L'exploration des génomes par l'outil ICEFinder révèle la forte prévalence et l'extrême diversité des ICE et des IME de streptocoques

Charles Coluzzi

Thèse présentée pour l'obtention du titre de
Docteur de l'Université de Lorraine
Spécialité « Ecotoxicologie, Biodiversité, Ecosystèmes »

Par **Charles COLUZZI**

L'exploration des génomes par l'outil ICEFinder révèle la forte prévalence et l'extrême diversité des ICE et des IME de streptocoques.

Genomic exploration using the ICEFinder tool reveals the strong predominance and extreme diversity of streptococcal ICEs and IMEs

Soutenance publique prévue le mercredi 20 décembre 2017

Composition du jury :

Rapporteurs :

Mr. Eduardo ROCHA, Directeur de recherche, Institut Pasteur
Mr. Benoit DOUBLET, Chargé de recherche (HDR), INRA

Examinateurs :

Mme. Hélène CHIAPELLO, Ingénieure de Recherche (HDR), INRA
Mme. Sophie PAYOT-LACROIX, Directeur de recherche, INRA
Mme. Nathalie LEBLOND-BOURGET, Professeur, Université de Lorraine (Co-Directrice de thèse)
Mr. Gérard GUEDON, Maître de conférences (HDR), Université de Lorraine (Co-directeur de thèse)

UMR1128 DynAMic,
INRA-Université de Lorraine, Faculté des Sciences et Technologies
Boulevard des Aiguillettes, BP 70239, 54506 Vandœuvre-lès-Nancy cedex.

REMERCIEMENTS

D’abord, je voudrais remercier l’ensemble des membres du jury d’avoir accepté de juger mon travail. Je remercie le Dr. Eduardo Rocha et le Dr. Benoit Doublet d’avoir accepté d’être les rapporteurs de mon manuscrit de thèse. Je remercie les Dr. Sophie Payot-lacroix et Dr. Hélène Chiapello d’avoir accepté d’examiner ce manuscrit.

J’adresse mes sincères remerciements au Pr. Pierre Leblond pour m’avoir accueilli au sein de son laboratoire et de m’avoir donné l’opportunité de réaliser cette thèse.

Je tiens particulièrement à remercier ma directrice de thèse, le Pr. Nathalie Leblond pour sa gentillesse, sa disponibilité, son encadrement ainsi que pour la grande confiance qu’elle m’a accordée au cours de ces trois dernières années.

Je remercie tout particulièrement, mon co-directeur de thèse, le Dr. Gérard Guédon avec qui j’ai eu la chance de partager un bureau au cours de ces années de thèse. Je te remercie pour le savoir et les connaissances que tu m’as transmis durant toutes ces heures passées à tes cotés. Merci également pour ton soutient inébranlable, ta disponibilité et ton enthousiasme.

Je remercie le Dr. Marie-Dominique Devignes pour son acceuil au sein de son laboratoire ainsi que pour ses conseils et nombreuses idées.

Un merci particulier, à mes « co-doctorants » avec qui j’ai traversé cette aventure. Un grand merci à *Nari, Greg, Maxime et Razak* pour toutes ces discussions et moments partagés.

Merci également à mes amis Nancéens d’abord, Mat et Petit frère pour avoir été présents à tous moments, jours et nuits. Merci aussi à mes amis Villeruptiens et Longoviciens pour avoir été là à chacun de mes retours sur Villerupt.

Enfin, à mes parents, à qui j’exprime toute ma gratitude pour leur soutient infaillible, tant financier que moral. Je tiens à vous remercier pour votre confiance ainsi que pour tous les sacrifices que vous avez faits au cours de ces 8 années afin de me permettre de réaliser cette thèse.

.

LISTE DES ABREVIATIONS :

att : site d'attachement (**att**achment site)

attB : site d'attachement bactérien (**B**acterial **att**achment site)

attI : site d'attachement de l'ICE (**I**CE **att**achment site)

attL : site d'attachement gauche (**L**eft **att**achment site)

attP : site d'attachement du prophage (**P**hage **att**achment site)

attR : site d'attachement droit (**R**ight **att**achment site)

CIME : élément mobilisable en *cis* (**ci**s-**M**obilizable **E**lement)

CP : protéine de couplage (**C**oupling **P**rotein)

HMM : Modèle de Markov caché (**H**idden **M**arkov **M**odel)

ICE : élément intégratif conjugatif (**I**ntegrative and **C**onjugative **E**lement)

IME : élément intégratif mobilisable (**I**ntegrative and **M**obilizable **E**lement)

IR : répétition inversée (**I**nverted **R**epeat)

IS : séquence d'insertion (**I**nsertion **S**equence)

kb : **k**ilo**b**ase

Mb : **m**éga**b**ase

MPF : pore de conjugaison (**M**ating **P**air **F**ormation)

oriT : **ori**gine de **T**ransfert

pb : **p**aire de **b**ases

R-M : **R**estriction-**M**odification

sso : origine simple brin (**s**ingle-**s**trand **o**rigin)

T4SS : système de sécrétion de type 4 (**T**ype **4** **S**ecretion **S**ystem)

SOMMAIRE

INTRODUCTION

Les transferts horizontaux de gènes jouent un rôle clé dans l'évolution des génomes bactériens. Ils se produisent non seulement entre souches de la même espèce mais aussi entre organismes très éloignés du point de vue phylogénétique (Hacker and Kaper, 2000; Juhas et al., 2009; Mazodier and Davies, 1991). Ils peuvent être décomposés en trois étapes : i) un fragment d'ADN doit transiter d'une cellule à une autre essentiellement par transformation, par l'intermédiaire d'une capside ou par conjugaison, ii) les gènes transférés doivent être transmis à la descendance (réplication et partition dans le cas des plasmides ou intégration dans un réplicon) et enfin, pour que ces gènes se maintiennent dans la population, (iii) les cellules les ayant acquis doivent rencontrer un succès évolutif. Ceci sera le cas, par exemple, si le transfert a entrainé l'acquisition de gènes conférant une fonction avantageuse pour l'organisme dans un environnement donné, telle que la résistance à des antibiotiques, la capacité à utiliser de nouveaux substrats ou à synthétiser de nouveaux métabolites (Gogarten and Townsend, 2005; Ochman et al., 2000). L'acquisition de gènes codant des fonctions avantageuses peut avoir non seulement pour conséquence d'améliorer la capacité de son hôte à survivre et à avoir des descendants dans son environnement habituel (valeur adaptative ou «fitness»), mais aussi peut entrainer la conquête de nouvelles niches écologiques. Elle peut ainsi engendrer l'acquisition ou la modification de la pathogénicité. Un des exemples les plus frappants est celui de l'émergence du pathogène responsable de la peste *Yersinia pestis*. Cette bactérie extrêmement virulente transmise par les puces est apparue à partir de *Yersinia pseudotuberculosis*, bactérie de pathogénicité beaucoup plus modérée provoquant des gastroentérites, transmise par voie alimentaire (Lesic and Carniel, 2005). Cette apparition est liée, entre autres, à l'acquisition par transfert horizontal d'éléments (2 plasmides et un îlot de pathogénicité) jouant un rôle déterminant dans la transmission par les puces. Ainsi, un ancêtre de *Y. pestis* a acquis par transfert horizontal un ilot de pathogénicité portant un système de capture du fer qui joue un rôle clé dans sa capacité à croitre dans le sang et à provoquer une septicémie. Cettte capacité est non seulement nécessaire à sa transmission par les puces qui se nourrissent du sang de l'hôte mais aussi impliqué dans l'extrême virulence de la bactérie (mort par septicémie).

La plupart des gènes acquis par transfert horizontal, notamment ceux acquis par conjugaison ou par l'intermédiaire d'une capside, sont regroupés dans des éléments mobiles et/ou des ilots génomiques. Les éléments mobiles, tels que les plasmides, les prophages ou les éléments transposables codent leur propre transfert et/ou maintien après transfert et peuvent ou non porter des gènes d'adaptation. Les ilots génomiques peuvent être définis comme des segments chromosomiques acquis par transfert horizontal et portant des gènes pouvant améliorer la capacité de son hôte à survivre et à avoir des descendants ; ils peuvent correspondre à des éléments mobiles ou des dérivés d'éléments mobiles.

Les transferts horizontaux ont des implications majeures pour l'humanité. Ainsi, selon l'OMS, les résistances aux antibiotiques constituent aujourd'hui l'une des plus graves menaces pesant sur la santé mondiale, la sécurité alimentaire et le développement. Ils ont fait l'objet pour la première fois d'une réunion des dirigeants mondiaux lors de l'Assemblée générale des Nations Unies à New York le 21 septembre 2016 (http://www.who.int/antimicrobial-resistance/events/UNGA-meeting-amr-sept2016/fr/). En effet, un nombre croissant de maladies infectieuses d'origine bactérienne, comme la pneumonie, la tuberculose ou la gonorrhée, deviennent de plus en plus difficiles à traiter, les antibiotiques utilisés pour les soigner perdant leur efficacité (http://www.who.int/mediacentre/factsheets/antibiotic-resistance/fr/). La prolifération de ces résistances aux antibiotiques étant, en grande partie, liée aux transferts horizontaux d'éléments génétiques mobiles et ilots génomiques, l'étude des mécanismes de leur transfert et de leur prévalence apparait comme un enjeu majeur.

Cette thèse s'inscrit dans la thématique du transfert horizontal de gènes chez les bactéries, plus particulièrement le transfert d'ADN par conjugaison. Elle a pour but d'étudier la diversité et la prévalence des éléments chromosomiques transférés par conjugaison au sein des streptocoques, un groupe bactérien qui présente un grand intérêt pour l'homme, car il est constitué exclusivement de bactéries commensales, pathogènes ou d'utilité industrielle.

1. Le genre *Streptococcus*.

Le genre *Streptococcus* appartient au phylum des firmicutes. Il regroupe un vaste ensemble de micro-organismes contenant 121 espèces (http://www.bacterio.net/streptococcus.html, 10 juin 2017) pouvant être classées en 8 groupes phylogénétiques distincts : mitis, sanguinis, anginosus, salivarius, downei, mutans, pyogenic et bovis (Richards et al., 2014).

1.1. Mode de vie

La quasi-totalité des streptocoques sont des bactéries commensales ou pathogènes de l'homme et d'autres animaux. Par exemple, *Streptococcus salivarius,* étudiée dans le laboratoire DynAMic, est une espèce commensale, présente en abondance au sein du microbiote buccal, au niveau des muqueuses (Mitchell, 2003). Les streptocoques incluent aussi de nombreuses espèces ayant un impact important en santé humaine, allant de la simple carie jusqu'à des maladies potentiellement mortelles comme des méningites (Köhler, 2007). Par exemple, *Streptococcus pyogenes* est responsable d'un large panel de maladies comme les pharyngites, l'impétigo, la scarlatine, le syndrome du choc toxique, ou encore la fasciite nécrosante (Carapetis et al., 2005; Ralph and Carapetis, 2013). De même, *Streptococcus pneumoniae*, une espèce commensale du nez et de la gorge, est aussi un agent pathogène responsable d'un grand nombre d'infections bactériennes, allant de simples otites à de sérieuses infections pouvant être mortelles comme des méningites ou des pneumonies (O'Brien et al., 2003). De même, *Streptococcus agalactiae,* également étudié dans le laboratoire DynAMic, est une bactérie commensale du tractus gastro-intestinal ; il peut provoquer des méningites, septicémies et des pneumonies chez l'adulte et le nouveau-né (Baker and Pritchard, 2000; Balter and Dowell, 2000; Dermer et al., 2004). Ces trois espèces bactériennes font partie de la liste émise par le centre de contrôle et de prévention des maladies (CDC) contenant les 18 espèces bactériennes dangereusement résistantes aux antibiotiques, allant de menace sérieuse pour *Streptococcus pneumoniae* à menace inquiétante pour *Streptococcus pyogenes* (résistance à l'érythromycine) et *Streptococcus agalactiae* (résistance à la clindamycine) (Huang et al., 2016). Le genre *Streptococcus* contient aussi des espèces pathogènes pour les animaux dont certaines provoquent d'importantes pertes économiques pour l'élevage. Ainsi, l'espèce *Streptococcus suis*, est responsable de cas sporadiques de méningites et de septicémies chez l'homme, et provoque d'importantes pertes de production dans les élevages intensifs de porcs (Gottschalk et al., 2010). De même, *S. uberis*, *S. agalactiae*, *S. dysgalactiae* subsp. *dysgalactiae* et *S. canis* peuvent provoquer des mammites chez la vache engendrant ainsi de grosses pertes économiques en réduisant la production de lait (Zadoks et al., 2011). Enfin deux espèces, *S. thermophilus*, objet d'étude au laboratoire DynAMic, et *S. macedonicus*, sont utilisées dans l'industrie laitière comme ferments lactiques dans la production de laits fermentés et de divers fromages.

1.2. Caractéristiques biologiques et génétiques

Les streptocoques sont des bactéries à coloration Gram+, aéro-anaérobies facultatives organisées en chainettes ou en diplocoques, préférant les milieux riches. Ils possèdent un chromosome circulaire ayant un faible pourcentage en G+C (34% à 40%) (Anisimova et al., 2007).

