	0	6871	72.64	-
S. alfalfae ACCC40021	ASM197502v1	8625867	0	7525	72.09	149241
S. ambofaciens ATCC 23877	ASM126788v1	8303940 [89658]	1	7413 [128]	72.19	202695
S. ambofaciens DSM 40697	ASM163286v1	8137876	0	7250	72.33	212696
S. atratus SCSIO ZH16	ASM333086v1	9641288	0	8700	69.59	288246
S. autolyticus CGMCC0516	ASM198397v1	10029028 [155632]	7	8243 [36]	71.21	-
S. avermitilis MA 4680	ASM976v2	9025608 [94287]	1	8038 [94]	70.7	50
S. bacillaris ATCC 15855	ASM326867v1	7888441	0	6953	71.95	-
S. bingchenggensis BCW 1	ASM9238v1	11936683	0	10216	70.75	139
S. brunneus CR22	ASM395571v1	10393987	0	9540	70.85	96011
S. cattleya NRRL 8057	ASM23730v1	6283062 [1809491]	1	5538 [1651]	73.01	-
Streptomyces sp. CB09001	ASM336979v1	7787608	0	7085	71.95	-
Streptomyces sp. CC0208	ASM344373v1	9320089	0	8531	70.59	-
Streptomyces sp. CCM MD2014	ASM77204v1	8274043	0	7501	72.13	14382
Streptomyces sp. CdTB01	ASM148456v1	9902731 [288836]	1	9157 [287]	71.53	400291
Streptomyces sp. CFMR 7	ASM127809v1	8207742 [99537]	1	7095 [95]	72.03	11739
S. chartreusis NRRL 3882	NRRL3882	8983317	0	8083	71.23	39831
S. clavuligerus F1D 5	Scla_1.0	6900908 [1158217]	2	5831 [1079]	72.53	34
Streptomyces sp. CLI2509	ASM228807v1	7088673 [144028]	1	6198 [149]	73.3	25387
Streptomyces sp. CMB StM0423	ASM284728v1	8029398	0	6930	73.14	9667
Streptomyces sp. CNQ 509	ASM101103v1	8039333	0	7028	73.07	32534
S. coelicolor A3(2)	ASM20383v1	8667507 [387340]	2	7920 [400]	72.0	21653
S. collinus Tu 365	ASM44487v1	8272925 [104361]	2	7182 [101]	72.55	631364
S. cyaneogriseus noncyanogenus NMWT 1	ASM93144v1	7762396	0	6681	72.86	-
Streptomyces sp. endophyte N2	ASM410448v1	8428700	0	7398	71.83	53557
Streptomyces sp. fd1 xmd	ASM200768v1	7929999	0	7236	72.51	-
S. formicae KY5	ASM255654v1	9611874	0	8317	71.38	35482
S. fulvissimus DSM 40593	ASM38594v1	7905758	0	7072	71.48	-
S. fungicidicus TXX3120	ASM366543v1	6740768 [926728]	1	5998 [841]	72.22	-
S. gilvosporeus F607	ASM208219v1	8482298	0	7624	70.95	-
S. glaucescens GLA O	ASM76121v1	7453200 [170574]	1	6575 [140]	72.91	14128
S. globisporus C 1027	ASM26134v2	7608611 [174988]	2	6925 [154]	71.54	18
S. globisporus TFH56	ASM314754v1	7488586 [177935]	2	6847 [155]	71.54	-
S. globosus LZH 48	ASM332537v1	6863360 [672390]	2	6196 [539]	73.65	-
Streptomyces sp. Go 475	ASM333084v1	8570609	0	7729	71.96	6902
S. griseochromogenes ATCC 14511	ASM154262v2	10764674	0	9839	70.76	-
S. griseorubiginosus 3E 1	ASM359523v1	9512378	0	8551	70.94	113251
S. griseoviridis F1 27	ASM399439v1	8963414	0	7785	72.38	58707
S. griseus NBRC 13350	ASM1060v1	8545929	0	7306	72.23	132910
Streptomyces sp. GSSD 12	ASM334496v1	8454852	0	7201	71.19	-
Streptomyces sp. HNM0039	ASM309751v1	7289495	0	6363	72.46	-
S. hundungensis BH38	ASM362781v1	8393044	0	7541	70.94	65122
S. hygroscopicus jinggangensis 5008	ASM24535v1	10145833 [237851]	2	9158 [243]	71.84	15
S. hygroscopicus XM201	ASM202187v1	12012215	0	10288	70.75	-
S. koyangensis VK A60T	ASM342892v1	7220839	0	6365	73.03	20855
Streptomyces sp. KPB2	ASM395005v1	8082236	0	7256	72.21	40616
S. lavendulae CCM 3239	ASM280384v1	8691711 [241081]	1	7854 [231]	72.62	237734
S. leeuwenhoekii sleC34	sleC34	7903895 [218596]	2	6857 [263]	72.68	388272
S. lincolnensis LC G	ASM334444v1	9513637	0	8622	71.06	-
S. lincolnensis NRRL 2936	ASM168535v1	8396100 [174091]	2	7724 [149]	69.75	-
S. lunaelactis MM109	ASM305455v1	7367863	0	6574	72.05	-
S. luteoverticillatus CGMCC 15060	ASM397071v1	8201357	0	7329	72.22	-
S. lydicus 103	ASM172948v1	9307519	0	8534	70.68	52505
S. lydicus A02	ASM95203v2	9125666	0	7970	70.8	8910
S. lydicus WYEC 108	ASM399437v1	8718751	0	8025	70.69	-
Streptomyces sp. M2	ASM410450v1	7868178 [848015]	3	7052 [812]	72.13	-
Streptomyces sp. Mg1	ASM41226v2	10677137	0	9604	70.63	131471
Streptomyces sp. MK45	ASM396353v1	8399509	0	7522	71.99	107206
Streptomyces sp. MOE7	ASM209033v1	9641634 [45805]	1	8456 [52]	70.78	-
Streptomyces sp. NEAU S7GS2	ASM317327v1	7641029	0	6992	71.91	107944
S. nigra 452	ASM307405v1	7990492	0	7105	70.46	-
S. niveus SCSIO 3406	ASM200917v1	9815884	0	8811	71.45	170209
S. noursei ATCC 11455	ASM170427v1	8180260	0	7384	72.41	141244
S. olivaceus KLBMP 5084	ASM176737v1	8809793	0	7593	71.11	170471
S. olivoreticuli ATCC 31159	ASM339113v1	7180417	0	6166	72.45	-
S. ongiicola HNM0071	ASM312236v1	9851971	0	8820	71.37	-
Streptomyces sp. P3	ASM303247v1	8550793	0	7375	72.06	239610

