الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la
□□عة زيان عاشور-الجلفة
Université Ziane Achour – Djelfa
كلية علوم الطبيعة و الحياة
Faculté des Sciences de la Nature et de la Vie
Département de Biologie

Mémoire

En vue de l'obtention du Diplôme de Master en Biologie

Option : ***Microbiologie Appliquée***

Analyse comparative de génomes microbiens d'agents PGPR appliquée à la recherche de biofertilisants

Présenté par : - OKACHA Meriem
- DAOUDI Nour Elhouda

Soutenu le :

Devant le jury composé de :

Président : BOUTAIBA Saad	MCA	Univ. Djelfa
Promoteur : BELAOUNI Hadj Ahmed	MAA	Univ. Djelfa
Examinateurs : BERRABAH Fathi	MCB	Univ. Djelfa
OUNISSI Mourad	MAA	Univ. Djelfa

Année universitaires : 2017/2018

Remerciement

Nous exprimons tout d'abord, nos profonds remerciement à DIEU tout puissant, qui nous a guidé sur le droit chemin et nous a donné le courage et la volonté de finir ce mémoire.

Au terme de ce travail, nous tenons à exprimer toutes nos profonde gratitudes et remerciements à notre promoteur BELAOUNI HADJ AHMED pour avoir accepté la responsabilité de nous diriger et pour ses conseils et ses critiques constructives.

Nous tenons à exprimer notre grande considération et nos profonds remerciements aux membres de jury qui nous ont fait l'honneur de juger notre travail.

Nous remercions ELOTTRI MILOUD , ASMAA , FELLA, FAIZA ,AMINE, FARID, Cherif et Karim et nos professeurs qui nous ont porté leur aide.

MERIEM *NOUR ELHOUDA*

DEDICACES

A nos très chers parents, nous n'oublions jamais vos sacrifices exprimés à notre égard, votre attention corrective et votre dévouement pour notre éducation.

A nos chères sœurs, nos chers frères, a nos chers amies, pour tous les moments que nous avons partagés.

A tous les membres des familles : ***OKACHA*** et ***DAOUDI***

Sommaire

LISTE DES ABREVIATION:

PGPR : Plant Growth Promoting Rhizobateria

AND : Acide désoxyribonucléique

ISR : Induced Systemic Resistance

ACC : aminocycopropane-1-carboxylic acid

G : Guanine

C : Cytosine

CDS: coding sequences

TP : Terminal Protein

TAP : Telomere Associated Protein

ARNt : Acide ribonucléique de transfert

ARNr : Acide ribonucléique ribosomique

TIR : Terminal Inverted Repeats

CDS : Coding Sequence

UTR: Untranslated region

TSS: Transcription start site

IPP : interaction protéine-protéine

NGS : Next (ou New) Generation Sequencing

SOLiD: Sequencing by Oligonucleotide Ligation and Detection

PCR : Polymerase chain reaction

FFP : Feature Frequency Profiles

UBCG: up-to-date bacterial core gene

BLAST: Basic Local Alignment Search

RAST : Rapid Annotation using the Subsystems Technology

PGAP : Prokaryotic Genome Annotation Pipeline

UPGMA: Unweighted pair group method with arithmetic mean

MEGA: Molecular Evolutionary Genetics Analysis

NJ : Neighbourg Joining

MSA : Alignement de Séquences Multiples

GSI :indice de support de gène

HMM : Hidden Markov Model

NCBI : National Center for Biotechnology Information

EMBL : (European Molecular Biology Laboratory

BRIG: BLAST Ring Image Generator

SAM : Sequence Alignment MAP

NRPS :non-ribosomal peptides synthetase

LISTE DES FIGURES :

INTRODUCTION

Certaines bactéries appartenant à la rhizosphère, nommées « PGPR », ou *rhizobactéries promotrices de la croissance des plantes*, sont capables d'exercer un effet bénéfique sur la croissance des plantes en augmentant la qualité des nutriments et en stimulant les mécanismes de défense inductibles chez l'hôte, entre autres, et de ce fait sont utilisées en agriculture pour la biofertilisation des sols. Parmi ces dernières, le genre *Streptomyces* se distingue par moult études confortant le caractère PGPR de plusieurs espèces au sein de ce genre, justifiant ainsi le choix du présent mémoire ayant porté sur un groupe de souches appartenant à ce genre, et ce précisément pour une analyse comparative de leurs génomes.

Les études de génomique comparative reposent sur le principe selon lequel toute séquence d'ADN est soumise à la sélection naturelle. Ainsi, des caractères communs entre deux organismes seront souvent codés par des séquences génomiques conservées entre ces deux espèces (Miller et *al*., 2004 ; Hardison et *al*., 2003). À l'inverse, les séquences qui codent pour des fonctions spécifiques à chacune des espèces auront des profils de séquences divergents. Ce processus dit « d'annotation des génomes » en séquences fonctionnelles, est indispensable à la compréhension des phénomènes biologiques au niveau moléculaire et cellulaire, et a bénéficié largement de l'automatisation des programmes informatiques pour autoriser la comparaison de larges séquences d'ADN et maintenant de génomes entiers.

L'objectif principale de notre étude consiste à analyser les génomes de souches candidates ou PGPR confirmés dans le but d'extraire des informations par rapport à leur caractère/potentiel PGPR appliqué au développement de biofertilisants, et ce en mettant en exergue la présence de particularité génétique en relation avec les mécanismes PGPR.

Notre travail est structuré en trois parties :

- Une première partie représente la recherche bibliographique qui débute par une généralité sur les PGPR et leur effet sur les plantes, puis une présentation des *Streptomyces* spp. ayant des traits PGPR, pour ensuite parler de la génomique et de l'annotation de génomes, en définissant par la même occasion les concepts de génomique comparative et les différentes approches suivies dans la phylogénomie.
- Une deuxième partie, « pratique », expose la démarche scientifique abordée lors de la réalisation de ce mémoire de fin d'étude, relatant le choix des souches (screening de *Streptomyces* spp. en relation avec les plantes), leur annotation, suivi par une étude

phylogénétique et phylogénomique des souches examinées par différentes approches, puis la comparaison des profiles génétiques de métabolisme secondaire, ainsi qu'une étude des particularités génétique des souches en question en extrapolant le tout sur le potentiel PGPR, en présentant à la fin de cette partie les résultats obtenues et leur discussion suite à l'analyse bioinformatique initiée au cours de cette investigation.

- Enfin, le tout est couronné par une conclusion retraçant les principaux aboutissements de notre analyse, ouvrant horizon à d'intéressantes perspectives.

Chapitre 1
Les PGPR

Chapitre 1 : Les PGPR

1.1. Définition

La rhizosphère est le lieu de multiples interactions entre microorganismes et racine, étant bénéfiques, nuisibles, ou neutre pour la plante (Bais et *al*., 2006). Certains microorganismes naturellement présents dans les sols sont bénéfiques pour la plante, ce qui améliore souvent la croissance végétale (Morgan et *al*., 2005). Ces microorganismes phytobénéfiques sont de deux types :

A. Ceux qui établissent une relation de symbiose (association à bénéfices réciproques) véritable avec la plante, avec comme exemples types la symbiose entre certains champignons du sol et la plupart des espèces végétales (Smith et Read, 1997), améliorant principalement la nutrition (phosphatée et azotée) de la plante, ainsi que sa capacité à résister au stress (dessiccation du sol, présence de métaux lourds en quantité importante) (Barea et *al*., 2002), ou encore la symbiose fixatrice d'azote entre l'α-protéobacteria *Rhizobium* et les légumineuses (Long, 1996), ou entre l'actinobactérie *Frankia* et les plantes actinorhiziennes (Huguet et *al*., 2005), une symbiose impliquant la formation de structures spécialisées au niveau des racines de la plante : les nodosités (Gage, 2004), améliorant de manière très significative la nutrition azotée du partenaire végétal.

B. Ceux qui restent à l'état libre dans le sol, souvent proches ou sur les racines, et parfois localisées à l'intérieur des racines (endophytes) (Gray et Smith, 2005), et qui établissent une relation facultative à bénéfices réciproques appelée *coopération ou symbiose associative*. Les bactéries PGPR (*Plant Growth-Promoting Rhizobacteria*) font partie de ce type de microorganismes, car elles n'établissent pas une relation de symbiose, mais favorisent la croissance des plantes auxquelles elles sont associées (Benmati, 2014).

Les PGPR ou «*Plant Growth Promoting Rhizobateria*» sont des bactéries qui se développent dans la rhizosphère, et qui ont un effet positif sur la plante, pour ces effets on les considère comme rhizobactéries promotrice de la croissance végétale (Dey et *al*., 2004 ; Herman et *al*., 2008 ; Microrsky, 2008). Ces bactéries sont utilisées en agriculture pour la biofertilisation des sols (Glick, 1995), et ce par plusieurs mécanismes agissant généralement de manière complexe et multiples, aboutissant à un effet positif, tel que la fixation par exemple de l'azote atmosphérique qui pourra être par la suite utilisé par les plantes, améliorant ainsi leur croissance lorsque l'azote du sol est limitant.

1.2. Les Interactions PGPR-Plante

1.2.1. La promotion directe

Ce mécanisme comprend la stimulation bactérienne des phytohormones (auxine ou cytokinine) par une modification de l'équilibre hormonal de la plante, et l'augmentation de la qualité de nutriments disponibles (fixation libre de l'azote, solubilisation du phosphate, etc.). Cela permet à la plante de développer un système racinaire abondant lui permettant notamment de coloniser une plus grande surface de sol, et améliorer l'état nutritionnel des plantes (Beauchamp, 1993; Kloepper, 1993, Ramos et *al*., 2009).

1.2.2. La promotion indirecte

Ce mécanisme repose sur la capacité des PGPR à réduire les effets nocifs pour la plante : la dégradation des xénobiotique dans les sols contaminés par la production des métabolites qui sont toxique aux pathogènes du sol, et l'hydrolyse des molécules libérées par des agents pathogènes par exemple quelques *Pseudomonas*, qui sont capables de décomposer l'acide fusarique (un composé responsable de la pourriture des racines causée par les champignons) (Ramos et *al*., 2009). L'effet phytobénéfiques indirect des bactéries PGPR résulte d'interactions entre PGPR et des pathogènes et/ou parasites de la plante, à l'occasion desquelles les effets négatifs de ces derniers sont diminués (Ramette et *al*., 2006 ; Rezzonico et *al*., 2007).

1.3. Mécanismes d'action des PGPR

Les PGPR peuvent favoriser la croissance des plantes hôtes par divers mécanismes :

A .La fixation d'azote (N_2) : Le sol contient de nombreuses espèces de bactéries pouvant transformer l'azote atmosphérique en ammoniac. Plusieurs de ces microorganismes vivent à la surface des racines des plantes ou même dans les tissus de certains végétaux. L'ammoniac est rapidement transformé en nitrates par les bactéries du sol (Benmati, 2014).

B. La résistance aux pathogènes du sol : Certaines souches de PGPR ont la capacité d´excréter des métabolites actifs contre différentes bactéries et champignons. Certaines de ces molécules sont de véritables antibiotiques, qui jouent un rôle important dans l´inactivation des facteurs de germination du pathogène ou la dégradation de leurs facteurs de pathogénicité comme les toxines (Benmati, 2014).

C. Induction de l'immunité : Certaines PGPR peuvent stimuler le système immunitaire des plantes et leur permettre une résistance contre certains virus, les champignons et même les bactéries pathogènes. Le phénomène est désigné ISR (*Induced Systemic Resistance*) ou résistance systémique induite (Benmati, 2014).

D. Tolérance aux stress : Certaines PGPR produisent des enzymes ACC désaminase qui facilitent le développement des plantes en réduisant leur production d'éthylène (Hydrocarbure gazeux incolore). Les PGPR produisant cet enzyme peuvent ainsi soulager la plante de plusieurs stress causés par des infections, l'absorption de métaux lourds, une salinité élevée et même la sécheresse (Macking, 2007).

L'ensemble de ces activités fait des PGPR une alternative biologique et écologique intéressante à considérer par rapport aux différents produits chimiques de synthèse existant (Benmati, 2014).

1.4. Les biofertilisants

1.4.1. Rôle et production de biofertilisants

Le biofertilisant est défini comme une substance qui contient des microorganismes vivants qui, lorsqu'ils sont appliqués aux semences, surfaces des plantes ou dans le sol, colonise la rhizosphère ou à l'intérieur de la plante et favorise ainsi la croissance en augmentant la disponibilité des éléments nutritifs primaires à la plante hôte. Cette définition est basée sur la logique que le terme biofertilisant est une contraction du terme fertilisant biologique «*biological fertilizer*» (Vessey, 2003). Par ailleurs, le biofertilisant devrait contenir des organismes vivants qui augmentent la teneur en éléments nutritifs de la plante hôte à travers leur existence permanente en association avec la plante (Vessey, 2003). De plus, le terme biofertilisant ne doit pas être utilisé de manière interchangeable avec les termes suivants : les engrais verts, fumier, cultures intercalaires ou organique complétée d'engrais chimiques (Vessey, 2003).

Cependant, une partie des PGPR ne sont pas considérés comme biofertilisants. Ainsi, Les bactéries qui favorisent la croissance des plantes par le contrôle de l'organisme nuisible sont les biopesticides, mais pas des biofertilisants. Il est intéressant de savoir que certains PGPR favoriser la croissance en agissant à la fois comme biofertilisants et biopesticides (Vessey, 2003).

1.4.2. Commercialisation des biofertilisants

Actuellement, diverses formulations commerciales de PGPR sont en vente. Des formulations bactériologiques de *Rhizobium* spp. sont disponibles dans plusieurs pays afin de favoriser la nodulation des légumineuses et de diminuer la fertilisation azotée des cultures. Depuis 1985 en Chine, des PGPR qui accroissent les rendements sont utilisées dans plusieurs cultures (Beauchamp,1993).

Chapitre 2
Les *Streptomyces*

Chapitre 2 : Les Streptomyces

2.1. Présentation du genre *Streptomyces*

2.1.1. La classe des Actinobactéries

Les Actinobactéries représentent l'une des classes les plus riches du règne bactérien avec 39 familles et 130 genres connus à ce jour. Cette classe regroupe les bactéries à Gram positif possédant un fort taux de G+C dans leur ADN, allant de 51% pour les *Corynebacterium* à plus de 70% chez les *Streptomyces* et les *Frankia*. Les représentants de ce groupe arborent des morphologies et des modes de vie variés. Ainsi on note la présence de coques (ex : *Micrococcus*), de bâtonnets (ex : *Arthrobacter*), de mycélium fragmenté (ex : *Nocardia*) ou de mycélium branché hautement différencié (ex : *Streptomyces*). On retrouve au sein de cette classe des pathogènes (*Mycobacterium* spp., *Nocardia* spp., *Tropheryma* spp., *Corynebacterium* spp.), des bactéries du sol (*Streptomyces* spp.), des commensales de plantes (*Leifsonia* spp.), des bactéries symbiotiques fixatrices d'azote (*Frankia*) ou encore, des bactéries de la flore intestinale (*Bifidobacterium* spp.) (Ventura et *al*., 2007). Ces organismes sont retrouvés dans de nombreux écosystèmes, terrestres et aquatiques, en particulier dans les sols où ces bactéries jouent un rôle dans la décomposition de la matière organique. Certains genres sont capables de produire une grande diversité de métabolites ayant des propriétés biologiques intéressantes, par exemple en médecine (antibiotiques, anticancéreux, etc…) ou en industrie (enzymes de dégradation). Du fait de la présence de ces métabolites, certaines espèces sont particulièrement étudiées (Drago, 2015). exemple de l'azote atmosphérique qui pourra être par la suite utilisé par les plantes, améliorant ainsi leur croissance lorsque l'azote du sol est limitant.

2.1.2. Ecologie des Streptomyces

Le genre *Streptomyces* regroupe les bactéries filamenteuses à Gram positif du sol, strictement aérobies et possédant un cycle de vie inhabituel chez des organismes procaryotes. Dans les sols, les *Streptomyces* existent principalement sous forme de spores (phase de dormance, résistance aux stress) et de mycélium végétatif (phase de multiplication et de colonisation des territoires). En laboratoire, il est possible de cultiver ces organismes en milieu liquide bien que les cellules n'existent pas sous forme planctonique. Il n'y a généralement pas de formation de spores en milieu liquide bien qu'il existe des exceptions pour certaines espèces (ex: *Streptomyces griseus*) (Kendrick et Ensign, 1983).

Outre les sols, les *Streptomyces* sont capables de coloniser la rhizosphère où ils apportent une résistance aux champignons pathogènes par la production de molécules antifongiques. Il a également été rapporté plusieurs cas de symbioses entre des bactéries du genre *Streptomyces* et des insectes de la classe des hyménoptères. On peut notamment citer le mutualisme avec les fourmis fongicultrices *Acromyrmex octospinosus* où la bactérie prévient le développement d'*Escovopsis weberi*, un champignon parasite des cultures fongiques (Seipke et *al*., 2011). Chez la guêpe européenne *Philanthus triangulum*, la présence de *Streptomyces* permet aux larves de résister aux infections (Kroiss et *al*., 2010).

2.2. Génome des *Streptomyces*

Les génomes de plusieurs espèces de *Streptomyces* ont été séquencés et assemblés. C'est le cas des génomes de *S. coelicolor* (Bentley et *al*., 2002), *S. avermitilis* (Ikeda et *al*., 2003) ou encore *S. griseus* (Ohnishi et *al*., 2008) pour ne citer que les plus étudiés. Les génomes de *Streptomyces* sont de grande taille, allant de 8 à 10 Mb, soit deux fois la taille des génomes de *E. coli* (4,6 Mb) ou de *B. subtilis* (4,2 Mb). Ils ont également la particularité d'être linéaires et de posséder un haut taux de G+C (aux alentours de 70%). La densité génique est élevée avec environ 1 gène tous les 1200 paires de bases. C'est notamment le premier exemple de bactérie contenant plus de gènes dans son génome que l'eucaryote *Saccharomyces cerevisiae* (7825 chez *S. coelicolor* contre 6607 chez *S. cerevisiae*) (Drago, 2015).

2.2.1. Organisation du chromosome des *Streptomyces*

La linéarité du chromosome a pu être mise en évidence chez plusieurs espèces : *S. lividans* (Lin et *al*., 1993), *S. coelicolor* (Redenbach et *al*., 1996) ou encore *S. ambofaciens* (Leblond et Decaris 1994). La réplication du chromosome s'effectue de façon bidirectionnelle à partir de l'origine de réplication oriC située au milieu de celui-ci. Une protéine TP (*Terminal Protein*) est covalemment liée aux extrémités 5'. Cette protéine permet, avec les protéines TAP (*Telomere Associated Protein*), la synthèse du dernier fragment d'Okazaki sur le brin retardé. (Drago, 2015).

Le chromosome des *Streptomyces* est décrit comme étant organisé en différentes parties. On distingue le « cœur », partie centrale contenant tous les gènes essentiels : métabolisme primaire, réplication de l'ADN, division cellulaire, ribosomes et ARNt. Les « bras », en revanche, contiennent des gènes non essentiels (en conditions de laboratoire), notamment impliqués dans le métabolisme secondaire. Cela dit, certains gènes du métabolisme

secondaire sont présents dans la partie centrale. Enfin, les TIR, régions inversées répétées (*Terminal Inverted Repeats*), sont comme leur nom l'indique des séquences identiques inversées situées aux extrémités du chromosome. Leur taille est très variable, pouvant aller de 174 pb chez *S. avermitilis* (Ikeda et *al*., 2003) à 550 kb chez *S. rimorus* (Pandza et *al*., 1997).

2.2.2. Impact du taux de G/C dans le génome :

L'ADN des *Streptomyces* est fortement enrichi en bases G/C, qui représentent entre 70 et 74% des nucléotides selon les espèces. Ce taux reste globalement le même dans les régions non codantes. Il est en revanche fortement perturbé dans les régions codantes. On observe ainsi que la couverture en G/C pour la première position des codons est de l'ordre de 70%, celle de la seconde position de 50% et la dernière de 90% (Wright et Bibb, 1992). Ce biais peut donc être utilisé pour rechercher les phases ouvertes de lecture dans les génomes de *Streptomyces* (Bibb et *al*., 1984). Du fait de ce taux de G/C élevé, les chances de rencontrer un codon TTA sont faibles, d'autant plus qu'il existe cinq autres codons leucine, tous plus riches en G/C. Un système de régulation de certains gènes basé sur l'utilisation du codon TTA leucine a été observé chez des Streptomyces. En effet, le gène bldA, codant l'unique ARNt leucine est indispensable à l'expression des gènes contenant un codon TTA. Ces gènes sont liés au métabolisme secondaire ou à des différentiations morphologiques (Chater et Chandra, 2008).

2.3. Richesse du métabolisme secondaire des *Streptomyces*

L'étude des bactéries du genre *Streptomyces* est principalement axée sur le métabolisme secondaire, extrêmement riche chez ces organismes et pouvant mener à la découverte de molécules intéressantes pour l'Homme. La mise en évidence de ce métabolisme secondaire prolifique s'est faite en deux temps au cours de l'Histoire. Tout d'abord, durant l'âge d'or de la découverte des antibiotiques : tester des *Streptomyces* pour leurs activités antibiotiques donna de très bons résultats et de nombreuses molécules encore en usage aujourd'hui furent découvertes. Ensuite, le séquençage de génomes de *Streptomyces* dans les années 2000 et leur analyse montra que ces organismes possèdent un métabolisme secondaire bien plus riche que ce qui était connu (Drago, 2015).

Chapitre 3
La génomique

Chapitre 3 : La génomique

3.1. Historique

En 1995, la «révolution génomique» a commencé avec l'achèvement de la première séquence du génome microbien (Fleischmann et *al*., 1995). Pour la première fois, la base génétique d'un isolat bactérien a été complètement caractérisée. À mesure que de nouvelles technologies de séquençage seront mises au point pour faciliter l'accès aux informations sur les séquences génomiques, la révolution génomique en cours devrait continuer à avoir un effet transformateur et avoir un impact sur de nombreux aspects de notre vie (Florian et *al*., 2011).

La recherche génomique vise à révéler et analyser les informations sur les séquences d'ADN et d'ARN. Chaque organisme sur terre est essentiellement défini par sa séquence génomique. La signature génomique est considérée comme la plus spécifique qui peut identifier sans ambiguïté la plupart des personnes sur terre. Il peut aider à distinguer des organismes étroitement apparentés, c'est-à-dire ceux qui présentent des phénotypes identique. Par exemple, avec la connaissance de séquences complètes du génome, il est maintenant possible de distinguer deux souches bactériennes associées à la même épidémie de maladie d'origine alimentaire. De plus, ces deux isolats pourraient même être classés et assignés à un arbre évolutif montrant leur relation. Jusqu'à présent, les coûts de séquençage élevés ont exclus l'application de l'analyse du génome entier comme outil médico-légal (Florian et *al*., 2011). Cependant, de nouvelles technologies de séquençage continuent d'être mises au point et offrent une sortie de données de séquence croissante à des coûts décroissants par cycle, dépassant la loi de Moore appliquée à la croissance des ressources informatiques (Moore, 1965).

En conséquence, la génomique devient la norme non seulement pour le domaine de la recherche, mais aussi pour la santé publique et la criminalistique microbienne. La génomique n'est pas un domaine de recherche en soi, mais est maintenant un outil de laboratoire universel. En tant que tel, une fois validé, il sera de plus en plus intégré à la boîte à outils de l'enquêteur en microbiologie (Florian et *al*., 2011).

3.2.Définition de la génomique

La génomique est l'étude des génomes. Elle concerne le séquençage du code génétique complet (ADN) des organismes vivants. C'est une science qui examine comment l'ensemble du génome interagit avec l'environnement. La génomique implique le séquençage de grandes quantités d'ADN et ainsi, produit une énorme quantité de données qui peuvent être conservées, organisées et consultées. Les avancées en génomique y sont pour beaucoup dans la croissance du domaine technique de la bioinformatique, qui consiste en l'application de la science et de la technologie informatiques à la gestion de l'information biologique (lien : www.explorecuriocite.org © Parlons sciences 2013).

3.3. Génomique structurale

La génomique structurale est définie comme l'analyse de la séquence et de la structure des éléments du génome: les gènes, les régulateurs et les éléments mobiles (Aburjaile et *al*., 2014). La génomique structurale vise à accroître nos connaissances de base des macromolécules biologiques tout en diminuant les coûts moyens de la détermination de la structure. Parallèlement, la bioinformatique structurale est liée aux analyses et à la prédiction de la structure tridimensionnelle (3D) des macromolécules biologiques, telles que : les protéines, l'ARN et l'ADN. Le terme structurel est le même qu'en biologie structurale, et la bioinformatique structurale est considérée comme une partie de la biologie structurale computationnelle, alors que la génomique structurale tend à décrire la structure 3D de chaque protéine codée par un génome donné. Cette approche fournit une méthode à haut débit pour la détermination de la structure compris des approches expérimentales et de modélisation. (Cassiana et *al*., 2018).

3.4. Annotation structurale

Dans le cadre de séquences nucléotidiques, l'annotation structurale permet d'identifier et localiser les différents éléments génétiques qui la composent (Gagniere, 2009). Pour une séquence génomique, cette localisation concerne des éléments relativement évidents tels que les gènes, les alternances introns/exons, les régions codantes (CDS, *Coding Sequence*) et les régions non traduites (UTR, *Untranslated region*), et peut être approfondie pour identifier les régions promotrices, régulatrices, les sites de début de transcription (TSS, *Transcription start site*), les sites de fixation d'histones, de réplication, etc. Pour des séquences de transcrits, il s'agit essentiellement de localiser la région codante (Gagniere, 2009).

3.4.1. Prédiction de gènes

Cette localisation débute systématiquement par la prédiction des gènes qui sont les éléments fondamentaux d'un génome. La notion de gène a évolué au fil du temps et reste encore difficile à définir. Le gène peut être défini comme une « région de séquence génomique, correspondante à une unité héréditaire, associée à des régions régulatrices, des régions transcrites et/ou d'autres régions fonctionnelles » (Pearson, 2006). Cependant la réalité est encore plus complexe, et la présence de produits fonctionnels différents partageant des régions génomiques chevauchantes redéfinit un gène en tant qu'« union de séquences génomiques codant pour un ensemble cohérent de produits fonctionnels potentiellement chevauchants » (Gerstein et *al*., 2007).

3.4.2. Autres éléments génétiques

Au-delà des gènes, il existe de nombreux autres éléments génétiques dans les génomes parmi lesquels on pourrait citer :

a- Les éléments de régulation d'expression. Ces sites de fixation des facteurs de transcription étant de très courte taille et souvent très variables, il faut utilise la génomique comparative avec les techniques de *phylogenetic footprinting*/ *shadowing* pour un résultat précis.

b - Les sites de début de transcription (TSS, *Transcription start site*). Ces derniers peuvent être localisés grâce à des expériences de *CAGE* (Gagniere, 2009) par des marqueurs de ces sites qui sont les 5'UTR et les TSS (Shimokawa et *al*., 2007).

c- Les CDS sont des séquences codantes dans le génome qui peuvent être localisées par des outils online tel que le serveur *Seed* (*RAST*) et le plus connu , celui de *NCBI*.

3.5. Génomique fonctionnelle

La génomique fonctionnelle est une branche qui intègre des études de biologie moléculaire et de biologie cellulaire et traite l'ensemble de la structure, la fonction et la régulation d'un gène. (Kaushik et *al*., 2018). La génomique fonctionnelle est l'étude de l'échelle du génome de la fonction de l'ADN (y compris les gènes et éléments), alors elle utilise des données du génome pour étudier l'expression génique et protéique à l'échelle du génome et leurs fonctions. Pour cela, des méthodes à haut débit sont utilisées pour comprendre la transcription et la traduction des gènes et meilleur interaction protéine-protéine (IPP). Les études de génomique fonctionnelle sont liées à la fonction des gènes dans le génome (Cassiana DeSousa et *al*., 2018).

Le rôle du gène et sa modulation dans une condition donnée est crucial pour mieux comprendre la biologie d'un organisme. Cette étude implique des études des transcrits et des protéines qui jouent un rôle essentiel dans le fonctionnement des processus biochimiques des cellules, associés avec la régulation (Cassiana DeSousa et *al*., 2018).

3.6. Annotation fonctionnelle

L'annotation fonctionnelle à pour but de prédire les fonctions de produit des gènes identifiés lors de l'annotation structurale (Byne, 2009). Elle consiste à attribuer une fonction biologique à chacun des gènes ou produits de gènes d'un génome ou d'une collection de transcrits. Alors elle permet de se forger rapidement une idée globale de cette fonction biologique réelle (Gagniere, 2009).

Cette définition soulève un point crucial de l'annotation : ***qu'est-ce que la fonction d'un gène ?***

La définition de la fonction biologique d'un gène reste vague et n'a jamais été clairement définie (Friedberg, 2006). Le sens biologique de la fonction est grandement dépendant du point de vue avec lequel on la décrit. Un gène n'a donc pas une fonction, mais des fonctions caractérisées différemment selon l'intérêt qui est porté au moment de l'annotation. C'est pourquoi, dans le contexte de l'annotation, le terme « *fonction* » englobe un ensemble de fonctions moléculaires, de localisations cellulaires, de domaines fonctionnels, de voies métaboliques, de signaux de localisation, ou toute autre caractéristique que l'on peut rattacher à un gène ou à son produit (Gagniere, 2009).

Il y'a deux façons d'annotation de génomes :

L'annotation automatique : s'appuie (essentiellement) sur des comparaisons des séquences à annoter avec les séquences présentes dans les banques de données.

L'annotation manuelle : basé sur l'expérience par des experts (des curateurs) qui valident ou invalident la prédiction en fonction de leurs connaissances ou de résultats expérimentaux.

Chapitre 4
La génomique comparative

Chapitre 4 : La génomique comparative

4.1. Définition

La génomique comparative est l'étude comparative de la structure et de la fonction des génomes de différentes espèces, dont les buts principaux sont de mieux comprendre comment les différentes espèces ont évolué, quels sont les effets de la sélection sur l'organisation et l'évolution des génomes, ainsi que de déterminer les fonctions des gènes et des régions non codantes du génome (Raphaël, 2010). La génomique comparative est une branche de la génomique qui vise à caractériser les similitudes et les différences des caractéristiques génomiques et à tracer leur gain et leur perte le long de différentes lignées évolutives, comprendre les forces évolutives telles que la mutation et la sélection qui régissent les changements de ces caractéristiques génomiques, et découvrir comment l'évolution génomique peut nous aider à combattre les maladies, restaurer la santé environnementale, gagner de l'argent, etc. (Xia, 2013). Les analyses comparatives en génomique peuvent se concentrer sur la similarité et les différences entre l'annotation ou entre la séquence de deux génomes ou plus (Herrero et *al*., 2015).

4.2. Core-genome et pane genome

La comparaison de plusieurs génomes au sein d'une espèce ou d'un genre bactérien permet de définir le génome central (ou « *core-genome* ») qui contient les gènes présents chez tous les organismes en question et le génome total (ou « *pan-genome* ») qui décrit le nombre total de gènes retrouvés au moins une fois dans une espèce ou un genre et le génome accessoire contenant les gènes présents dans deux ou plus des souches ou espèces, et les gènes uniques, spécifiques d'une seule souche ou espèce (Medini et *al*., 2005).

Les génomes bactériens de la même espèce contiennent un ensemble commun de gènes appelé génome central (*core-genome*) (Bryant, 2012 ; Darmon, 2014). Le *core genome* contient les gènes qui sont conservés dans toutes les lignées et le pan-génome est l'union de tous les ensembles de gènes de tous les génomes étudiés (Setubal et *al*., 2018).

.4.3Le séquençage de génomes complets

Récemment, une nouvelle génération de séquenceurs à très haut débit est apparue. Ces techniques permettent de séquencer, en quelques jours (heures), plusieurs giga-bases

d'ADN et sont regroupées sous le nom de méthodes « NGS » pour *Next* (ou *New*) *Generation Sequencing*.

Séquençages de nouvelle génération (NGS)

Un ensemble de nouvelles méthodes de séquençage qui permettent de réaliser du séquençage à très haut débit est apparu à partir de 2005. Les avantages de ces technologies sont nombreux : Pas d'étapes de clonage bactérien (et donc pas de biais inhérents à la construction des banques), rapidité (moins d'une semaine) et coûts beaucoup moins élevés (coût par paire de base *Solexa* environ 9000 fois moins cher que par le séquençage Sanger).

Ces nouveaux séquenceurs sont : *454* (*Roche*), *Solexa* (*Illumina*), *SOLID* (*Applied Biosystem*).

La technique 454 est une technique de séquençage basée sur l'amplification de l'ADN par PCR en émulsion et sur le pyroséquençage (luminescence par libération de pyrophosphate). Une présentation détaillée du principe est accessible sur le site internet de *Roche* (lien: http://454.com/resources-support/product-videos.as).

Les inconvénients liés à cette technique sont le taux d'erreur assez élevé, en particulier dans les régions homopolymériques, et la fenêtre de lecture étroite (la version *FLEX* utilisée pour le séquençage de *F. indicum* a une taille de lecture de 225 pb – aujourd'hui les séquences obtenues sont d'environ 700 pb).