Les streptocoques ont de relativement petits génomes, allant de 1,8 à 2,4 mégabases, qui contiennent de 1 700 à 2 400 gènes (Giovannoni et al., 2005). La variation du nombre de leurs gènes serait essentiellement due à des acquisitions de gènes par transfert horizontal et des pertes (Lefébure and Stanhope, 2007; Makarova and Koonin, 2007; Richards et al., 2014; Delorme et al., 2015). Le « core »-génome d'une espèce peut être défini comme l'ensemble des gènes présents dans l'ensemble des souches de l'espèce. Par exemple, pour *S. agalactiae*, une étude menée sur 8 génomes conclut que le core-génome de cette espèce est d'environ 1 800 gènes (Tettelin et al., 2005). Pour *S. thermophilus*, une étude menée sur un plus grand nombre de génomes, estime ce nombre à environ 1 500 gènes (Bolotin et al., 2004). Cette taille particulièrement faible du core-génome est en partie liée à la présence de nombreux pseudogènes récents, suggérant que le processus de perte est toujours en cours (Bolotin et al., 2004; Makarova and Koonin, 2007). Par opposition, le pangénome d'une espèce représente l'ensemble des gènes présent dans les génomes de cette espèce. En 2007, une comparaison de tous les génomes de streptocoques (drafts et complets) estimait le core génome à environ 600 gènes pour l'ensemble des streptocoques et le pangénome à au moins 6 000 gènes. De plus, 21% de ces gènes sont propres à une seule souche en moyenne, suggérant que les transferts horizontaux jouent un rôle important dans l'évolution des génomes de *Streptococcus* (Lefébure and Stanhope, 2007).

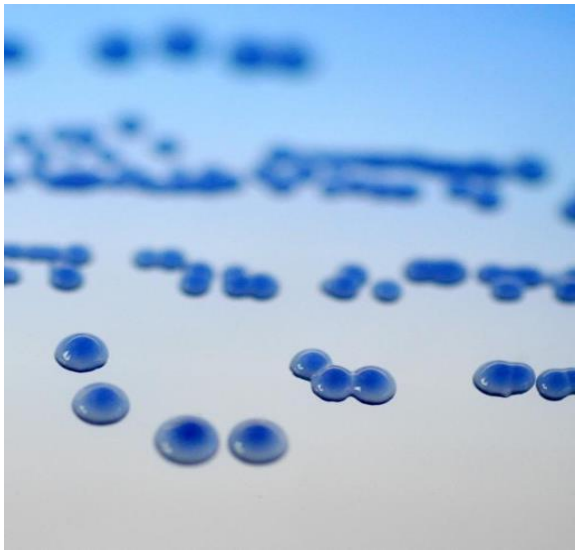

Figure 1 : Isolement de *S. salivarius* sur gélose.
Photo : P. Leblond, laboratoire DynAMic

2. Le transfert horizontal chez les streptocoques.

L'acquisition de gènes par transfert horizontal peut se faire par différents mécanismes, les trois principaux étant la transformation, le transfert par l'intermédiaire d'une capside (dont la transduction) et la conjugaison.

2.1. La transformation « naturelle »

La transformation « naturelle » correspond à l'acquisition active d'ADN libre dans le milieu par une cellule et à son intégration active au sein de son génome. La libération dans le milieu d'ADN peut survenir lors de la lyse cellulaire ou d'une sécrétion active d'ADN (Hamilton and Dillard, 2006; Claverys et al., 2007). Le mécanisme de transformation ne requiert pas de contact entre la cellule donatrice et réceptrice. Cependant, cette dernière doit être dans un état physiologique particulier dit état de compétence (Lorenz and Wackernagel, 1994). Dans cet état, l'expression de gènes chromosomiques dédiés va permettre l'entrée de l'ADN dans la cellule et son maintien par recombinaison homologue avec celui de la cellule réceptrice. La transformation naturelle est bien décrite chez diverses espèces de streptocoques comme *S. pneumoniae* (Straume et al., 2015) ou plus récemment *S. thermophilus* (Fontaine et al., 2010; Haustenne et al., 2015), mais n'est cependant pas démontrée chez tous les streptocoques.

2.2. Le transfert par une capside

La transfert d'ADN peut se faire par l'intermédiaire d'une capside qui le protège et lui permet de se disséminer. Les capsides assurent avant tout le transfert horizontal des éléments qui les codent, les prophages. Des fragments d'ADN ne codant pas de capside peuvent aussi être transférés par l'intermédiaire d'une capside (transduction). Ainsi l'ADN chromosomique peut être encapsidé par erreur et transféré à la place du génome du phage. Cet ADN devra alors être intégré au génome de la bactérie réceptrice par recombinaison homologue. De nombreux prophages ont été décrits chez les streptocoques. L'analyse des génomes montre ainsi qu'ils sont extrêmement répandus chez certaines espèces comme *S. pyogenes*, ou dans une moindre mesure chez *S. suis* et *S. agalactiae* (McShan and Nguyen, 2016; Ferretti et al., 2016).

2.3. La conjugaison

La conjugaison est un mécanisme qui nécessite un contact direct entre les deux cellules au cours duquel de l'ADN se transfère d'une cellule « donatrice » vers une cellule « réceptrice » (Lederberg and Tatum, 1946; Lederberg, 1998). Ce mécanisme nécessite la présence dans la bactérie donatrice d'un élément conjugatif qui code un pore de conjugaison permettant le transfert de l'ADN vers la bactérie réceptrice (Grohmann et al., 2003; Hayes et al., 2010). Deux types d'éléments sont capables de se transférer de manière autonome par conjugaison : les plasmides et les éléments intégratifs conjugatifs (ICE pour Integrative Conjugative Element). Après transfert, les plasmides et les ICE se distinguent par la manière de se maintenir dans les cellules réceptrices. Tandis que le maintien des plasmides repose sur une réplication autonome et sur une partition entre cellules au moment de la division cellulaire, celui des ICE repose essentiellement sur l'intégration de l'élément dans un réplicon.

Des analyses récentes démontrent que les plasmides et les ICE sont répandus dans diverses espèces de streptocoques notamment chez *S. suis, S. agalactiae* ou encore *S. salivarius* (Puymège et al., 2013; Goessweiner-Mohr et al., 2014; Huang et al., 2016; Dahmane et al., 2017)

3. Les éléments génétiques mobiles

L'acquisition de gènes par conjugaison et par transduction est, dans une très large majorité des cas, due à des éléments génétiques mobiles, c'est-à-dire aux éléments capables de se transférer entre cellules et/ou de se maintenir après transfert dans la descendance. Ces éléments peuvent porter 3 types de gènes ou séquences extra-géniques contribuant à leur succès dans la population bactérienne :

- des gènes ou séquences impliqués dans le transfert. Ils codent l'appareil de transfert de l'élément (capside, pore de conjugaison) ou assurent l'utilisation de l'appareil de transfert d'un autre élément. Ces gènes sont généralement inutiles voir nuisibles pour l'hôte (Vos et al., 2015).

- des gènes ou séquences impliqués dans le maintien de l'élément après transfert (réplication sous forme de plasmide, intégration dans un réplicon, partition…). Ces gènes sont généralement inutiles voir nuisibles pour l'hôte (Vos et al., 2015).

- des gènes d'adaptation (par exemple des résistances aux antibiotiques), qui peuvent contribuer au succès évolutif de l'hôte, et donc à celui de l'élément dans la population. La grande majorité des éléments conjugatifs en portent.

Jusqu'au début des années 2000, on distinguait 4 classes principales d'éléments acquis par transfert horizontal : les éléments transposables, les prophages, les plasmides et les ilots génomiques.

3.1. Les éléments transposables

Les éléments transposables sont des éléments capables de se déplacer, à différents endroits d'une même molécule d'ADN ou entre différentes molécules d'ADN (transposition). Ils ne codent pas leur transfert entre cellules.

Parmi eux, les séquences d'insertion (IS pour insertion sequence) ne codent que des fonctions impliquées dans leur mobilité intracellulaire. La plupart des IS codent une transposase à DDE et sont flanquées de séquences inversées répétées (IR pour inverted repeats) reconnues par leur transposase (Partridge and Hall, 2003; Siguier et al., 2015). La transposase catalyse la transposition de l'IS d'une position du génome vers une autre, généralement de manière peu spécifique (Nesmelova and Hackett, 2010). De nombreuses IS sont présentes au sein des génomes de streptocoques et ont un impact sur l'évolution de ces derniers. Par exemple, chez *S. pneumoniae,* la transposition d'IS*1515* dans le gène *ply*, induit l'inactivation de ce gène codant un facteur de virulence, et en conséquence modifie la pathogénicité de cette bactérie ce qui peut lui conférer un avantage sélectif lors de pressions fortes excercées par le système immunitaire (Garnier et al., 2007).

Contrairement aux IS, les transposons codent des fonctions qui ne sont pas impliquées dans leur mobilité et qui peuvent être utiles à l'hôte, comme des fonctions cataboliques ou des résistances aux antibiotiques. Parmi eux, les transposons composites sont des séquences d'ADN flanquées par 2 IS, la transposition de l'ensemble étant catalysée par la transposase codée par l'une de ces IS. Contrairement aux transposons composites, les transposons unitaires ne possèdent pas d'IS à leurs extrémités. La plupart d'entre eux codent, comme les IS, une transposase à DDE qui catalyse la transposition de l'élément par action au niveau des répétitions inversées qui le bordent (Roberts et al., 2008).

3.2. Les prophages

Après l'infection d'une bactérie par un phage dit tempéré, le génome du phage peut soit déclencher un cycle de multiplication du virus avec lyse de la bactérie, soit se transmettre à la descendance de la bactérie sous forme de prophage (lysogénisation). Globalement, la succession de phases lytiques et de lysogénisations implique que les prophages sont des éléments mobiles codant leur propre transfert horizontal entre bactéries par l'intermédiaire d'une capside. Si certains prophages sont transmis à la descendance sous forme « extra-chromosomique », la grande majorité doit s'intégrer dans le génome de la bactérie réceptrice pour se maintenir. Cette intégration est catalysée par une (parfois plusieurs) intégrase(s). Si les intégrases de quelques prophages sont des transposase à DDE, la plupart sont des intégrases à sérine ou des intégrases à tyrosine. Celles-ci vont reconnaitre une séquence spécifique du génome, appelée *attB* (Bacterial attachment site) et une séquence spécifique portée par l'élément, *attP* (Phage attachment site). L'intégrase va alors catalyser une recombinaison entre ces deux séquences et permettre l'intégration de l'élément dans le génome bactérien (Grindley et al., 2006; Fogg et al., 2014). L'élément intégré (prophage) est transmis à la descendance. En présence de stimuli spécifiques, le prophage code son excision, sa réplication et l'encapsidation de son génome. Enfin, la lyse de la cellule assure le plus souvent sa dissémination. Certains prophages intégrés sont également aussi capables de se maintenir temporairement sous forme excisée sans induire de phase lytique, durant certaines phases de vie de l'organisme hôte (Rabinovich et al., 2012). A côté des prophages « classiques » qui codent toutes les fonctions nécessaires à leur transfert, d'autre éléments, les prophages satellites n'en codent qu'une partie et, de ce fait, ont besoin de la présence d'un élément « helper » pour se transférer. Ces éléments détournent l'appareil de transfert des prophages pour se transférer eux-mêmes (Christie and Dokland, 2012). Les prophages ont un impact dans l'évolution des génomes de streptocoques (Toussaint and Rice, 2017), notamment de divers pathogènes (Bessen et al., 2015). Par exemple chez *S. pyogenes*, divers prophages codent des facteurs de virulence, en particulier des exotoxines. Ainsi le phage SF370.1 code 2 facteurs de virulence *speC* et *spd1* (McShan and Nguyen, 2016).

3.3. Les plasmides

Par définition, les plasmides sont des éléments qui se maintiennent uniquement sous forme extra-chromosomique. Tous portent des gènes impliqués dans leur maintien dans la

descendance par réplication. Beaucoup, en particulier les gros plasmides, codent aussi d'autres fonctions contribuant à leur maintien dans la descendance, notamment la résolution de dimères, la ségrégation des copies lors de la division cellulaire ou encore des systèmes poison-antipoison.

Les plasmides sont généralement classés en trois catégories en fonction de leur aptitude à se transférer par conjugaison: les plasmides conjugatifs, les plasmides mobilisables et les plasmides non conjugatifs (Smillie et al., 2010). Les premiers codent toutes les fonctions nécessaires à leur transfert et se transfèrent donc de manière autonome. Les seconds, quant à eux, ne codent que certaines des fonctions de transfert. Ils ne se transfèrent pas de façon autonome mais sont mobilisés en *trans* par des éléments conjugatifs auxquels ils ne sont pas physiquement liés. Enfin les plasmides non conjugatifs, quant à eux, ne portent aucune fonction de transfert par conjugaison. Cependant, ces plasmides peuvent être mobilisés en *cis* par des plasmides conjugatifs ou mobilisables. Ceci implique dans un premier temps, qu'ils se co-intègrent accidentellement avec un plasmide conjugatif ou mobilisable (par exemple en raison de la transposition réplicative intermoléculaire d'une IS portée par l'un des deux plasmides ou une recombinaison homologue entre IS portées par les deux plasmides). Ensuite, le co-intégrat peut être transféré par conjugaison puis se résoudre en ses 2 composants.

Enfin, beaucoup de plasmides portent des gènes d'adaptation (Smillie et al., 2010), notamment des gènes de virulence (Williamson et al., 1990) ou de résistances aux antibiotiques Au cours des dernières décennies, les plasmides ont été beaucoup étudiés et leur prévalence et diversité ainsi que les protéines impliquées dans leur transfert sont relativement bien connues (Frost et al., 2005; Alvarez-Martinez and Christie, 2009; Fronzes et al., 2009; de la Cruz et al., 2010; Goessweiner-Mohr et al., 2014; Waksman and Orlova, 2014; Cabezón et al., 2015; Christie, 2016).