S. pactum ACT12	ASM200522v1	7149446 [617085]	1	6520 [443]	72.73	14
S. parvulus 2297	ASM166004v1	8023114	0	7391	70.59	143511
S. peucetius caesius ATCC 27952	ASM277753v1	7346075	0	6561	69.94	-
S. pluripotens MUSC 135	ASM80224v2	7337497 [318607]	2	6583 [312]	70.99	800
S. pratensis ATCC 33331	ASM17611v2	8532592	0	7498	71.54	419828
S. pristinaespiralis HCCB 10218	ASM127807v1	9698948	0	8823	71.09	112391
S. puniciscabiei TW1S1	ASM173580v1	8353915 [982309]	3	7437 [832]	72.72	-
S. reticuli TUE45	TUE45	11142275	0	9796	71.28	19211
Streptomyces sp. RTd22	ASM165021v1	6543262	0	5602	74.78	39368
S. rubrolavendulae MJM4426	ASM175078v1	7614683	0	7110	71.49	39495
Streptomyces sp. S063	ASM283267v1	9083372	0	8241	71.26	134209
Streptomyces sp. S10 2016	ASM161179v1	7529075 [72789]	1	6546 [78]	72.29	12184
Streptomyces sp. S8	ASM209499v1	7472530	0	6503	73.15	-
S. scabiei 87 22	ASM9130v1	10148695	0	9009	71.45	18488
Streptomyces sp. SCSIO 03032	ASM212830v1	6287975	0	5506	73.52	-
Streptomyces sp. Sge12	ASM208045v1	7983613 [127085]	1	7257 [119]	72.17	162199
Streptomyces sp. SirexAA E	ASM17719v2	7414440	0	6663	71.75	-
Streptomyces sp. SM17	ASM291072v2	6975788 [204126]	3	6107 [186]	73.35	-
Streptomyces sp. SM18	ASM291077v2	7703166	0	6783	71.84	14611
Streptomyces sp. TN58	ASM194184v1	7585034	0	6936	72.3	193939
Streptomyces sp. Tue6075	ASM193163v1	7931832	0	6994	71.57	12088
S. venezuelae ATCC 15439	ASM144362v1	9054831	0	8130	71.74	48229
S. venezuelae NRRL B 65442	ASM188659v1	8222198 [158122]	1	7422 [158]	72.42	-
S. vietnamensis GIM4 0001	ASM83000v1	8867142 [286635]	1	8016 [271]	71.99	49821
S. violaceoruber S21	ASM208217v1	7916045	0	6979	72.65	4122
S. violaceusniger Tu 4113	ASM14781v3	10657107 [481206]	2	9087 [523]	70.88	-
Streptomyces sp. W1SF4	ASM395003v1	7272878 [795265]	2	6764 [656]	73.14	-
Streptomyces sp.WAC 01438	ASM394552v1	8138328 [62839]	1	7278 [72]	71.96	49769
Streptomyces sp. WAC 01529	ASM394554v1	8270461	0	7269	71.32	58922
Streptomyces sp. WAC 06738	ASM394550v1	8324535	0	7207	73.18	11391
Streptomyces sp. WAC00288	ASM294389v1	7467777 [385082]	4	6808 [379]	72.73	-
S. xiamenensis 318	ASM99378v2	5961401	0	5411	72.02	-
S. xinghaiensis S187	ASM22070v2	7137891	0	6223	73.14	-
Streptomyces sp. XZHG99	ASM294683v1	8541354 [168817]	2	8026 [210]	69.92	21003
Streptomyces sp. Z022	ASM367532v1	8085191 [72694]	1	7187 [55]	72.12	26248
Streptomyces sp. ZFG47	ASM326105v1	9269371 [969901]	1	8198 [772]	70.76	191546