La technique *Solexa* est basée sur l'amplification préalable des fragments d'ADN sur une lame. Pour déterminer la séquence, des nucléotides terminateurs réversibles marqués et fluorescents sont incorporés par amplification d'un brin complémentaire. Une image de la fluorescence est prise avant que la partie fluorescente, fixée à l'extrémité 3' de la base, soit enlevée chimiquement permettant la réalisation du cycle suivant. Egalement appelée séquençage par terminaison cyclique réversible, une présentation détaillée du principe de cette technique est accessible sur le site internet d'*Illumina* (lien : http://www.youtube.com/watch?v=77r5p8IBwJk).

Les avantages liés à la technique *Solexa* sont la rapidité du séquençage (jusqu'à un million de bases lues par seconde en 2012) et le très faible taux d'erreur. Récemment, la possibilité de réaliser des lectures « *paired-end* » (type de séquençage qui génère une paire de séquences (*reads*) séparées par une distance connue) et la possibilité de multiplexer les échantillons (plusieurs échantillons différents peuvent être séquencés en même temps) ont fait

de cette technologie une méthode de choix pour l'obtention de génomes bactériens complets *de novo* (c'est à dire s'il n'existe pas de séquence de référence) (Schuster, 2008).

La technologie *SOLID* est issue de la société *Agencourt Personal Genomics* et a été acquise par *Applied Biosystem* en 2006, elle-même fusionnée ensuite avec *Life Technologies*. Cette technique a été décrite pour la première fois dans les travaux de SHENDURE et al. en 2005. Cette technologie est basée sur une amplification par émulsion, suivie d'une étape de ligation.

L'un des avantages du séquençage par ligation est sa productivité. En effet, cette technique permet de générer jusqu'à 250 Gb par cycle en une semaine. La principale limite de cette technique de séquençage est la faible taille des amplicons lus qui est comprise entre 35 et 75 bp (Shokralla et *al*., 2012).

4.4. :La phylogénomie

4.4.1. Définition de la phylogénie

La phylogénie est une discipline émergeante. Elle permet d'étudier les espèces afin de les classer en fonction de leurs ressemblances phénotypiques (phylogénie phénétique) ou en fonction de leurs séquences géniques (phylogénie moléculaire). Les relations de parenté sont généralement représentées sous forme d'un arbre phylogénétique ou dendrogramme (Lechetta et *al*., 2005).

4.4.2. La phylogénétique bactérien avec analyse du gène de l'ARNr 16S

L'objectif de la phylogénie consiste à représenter statistiquement (arbre) les différentes taxons analysées pour visualiser clairement les relations phénotypiques ou/et génotypiques entre ces taxons.

Il y'a plusieurs façons d'évaluer la phylogénie d'un groupe donné de souches ou espèces. Une des façons les plus anciens et la plus commun est en comparant les séquences géniques de sous unité 16s de ribosome. Ces gènes qui code pour l'ARNr 16s ont d'abord été utilisé des années 1980 par Carl Woese comme outil d'étude de l'évolution bactérienne (Woese, 1987). Cependant, Les analyses basées sur la 16S sont souvent gênées par une résolution taxonomique limitée (Setubal et *al*., 2018), à cause de la forte similarité de gènes d'ARNr 16S au niveau de l'espèce (Land et *al*., 2015). Et aussi ces gènes ne sont probablement pas valables pour le genre *Streptomyces* (Labeda et *al*., 2017).

4.4.3. La phylogénomie

Au lieu de se concentrer sur le seul gène de l'ARNr 16S, il est maintenant possible d'établir un profil phylogénétique avec une analyse à l'échelle du génome à l'aide de génomes de référence, de groupes de protéines conservées, de génomes complets ou de protéomes. Ce genre d'approche a été nommé « *Phylogenomy* » ou phylogénomie. Toutefois, à l'échelle de génomes, de telles approches -vue la quantité importante de données analysées qui sont de l'ordre du millions de paires de bases- nécessitent une puissance de calcule considérable (Land et *al*., 2015).

C'est pourquoi les bioinformaticiens ont développées des programmes prenant en considération des algorithmes générant des matrices de distances sans alignements. En effet les génomes sont réduit à des abstraction par diverses approches, dans la quelle on calcule les distances entre les génomes entiers sans s'appuyer sur l'alignement entre les séquences (Setubal et *al*., 2018). Parmi les méthodes sans alignement il y'a le *polymorphisme d'un seul nucléotide* SNP (*single nucleotide polymorphysme*), qui constitue l'unité la plus petite possible pour l'analyse phylogénétique, qui a déjà été utilisée dans de nombreuses études phylogénomiques. Par ailleurs, l'approche des *profils de fréquence des fonctionnalités* (*Feature Frequency Profiles*) FFP constitue aussi une approche intéressante pour l'étude de la phylogénie au niveau génome (cette approche sera présentée dans la partie matériel et méthode).

Matériel et Méthodes

I. Objectif

L'objectif de cette étude est de faire usage des données génomiques de souches appartenant au genre *Streptomyces* pour faire ressortir les particularités génétiques en relation avec leur potentiel de promotion de la croissance des plantes.

Pour cela, l'approche (stratégie) employée est la suivante :

1. Screening bibliographique des espèces appartenant au genre *Streptomyces* et sélection de celle en relation avec les plantes.

2. Sélection des souches ayant des génomes complets.

3. Sélection de souches d'espèces du genre *Streptomyces*, sans relation établies avec les plantes ayant un génome complet pour servir de groupe de référence.

4. Etude phylogénétique (par analyse du gène de l'ARNr16S) et phylogénomique (phylogénie par analyse du génome complet) par approches *Feature Freqeuncy Profile* FFP et *up-to-date bacterial core gene* (UBCG).

5. Annotation des génomes par le serveur RAST (Rapid Annotation using the Subsystems Technology) et NCBI (NCBI Prokaryotic Genome Annotation Pipeline, « PGAP »).

6. Analyse comparative des séquences codantes, des gènes de protéines.

7. Détection de clusters de gènes de métabolites secondaires par outil online *Antismash*, et analyse comparative des profiles de métabolites secondaires des souches retenues.

8. Synthèse des données de l'analyse comparative.

II. Matériel et méthodes

Pour l'ensemble des analyses faites, une machine de type PC (4gb ram, 500gb SSD drive, CPU Core i7 5600) avec système d'exploitation Windows 10 (64 bits, x64) a été utilisé. Pour les applications sur Linux, le système d'exploitation Biolinux ver. 08.0.7. a été utilisé en l'installant sur une machine virtuel grâce à l'outil de virtualisation « Oracle VM VirtualBox » ver. 5.2.12 r122591 (Qt5.6.2).

1. Sélection des espèces *Streptomyces* à analyser

Le genre *Streptmyces* est un genre riche en taxa. En effet on recense plus de 600 espèces à ce jour. Ce travail s'intéresse uniquement à celle en relation avec les plantes, et plus précisément, ayant déjà été décrites comme étant bénéfique pour les plantes.

Pour cette partie, nous avons ciblé les moteurs de recherche bibliographique suivant pour ne retenir que les espèces ayant un effet PGPR :

- Pubmed
- Google Scholar
- Springer
- Science direct

En plus de ce premier critère, seules les souches ayant déjà eu leur génome séquencé en entier ont été retenues.

En plus de ces dernières, d'autres souches d'espèces *Streptomyces* spp. ont été choisie (ayant un génomes complet) pour servir de références aux analyses comparatives.

Une fois le choix établi, les séquences de génomes ont été obtenues à partir de la base de données (*genome database*) du serveur NCBI :

https://www.ncbi.nlm.nih.gov/genome/genomes/

Les séquences des génomes ont été téléchargée sou formats Fasta, en prenant soin de vérifier si il n'y a pas eu des erreurs de téléchargement en comparant le nombre de base affichées sur MEGA7 une fois les fichiers ouverts, et le nombre de bases avancé dans le résumé des génomes dans la page de téléchargement des séquences dans la base de donnée *Genome* du serveur NCBI. Nous avons séparés les souches en deux groupes en fonction de leur mode de vie/habitat :

Plant-related group : directement lié aux plantes (vérification bibliographique lors de la sélection des souches), ce groupe contient les souches :

- *Streptomyces coelicolor* $(A_3)_2$
- *Streptomyces lavendulae* sub sp. *laven* MTCC 706
- *Streptomyces griseochromogene* ATCC14511

- *Streptomyces lydicus* 101
- *Streptomyces hygroscopicus* sub. sp. *liminens* KCTC1717
- *Streptomyces aureofaciens* CCM 2339 = *Kitasatospora aureofaciens* CCM 2339
- *Streptomyces griseus* IFO13350
- *Streptomyces lividans* TK24

Non-related to plant group : n'ayant pas ou très peu d'évidence de leurs relations avec les plantes, ce groupe contient :

- *Streptomyces rapamycinicus* NRRL 5491
- *Streptomyces venezuelae* ATCC 15439
- *Streptomyces parvulus* 2297
- *Streptomyces albolongus* YIM101047 = *Kitasatospora albolonga* YIM101047
- *Streptomyces albus* SM 254 = DSM 398
- *Streptomyces ambofaciens* DSM 40697
- *Streptomyces atratus* Zh16
- *Streptomyces avermitilis* MA-4680 = NBRC 14893
- *Streptomyces collinus* Tu 365
- *Streptomyces formicae* KY5
- *Streptomyces glaucesens* GLA. O = DMS 40922
- *Streptomyces lunaelactis* MM109
- *Streptomyces peucetius* ATCC 27952

Enfin, la souche *E.coli* SE15 fut sélectionnée en tant qu'*outgroup* pour enraciner les arbres de phylogénie. Le tableau n°1 décrit ainsi les souches retenues.

2. Phylogénie par analyse du gène de l'ARNr 16S

Pour comprendre les liens de parenté génétique et génomique entre les souches de cette étude, nous avons procédé de plusieurs manières.

2.1. Etude phylogénétique par analyse du gène de l'ARNr 16S

Les séquences du gène de l'ARNr 16S de chaque souche ont été téléchargés à partir de la base de données *genome database* du serveur NCBI, en affichant l'annotation du génome de chaque souche, et en réglant l'affichage des données en optant pour un affichage des gènes

codant pour les ARNs seulement. Pour les souches dont l'annotation n'affiche pas les gènes des ARNs, les séquences de gènes des ARN ont été retrouvées dans la base de données *nucleotide* du serveur NCBI (https://www.ncbi.nlm.nih.gov/nuccore/).

Une fois les séquences téléchargées en format FASTA, il s'est avéré qu'il y'avait plusieurs copies du même gène chez chaque souche. Un alignement par MUSCLE (Algorithme UPGMA) sur progiciel MEGA7 (fig.1) nous à permis de différentier deux groupes de copies dans chaque souche :

a. Un groupe de copies surreprésentées : c'est-à-dire des copies conformes d'une séquence du gène de l'ARNr16S présente en sur nombre.
b. Un groupe de copies sous-représentées : c'est-à-dire, des copies conformes d'une séquence du gène de l'ARNr 16S, qui ne ressemble pas à celle surreprésentée, et qui est dupliquée en nombre relativement réduit par rapport à la séquence surreprésentées.

Exemple : Chez la souche *S. formycae* KY5, il y'aurait 6 copies du gène de l'ARNr 16S :

- Deux copies d'une séquence
- Quatre copies d'une autre séquence, différente de la précédente.

Ce polymorphisme du même gène au sein du même génome est appelé paralogie. Les copies du même gène sont appelé paralogues.

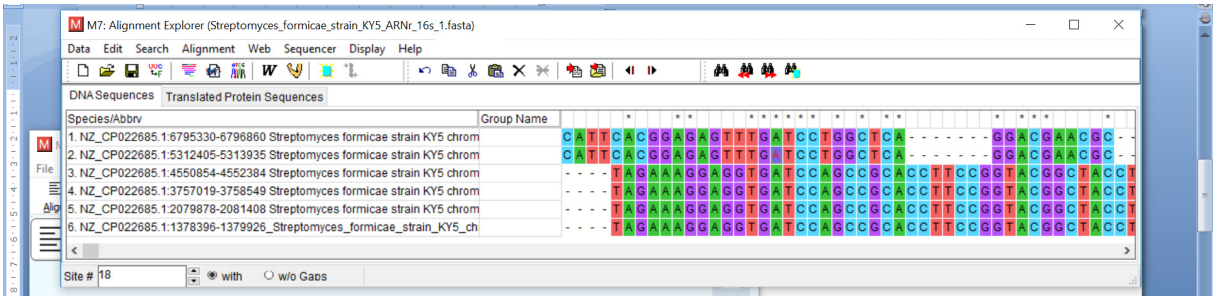

Figure 1. Alignement des 6 copies du gène de l'ARNr 16S de la souche *S. formycae* KY5. Les deux premières copies sont semblables (copies sous représentées, au nombre de 2), tout comme les quatre autre copies sont les mêmes (copies surreprésentées, au nombre de 6).

Une fois les copies sous représentées et sur représentées séparées, nous avons procédé à une analyse phylogénétique en se basant sur :

- Les copies sous représentées de chaque souches d'abord,
- Les copies surreprésentées de chaque souche

- Les séquences consensus de chaque souche.

Pour ce qui est de l'obtention des séquences consensus, ce fut par alignement de toutes les copies du gène de l'ARNr16S de chaque souche (par la même méthode que celle de l'exemple de *S. formycae KY5* décrit plus haut), pour ne garder ensuite que les base aux sites conservés chez toutes les copies.

Les arbres phylogénétiques ont été construits par méthode UPGMA, avec une validation statistique par bootstrap (1000 ré-échantillonnages). Pour un affichage décondensé, chaque arbre est par la suite affiché en mode *boosstrap consensus tree.*

2.2. Etude phylogénomique

L'étude phylogénomique à été faite par deux approches, se basant sur le génome complet pour faire une abstraction, permettant de réduire l'information à l'essentiel permettant de garder les signatures taxonomique de chaque taxa.

a. Par approche FFP

Dans le cas de la phylogénie prenant base sur les génomes, le plus grand challenge, est d'ordre computationnel, et plus précisément : le temps de calcule. En effet, l'étude de séquences de tailles importante de l'ordre du million de paires de bases, en passant par des alignements de séquences rend les calcule difficiles, voire impossible à conduire avec des machine de bureau. L'approche de la méthode FFP (*feature frequency profile*) ne se base pas sur des alignements. D'où son grand avantage par rapport aux méthodes classiques qui nécessitent des alignements, permettant ainsi de conduire des études de phylogénie sur des machines avec des puissances plutôt réduite, au lieu de stations de calcules onéreuses, justifiant ainsi notre choix.

Le FFP est une nouvelle méthode utilisée pour étudier la phylogénie du génome entier à partir de k-mers. Dans cette méthode, le nombre de caractéristiques d'une longueur particulière *l* apparaissant dans un génome particulier est compté et assemblé dans un vecteur FFP. Les FFP de différentes espèces sont ensuite comparés à l'aide de la divergence de Jensen – Shannon (JS). Un arbre phylogénétique *neighbourg joining* « NJ » peut ainsi être construit sur la base de la matrice de distance résultante. Comparée à la méthode classique basée sur l'alignement de séquences multiples (*MSA*), la méthode *FFP* sans alignement permet de comparer des régions géniques et non géniques du génome entier à une vitesse supérieure; il

peut incorporer une grande variété de caractéristiques génomiques dans chaque comparaison, notamment des délétions d'intron, des indels de séquence d'exon, des insertions d'éléments transposables, des transversions de base dans des séquences de codage et quelques modifications génomiques rares telles que des insertions d'éléments intercalés courts / longs (SINE / LINE) (Sims et al., 2009).

Le programme une fois installée sur *Biolinux*, il a été utilisé sous le *Terminal* par des ordres en *Command Line* (Fig.2).

Les séquences entières du génome des 22 souches retenues ont été converties au format multiFasta avant d'être téléchargées dans FFP –3.1910, où les différentes formes de partitions du génome ont été comparées entre souches, et des arbres NJ ont été construits à partir de la matrice de divergences de Jensen – Shannon de chaque type de partition du génome. En suivant les recommandations du programme, nous avons utilisé les outils de *ffpvocab* et *ffpre* pour trouver la plage de longueurs appropriée à utiliser (l = 15 a finalement été choisi dans l'analyse). Nous avons également procédé à l'amorçage (1000) pour évaluer l'analyse phylogénétique de la FFP. Le résultat de l'analyse a été importé dans dans MEGA 7 pour afficher l'arbre NJ finale.

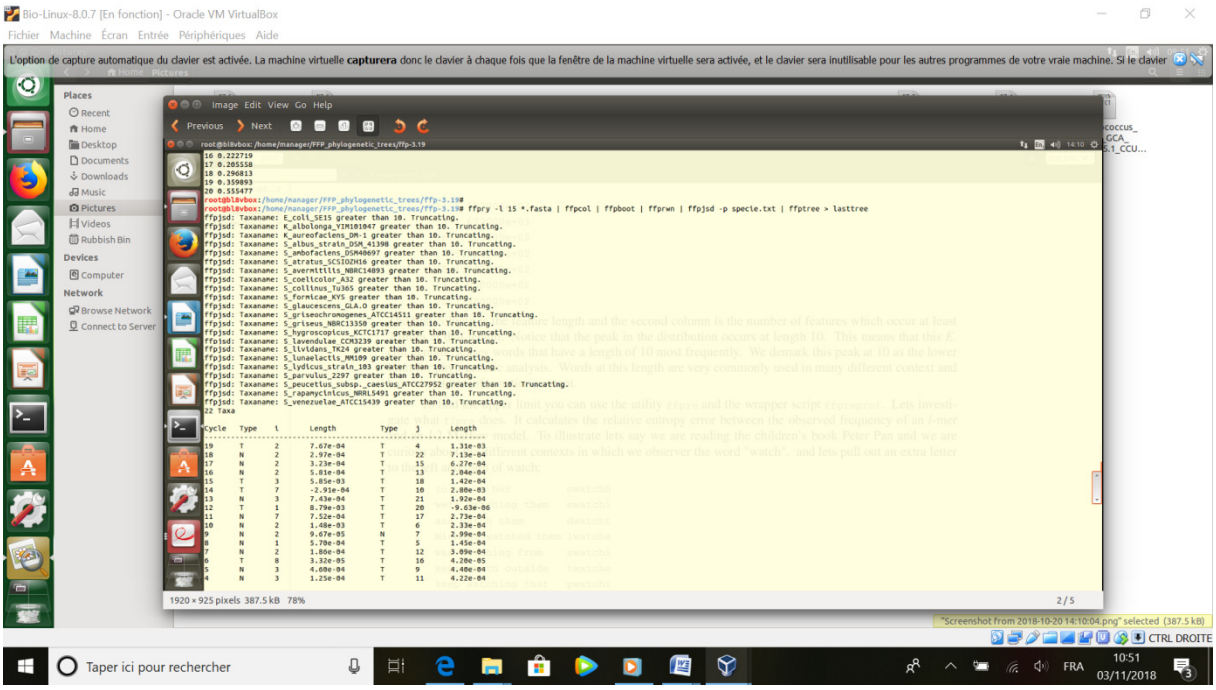

Figure 2. Usage du programme ffp-3.19 pour la phylogénomie, sous environnement Biolinux en virtualisation par VirtualBox.

b. Par approche UBCG

Comme le disent les concepteurs de cette méthode très récente (Na et al., 2018), le gène de l'ARNr 16S a joué un rôle essentiel dans la taxonomie bactérienne en fournissant un cadre phylogénétique universellement applicable. Cependant, ce gène unique ne contient qu'environ 1500 pb, ce qui limite de manière inhérente la résolution de l'analyse. Dans l'approche UBCG, qui est une approche multilocus, sont présentés l'ensemble des gènes centraux bactériens qui couvrent tous les phylums, nommés UBCG (*up-to-date bacterial core gene*). L'ensemble UBCG actuel a été calculé en utilisant des génomes complets de 1492 espèces couvrant 28 phylums, comprenant 92 gènes (Fig. 3) (Na et al, 2018).

Les étapes de cette approche se présentent comme suit :

- Extraction d'UBCG à partir d'assemblages génomiques des 22 souches
- Alignement multiple de 92 séquences de gènes
- Concaténation de 92 séquences de gènes
- Filtrage des positions d'alignements multiples
- Analyse phylogénétique avec RAxML et FastTree
- Calcul de l'indice de support de gène (GSI) qui indique combien de gènes supportent la branche dans l'arbre phylogénétique concaténé (nommé arbre UBCG).

Ce programme fut installé sur un système d'exploitation Biolinux, car nécessite un usage par *Command Line* sur le Terminal des systèmes d'exploitation de type linux (équivalent de l'invité de commandes sur systèmes d'exploitation Windows).

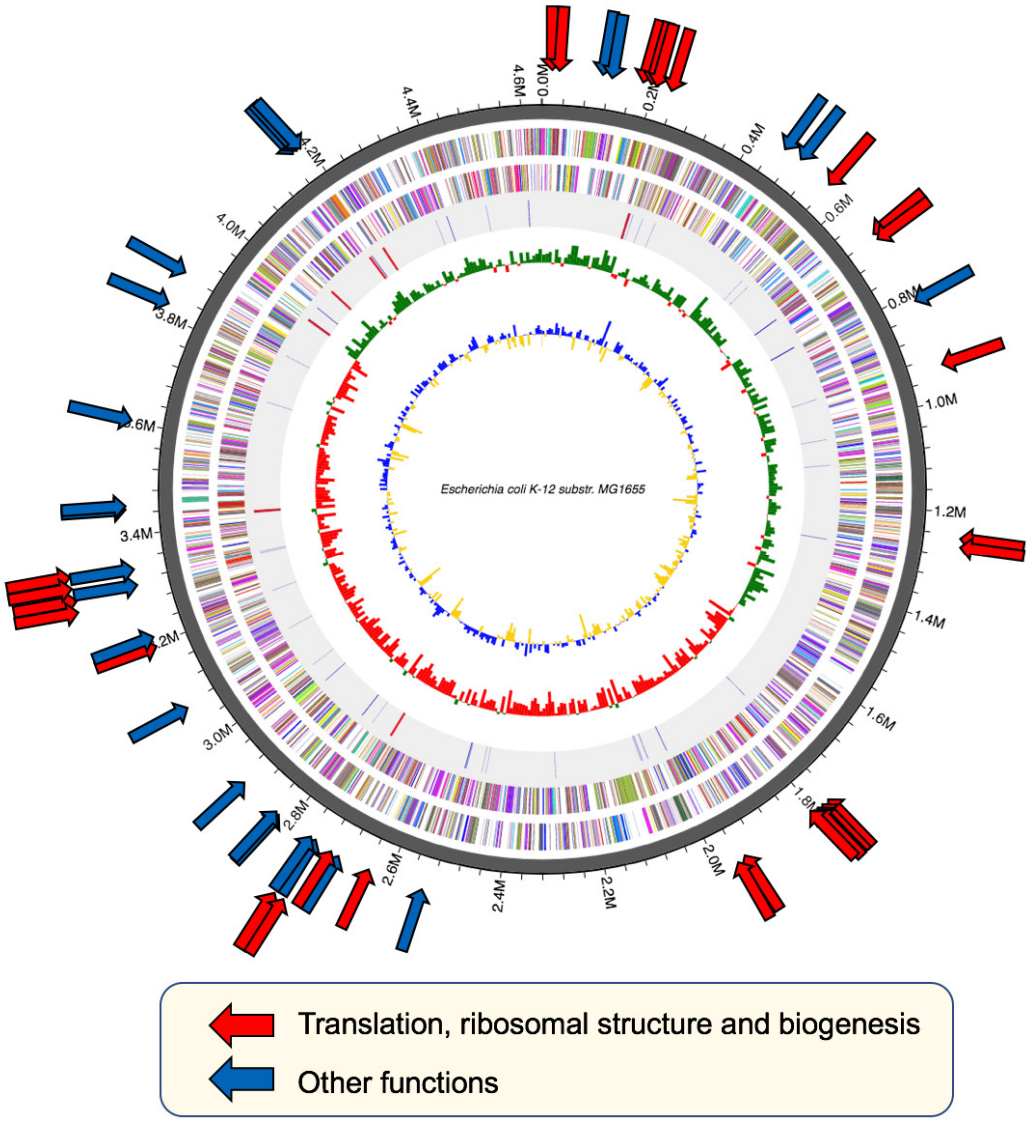

Figure 3: localisation des 92 gènes pris en considération lors de l'analyse UBCG dans le génome bactérien (à titre d'exemple) d'*E. coli* souche K12

3. Annotation des génomes

Pour éviter les disparités due à la nature de l'outil utilisé, cette étude à pris en compte deux moyens d'annotation des génomes :

- Par le serveur NCBI
- Par le serveur RAST

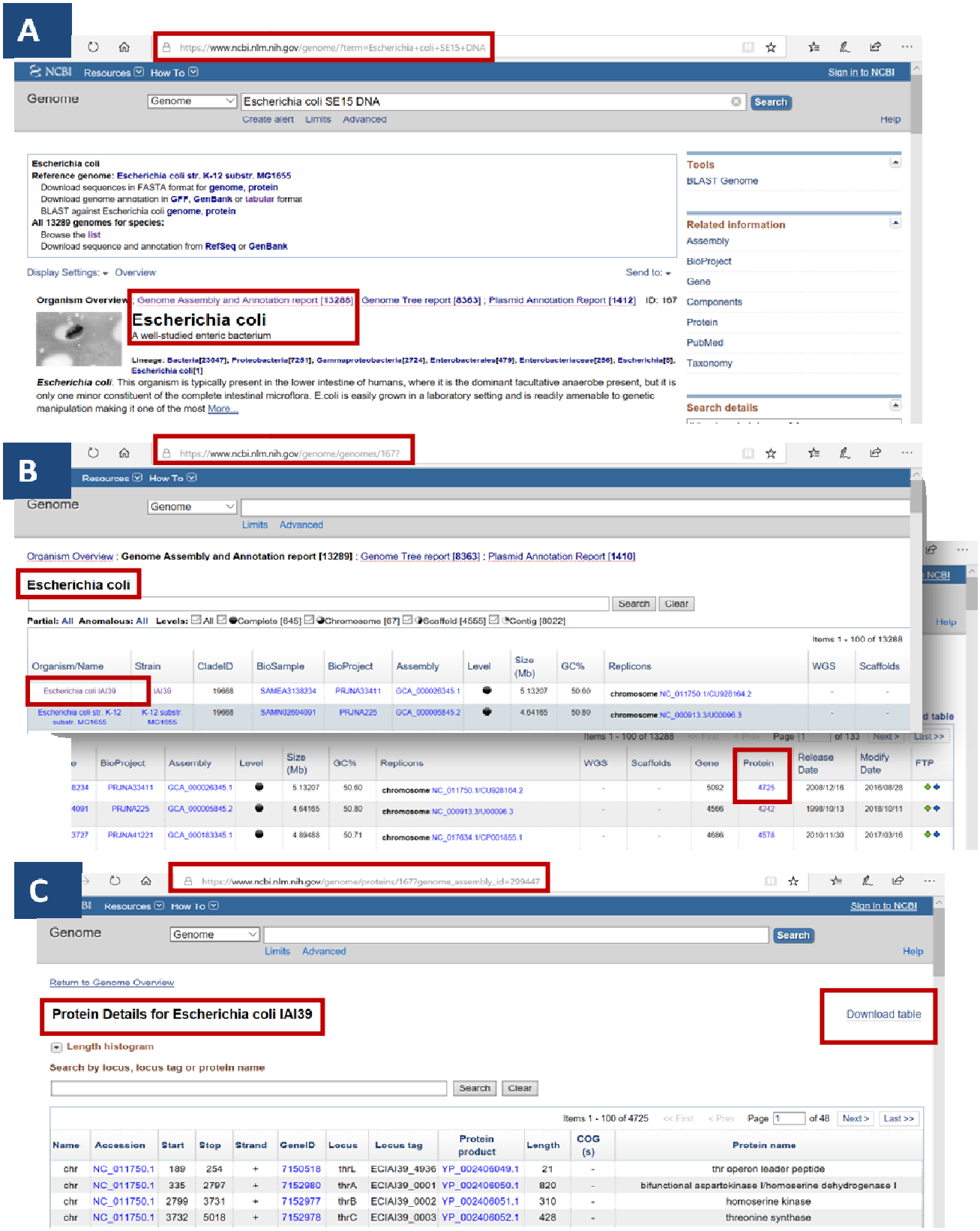

Figure 4: Les étapes d'extraction des tables d'annotation de protéines à partir du serveur NCBI. A : retrouver le génome de la souche en question, B, accéder au résumé de l'assemblage et de l'annotation et enfin C : téléchargement de la table de protéines encodées.

3.1. Annotation par NCBI (NCBI Prokaryotic Genome Annotation Pipeline, « PGAP »)

Le pipeline d'annotation du génome procaryote (PGAP) NCBI a été conçu pour annoter les génomes bactériens et archaéens (chromosomes et plasmides). L'annotation du génome est un processus à plusieurs niveaux qui inclut la prédiction de gènes codant pour des protéines, ainsi que d'autres unités génomiques fonctionnelles telles que les ARN structurels, les ARNt, les petits ARN, les pseudogènes, les régions de contrôle, les répétitions directes et inversées, les séquences d'insertion, les transposons et autres éléments mobiles. NCBI a développé un pipeline d'annotation automatique du génome procaryote qui combine des algorithmes de prédiction génique *ab initio* avec des méthodes basées sur l'homologie. La première version du pipeline NCBI Prokaryotic Genome Automatic a été développée en 2001 et est régulièrement mise à niveau pour améliorer la qualité des annotations structurelles et fonctionnelles (Haft et al., 2018, Tatusova., et al 2016). Des améliorations récentes utilisent des modèles de Markov cachés (HMM) avec un profil protéique organisé, y compris TIGRFAMS et de nouveaux HMM pour les protéines de résistance aux antimicrobiens, et des architectures de domaine complexe organisées pour une annotation fonctionnelle améliorée des protéines. Le pipeline d'annotations de NCBI dépend de plusieurs bases de données internes et n'est actuellement pas téléchargeable ni utilisable en dehors de l'environnement du serveur NCBI.

Dans le cas de cet outils, tout les génomes déjà présents dans la base de donnée GENOME sont déjà annoté, il ne reste plus qu'a téléchargé le format adéquat de donnée voulues. Dans la présente étude, seuls les tables de protéines ont été téléchargées, en formats .txt*, pour servir de matière première à l'analyse comparative des protéines encodés par les génomes des souches sélectionnées (Fig.4).

3.2. Annotation par RAST (Rapid Annotation using the Subsystems Technology)

RAST est un service entièrement automatisé pour annoter les génomes bactériens et archés. Le service identifie les gènes codant les protéines, les ARNr et les ARNt, attribue des fonctions aux gènes, prédit quels sous-systèmes sont représentés dans le génome, utilise ces informations pour reconstruire le réseau métabolique et permet aux utilisateurs de télécharger facilement le résultat. De plus, le génome annoté peut être parcouru dans un environnement prenant en charge l'analyse comparative avec les génomes annotés conservés dans l'environnement du serveur SEED. Le service rend normalement le génome annoté disponible dans les 12 à 24 heures suivant la soumission.

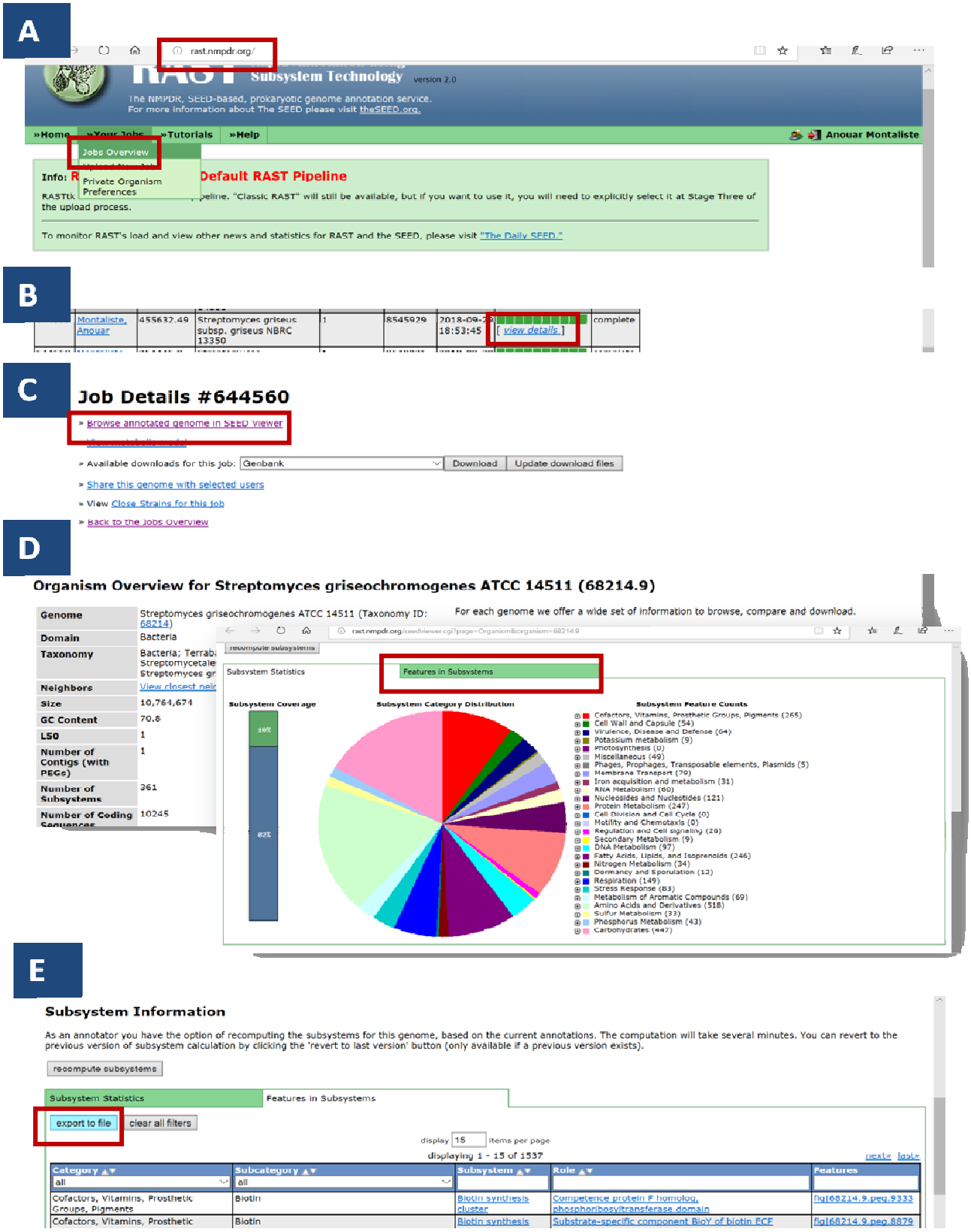

Figure 5 : les étapes d'obtention des tables d'annotation sur Rast (tous les fichiers) et des Sous-systèmes à part. A : accès au site, après création d'un compte, B : jobs d'annotation en cours, C : téléchargement des résultats globaux, D : visualisation des résultats et accès à la table des sous-systèmes, E : téléchargement de la table des sous-systèmes.