3.3.1. Les plasmides conjugatifs.

Du fait que les plasmides conjugatifs codent toutes les fonctions nécessaires à leur transfert par conjugaison, ils sont généralement de plus grande taille que les plasmides mobilisables. Chez les firmicutes, le mieux connu est le plasmide pIP501, isolé à l'origine chez *S. agalactiae* (Evans and Macrina, 1983). Il est capable d'assurer son transfert vers un large spectre de

bactéries Gram+ et aux moins quelques bactéries Gram- (Kurenbach et al., 2003). De plus pIP501 est aussi capable de mobiliser divers plasmides, dont pMV158, qui est le mieux connu des plasmides mobilisables de firmicutes (Schaberg et al., 1982; Thompson and Collins, 1988; Krah and Macrina, 1991; Kurenbach et al., 2002, 2003; Zúñiga et al., 2003; Abajy et al., 2007). D'une taille de 30,2 kb, pIP501 porte non seulement les 15 gènes nécessaires à son transfert par conjugaison, regroupés en 1 seul opéron (Kurenbach et al., 2003, 2006), mais aussi des gènes de résistance aux macrolides, lincosamide, chloramphénicol et streptogramine B (Kurenbach et al., 2002).

Le transfert par conjugaison des plasmides conjugatifs se déroule en 4 étapes qui sont schématisées dans la figure 2. Tout d'abord le plasmide est reconnu par une endonucléase, la relaxase, au niveau de son origine de transfert (*oriT*) qui est une séquence non-codante comportant généralement une ou plusieurs répétition(s) inversée(s) ainsi qu'un site *nic.* La relaxase initie le transfert en se liant à *oriT* et coupe le brin transféré au niveau du site *nic* (figure 2b). La relaxase reste fixée de manière covalente à l'extrémité 5' du brin transféré. Selon l'élément, la relaxase peut agir seule sur l'origine de transfert ou en interaction avec d'autres protéines codées par l'élément conjugatif pour constituer un complexe nucléoprotéique, le relaxosome. Ce complexe ADN-relaxase est ensuite reconnu par, une ATPase, la protéine de couplage (CP pour coupling protein), puis est transporté depuis la cellule donatrice vers la cellule réceptrice au travers du pore de conjugaison (MPF pour mating pore formation). Celui-ci inclut notamment une autre ATPase, la protéine VirB4 (figure 2c). Le complexe transmembranaire multiprotéique CP-MPF est un système de sécrétion de type 4 (T4SS pour Type 4 secretion system). La coupure par la relaxase, initie non seulement le transfert mais aussi la réplication par cercle roulant du plasmide (figure 2c). Enfin, le transfert se termine lorsque les deux copies du plasmide sont circularisées par la relaxase dans chacune des cellules

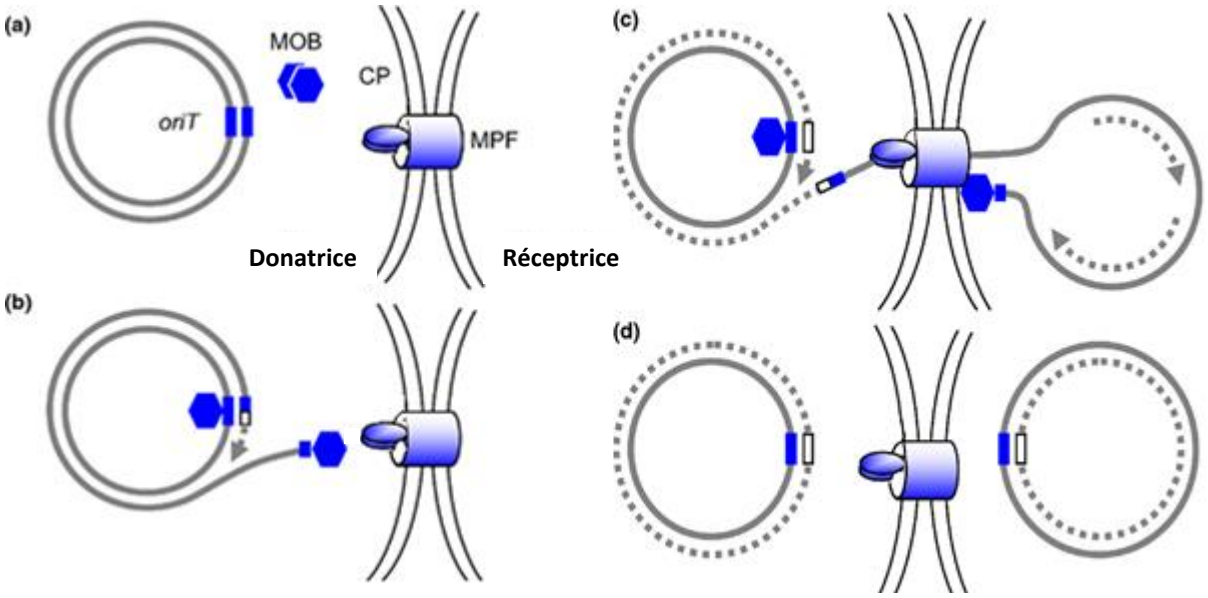

Figure 2 : Mécanisme du transfert de plasmide par conjugaison (d'après Bellanger et al., 2014). *Le plasmide est reconnu par la relaxase au niveau de son origine de transfert. La relaxase initie le transfert en coupant le brin transféré au niveau de l'*oriT *et reste fixée de manière covalente à l'extrémité 5' du brin transféré (a-b). Le complexe ADN-relaxase est ensuite reconnu par la protéine de couplage, puis transporté depuis la cellule « donatrice » vers la cellule « réceptrice » au travers du pore de conjugaison (b-c). La coupure par la relaxase, initie non seulement le transfert mais aussi la réplication par cercle roulant du plasmide. Dans la donatrice, la réplication utilise comme amorce l'extrémité 3' du brin coupé par la relaxase, et, dans la réceptrice, des amorces d'ARN synthétisée par une primase (c). Les deux copies du plasmide sont circularisées par la relaxase dans chacune des cellules (d). CP, protéine de couplage; MPF, pore de conjugaison; oriT, origine de transfert; MOB, relaxase. Les brins d'ADN en cours de synthèse sont représentés par des pointillés.*

3.3.2. Les plasmides mobilisables.

Tous les plasmides mobilisables portent leur propre *oriT.* Jusqu'à des temps récents, la quasi-totalité des plasmides mobilisables connus codaient leur propre relaxase (et souvent d'autres protéines du relaxosome) (Smillie et al., 2010). Cependant des études ont montré que certains plasmides mobilisables, en particulier les plasmides pIB485, pMW2 et pUSA300HOUMR de staphylocoques portaient leur propre *oriT*, apparentée à celle du plasmide mobilisateur, mais ne codaient aucune relaxase (F. G. O'Brien et al., 2015). L'analyse des séquences de nombreux autres plasmides de staphylocoques a révélé des *oriT* putatives, suggérant que de nombreux plasmides considérés comme non-mobilisables pourraient l'être (O'Brien et al., 2015; Pollet et al., 2016; Ramsay et al., 2016; Ramsay and Firth, 2017). Prises dans leur ensemble, les données récentes suggèrent que des stratégies variées sont utilisées par les plasmides mobilisables afin d'exploiter l'appareil de transfert d'un autre élément pour se transférer. Ces stratégies diffèrent selon les fonctions codées par le plasmide mobilisable (figure 3). Par exemple, si le plasmide mobilisable porte une *oriT*

mais ne code aucune protéine intervenant dans son transfert, sa mobilisation en *trans* nécessitera la présence dans la cellule d'un élément « helper » codant un MPF complet, une CP et une relaxase reconnaissant l'*oriT* du plasmide à mobiliser. Ainsi, le plasmide pWBG744/pIB485, du firmicute *S. aureus*, mobilisé en *trans* par pWBG749, ne code aucune protéine impliquée dans son transfert mais porte une *oriT* apparentée à celle du plasmide conjugatif pWBG749 de *S. aureus* (F. G. O'Brien et al., 2015; Frances G. O'Brien et al., 2015)(Figure 3 IV). D'autres plasmides, en plus de porter une *oriT* apparentée à celles d'éléments conjugatifs mobilisateurs, codent des protéines accessoires du relaxosome augmentant l'affinité de la relaxase de l'élément mobilisateur pour leur propre *oriT*, mais ne code aucune autre protéine impliquée dans leur transfert (Figure 3 V). Ainsi, la mobilisation en *trans* du plasmide pWBG745 de *S. aureus* par pWBG749 nécessite la présence du gène *smpO* de pWBG745 (Ramsay and Firth, 2017). Une grande majorité des plasmides mobilisables connus codent, quant à eux, leur propre relaxase (ainsi qu'éventuellement les autres protéines du relaxosome) et emprunteront le T4SS d'un élément conjugatif pour se transférer (Figure 3II). Ainsi le petit plasmide pMV158 de *S. agalactie* (5,5kb) est mobilisé en *trans* notamment par pIP501 ou pAMβ1 vers plus de 20 espèces différentes allant d'autres espèce de streptocoques comme *S. thermophilus* (Somkuti and Steinberg, 2007) à des espèces plus éloignées comme *Lactococcus lactis* (Farías et al., 1999) ou *S. aureus* (Li et al., 2013) voire vers des bactéries Gram- comme *E. coli* (Farías and Espinosa, 2000; Fernández-López et al., 2014). En plus de coder certaines fonctions nécessaires à son transfert comme sa propre relaxase, pMV158 code également la résistance à la tétracycline. Enfin, certains plasmides codent non seulement leur propre relaxase mais également leur propre CP, comme le plasmide CloDF13 d'*Enterobacter cloacae* (Cabezón et al., 1997). Dans ce dernier cas, il est généralement supposé que la relaxase interagirait avec la CP codée par l'élément qui remplacerait la CP de l'élément mobilisateur (Figure 3I).

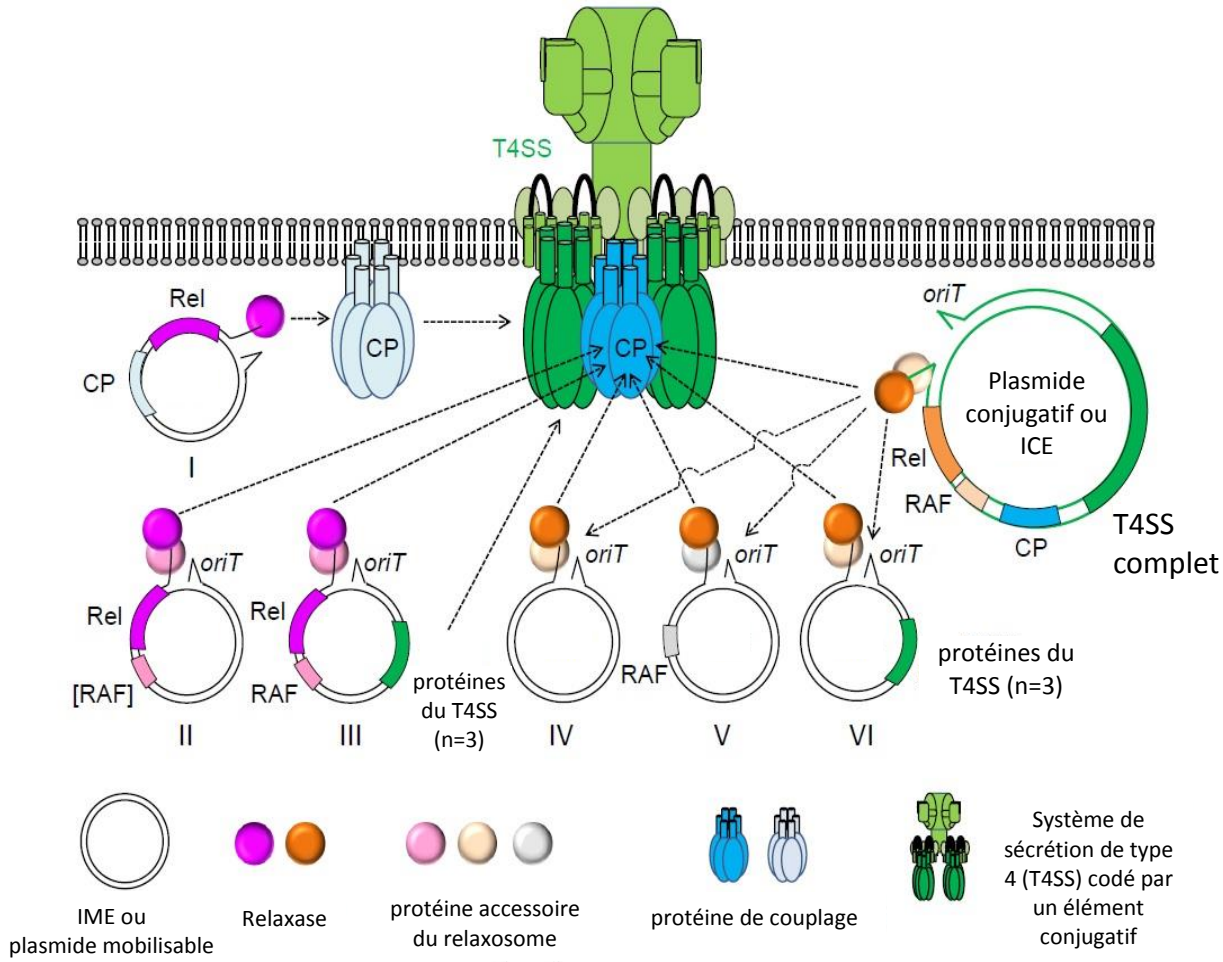

Figure 3 : Les diverses stratégies de mobilisation des éléments mobilisables en trans par des éléments conjugatifs. *Les éléments mobilisables peuvent exploiter les éléments conjugatifs pour se transférer en portant ou codant soit : I, leur propre* oriT*, relaxase et CP afin de recruter le T4SS ; II, leur propre* oriT*, relaxase et éventuellement une ou plusieurs RAF afin de recruter la CP; III, leur propre* oriT*, leur propre relaxase, des RAF et 3 protéines du T4SS afin de recruter la CP et le T4SS (prédit à partir des séquences pour quelques éléments intégratifs mobilisables ou IME uniquement); IV, uniquement une* oriT *afin de recruter la relaxase; V, une oriT et des RAF afin de recruter la relaxase d'un élément conjugatif ; VI, une* oriT *et 3 protéines du T4SS afin de recruter la relaxase, la CP et le T4SS (observé uniquement pour des IME). Les interactions entres les éléments sont représentées par des flèches en pointillées. Les plasmides conjugatifs et les éléments intégratifs conjugatifs (ICE) codent toutes les protéines nécessaires à leur transfert autonome par conjugaison, y compris une relaxase (Rel), des protéines accessoires du relaxosome (RAF), une protéine de couplage (CP) et un système de sécrétion de type 4 (T4SS). Les relaxases et les protéines accessoires du relaxosome sont représentées par des sphères colorées, la protéine de couplage par un hexamère protéique et le T4SS par un complexe multi-protéique. La CP en bleu ciel est codée par l'élément mobilisable et la CP en bleu foncé est codée par l'élément conjugatif.*