Annexe 1- Principales informations sur les 110 espèces de *Streptomyces* restantes après réduction de la collection 135. Name: nom scientifique ; Genome Assembly ID : identificateur unique utilisé dans la base de données Genome Assembly du NCBI ; Length (pb) : longueur du chromosome [longueur cumulative des plasmides] exprimée en pb ; # plasmid : nombre de plasmides ; # CDS : CDS dans le chromosome [CDS dans les plasmides] ; GC rate (%) : pourcentage de cytosine et de guanine dans la séquence ; TIR length (pb) : Taille d'un exemplaire d'une TIR exprimée en pb.

université PARIS-SACLAY

ECOLE DOCTORALE N°577
Structure et Dynamique des Systèmes Vivants (SDSV)

Titre : Dynamique des génomes du genre *Streptomyces*

Mots clés : *Streptomyces*, génomique comparée, synténie, transfert horizontal

Résumé : Les *Streptomyces* possèdent un grand chromosome linéaire (6-12 Mb) constitué d'une région centrale conservée, bordée des bras variables représentant jusqu'à plusieurs mégabases. Afin d'étudier l'évolution du chromosome au cours de l'évolution, un panel représentatif de souches et d'espèces de *Streptomyces* (110 génomes) dont les chromosomes sont complètement séquencés a été identifié. Le pangénome du genre a été modélisé et s'est révélé ouvert et présente un core-génome de 803 gènes.

L'évolution du chromosome de *Streptomyces* a été analysée en effectuant des comparaisons entre paires de chromosomes, grâce à des indices mesurant la conservation de l'ordre des gènes et du taux d'orthologie. Cette méthode, appliquée au niveau global, a également été appliquée au niveau local, permettant de mesurer l'intensité de la recombinaison le long du chromosome. Grâce à la profondeur phylogénétique offerte par le panel choisi, il a été possible de déduire que les bras chromosomiques ont évolué plus rapidement que la région centrale sous l'effet combiné des réarrangements et de l'ajout de nouvelles informations provenant du transfert horizontal.

Title : Genome dynamics of the *Streptomyces* genus

Keywords : *Streptomyces*, comparative genomics, synteny, horizontal transfer

Abstract : *Streptomyces* possess a large linear chromosome (6-12 Mb) consisting of a conserved central region flanked by variable arms covering several megabases. In order to study the evolution of the chromosome across evolutionary times, a representative panel of *Streptomyces* strain and species (110) whose chromosomes are completely sequenced was identified. The pangenome of the genus was modelized and shown to be open with a core- genome reaching 803 genes.

The evolution of *Streptomyces* chromosome was analyzed by carrying out pairwise comparisons, and by monitoring indexes measuring the conservation of gene order and the rate of orthology. This method, applied at the global level, was also applied at the local level, making it possible to measure recombination intensity along the chromosome. Using the phylogenetic depth offered by the chosen panel, it was possible to infer that the chromosomal arms evolved faster than the central region under the combined effect of rearrangements and the addition of new information from horizontal transfer.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France