Le *Genome viewer* inclus dans RAST prend en charge la comparaison détaillée avec génomes existants, détermination des gènes que le génome a des points communs avec des ensembles spécifiques de génomes (ou gènes qui distinguent le génome de ceux d'un ensemble de génomes existants), la possibilité d'afficher le contexte génomique autour de gènes spécifiques et la possibilité de télécharger les informations et les annotations pertinentes à volonté.

Cet outil a été sélectionné pour compléter l'outil d'NCBI, comme un complément d'annotation, qui en plus permet de visualiser de manière résumé les résultats de l'annotation, en rassemblant les résultats sous forme de sous-systèmes (Ramy et al, 2008).

L'outil RAST nécessite un compte sur le serveur. Une fois avoir eu l'accès au compte créer, les génomes à annoter peuvent être uploadés un par un. Après environ 24h, ont obtient les résultats de l'annotation, sous diverse formats de fichiers (GenBank, FASTA, GFF3, Excel, etc), qui seront par la suite employés pour l'étude comparative (Fig. 5).

4. Visualisation des génomes après annotation

Pour un rendu visuel simple et informatif, deux programmes sous environnement Java ont été utilisés : DNAplotter et BRIG.

4.1. visualisation individuel de chaque génome de souches retenues par DNAplotter

DNAPlotter peut être utilisé pour générer des images de cartes d'ADN circulaires et linéaires afin d'afficher des régions et des caractéristiques intéressantes. Les images peuvent être insérées dans un document ou imprimées directement. Comme cela utilise Artemis, il peut lire les formats de fichiers courants EMBL, GenBank et GFF3 (https://www.sanger.ac.uk/science/tools/dnaplotter) (Carver et al., 2009).

A partir des séquences .gbk* générées par l'annotation par RAST, nous avons pu illustrer les génomes *via* le programme DNAplotter, en affichant :

- Le squelette ADN (en noir, avec les positions des bases)
- Le contenu GC (GC plot, ocre/violet)
- Le GC-skew (gris/bleu ciel)
- Les CDS (*coding sequences*, séquences codantes) forward (en bleu)
- Les CDS reverse (en rouge)

- Les ARNr (en vert claire)
- Les ARNt (en rose vif)

Chaque génome a été représenté dans une image à part.

4.2. visualisation condensée des génomes des souches retenues par BRIG

BRIG est une application gratuite multi plateforme (Windows / Mac / Unix) capable d'afficher des comparaisons circulaires entre un grand nombre de génomes, en mettant l'accent sur la gestion des données d'assemblage du génome. L'application est disponible à l'adresse suivante: http://sourceforge.net/projects/brig (Alikhan et al., 2011).

Cette application à plusieurs fonctionnalités de :

- Illustration de la similitude entre une séquence de référence centrale et d'autres séquences sous forme d'anneaux concentriques.
- BRIG effectue toutes les comparaisons et analyses de fichiers BLAST automatiquement *via* une interface graphique simple.
- Les limites de contig et la couverture de lecture peuvent être affichées pour les génomes brouillons; des graphiques personnalisés et des annotations peuvent être affichés.
- En utilisant un ensemble de gènes défini par l'utilisateur comme entrée, BRIG peut afficher la présence, l'absence, la troncature ou la variation de séquence d'un gène dans un ensemble de génomes complets, de génomes de brouillon ou même de données de séquence brutes non assemblées.
- BRIG accepte également les fichiers de mappage-lecture au format SAM permettant la comparaison simultanée de régions génomiques présentes dans des données de séquence non assemblées provenant de plusieurs échantillons (Alikhan et al., 2011).

Dans le cas de notre étude, l'outil BRIG à été utilisé pour une comparaison visuelle des génomes des diverses souches en questions, en suivant les étapes décrites par le concepteur.

L'analyse passe par une configuration des informations affichées sur chacun des anneaux concentriques dans BRIG. Créez les 22 anneaux, pour chaque anneau:

1. Définir le texte de la légende pour chaque anneau
2. Sélectionner les séquences requises dans le pool de données et cliquer sur «ajouter des données» pour les ajouter à la liste des anneaux.

3. Choisir une couleur
4. Définir les seuils d'identité supérieure (90) et inférieure (70) lors de l'analyse par Blast.
5. Cliquer sur «ajouter un nouvel anneau» et répétez les étapes pour chaque nouvel anneau requis.

Remarque : Lorsque un fichier Genbank / EMBL est utilisé comme référence, les utilisateurs peuvent choisir d'utiliser la séquence protéique ou nucléotidique.

La dernière fenêtre nous permet de changer les options de BLAST, l'emplacement du fichier image et définir le titre de l'image, qui apparaîtra au centre des anneaux générés. Le Blast utilisé est une version offline, qui doit être installée indépendamment de la présence de l'exécutable du programme BRIG. Nous avons opté pour la version Blast-2.4.0+.

NB : pour des raisons d'affichage, le plus petit génome de notre sélection a été pris pour anneau central dans cette analyse, c'est-à-dire celui de la souche : *Kitasatospora aureofaciens* DM1 (6,8 MB). Pour la souche *S. hygroscopicus*, vu son chromosome divisé en deux partie, avant de lancer l'analyse par BRIG, les deux partie de son génome ont été concaténées pour avoir une seule séquence. Le programme BRIG ne répond pas aux fichiers dont le chemin contient des espaces. Pour cela, il faut prendre la peine de renommer les chemins de manière à ce qu'il n'y ait pas d'espaces.

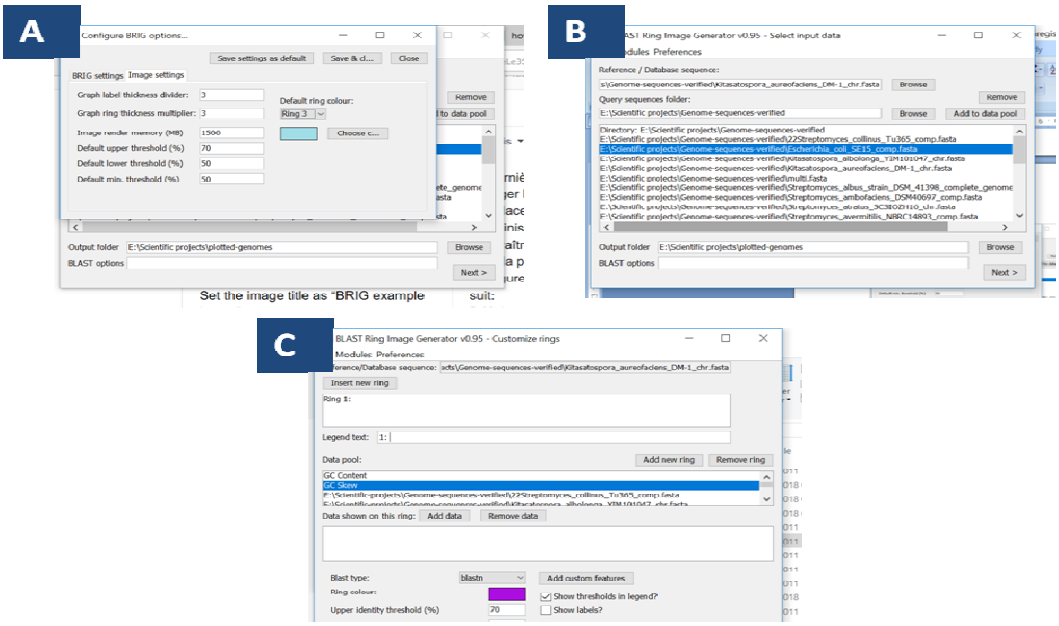

Figure 6 : Utilisation de l'outil BRIG pour une comparaison graphique des génomes. A : réglage des paramètres, B : choix des données à analyser, C : paramètre du rendu graphique des anneaux.

5. Détection des clusters de gènes de métabolites secondaire par outils antiSMASH

Pour ce qui est des gènes impliqués dans le métabolisme secondaire, l'outil online antiSMASH, meilleur plateforme de ce genre, fut sélectionné.

antiSMASH permet l'identification, l'annotation et l'analyse rapides à l'échelle du génome de groupes de gènes de biosynthèse de métabolites secondaires dans des génomes bactériens et fongiques. Il intègre et entretient des liaisons croisées avec un grand nombre d'outils d'analyse de métabolites secondaires *in silico* déjà publiés. antiSMASH est alimenté par plusieurs outils open source: NCBI BLAST +, HMMer 3, Muscle 3, Glimmer 3, FastTree, TreeGraph 2, Indigo-depict, PySVG et JQuery SVG (Tilmann et al., 2015) (Fig. 7).

Chaque génome des souches retenues ont été uploadés après dans le serveur du site d'antiSMASH, séparément. Une fois le job fini, le serveur alerte les utilisateurs *via* leur mail. Les résultats sont organisés sous forme de nombre (pour ainsi désigner l'ordre des clusters) et des couleurs différentes (pour des types de métabolites secondaires différents).

Concernant cette partie, nous nous sommes focalisés sur les clusters de gènes affichant un taux de similarité avec les gènes des bases de données interrogés d'au moins 60%. Les profils obtenus serviront par ailleurs à comparer les potentialités des souches sélectionnées.

NB : les paramètres par défaut de l'analyse par outil antiSMASH ont été utilisés sans modification.

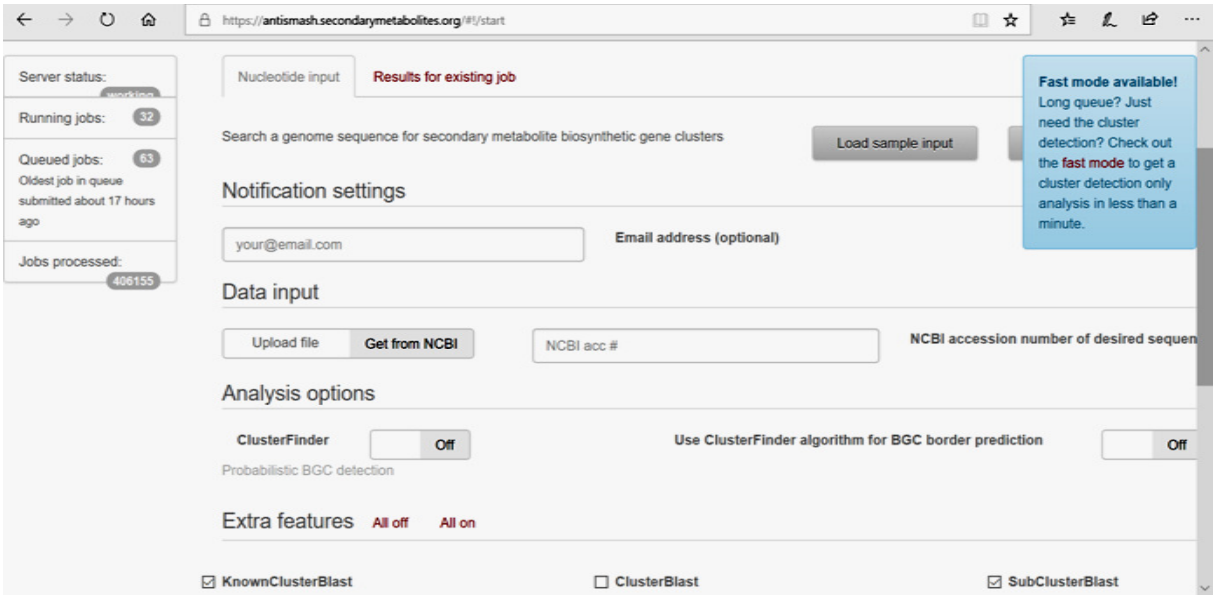

Figure 7 : Overview du serveur antiSMASH

6. Analyse comparative

L'analyse comparative a porté sur les résultats de :

- L'annotation par outil NCBI
- L'annotation par outil RAST

C'est en préparant des listes des traits génomiques de chaque souche de cette analyse, qu'on procède à un affichage des disparités et des ressemblances (quantitatives) par un diagramme de Venn, grâce à une application dédiée à cet effet.

Pour ce qui est des comparaisons d'ordre qualitatives, ce fut par le biais d'un curage manuel que l'on a pu faire la synthèse des points communs, et des divergences.

Rappel : Pour cette partie comparative, les souches ont été divisées en deux groupes :

Plant-related group : directement lié aux plantes (vérification bibliographique lors de la sélection des souches), ce groupe contient les souches :

- *Streptomyces coelicolor* $(A_3)_2$
- *Streptomyces lavendulae* sub sp. *laven* MTCC 706
- *Streptomyces griseochromogene* ATCC14511
- *Streptomyces lydicus* 101
- *Streptomyces hygroscopicus* sub. sp. *liminens* KCTC1717
- *Streptomyces aureofaciens* CCM 2339 = *Kitasatospora aureofaciens* CCM 2339
- *Streptomyces griseus* IFO13350
- *Streptomyces lividans* TK24

Non-related to plant group : n'ayant pas ou très peu d'évidence de leurs relations avec les plantes, ce groupe contient :

- *Streptomyces rapamycinicus* NRRL 5491
- *Streptomyces venezuelae* ATCC 15439
- *Streptomyces parvulus* 2297
- *Streptomyces albolongus* YIM101047 = *Kitasatospora albolonga* YIM101047
- *Streptomyces albus* SM 254 = DSM 398
- *Streptomyces ambofaciens* DSM 40697
- *Streptomyces atratus* Zh16

- *Streptomyces avermitilis* MA-4680 = NBRC 14893
- *Streptomyces collinus* Tu 365
- *Streptomyces formicae* KY5
- *Streptomyces glaucesens* GLA. O = DMS 40922
- *Streptomyces lunaelactis* MM109

  Streptomyces peucetius ATCC 27952

6.1. Comparaison des tables de séquences codantes (*CDS*) obtenues par annotation *via* RAST

Les résultats de l'annotation RAST une fois obtenue, le fichier .tsv contenant la table d'annotation est formaté sous Excel pour en extraire seulement les CDS (toutes séquences codantes), et les sauvegarder dans un autre fichier Excel (.xlsx*) prêt à l'usage en comparaison, et ce pour tous les génomes.

La comparaison est réalisée par outil VennPainter. VennPainter est un programme permettant de décrire des ensembles uniques et partagés de listes de gènes et de générer des diagrammes de Venn, en utilisant le framework Qt C ++. Le logiciel produit des diagrammes *Classic Venn*, *Edwards 'Venn* et *Nested Venn* et permet huit ensembles en mode graphique et 31 ensembles en mode de traitement de données uniquement. En comparaison, les programmes précédents produisent des diagrammes de type *Classic Venn* et *Edward's Venn* et permettent un maximum de six jeux. Le logiciel intègre des fonctionnalités conviviales et fonctionne sous Windows, Linux et Mac OS. Son interface graphique ne nécessite pas qu'un utilisateur possède des compétences en programmation. Les utilisateurs peuvent modifier le contenu du diagramme pour un maximum de huit jeux de données en raison de la sortie *Scalable Vector Graphics*. VennPainter peut fournir des résultats de sortie dans des formats verticaux, horizontaux et matriciels, ce qui facilite le partage des jeux de données nécessaires pour une identification plus poussée des gènes candidats. Les utilisateurs peuvent obtenir des listes de gènes à partir d'ensembles partagés en cliquant sur les numéros du diagramme. Ainsi, VennPainter est un programme puissant, multi-plateforme et puissant, facile à utiliser, qui fournit un outil plus complet d'identification de gènes candidats et de visualisation des relations entre gènes ou familles de gènes dans une analyse comparative (Lin et al., 2016) (Fig. 8).

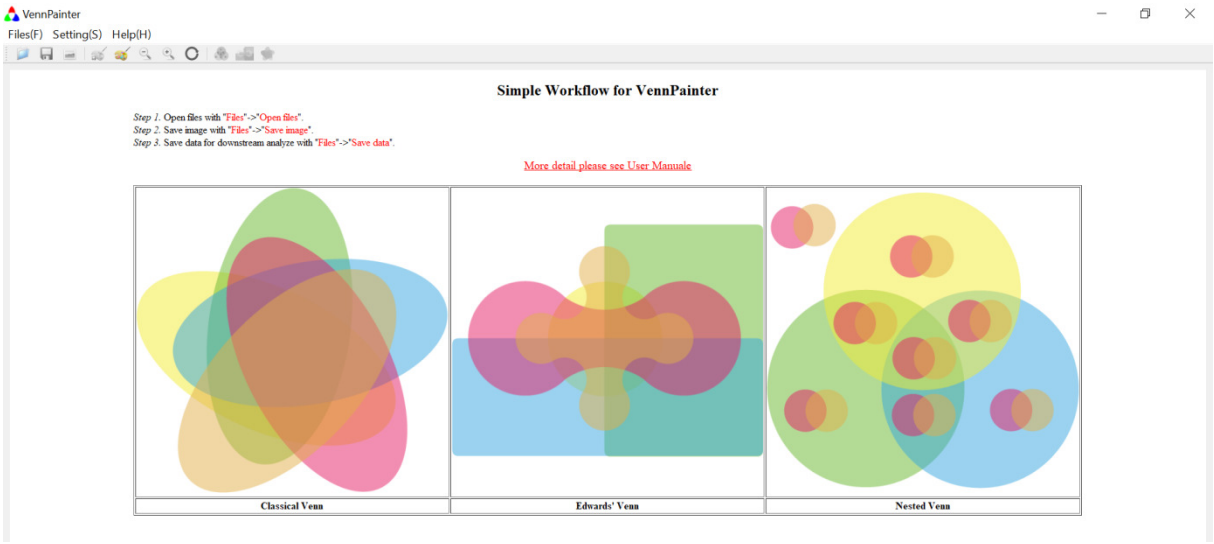

Figure 8 : Overview du programme VennPainter

VennPainter a été utilisé pour la comparaison des souches du groupe « plant-related ». Pour ce qui est de l'autre groupe, vu le nombre de souches (donc de listes) qui dépasse le 8 (limites supérieur de fichiers traitables par VennPainter), nous avons utilisé un outil online : *VennDiagram* (du portail *Bioinformatics and evolutionnary genomics*, de l'université de Ghent), qui permet d'aller jusqu'à 30 liste, toutefois sans génération de graphique (analyse des listes, et output sous forme de tableaux) (Fig. 9). Voici le lien de cet outil :

http://bioinformatics.psb.ugent.be/webtools/Venn/

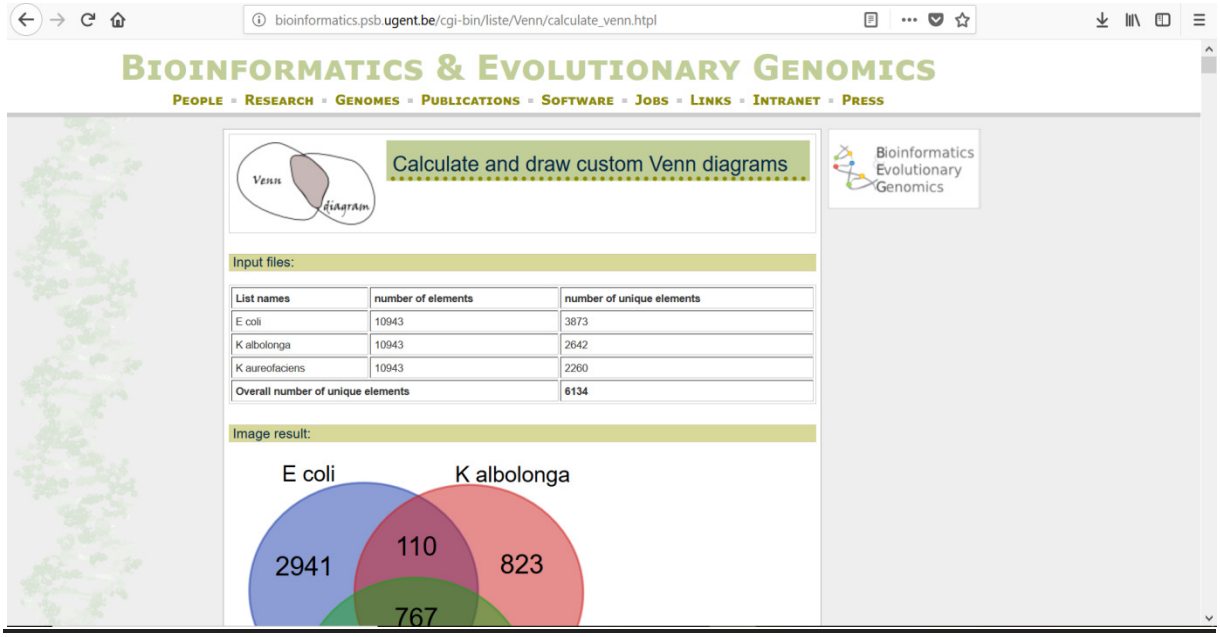

Figure 9 : Exemple de rendu sur l'outil online VennDiagram de l'universté de Ghent

6.2. Comparaison des tables de protéines encodées obtenues par annotation *via* NCBI

Selon le même principe et la même motivation, les tables de protéines obtenues à partir de l'annotation via NCBI ont été utilisé, après formatage et insertion dans l'outil VennPainter et *VennDiagram* pour souligner les ressemblances et les différences entres les génomes des souches au plan « protéines encodées ».

6.3. Comparaison des tables de sous-systèmes (*Subsystems*) obtenues par annotation via RAST

Un sous-système est un ensemble de rôles fonctionnels qu'un annotateur a décidé de considérer comme liés. Fréquemment, les sous-systèmes représentent la collection de rôles fonctionnels qui constituent une voie métabolique, un complexe (par exemple le ribosome) ou une classe de protéines (par exemple des protéines de transduction de signal à deux composants dans *Staphylococcus aureus*) (Ramy et al, 2008).

Pour avoir une idée sur les ressemblances métabolique entre les souches des deux groupes retenues, nous avons préparé les tables des sous-systèmes en format .txt* pour ensuite les analyser sous VennPainter et VennDiagram.

Tableau n°1 : Description des souches de retenues pour l'étude comparative des génomes.

Nom de l'espèce	Souche	Milieu d'isolement	Statut du génome	Numéro d'accession	Localisation géographique	Publication Du génome
Streptomyces collinus	Tu365	Le sol	Fini	PRJNA171216	Kouroussa (Guinée)	Rückert et al., 2013
*Escherichia coli**	SE15	les matières fécales d'un adulte	Fini	PRJDA19053	-	Toh et al., 2009
Kitasatospora albolonga	YIM101047	Sol	Fini	PRJNA381279	-	Labeda et al., 2017
Kitasatospora aureofaciens	DM-1	Sol	Fini	PRJNA381130	Chine: Hebei	Labeda et al., 2017
Streptomyces albus		Sol	Fini	PRJNA271625	Japon	Myronovskyi et al 2018
Streptomyces ambofaciens	DSM40697	Sol	Fini	PRJNA298666	Italie :Rome	Thibessard et al., 2016
Streptomyces atratus	SCSIOZH16	Sol	Fini	PRJNA436062	-	Li et al., 2018
Streptomyces avermitilis	NBRC14893	Sol	Fini	PRJNA189	-	Ikeda et al., 2013
Streptomyces coelicolor	A3(2)	Sol	Fini	PRJNA242	-	O'Rourke et al., 2008
Streptomyces formicae	KY5	Fourmis	Fini	PRJNA396953	Kenya	Qin et al 2017
Streptomyces glaucescens	GLA.O	Sol	Fini	PRJNA260814	-	Rockser et al., 2008
Streptomyces griseochromogenes	ATCC14511		Fini	PRJNA307132	-	Wu, Chen et Feng, 2017.
Streptomyces griseus subsp. griseus	NBRC13350		Fini	PRJDA20085	-	Ohnishi et al.,2008 Hirano et al.,2008
Streptomyces hygroscopicus_subsp. limoneus	KCTC1717	Sol	Fini	PRJNA302679	-	Hang et Shuang, 2012
Streptomyces lavendulae subsp. lavendulae	CCM3239	Sol	Fini	PRJNA407779	-	Bekeova et al., 2015
Streptomyces lividans	TK24	Sol	Fini	PRJNA224116	-	Cruz-Morales et al., 2013
Streptomyces lunaelactis	MM109	Grotte	Fini	PRJNA430192	-	Naome et al., 2018
Streptomyces lydicus	103	Sol	Fini	PRJNA341548	Chine: Tianjin	Jia et al., 2017
Streptomyces parvulus	2297	Écosystème de mangrove	Fini	PRJNA320204	-	Nishizawa, 2016
Streptomyces peucetius	ATCC27952	Sol	Fini	PRJNA394648	-	Dhakal et al., 2018
Streptomyces rapamicinicus	NRRL5491	Sol	Fini	PRJNA207502	-	Baranasic et al., 2013
Streptomyces venezuelae	ATCC15439	Sol	Fini	PRJNA298854	-	Jingxuan He et al., 2016

* : la souche *E. coli* SE15 a été retenue comme *outgroup* pour les analyses de phylogénie.

Résultats et discussion

Résultats et discussion

1. Visualisation des génomes des *Streptomyces* spp. analysées

L'outil DNAplotter nous a permis de représenter les génomes des souches retenues sous formes d'anneaux (de la Fig.10 à la Fig.32), permettant ainsi une abstraction des données basiques relative aux propriétés de chaque génome.

Celui de la souche *E. coli* SE 15 a été représenté par la même occasion, pour servir de référentiel, étant l'espèce bactérienne la plus étudiée.

Lors du paramétrage du programme, nous avons chois de représenter les séquences codantes forward en bleu et reverse en rouge ; les gènes codants pour les ARNr en rose, ceux des ARNt en vert.

Le plus petit des génomes étudié, est celui de la souche *Kitasatospora aureofaciens* DM1 (6,8 Mb), alors que le plus grand se trouve être celui de la souche *Streptomyces rapamycinicus* NRRL5491 (12,7 Mb), ce qui est exceptionnel pour un procaryote (plus que celui de *Saccharomyces cerevisae*), alors que la moyenne des génomes se trouve être aux alentours des 8-9 Mb.

Les génomes des souches représentées affichent des séquences codantes aussi nombreuses en *forward* qu'en *reverse*, révélant ainsi une richesse du génome, qui est une caractéristique connu du genre *Streptomyces*.

Lors du curage des bases de données génomiques, nous nous sommes aperçue que la souche *S. hygroscopicus* avait deux chromosomes, l'un étant plus grand que l'autre (8,6 et 1,8 Mb), ils sont représentés dans les figures 23 et 24.

Il est à retenir que les disparités des tailles des génomes semblent sans relation avec l'origine des souches.

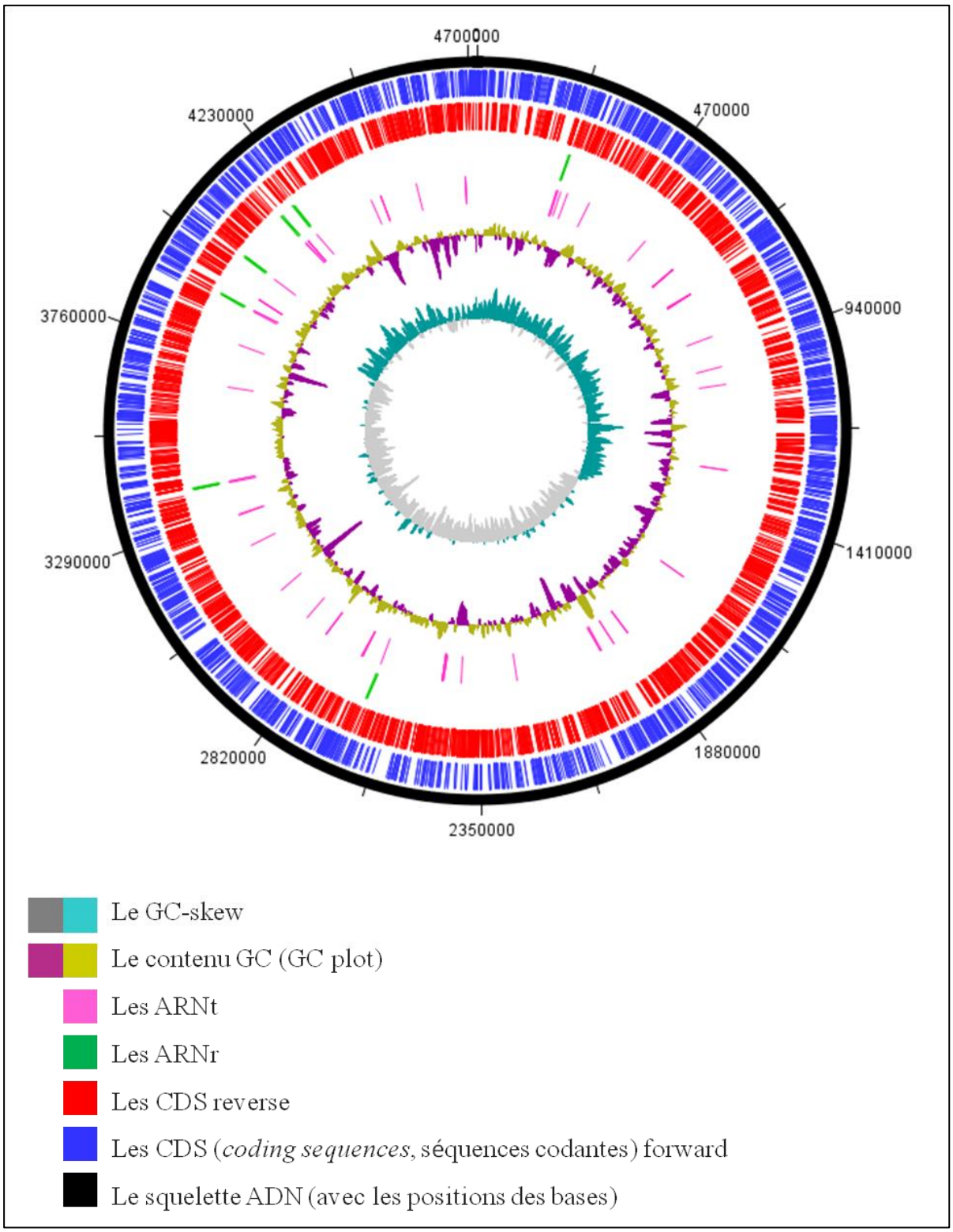

Figure 10: représentation du génome de la souche *E. coli* SE15 avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