3.4. Les ilots génomiques

Un ilot génomique peut être décrit comme une région chromosomique dont les propriétés suggèrent son acquisition par un ou plusieurs événements de transfert horizontal et qui porte des gènes susceptibles d'augmenter la valeur adaptative (« fitness ») de son hôte

(Bellanger et al., 2014). Ainsi, un ilot génomique est généralement présent dans le génome de certaines souches d'une espèce et absent du génome d'autres souches de la même espèce (et/ou d'espèces proches), en raison d'une acquisition récente par transfert horizontal. Les ilots présentent généralement des caractéristiques de séquence témoignant de cette acquisition par transfert horizontal, telles qu'un pourcentage en G+C ou une utilisation des codons différents de ceux du génome de l'organisme dans lequel ils se sont intégrés. Bien que le mécanisme d'acquisition par transfert horizontal n'ait été décrit que pour très peu d'ilots génomiques, la plupart d'être eux portent des gènes (ou des pseudogènes) codant des protéines de « mobilité » (intégrases, protéines de conjugaison ou de réplication), qui pourraient être impliquées dans le transfert et/ou maintien, soit de l'ilot lui-même, soit d'éléments mobiles portés par celui-ci. Globalement, les ilots génomiques sont probablement des éléments mobiles codant leur propre transfert et/ou maintien après transfert, des éléments en dérivant ou des combinaisons de tels éléments. Dans leur ensemble, les ilots génomiques présentent des caractéristiques très variables et restent, de ce fait, difficiles à détecter sans erreurs et à délimiter (voir paragraphe 5.1 p44 concernant les méthodes de détection). Beaucoup d'entre eux portent des gènes de conjugaison (ou pseudogènes) et pourraient appartenir à de nouvelles classes d'éléments mobiles définies dans le début des années 2000, les éléments intégratifs conjugatifs (ICE) ou mobilisables (IME pour Integrative and Mobilizable Element), ou des éléments en dérivant.

4. Les ICE et les IME

Les ICE sont définis comme des éléments codant leur propre excision, transfert par conjugaison et intégration après transfert, indépendamment des mécanismes impliqués (Burrus et al., 2002a). La plupart s'intègrent de façon site-spécifique. Cependant, comme beaucoup des premiers ICE identifiés présentaient une faible spécificité d'insertion, certains ICE sont parfois appelés transposons conjugatifs (Roberts et al., 2008).

Les IME, quant à eux, sont définis comme des éléments codant toutes les fonctions nécessaires à leur excision et intégration, mais contrairement aux ICE, ils ne portent qu'une petite partie des gènes et séquences nécessaires à leur transfert conjugatif, comme la séquence *oriT* ou le gène de la relaxase. Comme les plasmides mobilisables, les IME ne sont pas autonomes pour leur transfert et doivent exploiter l'appareil de conjugaison codé par un

élément conjugatif (ICE ou plasmide) pour se transférer (Burrus et al., 2002a; Bellanger et al., 2014). Comme la plupart des premiers IME identifiés présentaient une faible spécificité d'insertion, certains IME sont parfois appelés transposons mobilisables.

Les ICE et les IME connus ont des tailles très variables. Le plus petit et le plus grand des éléments dont le transfert ait été démontré sont respectivement l'IME M*TnSag1* de *S. agalactiae* (1,7 kb) et l'ICE PAISt de l'actinobactérie *Streptomyces turgidiscabies* (674 kb) (Kers et al., 2005; Achard and Leclercq, 2007; Bellanger et al., 2014). Les plus petits ICE connus du genre *Streptococcus* et de firmicutes mesurent 18 kb.

4.1. Une structure modulaire

Tous les éléments génétiques mobiles sont constitués d'un seul module ou d'une combinaison de modules (Burrus et al., 2002a; Toussaint and Merlin, 2002; Pavlovic et al., 2004). Un module est une région regroupant les gènes et les séquences intervenant dans une même fonction biologique. Les comparaisons d'éléments génétiques mobiles montrent que ceux-ci évoluent principalement par acquisition, perte et échanges de modules. Les ICE et les IME portent tous au moins deux modules distincts :

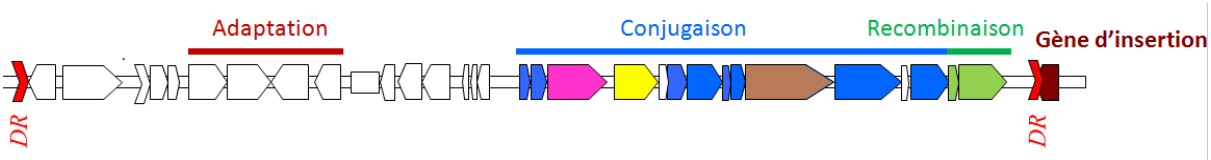

Figure 4 : Représentation schématique de la structure modulaire d'ICE*St3*. *Les chevrons rouges représentent les séquences répétées (DR). Les flèches vertes symbolisent les gènes de l'intégrase et de l'excisionase, la flèche jaune symbolise le gène de la relaxase, la flèche fuchsia le gène de la protéine de couplage, la flèche brune le gène de la protéine VirB4. Le gène cible de l'insertion est représenté en bordeaux.*

- Tous les ICE et tous les IME présentent un module d'intégration/excision (Bellanger et al., 2014) (figure 4, module en vert). Ils codent une ou plusieurs intégrase(s) catalysant l'excision et l'intégration de l'élément, et éventuellement une excisionase déplaçant l'équilibre de la réaction vers l'excision. Il porte également les séquences reconnues par les intégrases. Ce type de module est également toujours présent dans les éléments transposables ou les prophages intégrés.

- Tous les ICE portent un module de conjugaison, qui comporte une origine de transfert et code toutes les protéines de conjugaison (Bellanger et al., 2014)(figure 4,

module en bleu). Les rares IME décrits dans la littérature portent, quant à eux, un module de mobilisation qui comprend toujours une *oriT* et souvent une relaxase. Certains IME, exclusivement retrouvés chez les protéobactéries, codent aussi 3 protéines du T4SS mais ne codent cependant, ni la protéine de couplage, ni la protéine VirB4 (Carraro et al., 2017a, 2017b). D'autres IME putatifs de protéobactéries codent une relaxase, des protéines accessoires du relaxasome et 3 protéines du T4SS (Bellanger et al., 2014). Des modules de conjugaison et de mobilisation apparentés à ceux des ICE et des IME peuvent se retrouver également chez certains plasmides, comme pCW3 de *Clostridium perfringens*, pMV158 de *S. agalactiae*, pCF10 *d'Enterococcus faecalis*, ou encore pIP501 de *S. agalactiae* (Alvarez-Martinez and Christie, 2009; Bhatty et al., 2013).

- Tous les ICE et divers IME possèdent un module de régulation qui contrôle le processus de conjugaison en réponse à différents stimuli Par exemple, l'ICE Tn*916*, rencontré chez de nombreux firmicutes dont les streptocoques, code une résistance à la tétracycline ; l'expression des gènes de conjugaison est induite en présence de cet antibiotique (Showsh and Andrews, 1992). Ce module de régulation peut être apparenté à ceux retrouvés chez les prophages. Ainsi, l'expression de l'ICE ICE*Bs1* est régulée par le répresseur ImmR qui a 50% d'identité en acides aminés avec le répresseur du phage ϕ105 de *B. subtilis* (Auchtung et al., 2005). Un autre exemple est ICE*St3* de *S. thermophilus* dont la régulation implique un répresseur, de famille totalement différente de celui d'ICE*Bs1*, apparenté au répresseur cI du phage lambda (Bellanger et al., 2007).

- Tous les ICE et la quasi-totalité des IME portent un ou plusieurs modules d'adaptation pouvant conférer un avantage adaptatif à l'organisme qui le porte (Bellanger et 2014) (figure 4). La caractérisation des gènes portés par ces modules représente un enjeu majeur puisqu'ils comportent des gènes pouvant avoir une application industrielle comme thérapeutique. Ces modules étant extrêmement variables d'un élément à l'autre, ils rendent la délimitation des éléments difficile. Ces gènes peuvent être retrouvés dans d'autres éléments génétiques mobiles comme les plasmides. Ainsi, les enzymes du système de restriction-modification (R-M) codé par ICE*St1* de *S. thermophilus* sont apparentés aux enzymes du système R-M *Lla*KR2I codé par le plasmide pKR223 de *L. Lactis* (Burrus et al., 2001)

Aucun de ces modules n'est cependant caractéristique d'un type d'élément mobile. Ainsi, une IS, peut être définie comme un élément mobile constitué d'un module d'intégration/transposition unique. Cependant, divers ICE possèdent des modules de « transposition » apparentés à ceux des IS voire en dérivant (Guérillot et al., 2013; Bellanger et al., 2014). D'une façon générale, toutes les classes d'éléments génétiques mobiles dont les ICE et les IME sont caractérisées par une association typique de modules et non par des modules types (Burrus et al., 2002a). Cette caractéristique rend donc leur détection et leur caractérisation difficile : en effet, celle-ci repose sur la détection d'une association de modules, alors que ceux-ci peuvent être retrouvés seuls, être échangés entre éléments et être retrouvés dans différentes classes d'éléments génétiques mobiles.

4.1.1. Le module de recombinaison

Le module de recombinaison est le module responsable de l'excision et intégration de l'élément dans un réplicon, généralement le chromosome bactérien mais éventuellement dans un autre élément mobile comme un plasmide. Le module d'intégration comporte des sites flanquants, qui sont reconnus par la ou les intégrases et éventuellement une excisionase. Les intégrases peuvent appartenir à 3 familles non apparentées : les intégrases à tyrosine, les intégrases à sérine et les transposases à DDE. Les modules d'intégration retrouvés chez les ICE présentent une grande diversité de spécificité d'insertion (Bellanger, et al 2014).

4.1.1.1. Les intégrases à tyrosine

La majorité des ICE et des IME connus possèdent une intégrase appartenant à la superfamille des recombinases à tyrosine (Bellanger et al., 2014). Chez les firmicutes, la taille de ces intégrases varie entre 360 et 500 acides aminés. Toutes portent un domaine catalytique conservé du coté C-terminal et la plupart portent également un domaine de liaison à l'ADN variable du coté N-terminal (Grindley et al., 2006). Le modèle d'étude des intégrases à tyrosine est l'intégrase du phage lambda qui porte un domaine de fixation à l'ADN du côté N-terminal (PF09003) et un domaine catalytique du coté C-terminal (PF00589) (Nunes-Düby et al., 1998; Groth and Calos, 2004; Grindley et al., 2006; Piednoël et al., 2011; Fogg et al., 2014; Stark, 2017). Elles catalysent une recombinaison, dans la plupart des cas, site-spécifique, entre deux courtes séquences identiques ou presque identiques appartenant au site *attB* présent sur le chromosome bactérien et au site *attI* porté par l'ICE

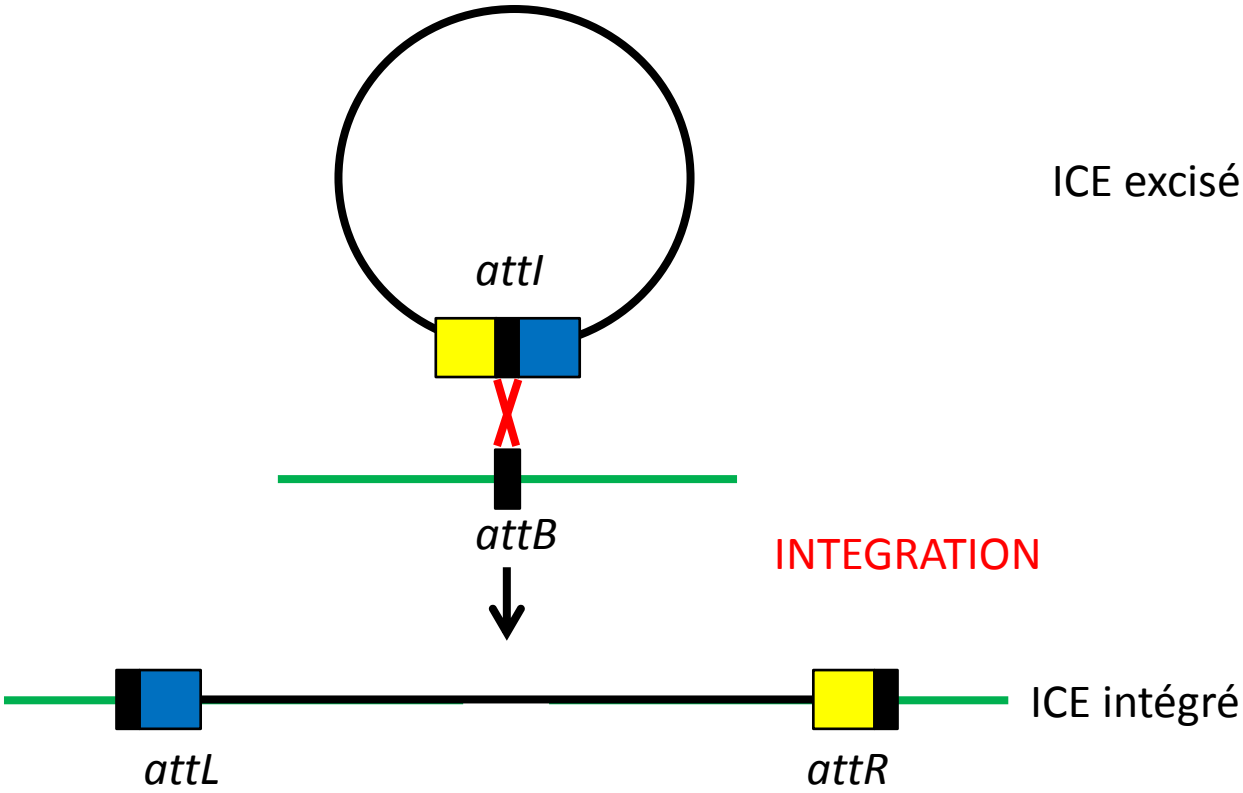

Figure 5 : Représentation schématique de l'intégration d'un ICE codant une intégrase site-spécifique. *L'intégrase codée par l'ICE catalyse la recombinaison site-spécifique entre les séquences identiques des sites* attI *et* attB *de deux molécules d'ADN conduisant à leur coïntégration et à la formation de sites* attL *et* attR*. Le trait vert représente le génome de l'hôte et le trait noir représente l'ICE. Le rectangle noir représente une séquence identique appartenant aux sites* attB*,* attI*,* attL *et* attR *; en jaune apparait le bras du site* attR*, séquence avec laquelle l'intégrase se lie avec une grande affinité et le bras correspondant dans le site* attI*; en bleu apparait le bras du site* attL*, séquence avec laquelle l'intégrase se lie avec une grande affinité et le bras correspondant dans le site* attI*.*

(Fogg et al., 2014; Stark, 2017). Du fait de la présence de sites de liaison de l'intégrase et de cofacteurs adjacents à la séquence directement impliquée dans la recombinaison (bras), le site *attI* est toujours beaucoup plus long (240-420 pb) que le site *attB* (25-40 pb). Ainsi, le site *attB* d'ICE*Bs1* du firmicute *B. subtilis* aurait une taille d'environ 17 bp et son site *attI* une taille d'environ 250 bp (Lee et al., 2007). Divers ICE et IME codant des intégrases à tyrosine site-spécifiques ont été identifiés chez les streptocoques. Ainsi, ICE*St3* de *S. thermophilus* code un intégrase à tyrosine catalysant une intégration site-spécifique dans l'extrémité 3' du gène *fda* codant la fructose-1,6-diphosphate aldolase (Bellanger et al., 2007). IME_Sag2603_tRNAlys de *S. agalactiae* qui s'intègre dans l'extrémité 3' de l'ARN de transfert lysine (Brochet et al., 2008) code également une intégrase à tyrosine. L'intégration site-spécifique d'un élément codant une intégrase à tyrosine provoque la formation de deux sites flanquant l'élément intégré, les sites *attL* et *attR* (L pour left et R pour right) (figure 5). Du

fait de la recombinaison entre séquences identiques ou presque identiques portées par les sites *attB* et *attI*, les site *attL* et *attR* incluent des séquences identiques pouvant aller de 6 à 100 pb flanquant l'élément, appelées répétitions directes ou « direct repeat » (DR). Certaines intégrases à tyrosine, quant à elles, ne catalysent pas d'intégrations site-spécifiques mais plutôt sites préférentielles. Ainsi, l'intégrase à tyrosine, des ICE de famille Tn*916* fréquemment rencontrés chez les streptocoques, catalyse une intégration dans divers sites ayant pour caractéristiques d'être riches en AT qui ne comportent généralement pas de courtes séquences identiques entre le site attI de l'ICE et le site chromosomique ciblé, ce qui par conséquent ne provoque pas la formation de DR (Roberts and Mullany, 2009). L'ICE Tn*916* peut donc être retrouvé intégré dans un large panel de sites, que ce soit dans des régions inter-géniques ou à l'intérieur de gènes (Mullany et al., 2012), pouvant ou non appartenir à d'autres éléments génétiques mobiles. Ainsi, l'ICE Tn*5253* de *S. pneumoniae* porte un ICE de la famille Tn*916* (Iannelli et al., 2014).

L'excisionase dirige l'équilibre de la recombinaison vers l'excision, d'une part en favorisant l'appariement des régions *attL* et *attR* et d'autre part en inhibant l'appariement entre les sites *attB* et *attP* (ou attI) (Groth and Calos, 2004; Fogg et al., 2014). Bien que généralement présente, l'excisionase ne semble pas toujours indispensable dans les modules d'intégration comportant une intégrase à tyrosine. Ainsi, dans le cas d'ICE*clc*, l'excision de l'élément a été constatée bien qu'aucune excisionase n'ait été mise en évidence (Miyazaki and van der Meer, 2013).

4.1.1.2. Les intégrases à sérine

Certains ICE et IME de firmicutes codent des intégrases appartenant à la superfamille des recombinases à sérine. Les intégrases à sérine ont toutes une taille de plus de 400 acides aminés et peuvent même comporter jusqu'à 770 acides aminés. Elles comportent toutes au moins 2 domaines caractéristiques : un domaine « résolvase » catalytique conservé en N-terminal (PF00239) suivi d'un domaine « recombinase » de fixation à l'ADN (PF07508). Le domaine « recombinase » est généralement suivi d'un petit domaine « zinc-ribbon » (PF13408) pouvant jouer un rôle dans l'appariement des deux brins liés par les intégrases (Yuan et al., 2008; Stark, 2014; Smith, 2015). La plupart sont retrouvées dans des phages (Groth and Calos, 2004; Stark, 2014). Comme les intégrases à tyrosine, elles assurent l'intégration site-spécifique des éléments en catalysant une recombinaison entre les sites

attI et *attB* entrainant la formation de sites *attL* et *attR* après intégration (figure 5). Cependant contrairement aux intégrases à tyrosine, les sites *attI* et *attB* sont tous les deux courts, d'environ 40-50 pb. Ainsi, le phage ϕBxb1 de mycobactéries présente un site *attP* de 48 pb et un site *attB* de 38 bp (Bai et al., 2011; Stark, 2017). Par ailleurs, l'intégration n'engendre la formation que de DR très courts, souvent de 2 pb. Enfin, chez certains prophages, la présence d'une excisionase semble être indispensable pour l'excision efficace des éléments par l'intégrase à sérine (Fogg et al., 2014; Ghosh et al., 2006; Khaleel et al., 2011; Stark, 2017). Cependant, chez le seul IME (Tn*4451* de *Clostridiodes difficile*) et le seul ICE (Tn*5397* de *Clostridiodes difficile*) codant une intégrase à sérine et dont l'excision est bien caractérisée, l'intégrase à est la seule protéine requise pour l'excision (Crellin and rood, 1997 ; Wang and Mullany, 2000). Des ICE et des IME codant des intégrases à sérine site-spécifiques ont été identifiés chez les streptocoques. Ainsi, ICE*Sp2905* de *S. pyogenes* code une intégrase à sérine qui catalyse l'intégration de l'élément dans le gène *rumA* (ARNr uracile méthyltransférase) (Giovanetti et al., 2012). C'est aussi le cas d'IME*Sp2907* de *S. pyogenes* qui code une intégrase à sérine catalysant une intégration dans le gène *traG*, un des gènes du module de conjugaison de l'ICE dans lequel il s'intègre (Mingoia et al., 2016).

4.1.1.3. Les transposases à DDE

Quelques ICE et IME de firmicutes codent des intégrases appartenant à la superfamille des transposases à DDE. Par exemple, ICE*6013* de Staphy*lococcus aureus* code une transposase à DDE appartenant à la famille IS*30* (Smyth and Robinson, 2009). Chez les streptocoques, les ICE de la famille Tn*GBS1* et Tn*GBS2* (Brochet et al., 2009) codent une transposase à DDE apparentée aux transposases d'IS de la famille IS*Lre2* dont la taille varie généralement entre 500 et 630 acides aminés (Guérillot et al., 2014; Siguier et al., 2015). Comme pour les intégrases à tyrosine et à sérine, cette transposase est chargée de l'excision et de l'intégration de l'élément. Elle catalyse cette intégration exclusivement 15-16 pb en amont de la séquence -35 de promoteurs variés, avec une forte préférence dans le cas de Tn*GBS2* pour un site particulier localisé en amont du promoteur du gène *nrdF*. Cependant, tous les autres ICE codant des transposases à DDE présentent des spécificités d'intégration très faibles (Bellanger et al., 2014). L'intégration va provoquer la duplication de la cible et l'élément se retrouve ainsi flanqué de répétitions directes de 8-9 pb. Pour l'excision, la

transposase reconnait des séquences répétées inversées situées aux extrémités de l'élément (Brochet et al., 2008; Guérillot et al., 2013).

Certains IME de streptocoques, dont MTn*Sag1* de *S. agalactiae,* codent également des transposases à DDE. Ce petit élément de 1,7 kb, mobilisable par l'ICE Tn*916*, code une transposase à DDE d'environ 330 acides aminés, apparentée à la famille IS*1595* (Achard and Leclercq, 2007). Cet élément, flanqué de répétitions inversées s'intègre préférentiellement dans des régions riches en AT.

4.1.2. Le module de conjugaison

Le module de conjugaison code toutes les séquences et fonctions nécessaires au transfert par conjugaison de l'ICE. Le module de conjugaison peut être constitué de sous-modules. Le sous-module MOB, chargé de la mobilisation de l'ADN transféré, contient l'origine de transfert de l'élément, le gène qui code la relaxase et éventuellement les gènes qui codent les autres protéines du relaxasome. Les gènes nécessaires à la mise en place du pore de conjugaison (discuté plus loin dans le manuscrit) sont généralement groupés en un sous-module MPF. Bien que ces 2 sous-modules puissent être associés en différentes combinaisons, on constate généralement une certaine cohérence dans leur association ; ainsi un type module MOB aura tendance à être associé avec le même type de module MPF et inversement (de la Cruz et al., 2010; Fernández-López et al., 2014; Ruiz-Masó et al., 2015).

4.1.2.1. Le transfert conjugatif

Transfert double brin d'ADN

Chez les bactéries Gram+, deux mécanismes distincts de conjugaison sont connus. Tous les éléments conjugatifs bien caractérisés des firmicutes et une partie de ceux des actinobactéries se transfèrent sous forme simple brin. Cependant, les ICE d'actinomycètes pluricellulaires (appelés AICE pour actinomycetes integrative and conjugative elements) et des plasmides conjugatifs de ces bactéries utilisent un système de transfert d'ADN double brin (Goessweiner-Mohr et al., 2014). Ce transfert est assuré par une seule protéine, la protéine TraSA apparentée aux protéines FtsK/SpoIIIE (translocase d'ADN double brin essentielle à la division cellulaire)(Reuther et al., 2006; te Poele et al., 2008; Thoma and Muth, 2016).