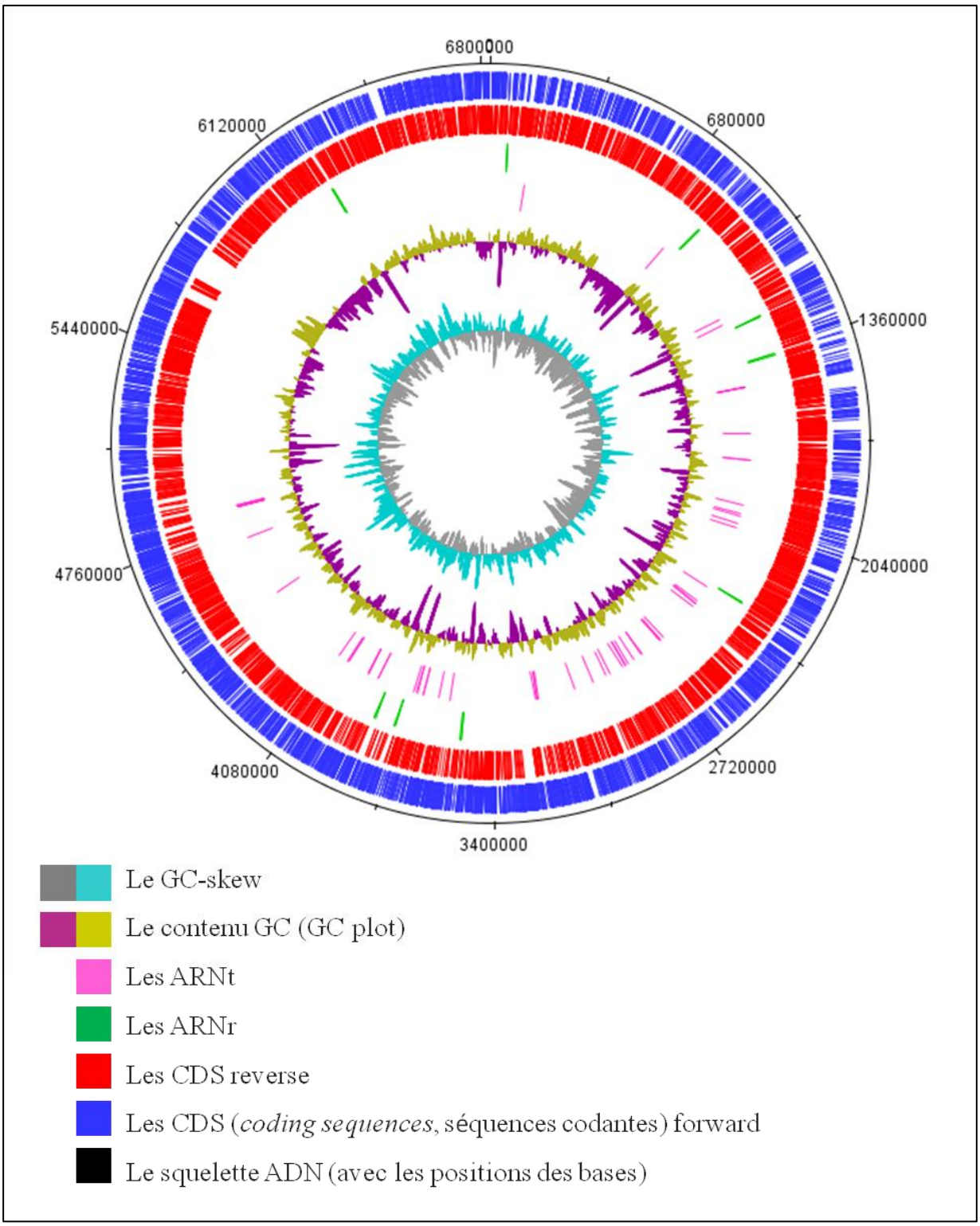

Figure 11: représentation du génome de la souche *Kitasatospora aureofaciens* DM-1 avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

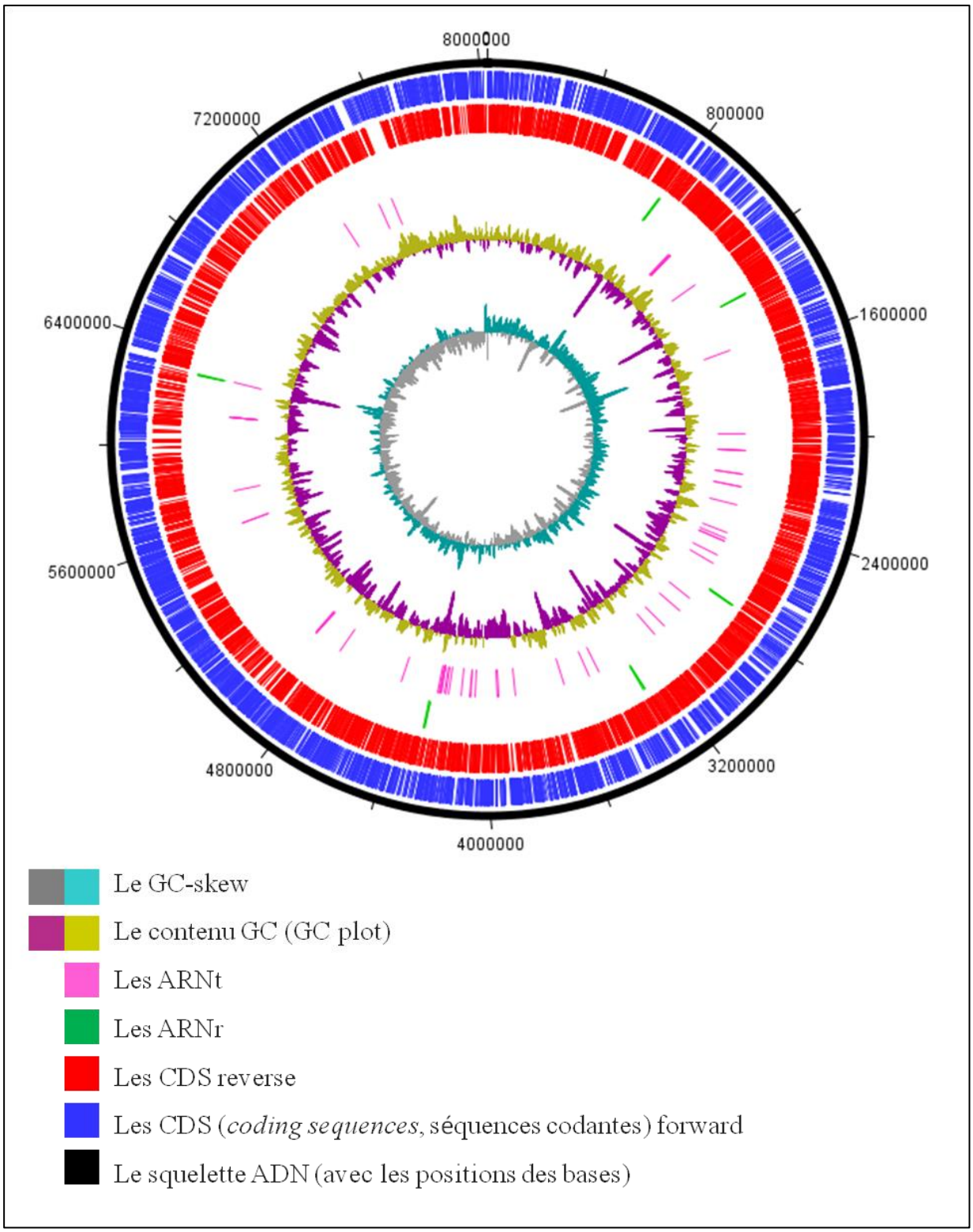

Figure 12: représentation du génome de la souche *Kitasatospora_albolonga* YIM101047 avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

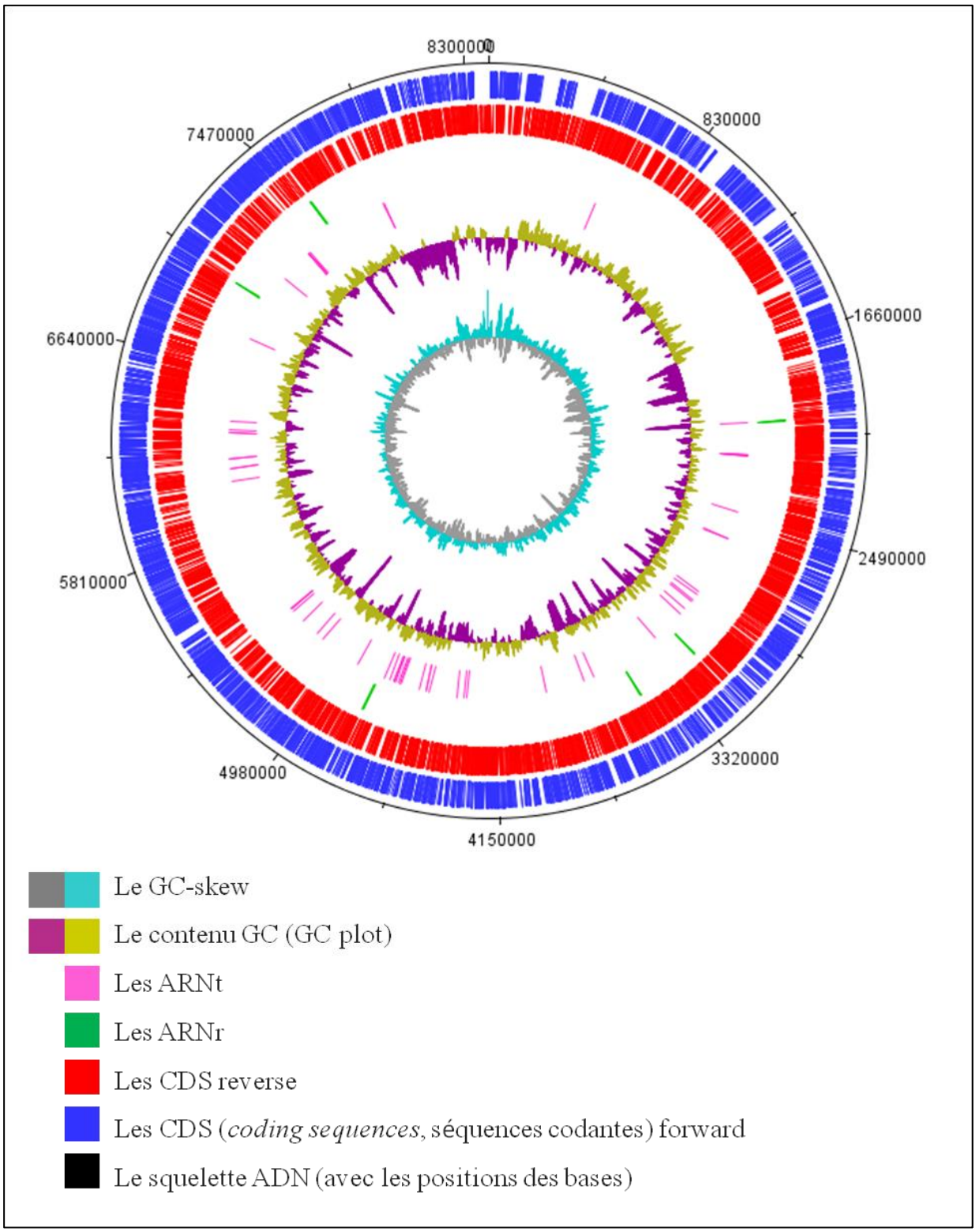

Figure 13: représentation du génome de la souche *Sreptomyces albus* DSM 41398 avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

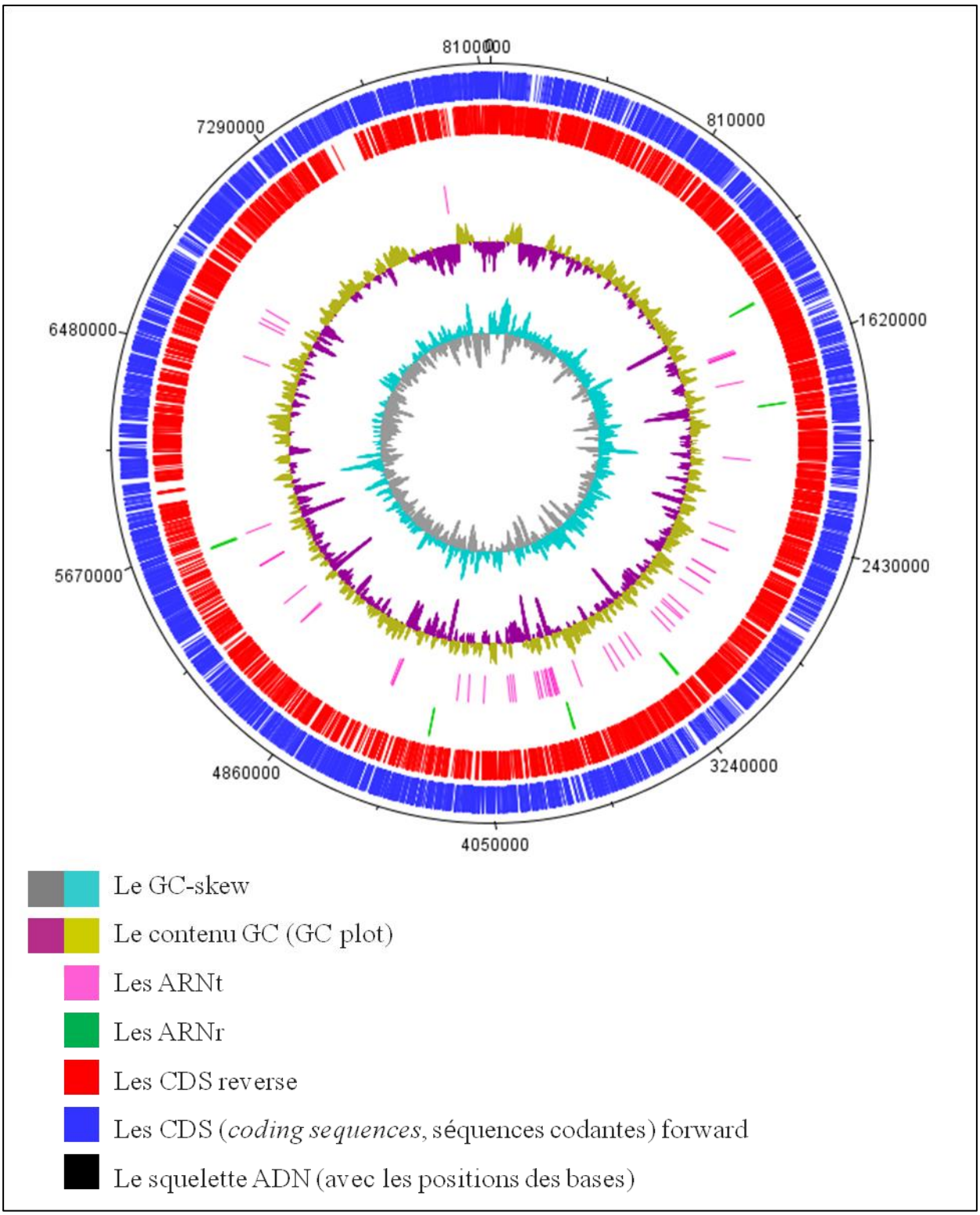

Figure 14: représentation du génome de la souche *S. ambofaciens* DSM40697 avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

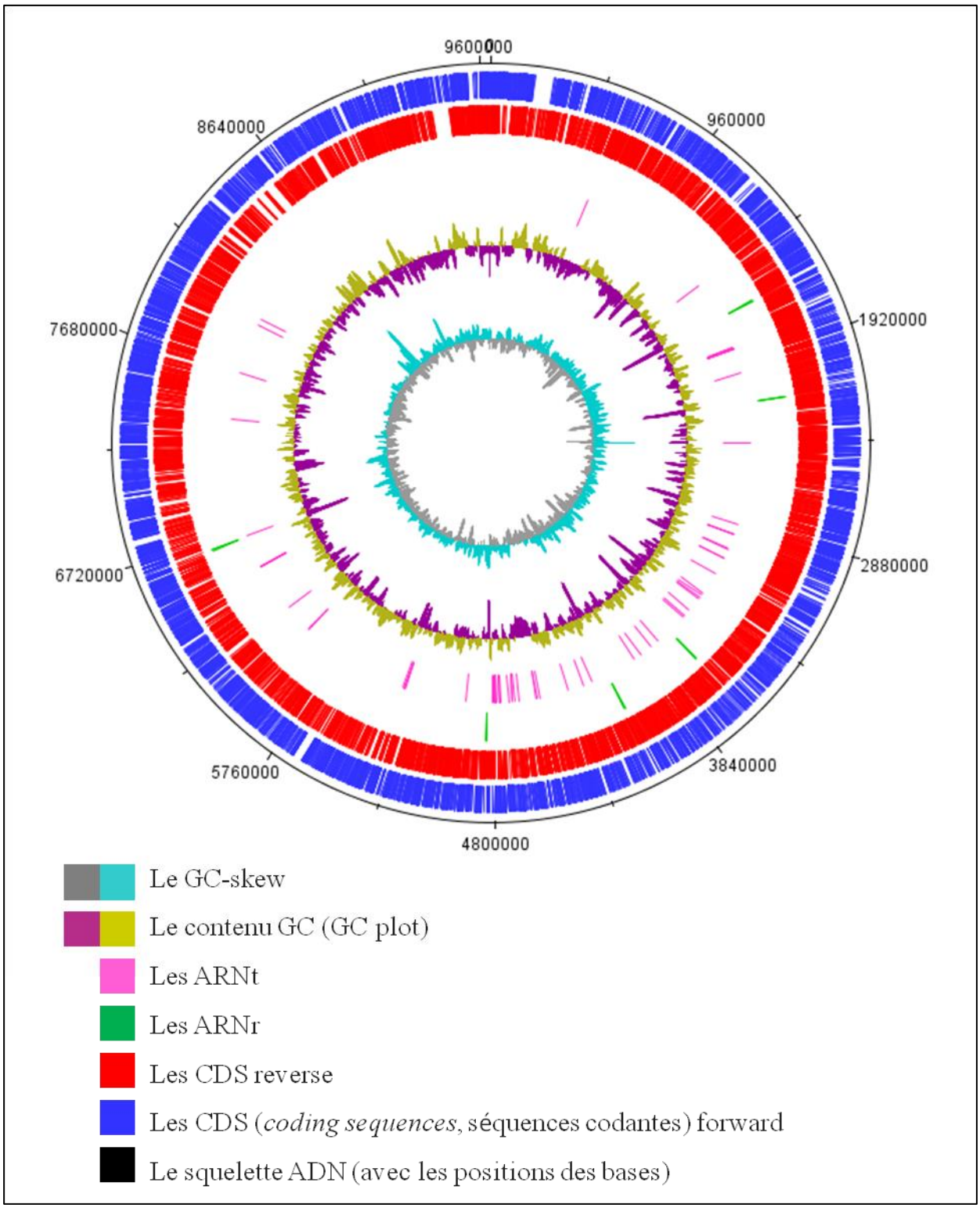

Figure 15: représentation du génome de la souche *S. atratus SCSIOZH16* avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

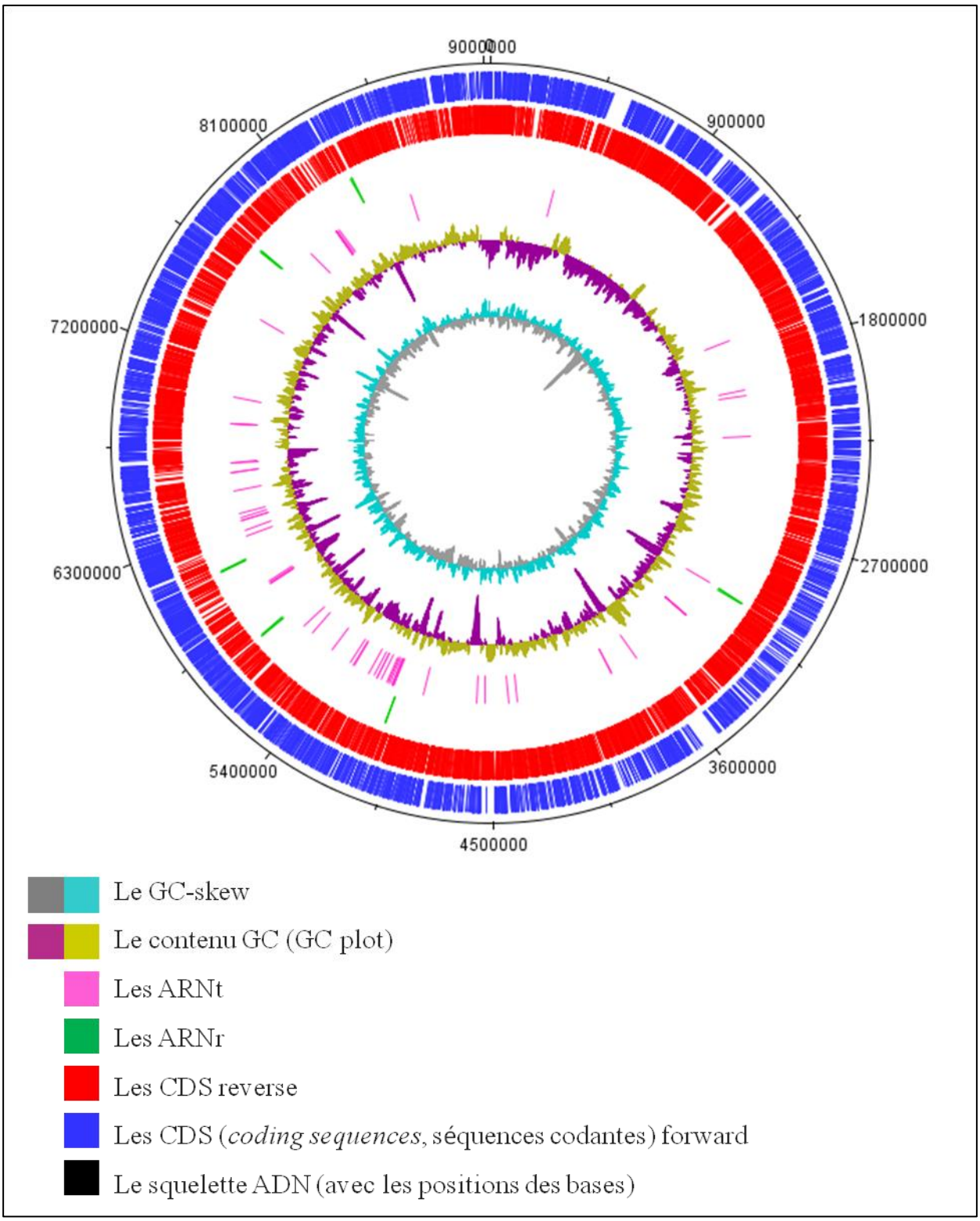

Figure 16: représentation du génome de la souche *S. avermitilis* NBRC14893 avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

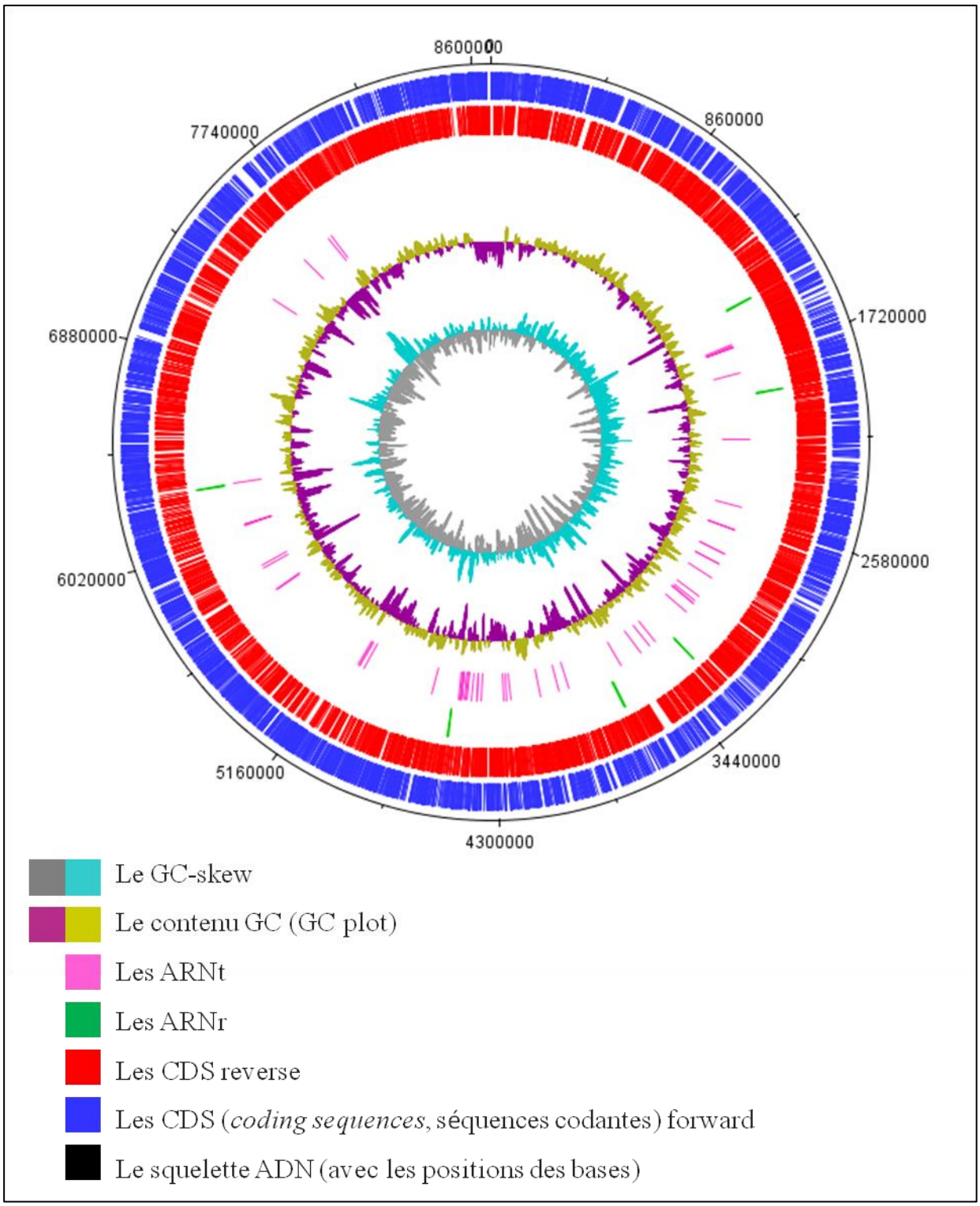

Figure 17: représentation du génome de la souche *S. celiocolor* (A3)2 avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

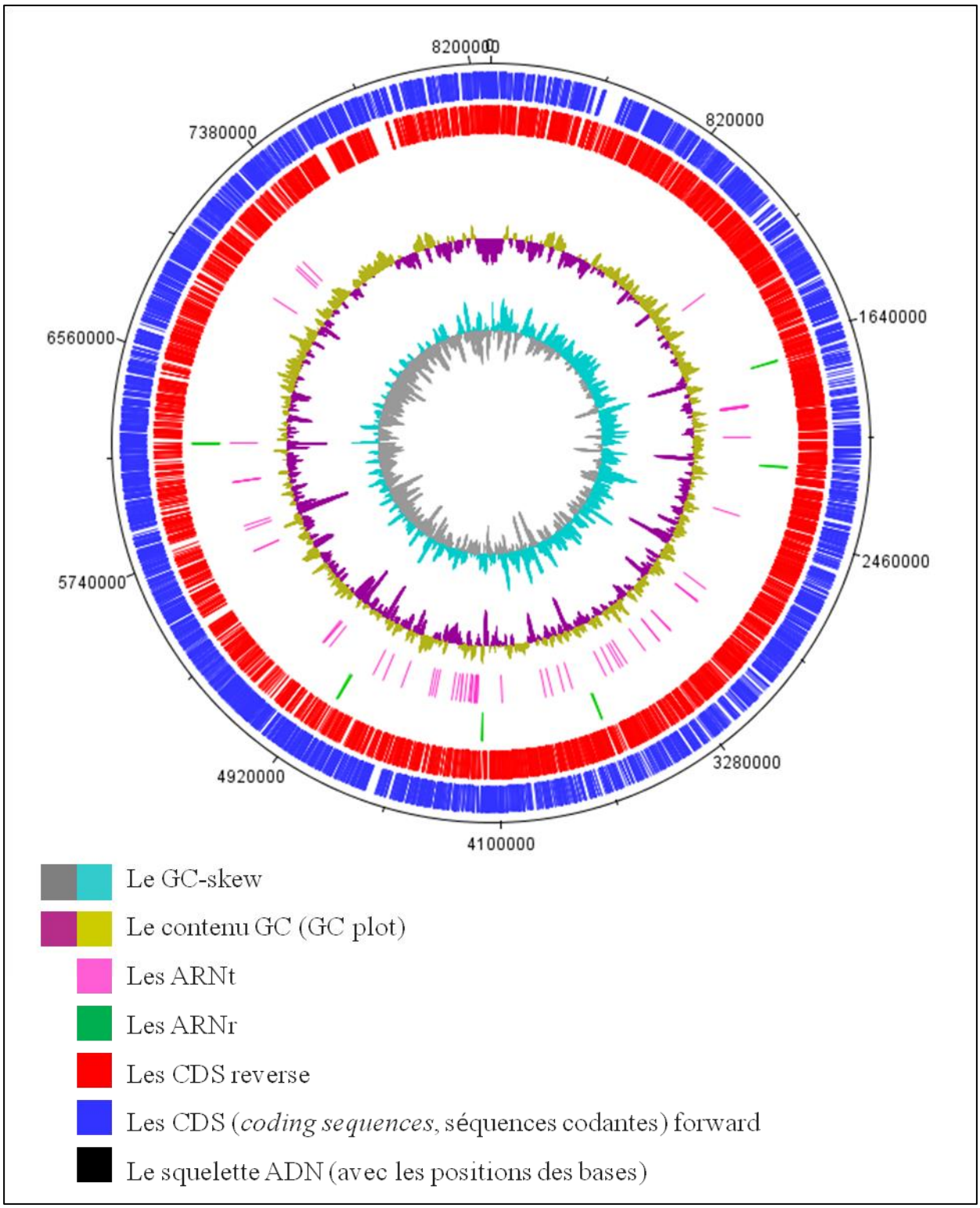

Figure 18: représentation du génome de la souche *S. collinus* Tu365 avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

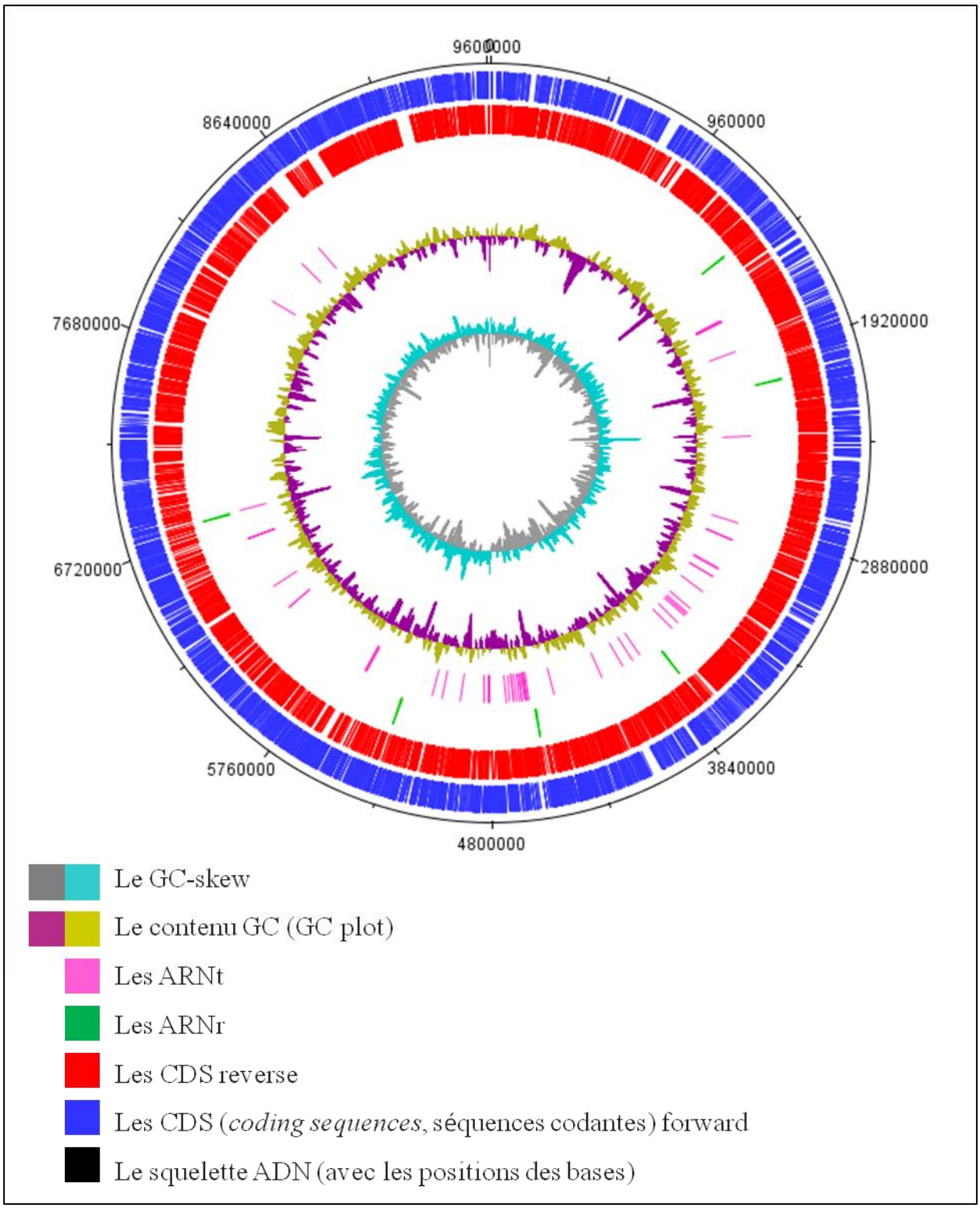

Figure 19: représentation du génome de la souche *S. formicae* KY5 avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