Transfert simple brin d'ADN

Le mécanisme de transfert conjugatif simple brin est bien connu chez divers plasmides conjugatifs ou mobilisables de protéobactéries, mais seulement certaines étapes de ce transfert ont été caractérisées chez les ICE et les IME. Cependant les données disponibles suggèrent que le mécanisme de transfert des ICE ressemble à celui des plasmides conjugatifs (Bellanger et al., 2014; Delavat et al., 2017). Ce transfert se déroulerait en 5 étapes, la première et la dernière étape, respectivement l'excision et l'intégration de l'élément, étant propres au transfert des ICE (Figure 6). Le transfert le mieux caractérisé chez les ICE de firmicutes est celui d'ICE*Bs1* de *B. subtilis* (Auchtung et al., 2016). Tout d'abord, les

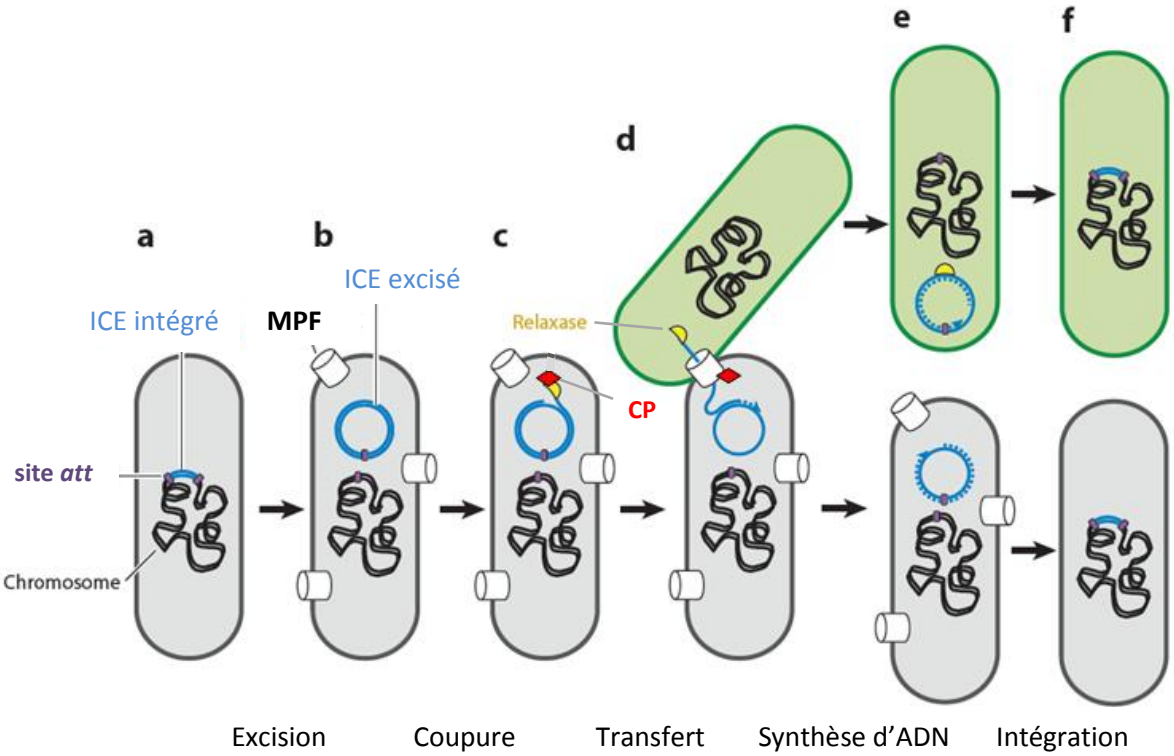

Figure 6 : Représentation schématique du cycle de vie des ICE d'après Johnson and Grossman (Annu. Rev. Genet. 2015). *Lorsque l'ICE est intégré, la plupart de ses gènes ne sont pas exprimés (a). Après activation, l'ICE s'excise sous forme double brin circulaire et le pore de conjugaison codé par l'ICE se forme (b). La relaxase codée par l'ICE coupe le brin qui va être transféré au niveau de son* oriT *et se fixe à son extrémité 5' de manière covalente formant ainsi un complexe ADN-relaxase (c). Le complexe ADN-relaxase est transporté depuis la bactérie donatrice vers la bactérie réceptrice au travers du pore de conjugaison (d). Dans la bactérie réceptrice, l'ADN simple brin est circularisé par la relaxase et le brin complémentaire est synthétisé. Dans la bactérie donatrice, le brin complémentaire de l'ADN simple brin non-transféré est synthétisé par réplication par cercle roulant (e). Les ICE double brin s'intègrent dans une molécule d'ADN dans les bactéries donatrice et réceptrice (f). La bactérie donatrice est représentée en fond gris et la réceptrice en fond vert. L'ADN de l'ICE est représenté en bleu et le chromosome bactérien en noir. Le rectangle rouge représente la protéine de couplage (CP), le demi-cercle jaune la relaxase et le cylindre blanc le pore de conjugaison (MPF). Les sites d'attachements (*att*) sont représentés en violet.*

dommages à l'ADN ou la présence d'un grand nombre de bactéries ne possédant pas ICE*Bs1* induisent l'excision de l'ICE par recombinaison site spécifique (Figure 6a-b) et l'expression des gènes impliqués dans la formation du pore du conjugaison (Auchtung et al., 2005, 2007; Bose et al., 2008; Bose and Grossman, 2011). Après excision de l'ICE sous forme double brin circulaire, la relaxase coupe l'un des deux brins au niveau du site *nic* de l'origine de transfert. La relaxase serait alors fixée de manière covalente à l'extrémité 5' du brin coupé, formant un complexe ADN-relaxase. Cette coupure initierait la réplication par cercle roulant de l'élément à partir de l'extrémité 3' dans la cellule donatrice (Figure 6b-c) (Lee et al., 2007, 2010, 2010; Thomas et al., 2013). Avant transfert, le complexe ADN simple brin-relaxase serait recruté au niveau de la membrane cellulaire par la protéine de couplage putative ConQ, et serait transporté via le pore de conjugaison de la cellule donatrice vers la cellule réceptrice (Figure 6c-d) (Gomis-Rüth et al., 2004; Iyer et al., 2004; Lee et al., 2007; Alvarez-Martinez and Christie, 2009; Lee et al., 2010). Après transfert, la molécule simple brin est alors circularisée dans la bactérie réceptrice par la relaxase et est répliquée. Cette réplication est facilitée par la présence de séquences *sso* (single-strand origin) présentes sur l'ICE et reconnues par des protéines de l'hôte (Figure 6d-e) (Lee et al., 2007; Wright et al., 2015). Enfin, une fois l'ICE sous forme double brin, l'intégrase à tyrosine, catalyse son intégration dans le génome des bactéries donatrice et réceptrice (Figure 5e-f). Contrairement aux plasmides, qui peuvent se maintenir sous forme extra-chromosomique, cette étape est nécessaire au maintien sur le long terme des ICE (Lee et al., 2007; Auchtung et al., 2016).

Bien que le transfert des IME soit très peu étudié, il est probable que celui-ci soit similaire à celui des ICE. Cependant, aucun des IME de firmicutes décrits avant ce travail, ne code de protéines du pore de conjugaison ou de protéine de couplage ; certains sont même dépourvus de relaxase. Le transfert de l'IME est donc dépendant des protéines fournies par d'autres éléments conjugatifs présents dans l'organisme (Bellanger et al., 2014). Néanmoins, l'étude de SGI1, un IME de *Salmonella enterica* codant des homologues distants des protéines Tra_G, Tra_N et Tra_H des plasmides IncA/C, démontre que, non seulement l'expression de ces protéines est induite par l'élément mobilisateur, mais qu'en plus ces protéines remplacent les protéines homologues du plasmide dans le pore de conjugaison. Ceci a pour conséquence de favoriser le transfert de SGI1 au détriment de celui des plasmides de type IncA/C (Carraro et al., 2017a). Dans ce cas, l'IME ne serait donc pas un

élément mobilisé, passif, comme le terme mobilisable pourrait le laisser entendre, mais plutôt un élément actif dans son transfert, piratant à son avantage le pore de conjugaison de l'élément mobilisateur.

4.1.2.2. Le module MOB des ICE et des IME

Le module MOB est impliqué dans la prise en charge de l'ADN transféré et parfois intervient aussi dans la réplication par cercle roulant extra-chromosomique de l'élément. Il contient au minimum l'*oriT* de l'élément, une séquence qui comporte des répétitions directes et inversées et en particulier qui génère généralement au moins une structure tige-boucle au niveau de laquelle la relaxase réalise la coupure simple brin (Carballeira et al., 2014). Ainsi l'*oriT* d'ICE*Bs1* comporte une répétition inversée constituant le site *nic* (**ACCCCCC**CACGCTAACA**GGGGGGGT**) (Lee and Grossman, 2007) qui est reconnue et coupée par sa relaxase.

Le module MOB des ICE et d'une grande partie des IME code une relaxase. Les relaxases, encore appelée MOB, font partie des endonucléases HUH, caractérisées d'une part par le motif conservé HUH consistant en deux résidus Histidine (H) séparés par un résidu hydrophobe et d'autre part par le motif Y comportant 1 ou 2 tyrosine(s) catalytique(s). Ces endonucléases englobent également les protéines impliquées dans l'initiation de la réplication par cercle roulant de nombreux petits plasmides ou de virus (Chandler et al., 2013). Ainsi, bien que toutes les enzymes appartenant au groupe des endonucléases HUH aient la même fonction biochimique, celle du clivage d'un brin d'ADN, elles sont impliquées dans des fonctions biologiques différentes telles que le maintien par réplication ou le transfert par conjugaison simple brin. Les relaxases retrouvées chez les éléments conjugatifs et mobilisables (plasmides conjugatifs, plasmides mobilisables, ICE et IME confondus) ont des tailles très variables pouvant aller d'environ 90 acides aminés à plus de 1 900 acides aminés. Les comparaisons et analyses phylogénétiques ont classé initialement les relaxases connues en 6 familles MOB (figure 7) (MOB_F, MOB_H, MOB_C, MOB_Q, MOB_P et MOB_V) (Francia et al., 2004; Garcillán-Barcia et al., 2009), toutes présentes chez les protéobactéries où elles ont été étudiées. Dans ce manuscrit, ces 6 familles MOB sont appelées relaxases canoniques car leur fonction biologique première est le transfert d'ADN. Des études ont révélé plus récemment de nouvelles familles de relaxases chez les firmicutes. Ainsi, une étude extrêmement récente suggère que la relaxase du plasmide pLS20 de *B. subtilis* est le

prototype d'une nouvelle famille de relaxase répandue chez les firmicutes, la famille MOB_L qui serait lointainement apparentée aux relaxases des familles MOB_P, MOB_V et MOB_Q (figure7)(Ramachandran et al., 2017). Des études récentes ont également montré que les relaxases de divers ICE de firmicutes appartenaient à une autre famille, MOB_T qui est apparentée aux initiateurs de réplication par cercle roulant de type Rep_Trans (domaine PF02486) impliqués dans la réplication par cercle roulant de divers petits plasmides (Garcillán-Barcia et al., 2009; Guglielmini et al., 2011). Des études récentes ont par ailleurs démontré que les relaxases MOB_T d'ICE*Bs1* de *Bacillus subtilis* et de Tn*916* étaient, non seulement impliquées dans l'initiation du transfert conjugatif de l'ICE et de sa réplication concomitante, mais aussi dans l'initiation de la réplication par cercle roulant de l'élément assurant son maintien sous forme excisée (Wright et al., 2015; Delavat et al., 2017). Par ailleurs, une autre étude a montré que des "initiateurs de réplication par cercle roulant" appartenant à une autre famille (Rep_1) étaient non seulement impliqués dans le maintien par réplication des trois petits plasmides de firmicutes qui les codent mais aussi, en tant que relaxases, dans leur mobilisation par ICE*Bs1* (Lee et al., 2012). Les familles MOB_F et MOB_H ne sont présentes dans aucun des systèmes conjugatifs connus des firmicutes. La famille MOB_Q a été établie à partir des relaxases du plasmide mobilisable RSF1010 d'*E. coli* (Rawlings and Tietze, 2001), du plasmide pTi d'*Agrobacterium*, du plasmide p42a de *Rhizobium* et du plasmide conjugatif pIP501 de *S. agalactiae* (Garcillán-Barcia et al., 2009).

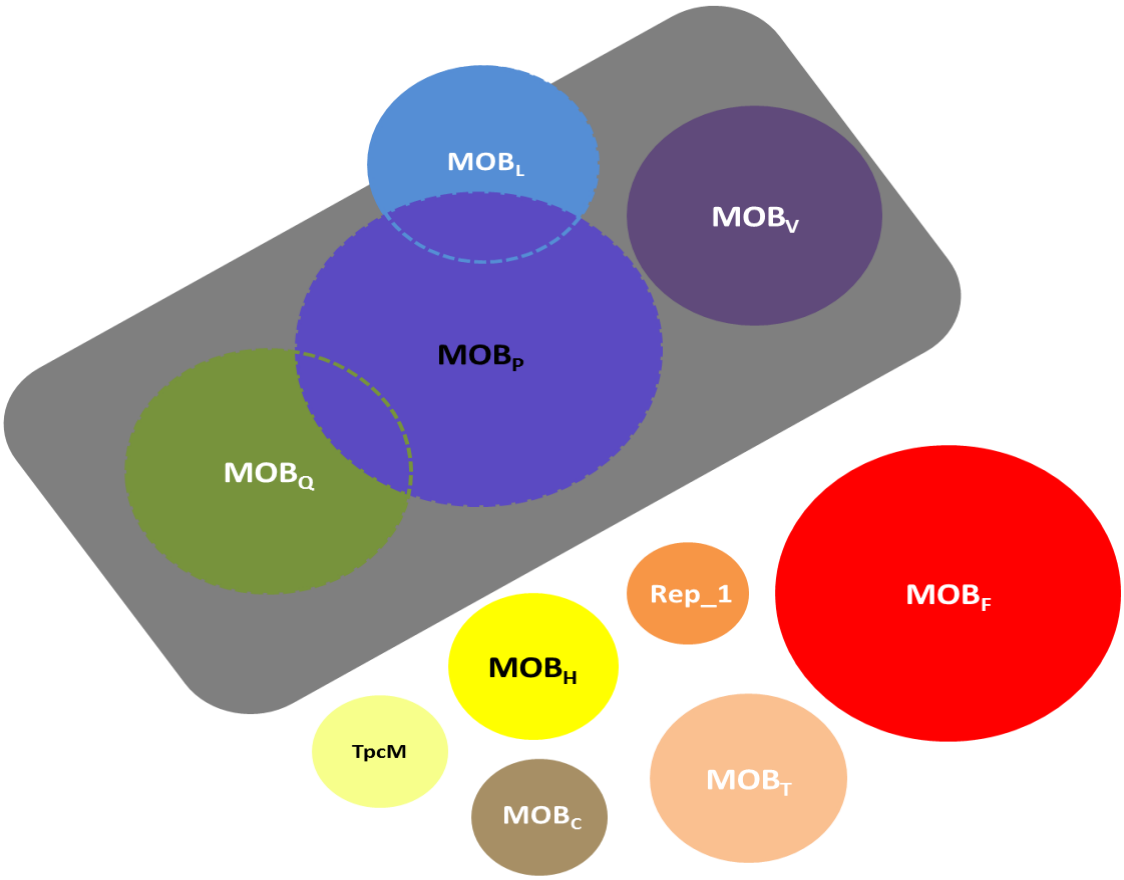

Figure 7 : Représentation schématique des relations de parentés entre les familles de relaxases (adapté de Garcillán-Barcia et al., 2009). *Les cercles se recoupent lors d'une parenté proche. Les relaxases dans le rectangle gris n'ont qu'une seule tyrosine catalytique. TcpM représente les relaxases apparentées à la relaxase de pcW3 ; Rep_1 représente les relaxases apparentées aux « initiateurs de la réplication par cercle roulant » retrouvées chez de petits plasmides de firmicutes. Les familles MOB$_T$ et MOB$_L$ représentent les 2 familles ajoutées récemment à la classification originale.*

Les relaxases de ces familles comportent un domaine relaxase en N-terminal avec une seule tyrosine catalytique et un domaine primase en C-terminal (Francia et al., 2004). Des relaxases de famille MOB$_Q$ ont été identifiées chez divers plasmides conjugatifs et quelques IME de streptocoques, comme IME*Sp2907* de *S. pyogenes* (Bellanger et al., 2014; Mingoia et al., 2016). La famille MOB$_P$, qui contient le plus de représentants, est relativement proche phylogénétiquement de la famille MOB$_Q$ (figure7). Leurs structures sont certainement identiques à celle de la relaxase Mob-A du plasmide RSF1010 appartenant à la famille MOB$_Q$. Elles comportent un domaine relaxase en N-terminal avec 1 seul tyrosine catalytique. Certaines d'entre elles comportent également un domaine primase en C-terminal. Des relaxases de famille MOB$_p$ ont été identifiées chez divers plasmides conjugatifs, plasmides

mobilisables, quelques IME et de nombreux ICE de streptocoques. Ainsi, on retrouve dans la famille MOB_P les relaxases des ICE Tn*GBS2* de *S. agalactiae* NEM316 (Brochet et al., 2009), *vanG-1* et *vanG-2* de *S. agalactiae* (Srinivasan et al., 2014) et les ICE apparentés aux ICE Tn*1549* d'*E. faecalis* (VanB, ICE*Se1*, ICE*Se2*)(Garnier et al., 2007; Heather et al., 2008; Hegstad et al., 2010) et Tn*5252* de *S. pneumoniae* BM6001 (ICE*Sp2905*, Tn*5253*, ICE-r) (Vijayakumar et al., 1986; Ayoubi et al., 1991; Brenciani et al., 2011). La famille MOB_L contient des relaxases de plasmides conjugatifs de firmicutes (Ramachandran et al., 2017). Par ailleurs, cette étude a montré que de nombreuses relaxases de cette famille sont codées par des gènes chromosomiques de firmicutes, mais n'a cependant pas identifié la nature des éléments qui les codaient ni même proposé qu'ils puissent être des ICE ou des IME. La famille MOB_v a été établie à partir de la relaxase MobM du plasmide mobilisable pMV158 de *S. agalactiae* (Guzmán and Espinosa, 1997). Cette famille comporte, entre autres, les relaxases du plasmide pBR1 de *Bordetella* (Szpirer et al., 2001) et de l'IME Tn*4555* de *Bacteroides* (Smith and Parker, 1998). Cette famille semble être lointainement apparentée à la famille MOB_P (figure7). La structure des relaxases de cette famille reste à déterminer ainsi que l'identification de leur(s) tyrosine(s) catalytique(s). Des relaxases de la famille MOB_v ont été identifiées chez divers plasmides mobilisables et quelques IME de streptocoques (Francia et al., 2004; Garcillán-Barcia et al., 2009; Bellanger et al., 2014; Lorenzo-Diaz et al., 2016). La famille MOB_C a été établie à partir des relaxases du plasmide mobilisable CloDF13 d'*E. coli* (Núñez and De La Cruz, 2001) et du plasmide conjugatif pAD1 du firmicute *Enteroccocus faecalis* (Francia and Clewell, 2002). Contrairement aux relaxases des autres familles MOB, les relaxases de la famille MOB_C ne se fixeraient pas de manière covalente au brin d'ADN lors du transfert, comme cela l'a été démontré pour le transfert de CloDF13 (Núñez and De La Cruz, 2001). Des relaxases de famille MOB_C ont été identifiées chez divers plasmides conjugatifs, plasmides mobilisables et ICE de firmicutes (Francia et al., 2004; Garcillán-Barcia et al., 2009), mais non chez des ICE ou IME de streptocoques. La famille MOB_T est une famille de relaxases de firmicutes récemment décrite et dont les structures restent à déterminer. Il semblerait que les relaxases de ces familles comportent un domaine catalytique avec une seule tyrosine catalytique et qu'elles se fixeraient de manière covalente au brin d'ADN transféré (Rocco and Churchward, 2006). On retrouve, au sein de la famille MOB_T, les relaxases d'ICE de divers streptocoques appartenant aux familles ICE*St3* et Tn*916* (Bellanger et al., 2014, 2009; Brochet et al., 2008; Puymège et al., 2015; Dahmane et al., 2017) ainsi que

les relaxases putatives d'IME de *S. agalactiae* (Douarre et al., 2015; Puymège et al., 2015). Les « initiateurs de réplication par cercle roulant » famille Rep_1 (dont certains au moins auraient aussi la fonction de relaxase) ne sont codés que par des petits plasmides de firmicutes (Lorenzo-Díaz et al., 2014). Globalement, avant ce travail, les relaxases codées par les ICE et IME de streptocoques connus n'appartenaient qu'à 4 familles, les familles MOB_T, MOB_P, Mob_Q, et MOB_V. Les relaxases décrites jusque maintenant chez les ICE et les IME portent toujours un domaine relaxase à proprement parler composé de deux sites catalytiques conservés, un site fixateur d'ions métalliques (généralement un motif HUH) et un motif Y avec une ou deux tyrosine(s) catalytique(s). Cependant, en plus du domaine relaxase ces protéines peuvent aussi porter d'autres domaines, comme des domaines hélicases (domaines qui déroulent et séparent l'ADN double brin), primases (domaines intervenant dans la réplication de l'ADN en synthétisant de petits fragments d'ARN utilisés par l'ADN polymérase) ou des domaines de fixation à l'ADN (Chandler et al., 2013). Il est à noter que les relaxases d'IME et d'ICE, bien que pouvant appartenir à la même superfamille, sont toutes phylogénétiquement éloignées.

Enfin, le module MOB peut parfois également coder d'autres protéines pouvant interagir avec la relaxase et/ou faisant partie du relaxosome. Ces protéines augmenteraient l'efficacité de coupure de la relaxase ou son affinité pour l'*oriT*. Dans le cas des éléments mobilisables, ces protéines pourraient augmenter la diversité des éléments conjugatifs capables de les mobiliser. Ainsi le module MOB des plasmides ColE1 d'*E. coli* et pC221 de *S. aureus* codent des protéines se fixant à l'*oriT* et qui sont nécessaires à sa coupure par la relaxase. Aussi, les plasmides mobilisables de la famille pCERC7 d'*E. coli* codent des protéines accessoires du relaxasome homologues à NikA du plasmide conjugatifs R64. Ces protéines homologues se fixent sur les origines de transfert des plasmides mobilisables et interagiraient avec la protéine NikB du plasmide R64. Elles faciliteraient ainsi la mobilisation des plasmides « pCERC7-like » (Furuya and Komano, 1995; Lanka and Wilkins, 1995; Fernández-López et al., 2014; Ruiz-Masó et al., 2015; Frances G. O'Brien et al., 2015; Moran et al., 2016).

4.1.2.3. La protéine de couplage et le module MPF des ICE

Les protéines de couplage des T4SS de bactéries Gram- appartiennent toutes à la famille VirD4. Cependant, les T4SS de Firmicutes incluent des CP appartenant soit à la famille VirD4

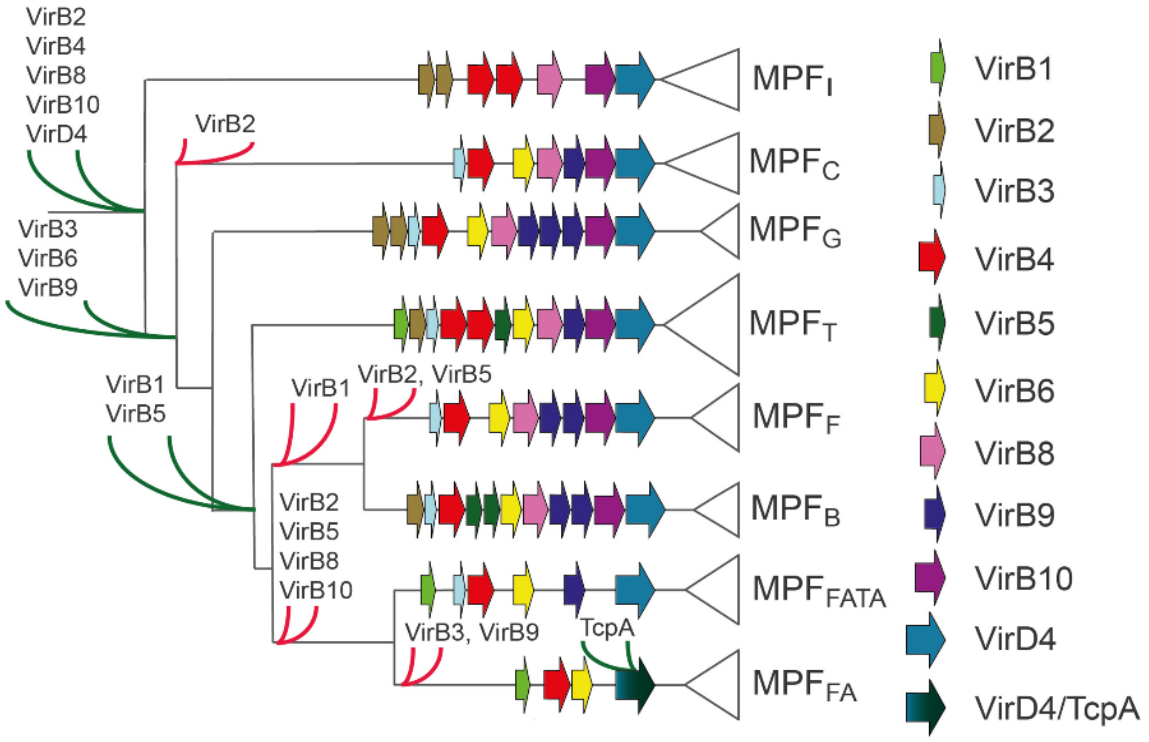

Figure 8 : Représentation des parentés entre MPF et des protéines les plus conservées au sein des différents types de MPF (Guglielmini et al., 2014). *L'arbre représenté suit la phylogénie de la protéine VirB4. Les traits verts représentent les gains de protéines et les traits rouges les pertes. La flèche bicolore pour VirD4/TcpA signifie que certains MPF$_{FA}$ possèdent une protéine de couplage de typeVirD4 tandis que d'autres possèdent une protéine de couplage de type TcpA.*

soit à la famille TcpA récemment identifiée. Les CP de type TcpA sont apparentées aux protéines FtsK (translocase d'ADN double brin impliquée dans la division cellulaire) et aux translocases TraSA impliquées dans le transfert conjugatif double brin des AICE et plasmides conjugatifs des actinomycètes (Guglielmini et al., 2013).

Les MPF retrouvés dans l'ensemble des bactéries sont très divers. La seule protéine suffisamment conservée au sein de ces modules pour avoir des homologues reconnaissables dans tous les T4SS est la protéine VirB4 (Guglielmini et al., 2011, 2014). En se basant sur le contenu en protéines et les relations phylogénétiques, les modules MPF peuvent être subdivisés en 8 classes (MPF$_F$, MPF$_I$, MPF$_G$, MPF$_T$, MPF$_B$, MPF$_C$, MPF$_{FA}$ et MPF$_{FATA}$) (figure 8). Aucun des groupes présents chez les bactéries Gram- n'existe chez les systèmes conjugatifs des firmicutes. Ceux-ci comprennent exclusivement les MPF$_{FA}$, MPF$_{FATA}$ et des systèmes

n'ayant pas fait l'objet de classement (comme ceux de l'ICE Tn*GBS1*) (Guglielmini et al., 2014).