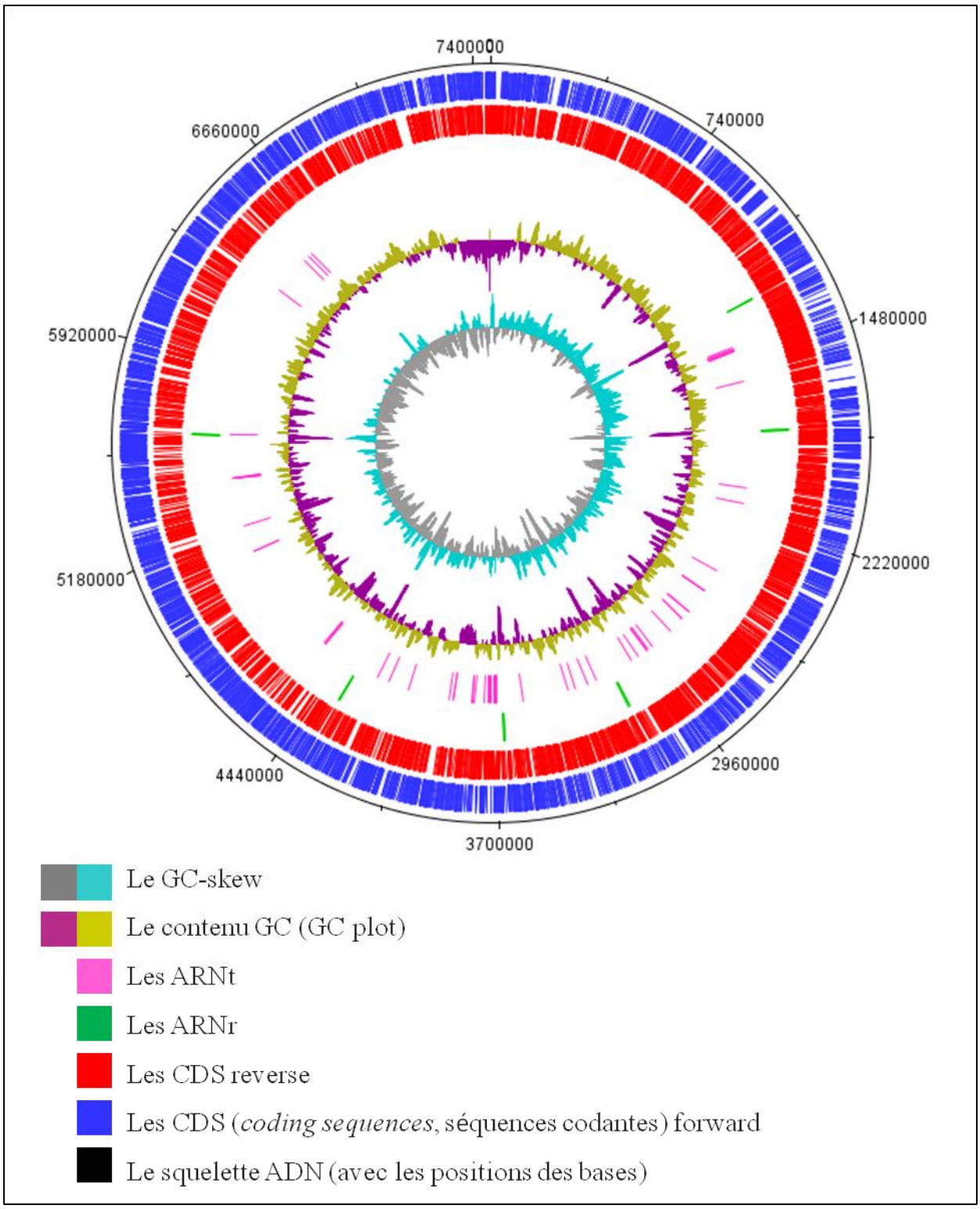

Figure 20: représentation du génome de la souche *S. glaucesens* GLA.O avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

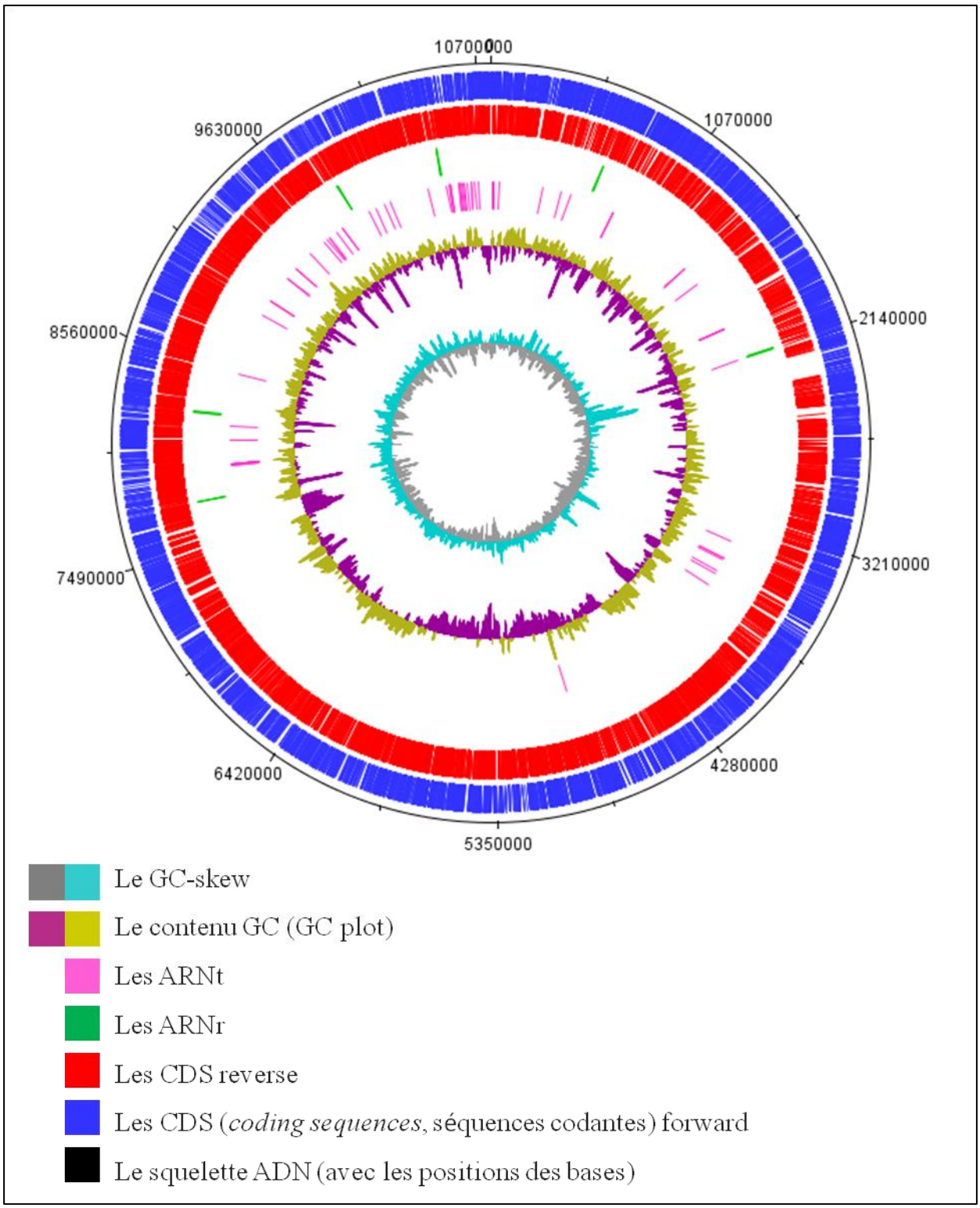

Figure 21: représentation du génome de la souche *S. griseochromogenes* ATCC14511 avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

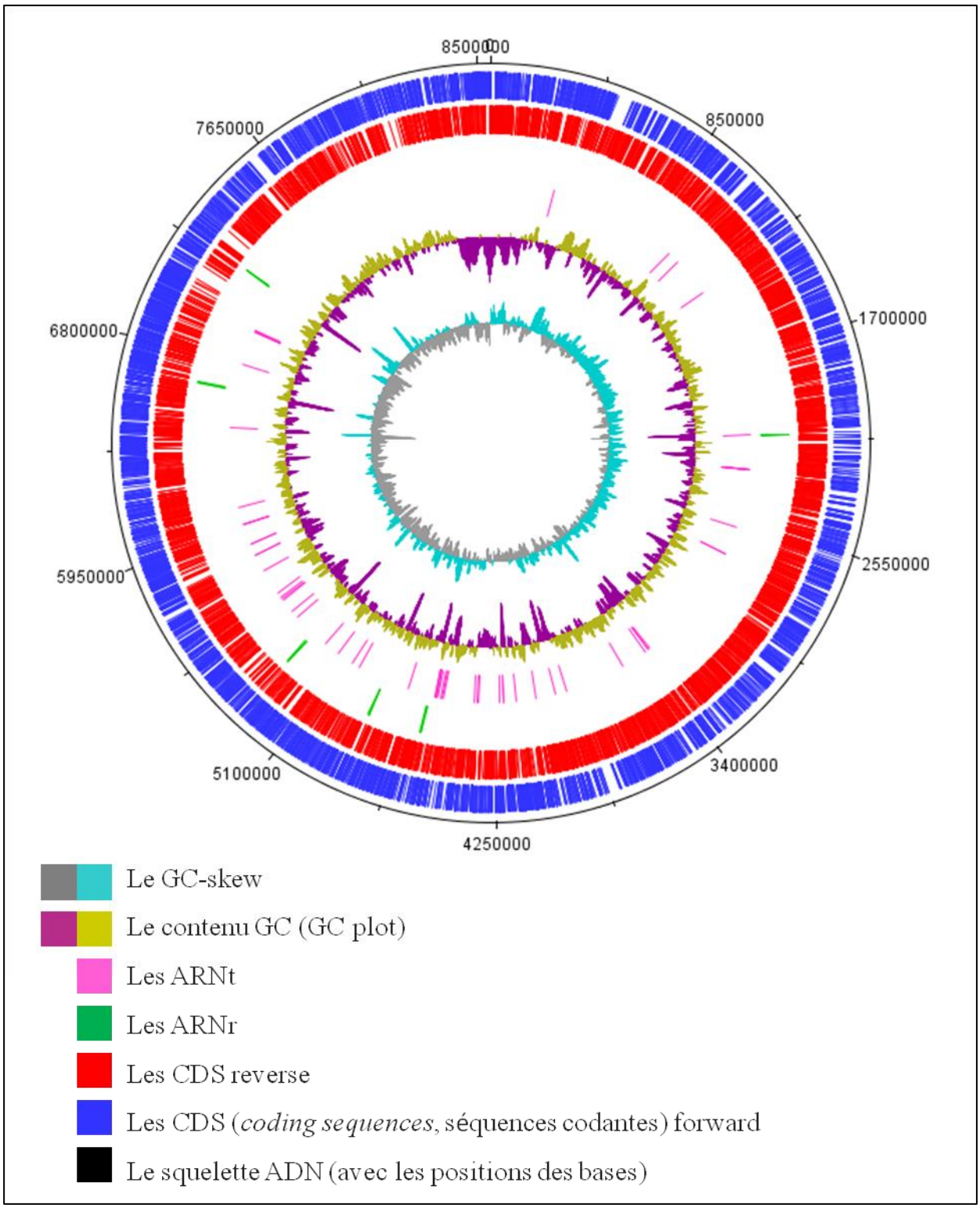

Figure 22: représentation du génome de la souche *S. griseus* NBRC13350 avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

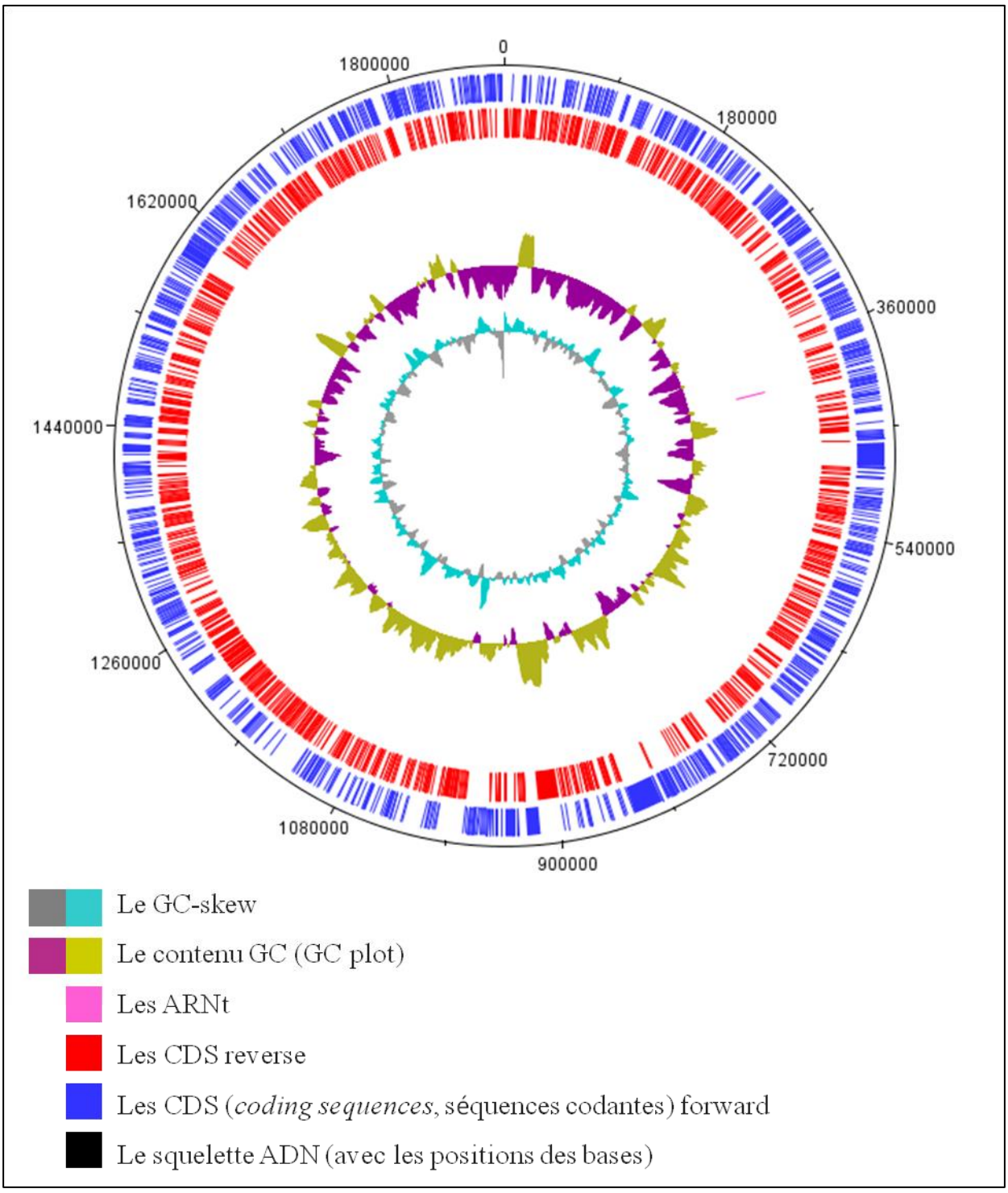

Figure 23: représentation du génome de la souche *S. hygroscopicus* KCTC1717 chromosome 1, avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNt (en rose vif).

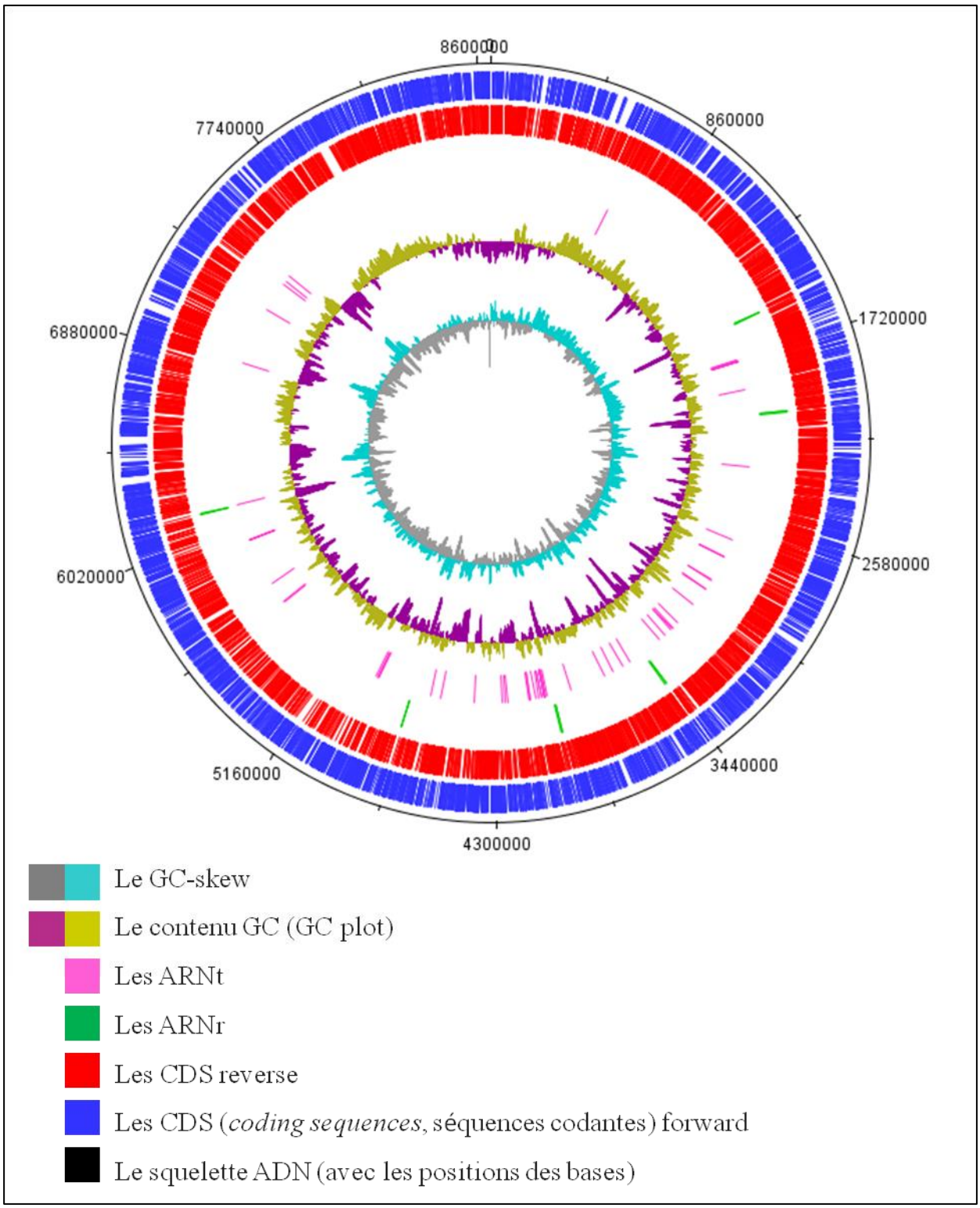

Figure 24: représentation du génome de la souche *S. hygroscopicus* KCTC1717 chromosome 2, avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

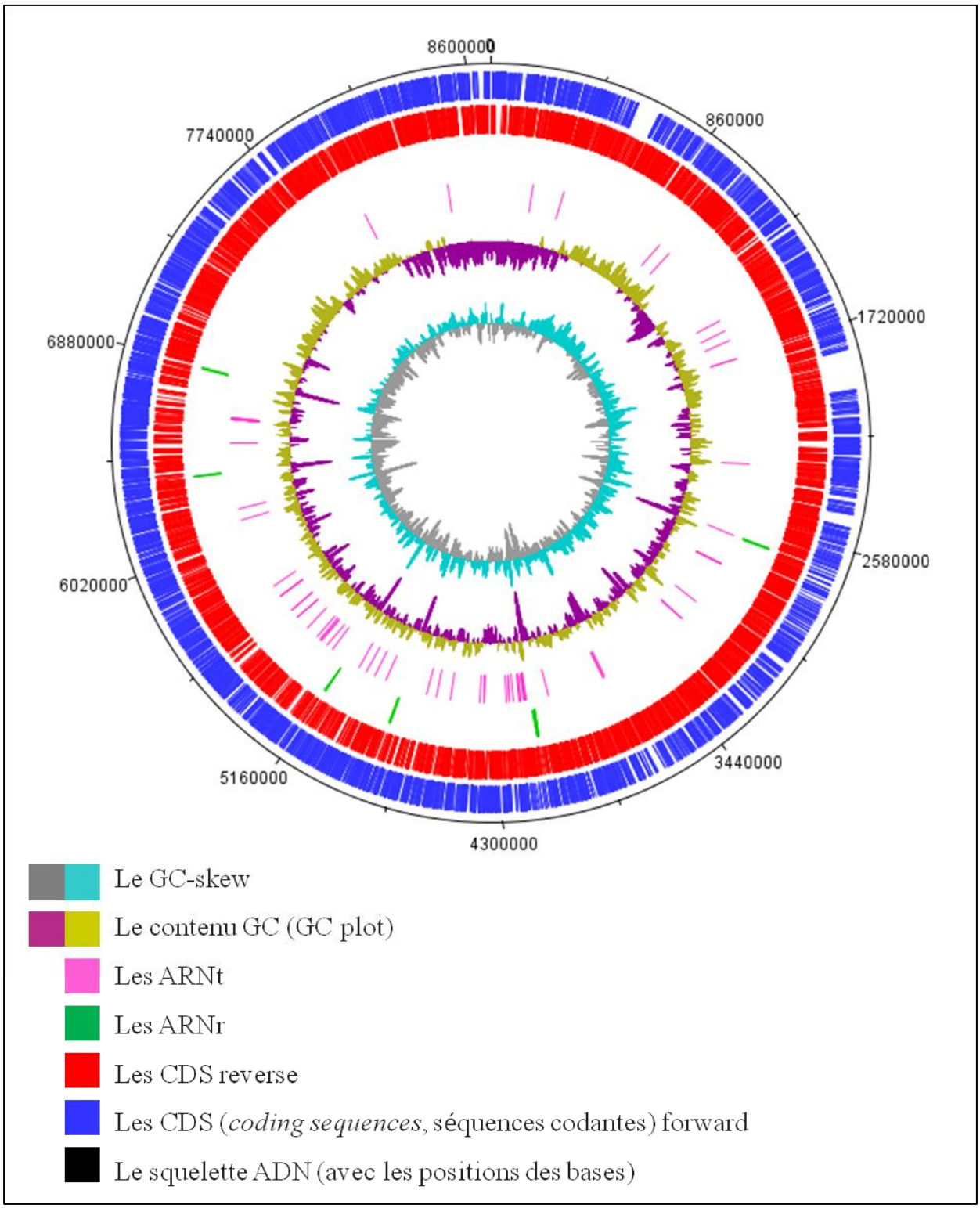

Figure 25: représentation du génome de la souche *S. lavendulae* CCM3239, avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

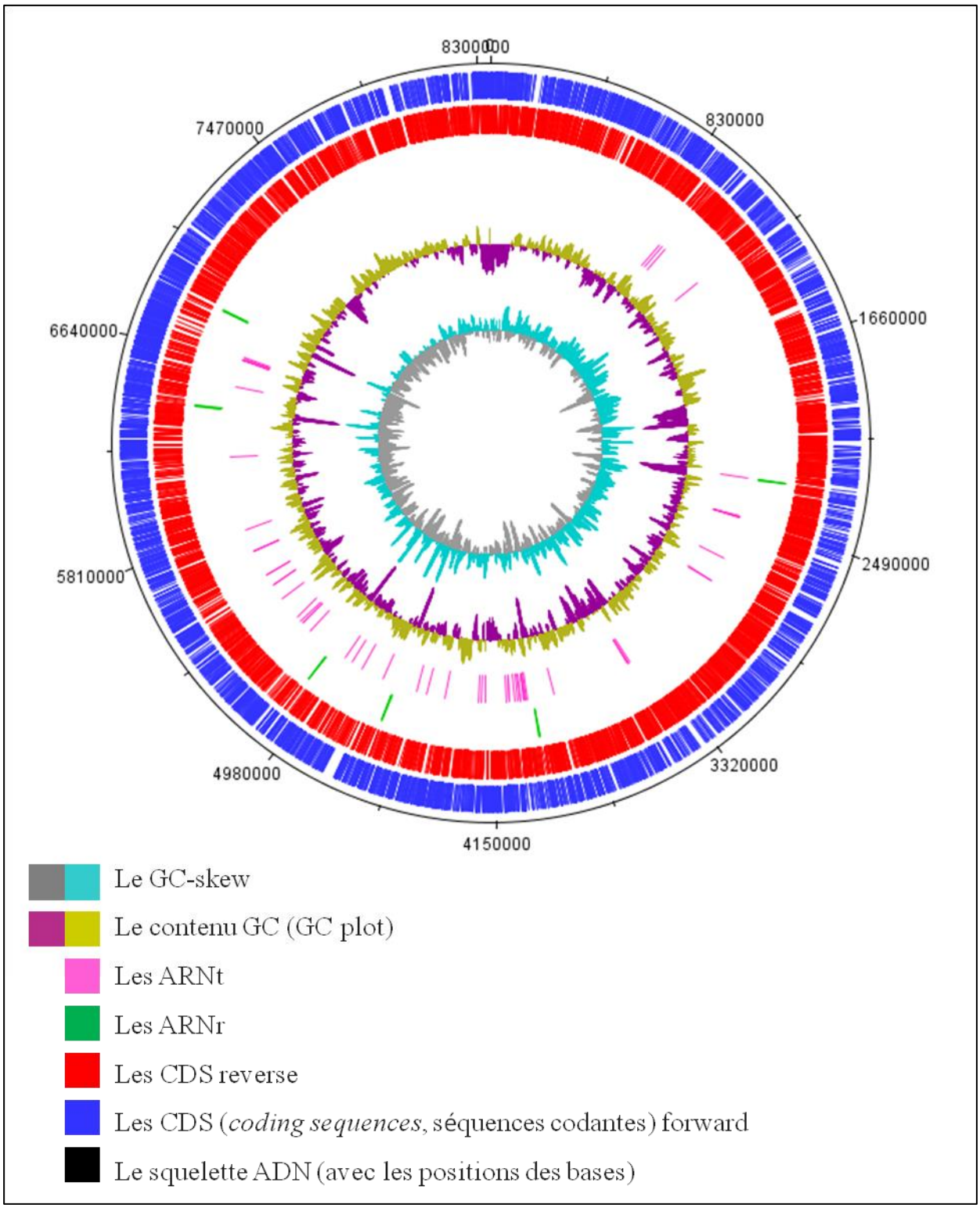

Figure 26: représentation du génome de la souche *S. lividans* TK24, avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

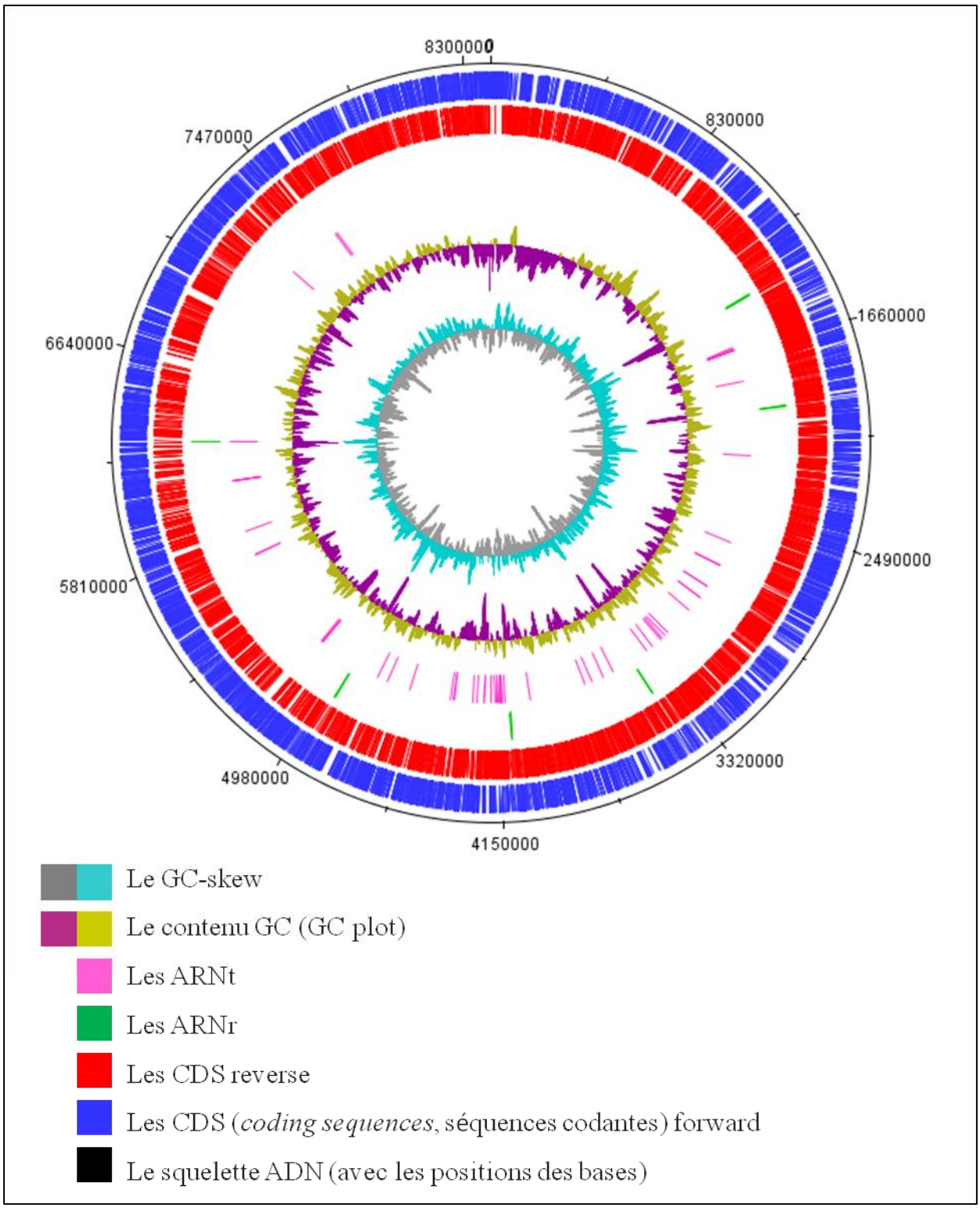

Figure 27: représentation du génome de la souche *S. lunaelactis* MM109, avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

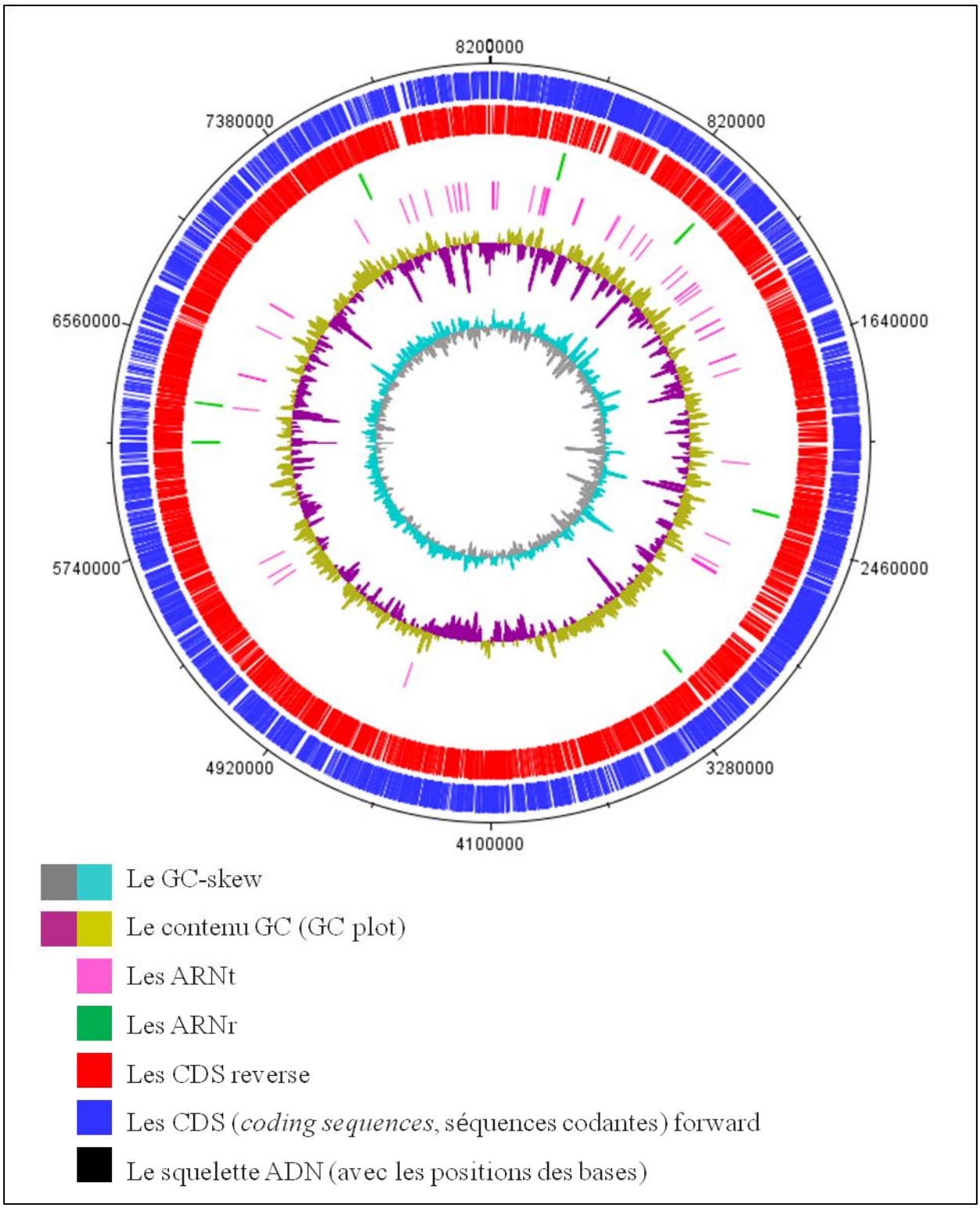

Figure 28: représentation du génome de la souche *S. lydicus* 103, avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

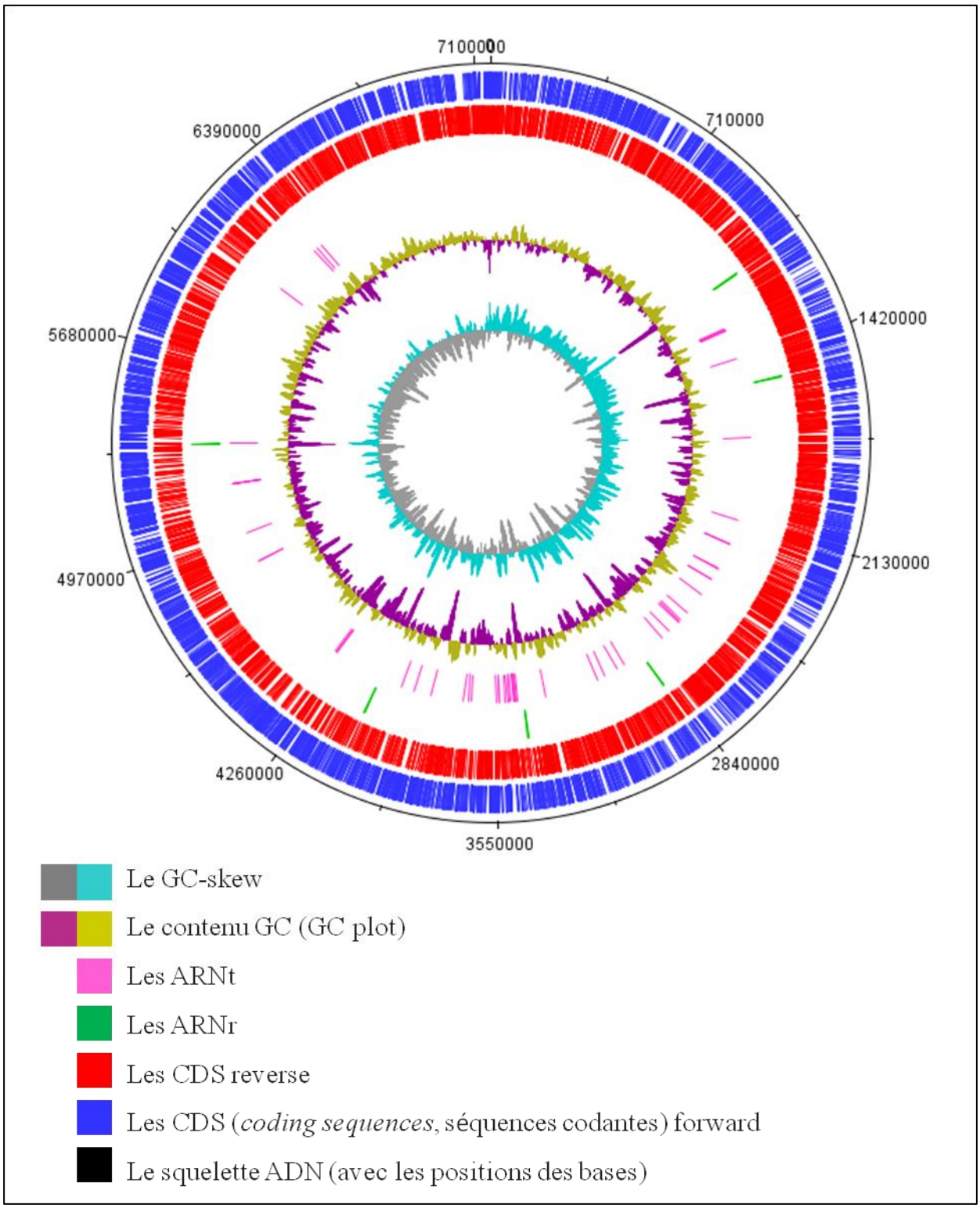

Figure 29: représentation du génome de la souche *S. parvalus* 2297, avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

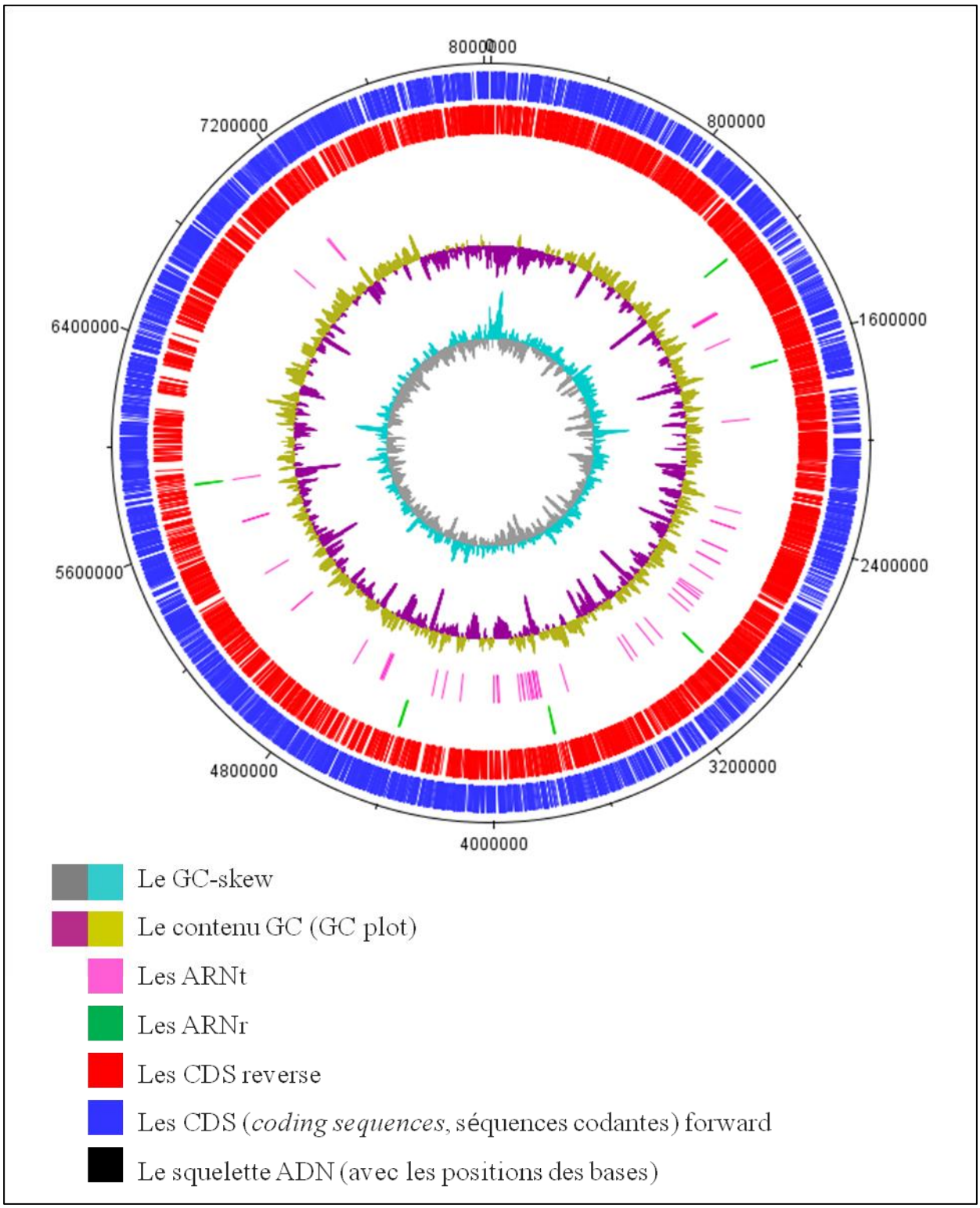

Figure 30: représentation du génome de la souche *S. peucetius* ATCC27952, avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

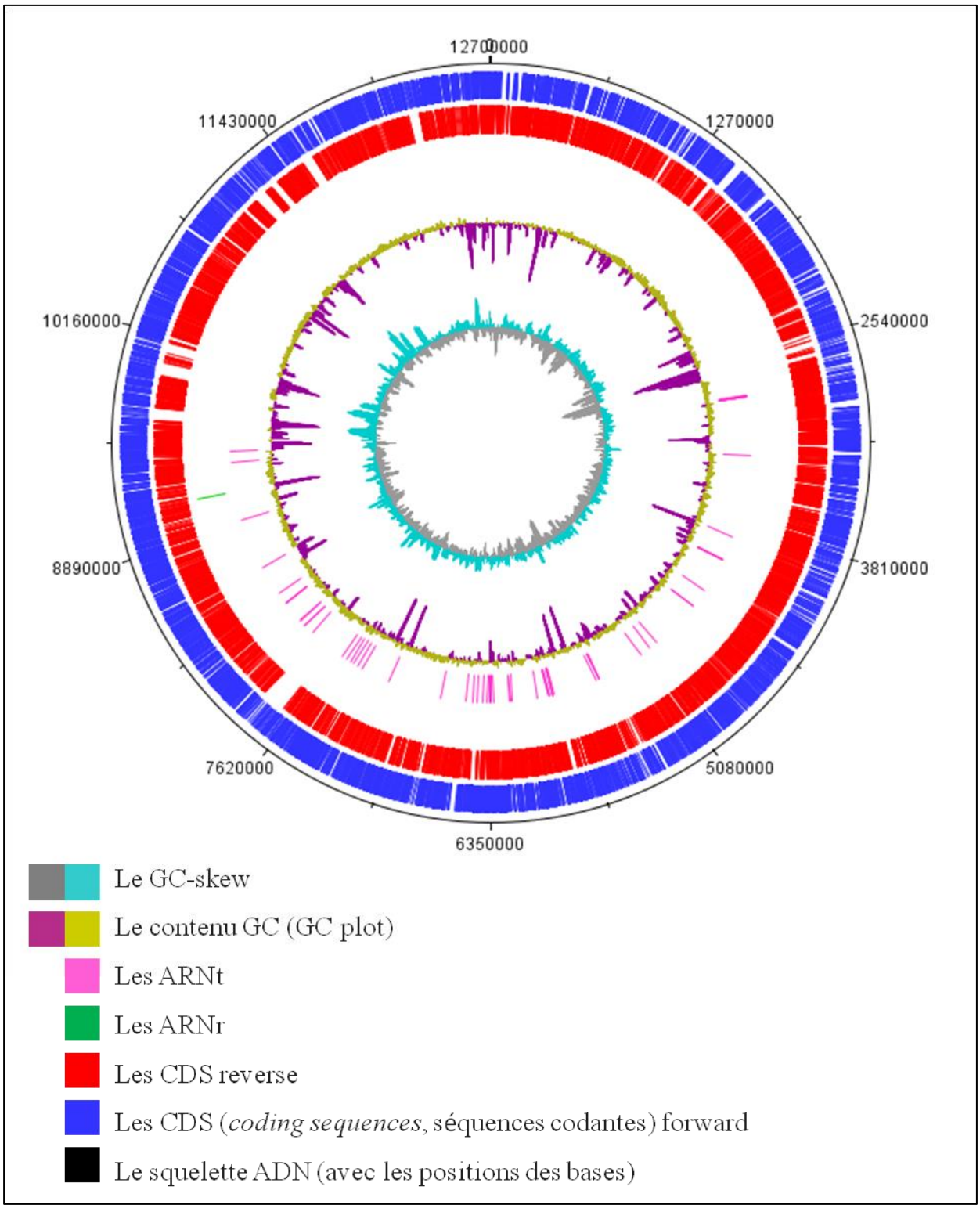

Figure 31: représentation du génome de la souche *S. rapamycinicus* NRRL5491, avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

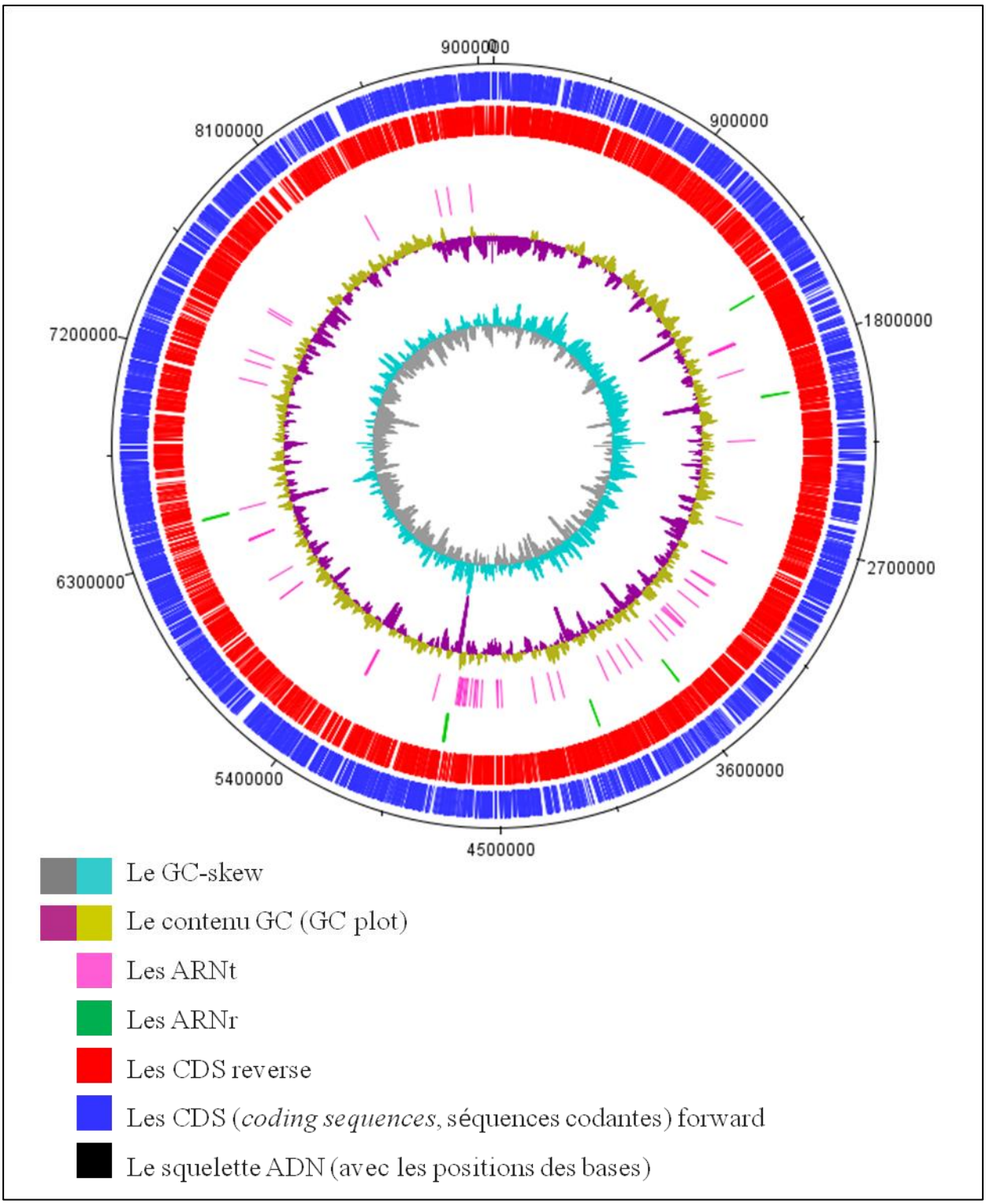

Figure 32: représentation du génome de la souche *S. venuzualae* ATCC15439, avec le programme DNAplotter. La clé de la figure présente les légendes : le squelette ADN (en noir, avec les positions des bases); le contenu GC (GC plot, ocre/violet) ; le GC-skew (gris/bleu ciel) ; les CDS (*coding sequences*, séquences codantes) forward (en bleu) ; les CDS reverse (en rouge) ; les ARNr (en vert claire) ; les ARNt (en rose vif).

2. Phylogénie par analyse du gène de l'ARNr 16S

2.1. Etude phylogénétique par analyse du gène de l'ARNr 16S

L'arbre phylogénétique des copies sous représenté est illustré dans la figure 33 ci-dessous. Celui des copies surreprésentées est illustré dans la figure 34 et enfin, les séquences consensus ont été phylogénétiquement représentées dans la figure 35.

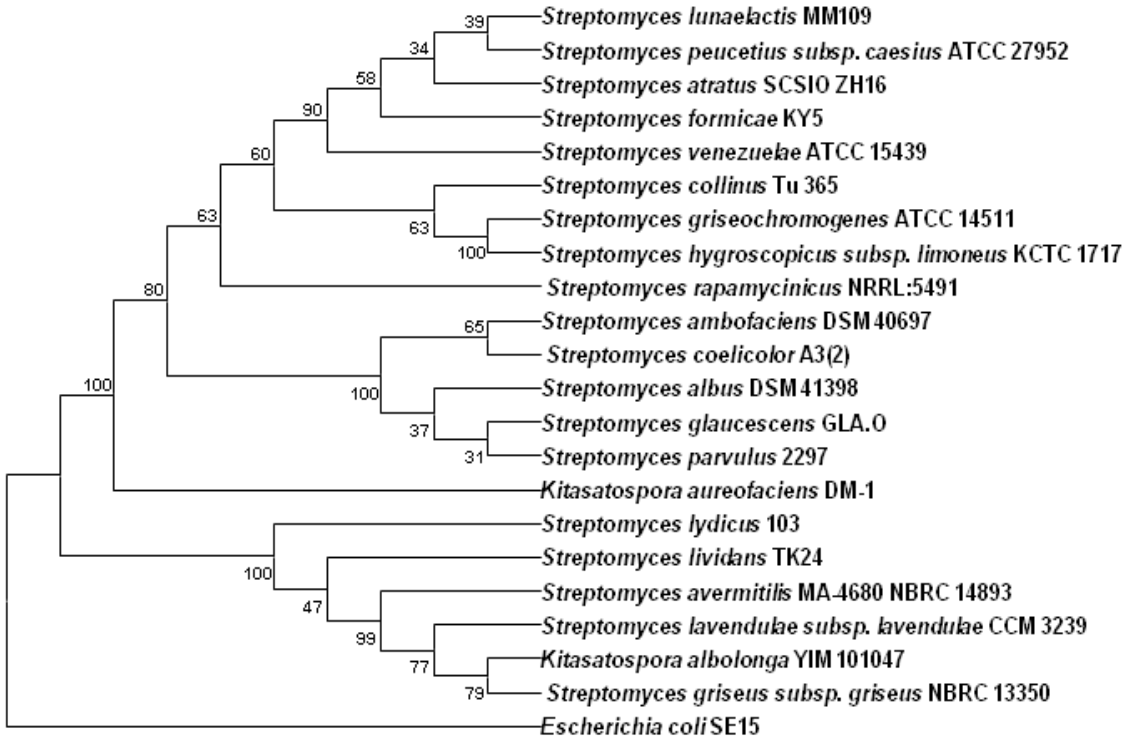

Figure 33: Arbre phylogénétique à partir des séquences du gène de l'ARNr16S (copies sous représentées) (boostrap consensus tree). L'histoire évolutive a été déduite par la méthode *UPGMA*. L'arbre de consensus de bootstrap déduit de 1 000 réplicats est considéré comme représentant l'historique de l'évolution des taxons analysés. Le pourcentage d'arbres de réplication dans lesquels les taxons associés se sont regroupés dans le test d'amorçage (1 000 réplicats) est indiqué à côté des branches. Les distances d'évolution ont été calculées à l'aide de la méthode de *Maximum Composite Likelihood* et sont exprimées en unités du nombre de substitutions de base par site. L'analyse a porté sur 22 séquences de nucléotides. Toutes les positions contenant des lacunes et des données manquantes ont été éliminées. Des analyses évolutives ont été menées dans MEGA7.

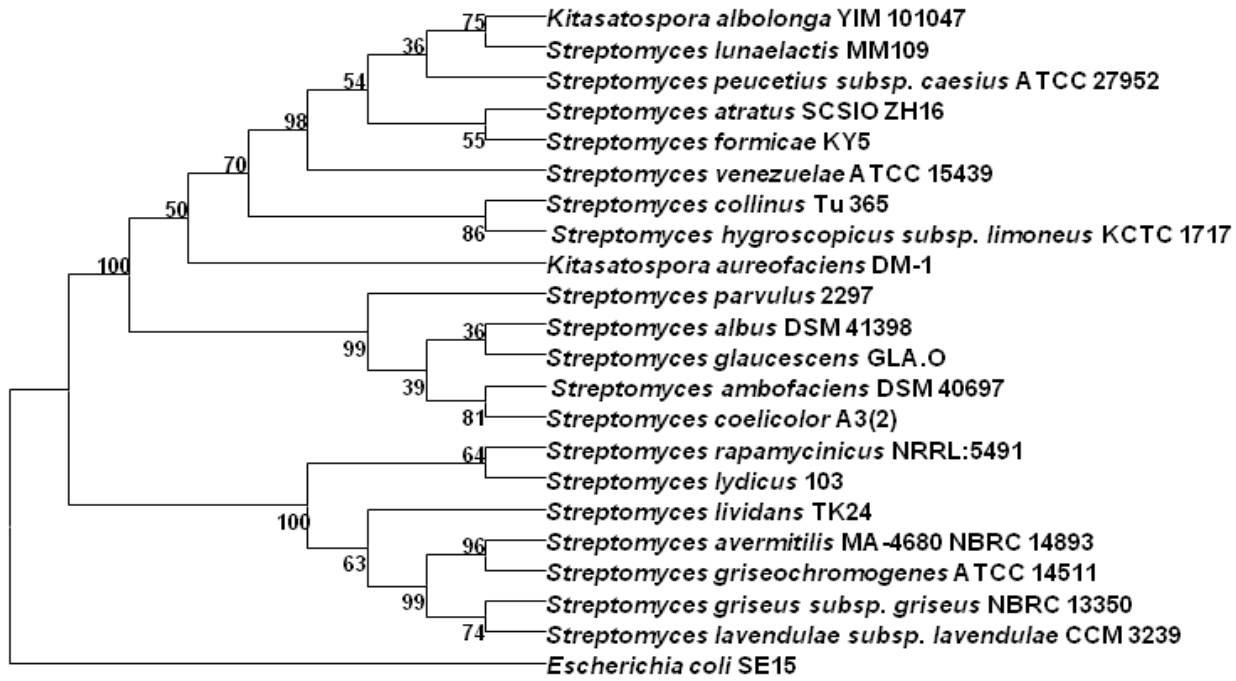

Figure 34: Arbre phylogénétique à partir des séquences du gène de l'ARNr16S (copies surreprésentées) (bootstrap consensus tree). L'histoire évolutive a été déduite par la méthode *UPGMA*. Les mêmes paramètres que ceux de l'arbre précédent ont été pris en considération.

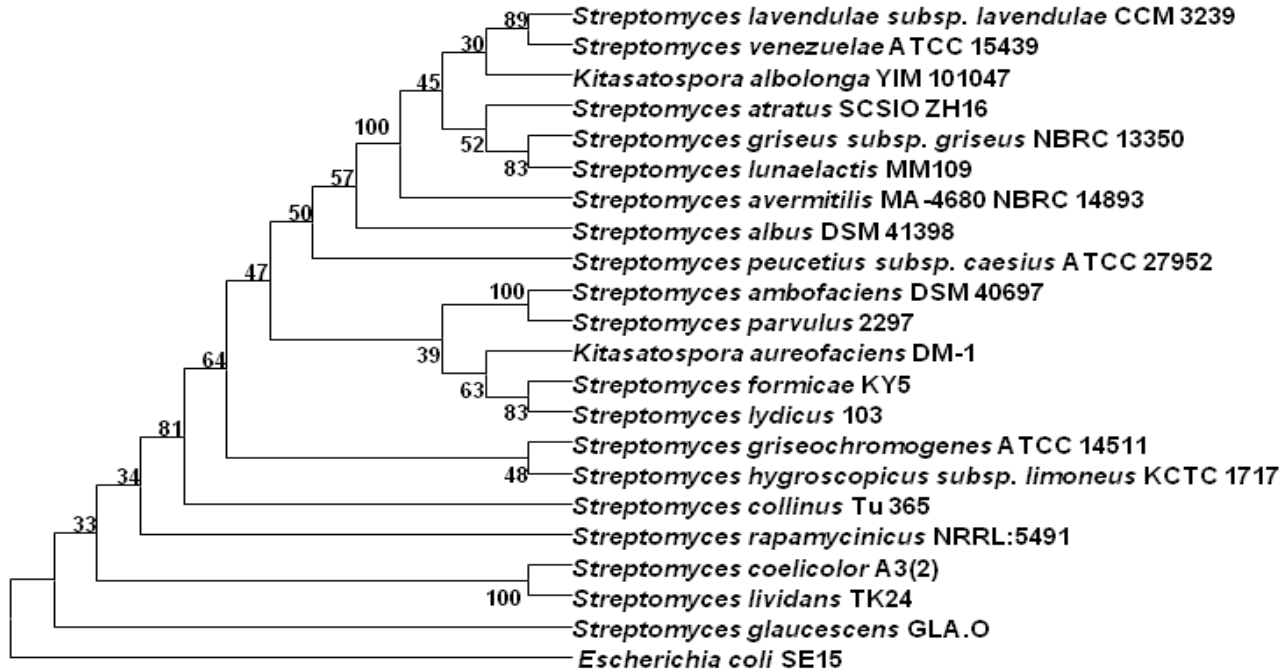

Figure 35 : Arbre phylogénétique à partir des séquences du gène de l'ARNr16S (séquences consensus) (bootstrap consensus tree). L'histoire évolutive a été déduite par la méthode *UPGMA*. L'arbre Les mêmes paramètres que ceux des arbres précédents ont été pris en considération.

L'analyse phylogénétique des souches *Streptomyces* spp. retenues, à aboutie à des résultats divergents, en fonction de l'approche, par rapport aux séquences des ARNr16S employées.

En effet, les trois arbres n'affichent pas les même positions pour les mêmes taxa, d'une part, et d'autre part les clades ne sont pas les mêmes, avec des valeurs de bootsrap hétérogènes, d'un arbre à l'autre.

Entre l'arbre des copies sous-représentées, et surreprésentées, la souche *S. griseochromogenes* n'est pas dans le même clade (tantôt avec *S. hygroscopicus,* tantot avec *S. avermitilis*).

Toutefois, c'est bien l'arbre inféré à partir des séquences consensus (par alignement) qui a donné la meilleur résolution par analyse du gène de l'ARNr 16S.

2.2. Etude phylogénomique

a. Par approche FFP

L'approche FFP est de nos jours une approche de choix, car elle prend base sur le génome complet d'une part, et d'autre part elle ne nécessite pas de fastidieux alignement.

Les *Streptomyces* sont connues pour être un groupe taxonomique non résolu par ARNr16S. En effet, la seule analyse de ce gène est insuffisante pour inférer correctement les positions taxonomique des membres de ce très large genre.

La figure 36 représente les résultats de l'analyse par FFP conduite sous Biolinux pour l'occasion (car c'est un programme fonctionnant en commande Line).

Il s'est avéré que les résultats obtenus par cette approche sont bien meilleure que la précédente, en terme de résolution des positions des taxa. En effet, la délinéation entre les souches est claire, sans ambigüité, par rapport à ceux des arbres par ARNr 16S.

Par exemple, *S. griseochromogenes* se détache nettement de *S. avermitilis* ou de *S. hygroscopicus*.

Seules les souches des espèces *S. lydicus, K. aureofaciens, S avermitilis, S. formicae, S. albus, K. albolonga* nécessitent une meilleure approche.

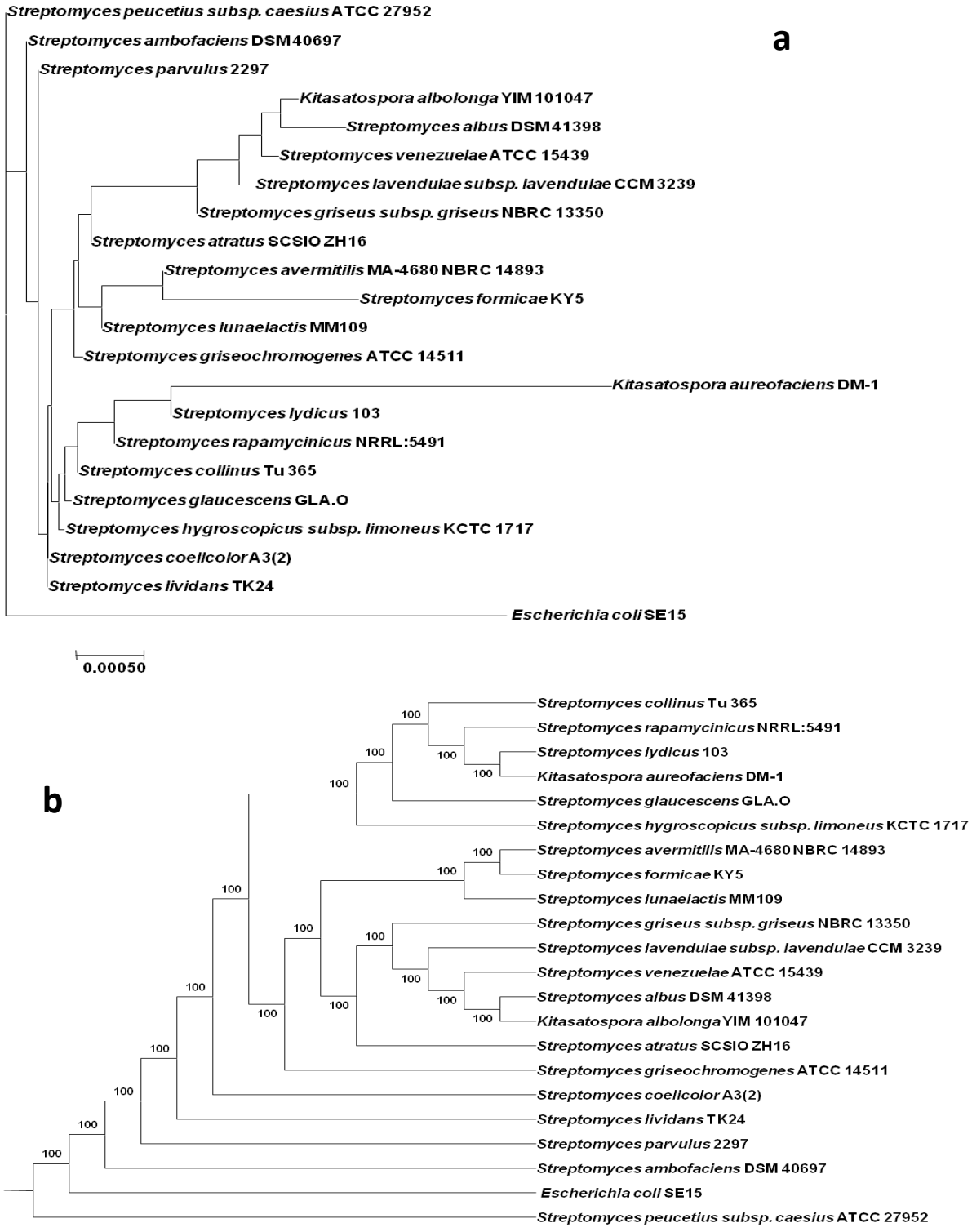

Figure 36: Arbre phylogénomique par algorithme Neighbor Joining NJ (basé sur le génome complet des souches sélectionnées) par approche FFP. a : arbre original, b : arbre consensus.

L'arbre a été construit à l'aide de l'algorithme NJ basé sur les caractéristiques FFP des données du génome entier. *Escherichia coli* SE15 a été utilisé comme groupe externe dans l'analyse. Les valeurs de confiance du bootstrap ont été générées en utilisant 1000 permutations.

b. Par approche UBCG

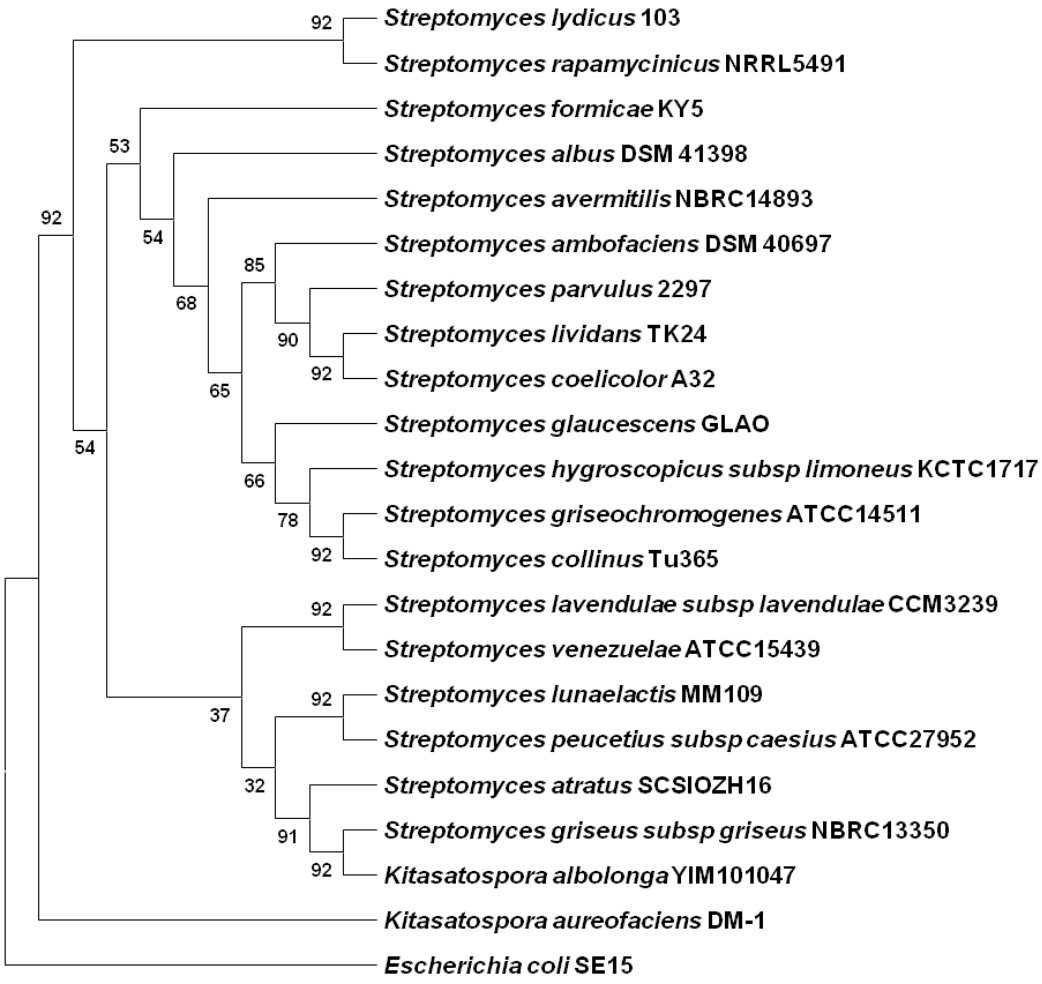

Figure 37: Arbre phylogénomique par algorithme maximum likelihood (basé sur 92 gènes, voir annexe) par approche UBCG (arbre consensus).

Les chiffres sur les nœuds indiquent le nombre de gènes qui supporte la position en question.

L'approche UBCG est une autre alternative aux approches classiques basées sur un seul gène. Au lieu de ça, ce sont 92gènes utilisés pour une meilleures évaluation des relations évolutives entre différents taxa.

Toutefois, dans le cas de notre data set, cette méthodes n'apporta pas les résultats escomptés. En effet, la résolution demeure assez floue, si comparée à ca précédente, par FFP.

Si cette méthode fit déjà ses preuves dans d'autres études ayant portées sur d'autres groupes taxonomiques, il semblerait qu'elle ne soit pas la meilleure solution pour les *Streptomyces*.

3. Annotation des génomes et analyse comparative

3.1. Comparaison des tables de séquences codantes (*CDS*) obtenues par annotation *via* RAST

Le diagramme de Venn généré par cette analyse nous à permis de mettre en évidence un nombre assez important de CDS partagé par les souches du groupe en relation avec les plantes. En effet, 3133 séquences codantes seraient en commun entre les souches Streptomyces coeliocolor A3(2); *Kitasatospora aureofaciens* DM-1; *Streptomyces griseus subsp. griseus* NBRC13350; *Streptomyces lividans* TK24 ; *Streptomyces griseochromogenes* ATCC14511; *Streptomyces lavendulae subsp. lavendulae* CCM3239 ; *Streptomyces hygroscopicus subsp. limoneus* KCTC1717 ; *Streptomyces lydicus* 103. En plus d'avoir les chiffre, le programme VennPainter nous a permis aussi d'avoir le contenu de la liste en commun.

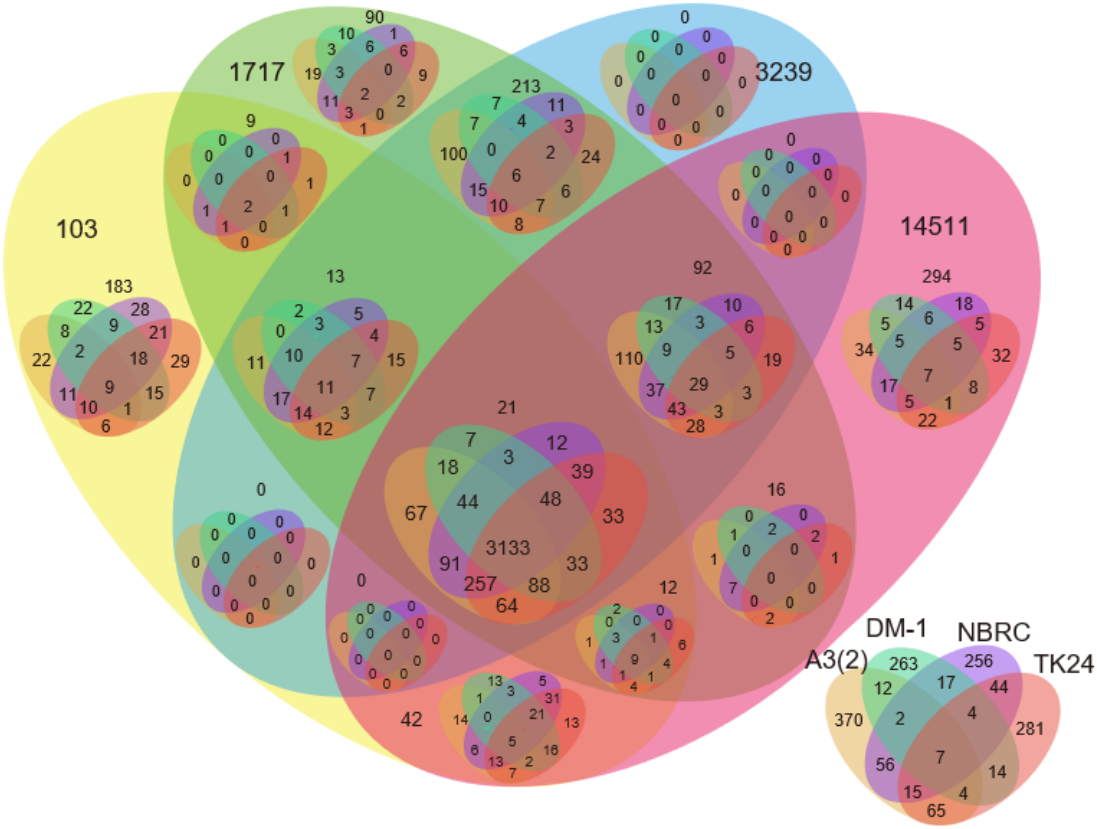

Figure 38: Diagramme de Venn représentant les CDS (Annotation RAST) en communs et uniques au sain du groupe plant-related. Les nombres représentent les séquences codantes en commun entre les souches représentées. A3(2) : Streptomyces coeliocolor A3(2); DM-1 : *Kitasatospora aureofaciens* DM-1; NRBC : *Streptomyces griseus subsp. griseus* NBRC13350; TK24 : *Streptomyces lividans* TK24 ; 14511 : *Streptomyces griseochromogenes* ATCC14511; 3239 : *Streptomyces lavendulae subsp. lavendulae* CCM3239 ; 1717 : *Streptomyces hygroscopicus subsp. limoneus* KCTC1717 ; 103 : *Streptomyces lydicus* 103.

3.2. Comparaison des tables de protéines encodées obtenues par annotation *via* NCBI

Cette comparaison nous a permis d'avoir la liste de 751 protéines partagées entre les souches du groupe en relation avec les plantes, et c'est en scrutant (comme perspective) le contenue de ce pool protéique qu'il sera possible d'établir des liens entre protéome et habilité à promouvoir la croissance des plantes. Par ailleurs il est intéressant de signaler que la souche *S. coeliocolor* se distingue par un set de 4257 protéines uniques, ce qui en fait la souche la plus distinguées des autres souche du groupe en relation avec les plantes.

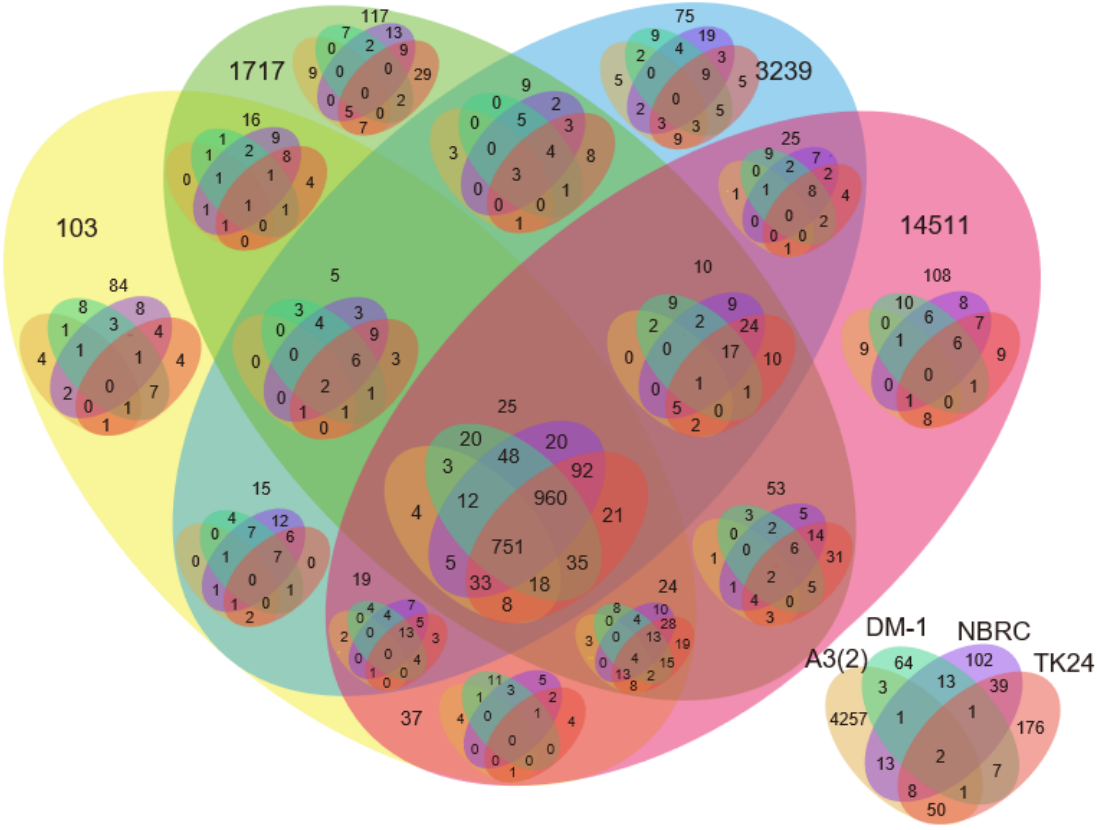

Figure 39: Diagramme de Venn représentant les proteines encodés (annotation NCBI) en communs et uniques au sain du groupe plant-related. Les nombres représentent les gènes codants pour des protéines en commun entre les souches représentées. A3(2) : Streptomyces coeliocolor A3(2); DM-1 : *Kitasatospora aureofaciens* DM-1; NRBC : *Streptomyces griseus subsp. griseus* NBRC13350; TK24 : *Streptomyces lividans* TK24 ; 14511 : *Streptomyces griseochromogenes* ATCC14511; 3239 : *Streptomyces lavendulae subsp. lavendulae* CCM3239 ; 1717 : *Streptomyces hygroscopicus subsp. limoneus* KCTC1717 ; 103 : *Streptomyces lydicus* 103.

3.3. Comparaison des tables de sous-systèmes (*Subsystems*) issues de l'annotation RAST

Un grand nombre de sous-systèmes (1523) s'est avéré en commun pour les souches, ce qui est logique vu leur parenté générique. En effet un grand nombre de fonction sont partagées par les souches de *Streptomyces* ayant une relation avec les plantes, ce qui sous entend un nombre important d'enzymes et de voie métabolique de manière générale. Il serait intéressant à présent que les listes des sous-systèmes en communs étant connu, de revenir au serveur RAST pour des constructions métaboliques qui aiderait à mettre en exergue les chevauchement de métabolisme dans le but de mettre en évidence de nouvelle voie en relation avec l'efficacité dans la promotion de la croissance des plantes, et ainsi identifier de nouveaux mécanismes.

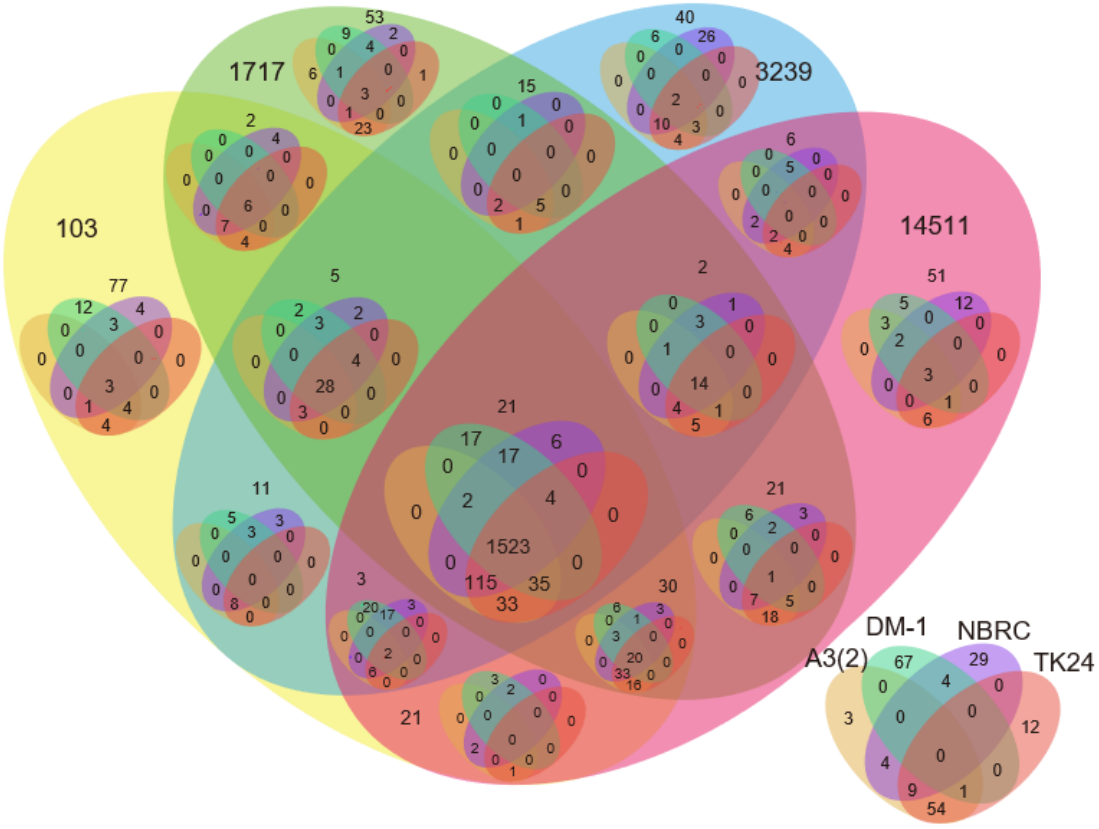

Figure 40: Diagramme de Venn représentant les sous-systèmes (Annotation RAST) - organisés selon le rôle- en communs et uniques au sain du groupe *plant-related*. Les nombres représentent les gènes codants pour des sous-systèmes en commun entre les souches représentées. A3(2) : Streptomyces coeliocolor A3(2); DM-1 : *Kitasatospora aureofaciens* DM-1; NRBC : *Streptomyces griseus subsp. griseus* NBRC13350; TK24 : *Streptomyces lividans* TK24 ; 14511 : *Streptomyces griseochromogenes* ATCC14511; 3239 : *Streptomyces lavendulae subsp. lavendulae* CCM3239 ; 1717 : *Streptomyces hygroscopicus subsp. limoneus* KCTC1717 ; 103 : *Streptomyces lydicus* 103.

4. Visualisation condensée des génomes des souches retenues par BRIG

Le but de cette partie est de voir a quelle point les séquences des génomes des deux groupes de souches sont semblables, et ce visuellement garce à une illustration des résultats de Blast. Pour ce qui est du groupe en relation avec les plantes, il en ressort 5 régions de faible similarité (par rapport au génome de référence) (fissures entre les anneaux).

Un nombre si peu révèle une grande similarité séquentielle entre les génomes constituant les anneaux affichés. L'outil Brig permet ainsi de voir efficacement si un set de génome est ou pas homogène, et c'est le cas des génomes de ce groupe.

Pour ce qui est du groupe non-related to plants, celui-ci en raison de la diversité des habitats des souches sélectionnées, serait d'une certaine hétérogénéité au niveau séquences. En effet, les anneaux de l'illustration par Brig affichent plus de 10 zones de faibles similarités (par rapport au génome de référence).

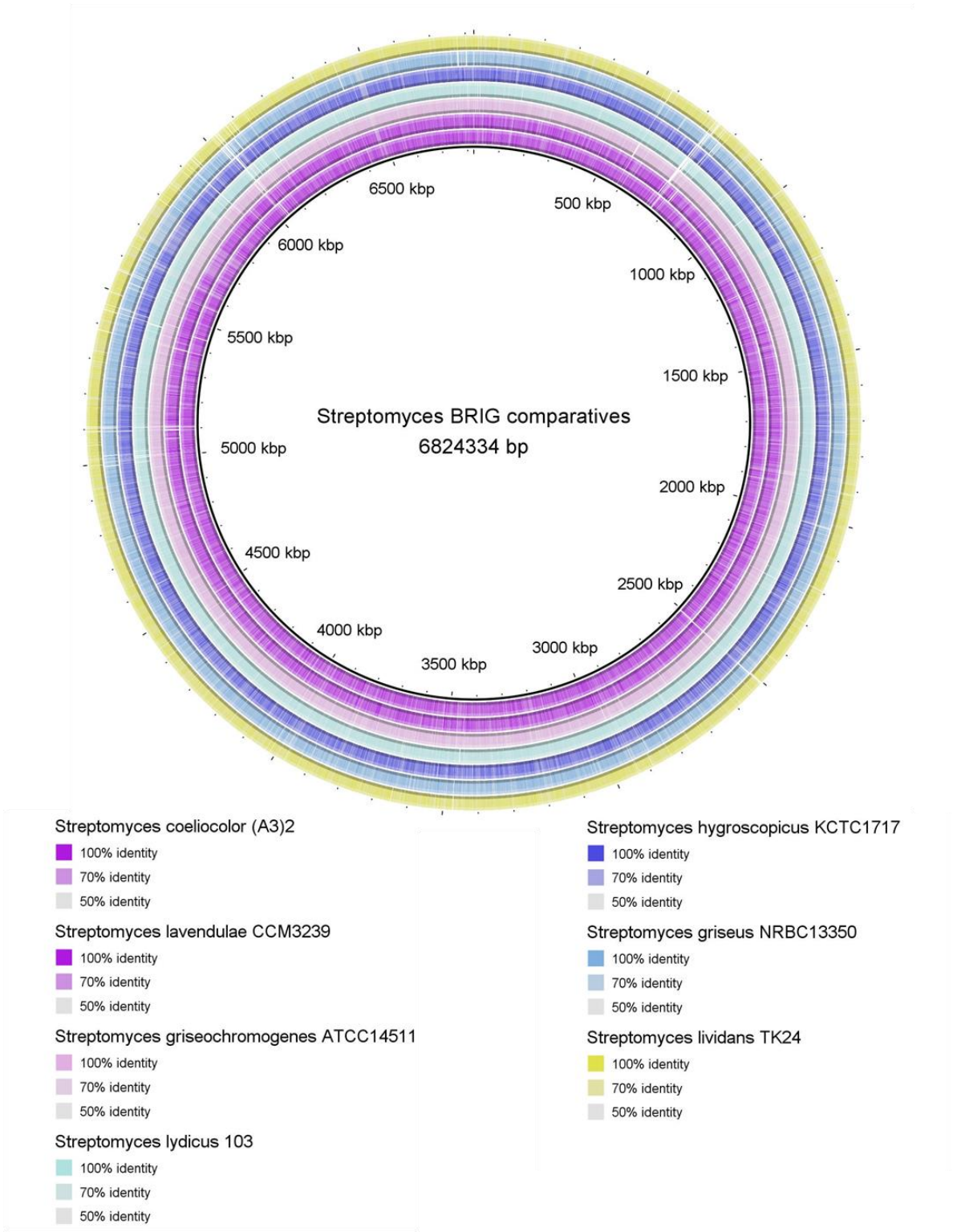

Figure 41: représentation condensée des génomes des souches du groupe « *plant-related* » par outil Brig. Au centre, en noir : le génome de la souche *K. aureofaciens DM-1*, au tour de ce dernier, les autres génomes des souches du même groupe.

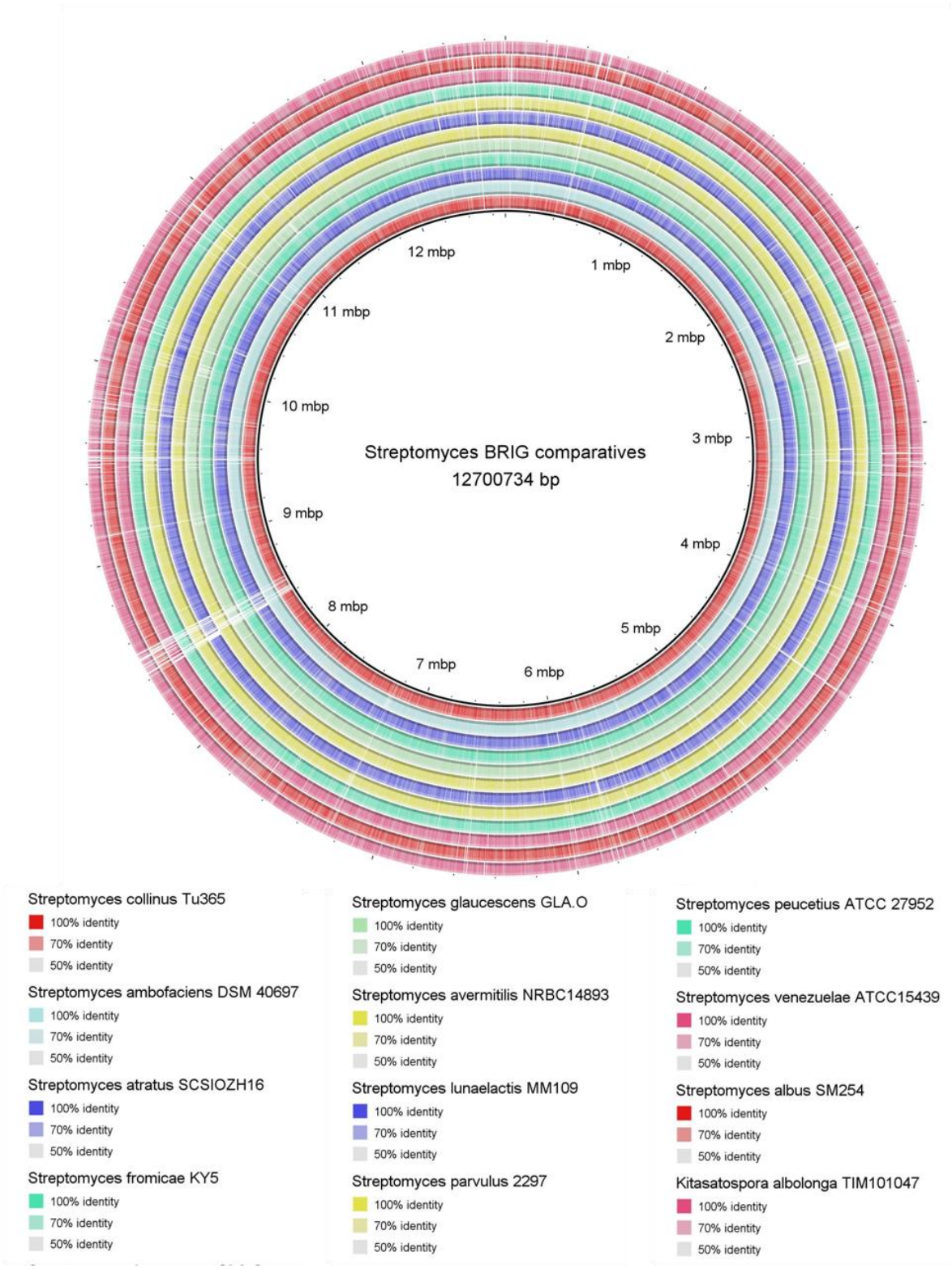

Figure 42: représentation condensée des génomes des souches du groupe « *Non related to plant* » par outil Brig. Au centre, en noir : le génome de la souche *S. rapamycinicus* NRRL 5491 , au tour de ce dernier, les autres génomes des souches du même groupe.

5. Détection des clusters de gènes de métabolites secondaire par outils antiSMASH

Cette analyse des clusters de gènes de métabolites secondaires laissa entrevoir clairement la richesse du métabolisme secondaire des *Streptomyces*. En contraste avec celui d'*E. coli*, qui n'affiche que 25 clusters (figure 43) pour la totalité de son génome, celui des *Streptomyces* va de 63 (*S. glaucescens*) à 144 (*S. griseochromogenes*).

En effet, une telle richesse explique aussi leur versatilité : pour autant de potentialités génomiques, on leur connait des traits aussi nombreux que diversifiés, de la bioremédiation, à la production pigments en passant par la promotion de la croissance des plantes.

Dans ce pool, il a été possible de détecter des clusters de gènes :

Des acide gras

Des saccharides ;

Putatif ;

De bactériocine/ lassopeptide

D'association de métabolites secondaires ;

De terpène ;

De butyrolactone ;

De Lantipeptide ;

De Thiopeptide ;

De Siderophores ;

De pks de type 1, 2 et 3;

D'Indole ;

De mélanine ;

D' NRPS ;

D'Ectoine ;

Et de Hserlactone

Pour ce qui est de l'autre groupe, ce sont surtout les clusters de gènes en relation avec les antibiotiques qui sont les plus présents.

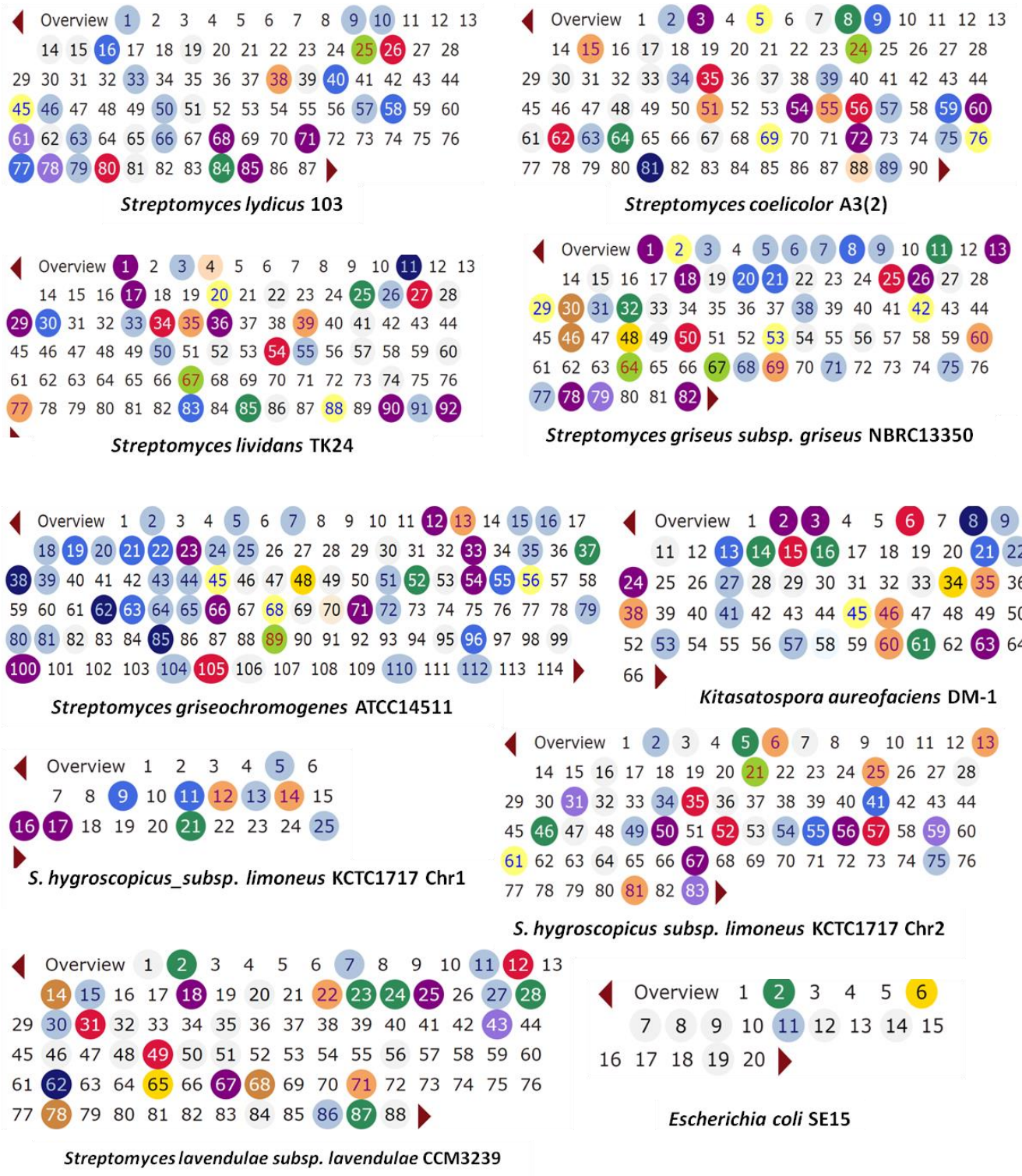

Figure 43: Comparaison entre els profiles de clusters de gènes de métabolites secondaires chez les souches du groupe « plant-related ». Légendes des couleurs : Gris : acide gras/ saccharide ; Blanc : cluster putatif ; Bleu : bactériocine/ lassopeptide ;Bleu pale : association de métabolites secondaires ; Bleu foncé : autre ; Violet : terpène ; Violet claire : butyrolactone ; Jaune claire: Lantipeptide ; Jaune : Thiopeptide ; Rouge : Siderophore ; Orange : Type 1 pks , Type 2 pks, Type 3 pks ; Orange claire : Indole ; Marron : mélanine ; Vert : NRPS ; Vert claire : Ectoine ; Caramel : Hserlactone

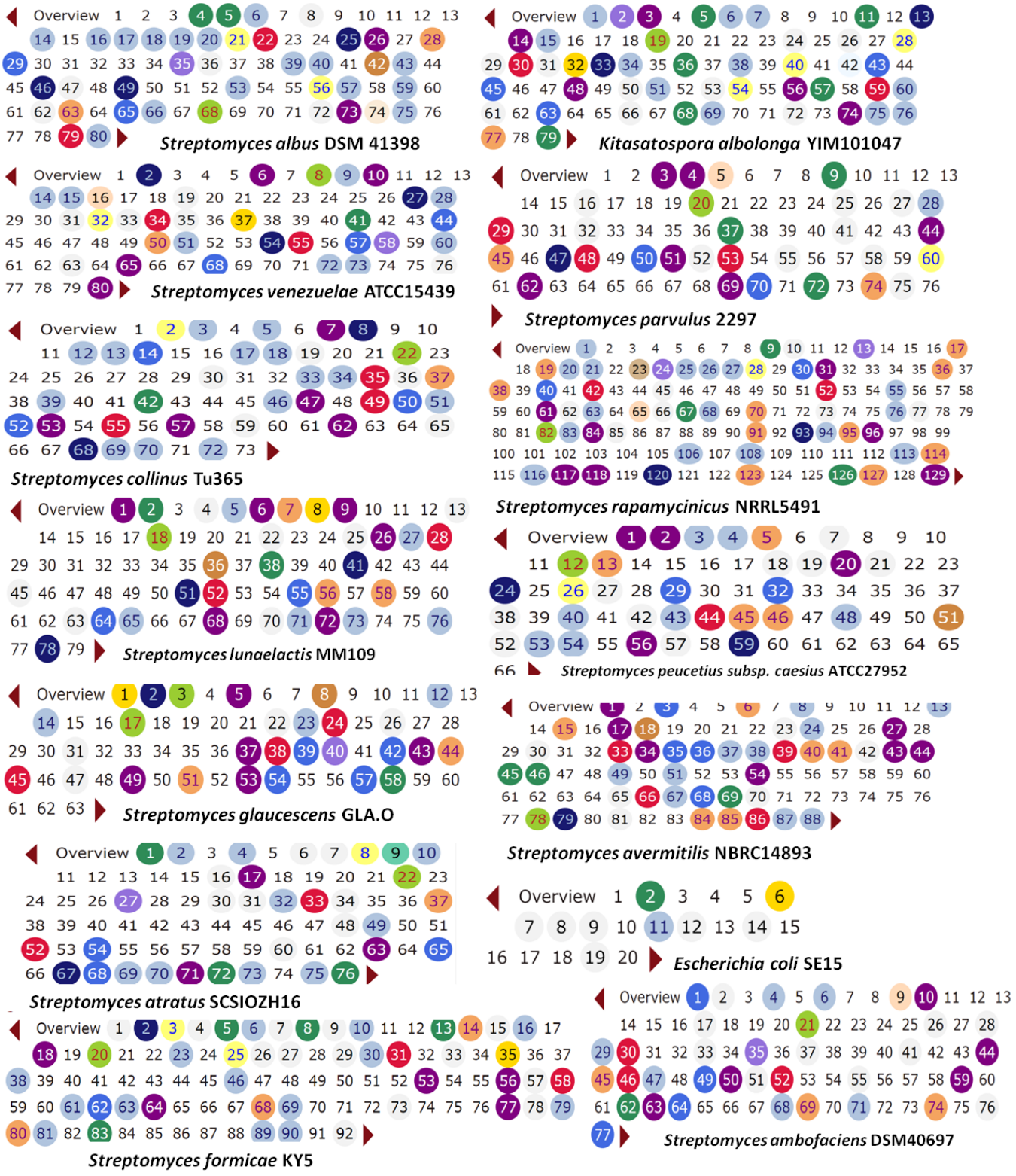

Figure 44: Comparaison entre els profiles de clusters de gènes de métabolites secondaires chez les souches du groupe « *Not related to plant*». Légendes des couleurs : Gris : acide gras/ saccharide ; Blanc : cluster putatif ; Bleu : bactériocine/ lassopeptide ;Bleu pale : association de métabolites secondaires ; Bleu foncé : autre ; Violet : terpène ; Violet claire : butyrolactone ; Jaune claire: Lantipeptide ; Jaune : Thiopeptide ; Rouge : Siderophore ; Orange : Type 1 pks , Type 2 pks, Type 3 pks ; Orange claire : Indole ; Marron : mélanine ; Vert : NRPS ; Vert claire : Ectoine ; Caramel : Hserlactone

Conclusion

Lors de cette étude, il a été possible de combiner moult approches et outils de bioinformatiques afin de faire une analyse comparative entre un set de 21 souches appartenant au genre *Streptomyces,* en les divisant en deux groupe : l'un en relation avec les plantes, l'autre d'origines diverses non liées aux plantes, avec des génomes allant de celui de la souche *Kitasatospora aureofaciens* DM1 (6,8 Mb), à celui de la souche *Streptomyces rapamycinicus* NRRL5491 (12,7 Mb).

Par ailleurs l'analyse phylogénétique en se basant sur les gènes de l'ARNr 16S à permit de confirmer que ce gène été insuffisant pour assurer une bonne résolution du groupe en question, et ce par les trois approches investiguées (copies sous représentées, copies sur représentées, séquences consensus).

En outre, l'approche par FFP qui ne nécessite pas d'alignement, a aboutit à la meilleure représentation taxonomique des espèces des souches étudiées, en comparaison avec celle du gène de l'ARNr16 seul.

Par ailleurs, l'approche par UBCG a faillit à décrire avec précision les différents taxa et leur relations évolutives.

L'analyse comparative des outputs de l'annotation à permit de mettre en exergue un pool commun de 3133 séquences codantes partagé par les souches du groupe en relation avec les plantes, ainsi que 751 protéines et un grand nombre de sous-systèmes (1523).

L'outil Brig a aboutit à la visualisation d'une grande similarité séquentielle entre les souches du groupe en relation avec les plantes, alors que celui des souches ayant des modes de vies différents, a démontré une similarité moindre entre les souches de ce groupe, si comparé au premier groupe.

L'outils antiSMASH a révélé la richesse du patrimoine de métabolites secondaires chez les *Streptomyces*, avec la détection de clusters de gènes d'acide gras/saccharides ; métabolites putatifs ; de bactériocine/ lassopeptide ; d'association de métabolites secondaires ; de terpène ; de butyrolactone ; de lantipeptide ; de thiopeptide ; de siderophores ; de pks de type 1, 2 et 3; d'Indole ; de mélanine ; d' NRPS ; d'Ectoine ; et de Hserlactone.

L'ensemble de ces résultats constitue une très bonne base pour des études poussées quant aux particularités de chaque souche séparément, ou en *comparaison croisée.*

Il serait idéal de prendre les listes générées par les analyses comparatives et de les étudier amplement pour leur contenu en gènes qui seraient liés à des mécanismes PGPR inconnue ou encore mal connu.

Références Bibliographiques :

- Aburjaile, F.F., Santana, M.P., Viana, M.V.C., Silva, W.M., Folador, E.L., Silva, A., Azevedo, V.,(2014). In tech:Genomics, SMGroup. Bioinformatics 2*0,* 170-179.
- Bais H.P., Weir.T.L., Perry L.G., Gilroy S et Vivanco J.M (2006). The role of root exudates inrhizosphere interactions with plants and other organisms., Annu., Rev., Plnt Biol.,57: 233 266.
- Bally. R, Elmerich. C (2007). Biocontrol of plant diseases by associative and endophyticnitrogen-fixing bacteria., In: C. Elmerich, W.E. Newton (eds). Associative and Endophytic Nitrogen-Fixing Bacteria and Cyanobacterial Associations. Springer., 171-190
- Beauchamp, C. J (1993). Mode of action of plant growth-promoting rhizobacteria and their potential use as biological control agents.,Phytoprotection.,71:19-27
- Benmati.M ,(2014). PGPR, paranodules, stimulation de la croissance et tolérance au déficit hydrique chez le blé dur (Triticum durum Desf.) : Aspects moléculaires et génétiques ., Thèse de Doctorat ,UNV . Constantine., Faculté des sciences de Nature et de la vie .Département de Biologie végétale et d'écologie ,Algeria,185pge.
- Bentley SD, Chater KF, Cerdeño-Tárraga AM, 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 CH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream MA, 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(6885):141
- Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., *et al.*

(2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project., Nature *447*,799-816.

- Bryant J, Chewapreecha C, Bentley SD(2012).Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future Microbiol*,7 2:83-96.

- Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., Blake, J. A., et Group, T. M. G. D. (2008). The Mouse Genome Database (MGD): mouse biology and model systems. NucleicAcidsRes. *36*,D724–D728.
- Cassiana S. De Sousa; Syed S. H, Anne C. Pinto, Wanderson M.S, SintiaS.A;Siomar. S,Marcela S.P. Azevedo;Clarissa S. R, Debmalya.B; Vasco. A.,(2018) ,Microbial Omics: Applications in Biotechnology., DansDebmalya.B et Vasco Azevedo., *Omics Technologies and Bio-Engineering.;.* (vol(2)., 3_20) Horizonte, Minas Gerais, Brazil.
  Darmon E, Leach DR., (2014).Bacterial genome instability. *Microbiol Mol Biol Rev*,78:1-39.
- Dickmeis, T., et Müller, F. (2005). The identification and functional characterisation of conserved regulatory elements in developmental genes., Brief Funct Genomic Proteomic *3*, 332-350.
- Dobbelaere S, Vanderleyden J, Okon Y (2003). Plant growth promoting effects of diazotrophs in the rhizosphere.,Plant Sci., 22:107-149.
- Drago H.,2015-Métabolisme secondaire de *Streptomuces ambofaciens :* Explorationgénomique et étude de groupe de gènes dirige la synthèse sphydrophurane. Thèse de doctorat, Uni. Pari sud ,193p
- Durães Sette L, Mendonça Alves Da Costa LA, Marsaioli AJ, Manfio GP. (2004). Biodegradation of alachlor by soil streptomycetes. Appl Microbiol Biotechnol. 64(5):712-7.
- Elsliger. M-A., Wilson. IA (2013).structruralgenomic DansS. Maloy.,K.HughesBrenner's.,Encyclopedia of Genetics,2Ed,vol(6),575_579.

- Freireb DM, ASoaresa RM, FLeitec SG ,RCoelhoa RR(2003).Production and partial characterization of thermophilic proteases from Streptomyces sp. isolated from Brazilian cerrado soi
- **Fricke. F, Cebula.T , Ravela .J(2011).Genomics** Dans, BudowleB., Schutzer E.S, RogerG. Breeze. Morse A.S., KeimP.S,.**Microbial Forensics,(2éd, 479_492)** Fort Worth, Texas **.**
- Friedberg,I. (2006). Automated protein function prediction, the genomic challenge., Brief. Bioinformatics *7*, 225- 242.BiologyAltschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs., Nucleic Acids Res *25*, 3389-3402. G.E. Moore, (1965).Cramming more components onto integrated circuits, Electron. Mag. 38 p 4–7
- Gage D.J (2004) Infection and invasion of roots by symbiotic, nitrogen-fixing rhizobia during nodulation of temperate legumes. Microbiol Mol.,Biol.Rev., 68 (2): 280-300.
- Gagniere.N,(2009).Développement d'une suite logicielle pour l'analyse et l'annotation intégrative automatiques de transcrits et de protéines., Application aux banques d'ADNc de l'annélide polychète *Alvinellapompejana* .,Thèse Docteura, Institut de Génétique et de Biologie Moléculaire et Cellulaire ., Université de Strasbourg,page222 .
- Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A. H., Coudert, E., Lima, T., Kersey, P., Pagni, M., Sigrist, C. J. A., Lachaize, C., *et al.* (2003).Automated annotation of microbial proteomes in SWISS-PROT., ComputBiolChem*27*, 49-58.
- Gilchrist, D. A., Fargo, D. C., et Adelman, K. (2009). Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation., Methods *48*,398-408.
- Glick B.R (1995). The enhancement of plant growth by free-living bacteria. Can.J. Microbiol.,41:109-117.

- Gray and L.M. Smith (2005). Influence of land use on postmetamorphic body size of play alake amphibians. Journal of Wildlife Management., 69:515-524.
- Hardison RC.(2003).Comparative genomics. *PLoS Biol*. Nov;1(2):E58. PMID: 14624258
- Herman MAB, Nault BA, Smart CD (2008). Effects of plant growthpromoting rhizobacteria on bell pepper production and green peach aphid infestations in New York. Crop. Protect. 27: 996-1002.Hope SM, Li CY (1997). Respiration, nitrogen fixation., and mineralizable nitrogen spatial and temporal patterns within two Oregon Douglas-fir stands. Can. J., Res., 27:501-509.
- Herrero J., Muffato M. , Beal K., Fitzgerald., Gordon L., Pignatelli M L., Vilella A J.,Searle . S. M. J., Amode R. (2015). Ensembl comparative genomics resources. *Database* update.2016 :2-3.
- Huguet, V., Land, E.O., Casanova, J.G., Zimpfer, J.F., and Fernandez, M.P (2005). Geneticdiversity of Frankia microsymbionts from the relict species Myrica faya (Ait.) and Myrica rivas-martinezii (S.) in Canary Islands and Hawaii. Microb., Ecol., 49: 617–625.
- Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Omura S. (2003). Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. Nat Biotechnol. 21(5):526-31.
- Kaushik.S.,Kaushik.S.,Sharma.D(2018).functional genomic. DansS.Ranganathan., Gribskov.M.,Nakai.K.,Schönbach.C.,EncyclopediaofBioinformatics and Computational Biology.;vol(2):118 _133.
- Kendrick KE, Ensign JC. (1983). Sporulation of *Streptomyces griseus* in submerged culture. J Bacteriol. 155(1):357-66.
- Kloepper JW (1993). Plant growth-promoting rhizobacteria as biological control agents. In: Metting FB Jr., (ed) Soil microbial ecologyapplication in agricultural and environ mental management. Marcel,Dekker., Inc. New York., 255-274.

- Kroiss J, Kaltenpoth M, Schneider B, Schwinger MG, Hertweck C, Maddula RK, Strohm E, Svatos A. (2010). Symbiotic Streptomycetes provide antibiotic combination prophylaxis for wasp offspring. Nat Chem Biol. 6(4):261-3.
- Labeda, D. P., Dunlap, C. A., Rong, X., Huang, Y., Doroghazi, J. R., JU, K. S. and Metcal, W. W(2017). Phylogenetic relationships in the family *Streptomycetaceae* using multi-locus sequence analysis. *Antonie van Leeuwenhoek*, 110, 563-583.
- Land M., Hauser, L., Jun S.R., Nookaew, I., Leuze M.R., Ahn, T.-H., Karpinets T., Lund, O., Kora G., Wassenaar, T(2015).Insights from 20 years of bacterial genome sequencing, *Funct. Integ*r. *Genomics* 15,
- Long S.R (1996). Rhizobium symbiosis: Nod factors in perspective., Plant Cell 8, 1885–1898.
- Macking.H (2007). Phytoremediation of contamined soil on plant efficiency, rhizosphere bacteria and the physical effects of chemical agents. Korea society for Applied Microbiology and Biotechnology., 35: 26-271.
- Marcotte, C. J. V., et Marcotte, E. M. (2002). Predicting functional linkages from gene fusions with confidence.,Appl., Bioinformatics *1*, 93-100.
- Medini D., Donati C., Tettelin H., Masignani V., and Rappuoli R (2005).The microbial pan-genome," Current opinion in genetics & development, vol. 15, no. 6, pp. 589–94.
- Morgan, J.A.W., Bending, G. D. et White, P. J (2005). Biological costs and benefits to plantmicrobe interactions in the rhizosphere. Journal of Experimental Botany., 56: 1729–1739.
- Muller C, Denis M, Gentzbittel L, Faraut T, (2004).The Iccare web server: an attempt to merge sequence and mapping information for plant and animal species. Nucleic Acids Res. Jul 1;32(Web Server issue):W429-34. PMID: 15215424.
- Naome A., Maciejewska M., Calusinska M., Martinet L., Anderssen S., Adam D., Tenconi E., Deflandre B., Coppieters W., Karim L., Hanikenne M., Baurain D., Delfosse P., van Wezel G.P., and Rigali S.,(2018).Complete genome sequence of

Streptomyces lunaelactis MM109T, isolated from cave moonmilk deposits." Genome Announc. 6:e00435-18.

- Ohnishi Y, Ishikawa J, Hara H, Suzuki H, Ikenoya M, Ikeda H, Yamashita A, Hattori M, Horinouchi S. (2008). Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. J Bacteriol. 190(11):4050-60.
- Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., etMaltsev, N. (1999). The use of gene clusters to infer functional coupling., Proc. Natl. Acad. Sci., U.S.A 9*6*, 2896-2901.
- Pearson, H. (2006). Genetics: What is a gene? Nature 4*41*, 398-401.
- R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, et al., (1995).Whole-genome random sequencing and assembly of *Haemophilusinfluenzae*Rd, Science 269.,p496–512.

- Ramette, A., Moënne-Loccoz, Y., and Défago, G (2006). Genetic diversity and biocontrolpotential of fluorescent pseudomonads producing phloroglucinols and hydrogen cyanide from Swiss soils naturally suppressive or conducive to Thielaviopsis basicola-mediated black root rot of tobacco. FEMS Microbiol. Ecol. 55: 369–381. ; Rezzonico, F., Zala, M., Keel, C., Duffy, B., Moënne-Loccoz, Y. and Défago, G (2007). Is the ability of biocontrol fluorescent pseudomonads to produce the antifungal metabolite 2,4-diacetylphloroglucinol really synonymous with higher plant protection? New Phytologist.,173: 861- 872.
- Ramos-Solano B, Barriuso-Maicas J, Gutierrez-Mañero J (2009). Biotechnology of the Rhizosphere. In: Kirakosyan A, Kaufman PB (eds.) Recent Advances in Plant Biotechnology. 137, Springer Science & Business Media., pp. 137-162.
- Ranea, J. A. G., Yeats, C., Grant, A., etOrengo, C. A. (2007). Predicting protein function with hierarchical phylogenetic profiles: the Gene3D Phylo-Tuner method applied to eukaryotic genomes.,PLoSComput., Biol 3, e237
- Raphaël H(2010).Développement en phylogénomique :Comparaison de génomes et Estimation de grande phylogénie. Thèse de doctorat., Université Namur Belgique, 115p.

- Schuster S (2008).Next-generation sequencing transforms today ’s biology,” *Nature Methods*, vol. 5, no. 1, pp. 16–18.
- Seipke RF, Barke J, Brearley C, Hill L, Yu DW, Goss RJ, Hutchings MI. (2011). A single *Streptomyces* symbiont makes multiple antifungals to support the fungus farming ant Acromyrmex octospinosus. PLoS One. 6(8): e22028.
- Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., Richter, A. R., et White,O. (2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes., Nucleic Acids Res. 3*5*, D260–D264.
- Setubal J. C., Stoye J., Stadler P.F (2018). Comparative Genomics :Methods and Protocols,Ed. Spring St, New York,473p.
- Shimokawa, K., Okamura-Oho, Y., Kurita, T., Frith, M. C., Kawai, J., Carninci, P., etHayashizaki, Y. (2007). Large- scale clustering of CAGE tag expression data.BMC Bioinformatics 8, 161.
- Shokralla S, Spall JL, Gibson JF, Hajibabae IM(2012).Next generation sequencing technologies for environmental DNA research. Mol Ecol. 21(8), 794-805.
- Sjölander, K. (2004). Phylogenomic inference of protein molecular function: advances and challenges.,
- Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, van Sinderen D. (2007). Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. Microbiol Mol Biol Rev. 71(3):495-548.
- Verma, J.P., J. Yadav and K.N. Tiwari, (2010). Application of Rhizobium sp. BHURC01 and plant growth promoting rhizobactria on nodulation. plant biomass and yields of Chickpea (Cicer arietinum L.). Int. J. Agric. Res., 5: 148-156.
- Vessey J K (2003). Plant growth promoting rhizobacteria as biofertilizers. Plant Soil .255:571-586.
- Woese C.R. ;(1987).Bacterial evolution. *Microbiol*.512:211-271.
- Xia X(2013).Comparative Genomics, Ed.Springer Heidelberg, New York Dordrecht London, 67p.
- **Liens:**

- www.explorecuriocite.org © Parlons, sciences 2013.
- https://explorecuriocite.org/Portals/2/Themes/Biotechnology/BLM%206-Backgrounder-
- What%20is%20Genomics%20NEWFR.pdf
- "Illumina Solexa Sequencing Overview." [Online]. Available: http://www.youtube.com/watch?v=77r5p8IBwJk.
- Roche, "454 Sequencing Systems Technology Overview." [Online]. Available: http://454.com/resources-support/product-videos.as.
- .https://doi.org/10.1016/j.enzmictec.2003.11.015 science direct

- http://www.youtube.com/watch?v=77r5p8IBwJk.

- https://explorecuriocite.org/Portals/2/Themes/Biotechnology/BLM%206-Backgrounder-What%20is%20Genomics%20NEWFR.pdf.

Annexe A

Description de quelques outils utilisés

NCBI

Le *National Center for Biotechnology Information* (NCBI, « Centre américain pour les informations biotechnologiques ») est un institut national américain pour l'information biologique moléculaire. Ce serveur fourni un set d'outils bioinformatique très complet et versatile accessible online, dont le plus connu de tous : l'outil BLAST. https://www.ncbi.nlm.nih.gov

BLAST

BLAST (acronyme de *basic local alignment search tool*) est une méthode de recherche heuristique utilisée en bio-informatique. Il permet de trouver les régions similaires entre deux ou plusieurs séquences de nucléotides ou d'acides aminés, et de réaliser un alignement de ces régions homologues.

https://**blast**.ncbi.nlm.nih.gov/**Blast**.cgi

RAST

RAST **(***Rapid Annotations using Subsystems Technology*) est un service entièrement automatisé pour l'annotation des génomes bactériens et archéens. Il identifie les séquences codant des protéines, les gènes d'ARNr et ARNt. Il attribue également des fonctions aux gènes et prédit les sous-systèmes représentés dans le génome et utilise ces informations pour reconstruire le réseau métabolique.

http://rast.nmpdr.org/

AntiSMASH

AntiSMASH est un serveur web et un logiciel autonome permettant de prédire les clusters de gènes de métabolites secondaires dans les génomes bactériens. http://antismash.secondarymetabolites.org.

BioEdit

BioEdit est un éditeur d'alignement des séquences biologiques sous environnement Windows. Une interface intuitive de documents multiples avec des fonctions pratiques rend l'alignement et la manipulation des séquences relativement facile sur votre ordinateur de bureau.

https://perl.developpez.com/telecharger/detail/id/1909/BioEdit

MEGA

Molecular Evolutionary Genetics Analysis (MEGA) est un logiciel permettant d'effectuer une analyse statistique de l'évolution moléculaire et de construire des arbres phylogénétiques (lien : https://www.megasoftware.net/). Au cours de notre étude nous avons utilisé le MEGA version 7.

Annexe B

Liste des gènes pris en considération par l'analyse UBCG

Gene	COG catégorie	COG ID	profile HMM	Fonction
alaS	J	COG0013	TIGR00344	Alanine-tRNA ligase
argS	J	COG0018	TIGR00456	Arginine-tRNA ligase
aspS	J	COG0173	TIGR00459	Aspartate-tRNA ligase
cgtA	DL	COG0536	TIGR02729/PF01018	GTPase ObgE/CgtA
coaE	H	COG0237	TIGR00152	Dephospho-CoA kinase
cysS	J	COG0215	TIGR00435	Cysteine-tRNA ligase
dnaA	L	COG0593	TIGR00362/PF00308	Chromosomal replication initiator protein DnaA
dnaG	L	COG0358	TIGR01391	DNA primase
dnaX	L	COG2812	TIGR02397	DNA polymerase III subunit gamma
engA	R	COG1160	TIGR03594	GTPase Der
ffh	U	COG0541	TIGR00959	Signal recognition particle protein
fmt	J	COG0223	TIGR00460	Methionyl-tRNA formyltransferase
frr	J	COG0233	TIGR00496	Ribosome-recycling factor
ftsY	U	COG0552	TIGR00064	Signal recognition particle receptor FtsY
gmk	F	COG0194	TIGR03263/PF00625	Guanylate kinase
hisS	J	COG0124	TIGR00442	Histidine-tRNA ligase
ileS	J	COG0060	TIGR00392	Isoleucine-tRNA ligase 1
infB	J	COG0532	TIGR00487	Translation initiation factor IF-2
infC	J	COG0290	TIGR00168	Translation initiation factor IF-3
ksgA	J	COG0030	TIGR00755	Ribosomal RNA small subunit methyltransferase A
lepA	J	COG0481	TIGR01393	Elongation factor 4
leuS	J	COG0495	TIGR00396	Leucine-tRNA ligase
ligA	L	COG0272	TIGR00575	DNA ligase
nusA	K	COG0195	TIGR01953	Transcription termination/antitermination protein NusA

nusG	K	COG0250	TIGR00922	Transcription termination/antitermination protein NusG
pgk	G	COG0126	PF00162	Phosphoglycerate kinase
pheS	J	COG0016	TIGR00468	Phenylalanine-tRNA ligase alpha subunit
pheT	J	COG0073	TIGR00472	Phenylalanine-tRNA ligase beta subunit
prfA	J	COG0216	TIGR00019	Peptide chain release factor 1
pyrG	F	COG0504	TIGR00337	CTP synthase
recA	L	COG0468	TIGR02012	DNA recombination and repair protein
rbfA	J	COG0858	TIGR00082	30S ribosome-binding factor
rnc	K	COG0571	TIGR02191	Ribonuclease 3
rplA	J	COG0081	TIGR01169	50S ribosomal protein L1
rplB	J	COG0090	TIGR01171	50S ribosomal protein L2
rplC	J	COG0087	TIGR03625/PF00297	50S ribosomal protein L3
rplD	J	COG0088	TIGR03953	50S ribosomal protein L4
rplE	J	COG0094	PF00281	50S ribosomal protein L5
rplF	J	COG0097	TIGR03654/PF00347	50S ribosomal protein L6
rplI	J	COG0359	TIGR00158/PF01281	50S ribosomal protein L9
rplJ	J	COG0244	PF00466	50S ribosomal protein L10
rplK	J	COG0080	TIGR01632	50S ribosomal protein L11
rplL	J	COG0222	TIGR00855	50S ribosomal protein L7/L12
rplM	J	COG0102	TIGR01066	50S ribosomal protein L13
rplN	J	COG0093	TIGR01067	50S ribosomal protein L14
rplO	J	COG0200	TIGR01071	50S ribosomal protein L15
rplP	J	COG0197	TIGR01164	50S ribosomal protein L16
rplQ	J	COG0203	TIGR00059	50S ribosomal protein L17
rplR	J	COG0256	TIGR00060	50S ribosomal protein L18
rplS	J	COG0335	TIGR01024	50S ribosomal protein L19
rplT	J	COG0292	TIGR01032	50S ribosomal protein L20
rplU	J	COG0261	TIGR00061	50S ribosomal protein L21

rplV	J	COG0091	TIGR01044	50S ribosomal protein L22
rplW	J	COG0089	PF00276	50S ribosomal protein L23
rplX	J	COG0198	TIGR01079	50S ribosomal protein L24
rpmA	J	COG0211	TIGR00062	50S ribosomal protein L27
rpmC	J	COG0255	TIGR00012	50S ribosomal protein L29
rpmI	J	COG0291	TIGR00001	50S ribosomal protein L35
rpoA	K	COG0202	TIGR02027	DNA-directed RNA polymerase subunit alpha
rpoB	K	COG0085	TIGR02013	DNA-directed RNA polymerase subunit beta
rpoC	K	COG0086	TIGR02386	DNA-directed RNA polymerase subunit beta’
rpsB	J	COG0052	TIGR01011	30S ribosomal protein S2
rpsC	J	COG0092	TIGR01009	30S ribosomal protein S3
rpsD	J	COG0522	TIGR01017	30S ribosomal protein S4
rpsE	J	COG0098	TIGR01021	30S ribosomal protein S5
rpsF	J	COG0360	TIGR00166/PF01250	30S ribosomal protein S6
rpsG	J	COG0049	TIGR01029	30S ribosomal protein S7
rpsH	J	COG0096	PF00410	30S ribosomal protein S8
rpsI	J	COG0103	PF00380	30S ribosomal protein S9
rpsJ	J	COG0051	TIGR01049	30S ribosomal protein S10
rpsK	J	COG0100	TIGR03632	30S ribosomal protein S11
rpsL	J	COG0048	TIGR00981	30S ribosomal protein S12
rpsM	J	COG0099	TIGR03631	30S ribosomal protein S13
rpsO	J	COG0184	TIGR00952	30S ribosomal protein S15
rpsP	J	COG0228	TIGR00002	30S ribosomal protein S16
rpsQ	J	COG0186	TIGR03635	30S ribosomal protein S17
rpsR	J	COG0238	TIGR00165	30S ribosomal protein S18
rpsS	J	COG0185	TIGR01050	30S ribosomal protein S19
rpsT	J	COG0268	TIGR00029	30S ribosomal protein S20
secA	U	COG0653	TIGR00963	Protein translocase subunit SecA
secG	U	COG1314	TIGR00810	Protein-export membrane protein SecG

secY	U	COG0201	TIGR00967	Protein translocase subunit SecY
serS	J	COG0172	TIGR00414	Serine-tRNA ligase
smpB	O	COG0691	TIGR00086	SsrA-binding protein
tig	O	COG0544	TIGR00115	Trigger factor
tilS	J	COG0037	TIGR02432	tRNA(Ile)-lysidine synthase
truB	J	COG0130	TIGR00431	tRNA pseudouridine synthase B
tsaD	J	COG0533	TIGR03723	tRNA N6-adenosine threonylcarbamoyltransferase
tsf	J	COG0264	TIGR00116/PF00889	Elongation factor Ts
uvrB	L	COG0556	TIGR00631	UvrABC system protein B
ybeY	J	COG0319	TIGR00043	Endoribonuclease YbeY
ychF	J	COG0012	TIGR00092	Ribosome-binding ATPase YchF

Résumé

L'objectif principale de ce travail consiste à analyser les génomes de souches candidates (un set de 21 souches) ou PGPR confirmés appartenant au genre *Streptomyces* , dans le but d'extraire des données PGPR. La démarche scientifique abordée lors de la réalisation s'articule autour du choix des souches (screening de *Streptomyces* spp. en relation avec les plantes), leur annotation, suivi par une étude phylogénétique et phylogénomique des souches examinées par différentes approches, puis la comparaison des profiles génétique de métabolisme. L'analyse phylogénétique en se basant sur les gènes de l'ARNr 16S à permit de confirmer que ce gène été insuffisant pour assurer une bonne résolution du groupe en question. L'approche par FFP qui ne nécessite pas d'alignement, a aboutit à la meilleure représentation taxonomique des espèces des souches étudiées. L'approche par UBCG a faillit à décrire avec précision les différents taxa et leur relations évolutives. L'analyse comparative des tables de l'annotation à permit de mettre en évidence un pool commun de 3133 séquences codantes partagé par les souches du groupe en relation avec les plantes, ainsi que 751 protéines et un grand nombre de sous-systèmes (1523). L'outil Brig a aboutit à la visualisation d'une grande similarité séquentielle entre les souches du groupe en relation avec les plantes. L'outil online antiSMASH a révélé la richesse du génomes en gènes de métabolites secondaires chez les *Streptomyces*, avec la détection de clusters de gènes d'acide gras/saccharides ; métabolites putatifs ; de bactériocine/ lassopeptide ; d'association de métabolites secondaires ; de terpène ; de butyrolactone ; de lantipeptide ; de thiopeptide ; de siderophores ; de pks de type 1, 2 et 3; d'Indole ; de mélanine ; d' NRPS ; d'Ectoine ; et de Hserlactone.

Mots clés : Génomique, analyse comparative, *Streptomyces* spp., phylogénomie, annotation, métabolisme secondaire.

Summary:

The main objective of this work is to analyze the genomes of candidate strains (a set of 21 strains) or confirmed PGPRs belonging to the genus Streptomyces, with the aim of extracting PGPR data. The scientific approach approached during the production is based on the choice of strains (screening of Streptomyces spp., In relation to the plants), their annotation, followed by a phylogenetic and phylogenomic study of the strains examined by different approaches, then the comparison. Genetic profiles of metabolism.Phylogenetic analysis based on 16S rRNA genes confirmed that this gene was insufficient to ensure good resolution of the group in question. The FFP approach, which does not require alignment, has resulted in the best taxonomic representation of the species of the strains studied. The UBCG approach has failed to accurately describe the different taxa and their evolutionary relationships. The comparative analysis of the annotation tables allowed highlighting a common pool of 3133 coding sequences shared by the strains of the group in relation to the plants, as well as 751 proteins and a large number of subsystems (1523). The Brig tool led to the visualization of a great sequential similarity between the strains of the group in relation to the plants. The antiSMASH online tool revealed the genome richness in secondary metabolite genes in Streptomyces, with the detection of fatty acid / saccharide gene clusters; putative metabolites; bacteriocin / lasso peptide; association of secondary metabolites; terpene; butyrolactone; antipeptide; thiopeptide; siderophores; pks type 1, 2 and 3; Indole; melanin; NRPS; d'Ectoine; and Hserlactone.

Key words: Genomics, comparative analysis, Streptomyces spp., Phylogeny, annotation, secondary metabolism.

ملخص:

الهدف الرئيسي من هذا العمل هو تحليل جينومات السلالات المرشحة (مجموعة مكونة من 21 سلالة) أو المؤكد أنها PGPR والمنتمية إلى جنس ال *Streptomyces* ، لغرض استخراج بيانات PGPR. المنطلق العلمي المنتهج أثناء التنفيذ هو الاعتماد على اختيار السلالات (غربلة *Streptomyces* spp المتعلقة بالنباتات)، شرحهم، تليها دراسة phylogenetique وphylogénomiqueللسلالات المدروسة بأساليب مختلفة ثم مقارنة مواصفات الجينات فيما يخص عملية التمثيل الغذائي .

سمح تحليل تطور الجينات المعتمد على جينات ARNr 16sأن هذا الجين غير كافٍ لضمان الحل الجيد للمجموعة المعنية .وقد أدت طريقة FFP التي لا تتطلب المحاذات، إلى أفضل تمثيل تصنيفي لأنواع السلالات التي تمت دراستها.بينما فشلتطريقة UBCGفي الوصف بدقة الأصناف المختلفة وعلاقاتهم التطورية

سمح التحليل المقارن لجداول الشروحات بتسليط الضوء على مجموعة مشتركة مكونة من 3133 تسلسل مُشفر تتقاسمها سلالات المجموعة المتعلقة بالنباتات ، بالإضافة إلى 751 بروتين وعدد كبير من الأنظمة الفرعية (1523).

أدت برنامج Brig إلى تبيان تشابه كبير متسلسل بين سلالات المجموعة المتعلقة بالنباتات.كشف برنامج antiSMASH عن ثراء الجينوم من جيناتات المسؤولة عن المستقلبات الثانوية في Streptomyces.مع الكشف عن مجموعات الجينات المسؤولة عن الأحماض الدهنية / السكريتيد؛ المستقلبات المفترضة bacteriocin / lassopeptide ؛ رابطة المستقلبات الثانوية ؛ تربين؛ بيوتيرولاكتون؛ lantipeptide؛ thiopeptide؛ حاملة الحديد؛ الأنواع 1, 2 و 3 للـ pks؛ الإندول؛ الميلانين؛ NRPS؛ ectoineو Hserlactone.

الكلمات المفتاحية: الجينوميات ، التحليل المقارن ، Streptomyces spp.، Phylogenyannotation ، الأيض الثانوي