- *protéines impliquées dans le transfert d'ADN simple brin de bactéries Gram-*

Les systèmes de conjugaison n'ont bien été étudiés que chez divers plasmides de protéobactéries. Les protéines impliquées dans le transfert d'ADN sont appelées protéines VirB1 à VirB11 pour celles qui constituent le pore de conjugaison et VirD4 pour la protéine de couplage. Ces protéines ont été les mieux décrites chez le plasmide pKM101 d'*E. coli*. Ces protéines peuvent être séparées en 3 classes :

- Les protéines du pilus VirB2 et VirB5 constituent la partie externe du système de conjugaison. Celles-ci permettent de rapprocher les cellules donatrice et réceptrice et de transférer l'ADN vers la cellule réceptrice. VirB1 est impliquée dans la synthèse du pilus et dans sa stabilité.
- VirB7, VirB9, VirB10 se situent au niveau de la paroi et la membrane externe et elles constituent la « partie haute » du système (figure 9). Les protéines de translocation VirB3, B4, B6, B8, B10 sont localisées au niveau de la membrane interne et constituent la « partie basse » du système (figure 9). VirB7 guiderait le complexe lors de son insertion dans la membrane externe une fois celui-ci assemblé tandis que VirB3 et VirB6 aideraient à le stabiliser. La protéine VirB10 interagirait avec les ATPases du T4SS (VirB4, VirB11 et VirD4). Enfin, les fonctions

des protéines VirB8 et VirB9 ne sont pas encore définies.

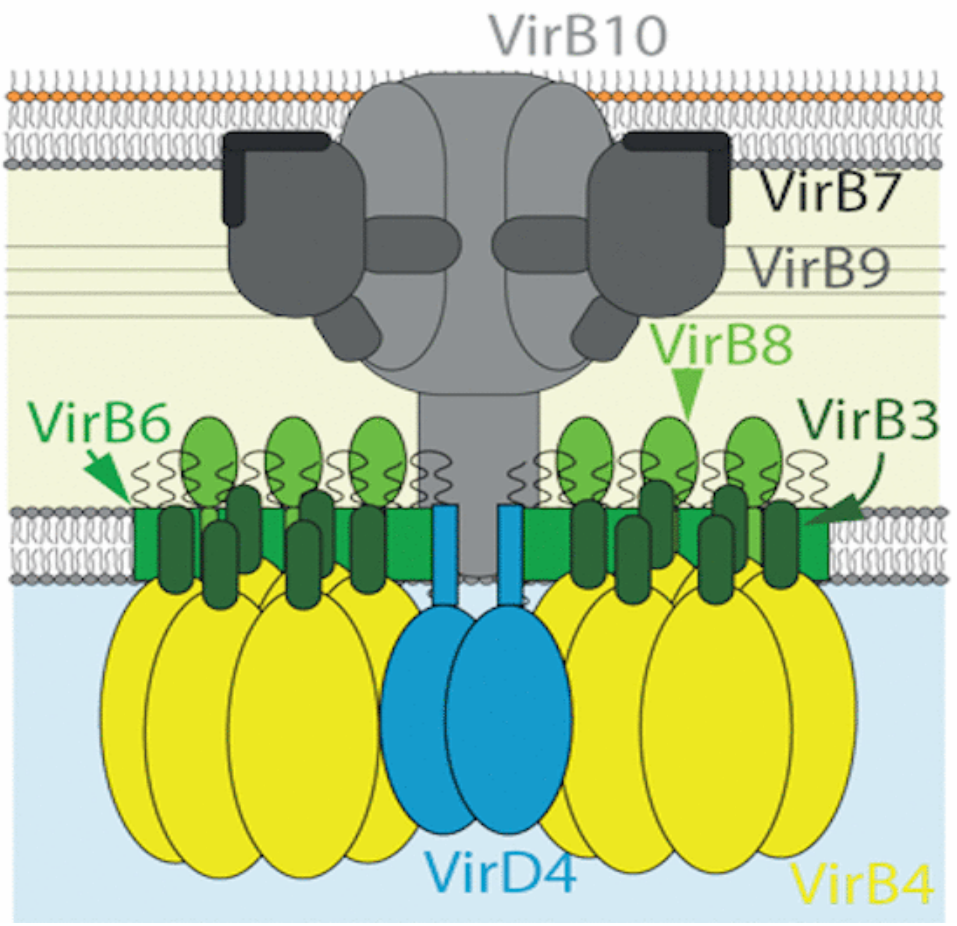

Figure 9 : Représentation schématique d'une vue en coupe du T4SS du plasmide R388 chez *E. coli* (d'après Adam Redzej et al,. 2017). *Le complexe T4SS contenant les 8 protéines VirB3-10 est fixé à la protéine VirD4 qui se situe au centre du complexe. La protéine VirB5 située au niveau central de la partie « haute » du T4SS n'est pas représentée.*

- VirB4, VirB11 et VirD4 (la protéine de couplage) sont les 3 ATPases fournissant de l'énergie au système. VirB4 est située dans la membrane interne et interagirait avec VirB3, VirB6 et VirB8 (figure 9). Deux dimères de VirD4 seraient localisés au niveau de la membrane interne, de part et d'autre du T4SS entre les deux hexamères de VirB4, cependant la structure et la position de VirD4 pendant le transfert restent inconnues (Redzej et al., 2017). La protéine de couplage VirD4 conduirait le relaxosome fixé alors à l'ADN à transférer, jusqu'au T4SS, permettant le passage du complexe ADN-relaxase. Il a été proposé que VirD4 pourrait assurer le transfert de l'ADN et VirB4 celui de la relaxase (Redzej et al., 2017). La position exacte de VirB11, protéine présente dans une partie seulement

des T4SS de protéobactéries, est incertaine (Walldén et al., 2012; Ilangovan et al., 2015)

- *protéines impliquées dans le transfert d'ADN simple brin des bactéries Gram+*

Aucune structure de T4SS des bactéries Gram+ n'est actuellement connue. Cependant les comparaisons de séquences entre les protéines des MPF de bactéries Gram+ et de bactéries Gram- révèlent des homologues des protéines de la partie « basse », ancrées dans la membrane interne chez les firmicutes, à l'exclusion de VirB11 (Alvarez-Martinez and Christie, 2009; Goessweiner-Mohr et al., 2013; Leonetti et al., 2015). On retrouve un homologue de l'hydrolase de peptidoglycane VirB1 et de l'ATPase VirB4. De plus, la plupart des éléments conjugatifs de Gram+ se transférant sous forme simple brin (si ce n'est tous) codent des protéines avec des structures similaires ou prédites comme similaires aux protéines VirB3, VirB6 et VirB8. Ces protéines sont notamment codées par le plasmide pIP501 (MPF_{FATA}) et par ICE*Bs1* de *Bacillus subtilis* (MPF_{FA}). Les éléments de type FATA codent des CP de famille VirD4 tandis que la plupart des éléments de type FA codent des CP de famille TcpA. Cependant, les protéines de la partie « haute », qui sont ancrées dans la membrane externe, seraient toutes absentes des T4SS de firmicutes (Bhatty et al., 2013; Goessweiner-Mohr et al., 2013). Cette différence avec les systèmes de protéobactéries semble logique puisque les Firmicutes (à l'exception des négativicutes dont les systèmes de conjugaison ne sont pas connus) sont des bactéries Gram+ et ne possèdent donc pas de membrane externe.

Le modèle d'étude le plus abouti pour les T4SS de Gram+ est celui du plasmide pIP501 pour lequel quelques interactions protéiques ont été mis en évidence (indiquées par des flèches, figure 10), les données concernant les T4SS d'ICE ne sont, quant à elles, pour l'instant que très fragmentaires. Les T4SS de Gram- forment un pilus lors du transfert, cependant à ce jour, aucun T4SS de Gram+ formant de pili n'a été observé. Les seules protéines de T4SS de Gram+ prédites comme étant localisées du côté extérieur de la membrane cytoplasmique sont les protéines homologues de VirB1 et VirB8. VirB6 étant conservée au sein des T4SS de Gram+ et Gram-, il est envisageable que le complexe VirB1/VirB8 s'associe à VirB6 pour former le complexe protéique traversant la membrane cytoplasmique, comme cela est observé pour les systèmes de Gram- (Bhatty et al., 2013; Goessweiner-Mohr et al., 2013). De plus, contrairement aux T4SS de Gram+, VirB1 semble être indispensable pour le transfert conjugatif des éléments de Gram+ (Arends et al., 2013; DeWitt and Grossman, 2014). Les

protéines VirD4 et VirB4, fournissant de l'énergie au système sont toujours présentes au sein des T4SS de bactéries Gram-. Bien que les données au sujet des interactions entre et avec ces sous-unités au sein des T4SS de Gram+ soient limitées, les informations disponibles à ce jour suggèrent que les protéines homologues au sein des T4SS de Gram+ joueraient le même rôle (Abajy et al., 2007; Teng et al., 2008; Steen et al., 2009; Porter et al., 2012).

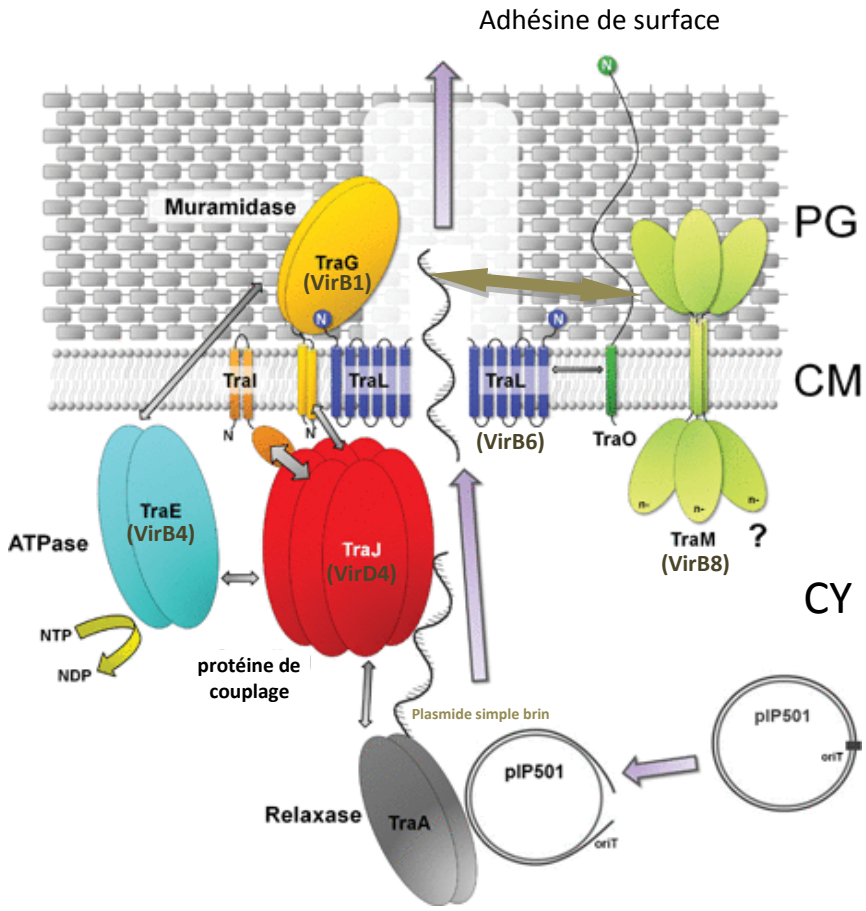

Figure 10 : Modèle de localisation et d'interaction des protéines du T4SS du plasmide pIP501 de *S. agalactiae* (adapté de Goessweiner-Mohr et al., 2014). *La protéine TraG (homologue de VirB1) s'associe à TraM (homologue de VirB8) pour former le corps du pore de conjugaison. TraG est aussi ancrée dans la membrane cytoplasmique où elle interagit avec TraL, homologue de VirB6, afin d'y former un canal. TraG interagit aussi avec TraE, homologue de VirB4, cette dernière fournissant l'énergie au T4SS. La protéine de couplage TraJ, de famille VirD4, va interagir avec TraG et TraA (la relaxase) afin de mettre en relation le relaxosome avec le pore de conjugaison. Enfin, TraL, homologue de VirB6, interagit avec TraO qui intervient dans l'adhésion entre la cellule donatrice et la cellule réceptrice. Les flèches représentent les interactions connues entre les protéines. Les extrémités N-terminale des protéines sont marquées par un N. PG, peptidoglycane; CM membrane cytoplasmique; CY pour cytoplasme*

Malgré le fait que les interactions de certaines protéines du T4SS de firmicutes soient bien définies et que des prédictions par homologies avec les T4SS de protéobactéries puissent

être faites, les données que nous avons à ce jour sont fragmentaires. La structure et le mécanisme de fonctionnement des T4SS de firmicutes restent à élucider.

4.1.3. Les modules d'adaptation portés par les ICE et IME

Un module d'adaptation peut être défini comme l'ensemble des gènes n'intervenant pas dans le cycle de vie de l'ICE ou de l'IME, et pouvant conférer un avantage sélectif pour l'organisme hôte dans certaines situations.

Les toutes premières études portant sur les ICE ont découlé de l'étude du transfert de résistances aux antibiotiques (Mays et al., 1982; Rashtchian et al., 1982; Magot et al., 1983). Ainsi le premier ICE clairement identifié, Tn*916*, code une résistance à la tétracycline (Franke and Clewell, 1981). Depuis, de nombreux autres ICE portant des résistances aux antibiotiques variées ont été identifiés (Bellanger et al., 2014). Par ailleurs, l'ICE SXT de *V. cholerae* et ses apparentés porte un intégron qui transporte les résistances au sulfaméthoxazole, à la triméthoprime et à la streptomycine (Waldor et al., 1996). Aussi, l'ICE CTnDOT de *Bacteroides thetaiotaomicron* code non seulement une résistance à la tétracycline mais aussi une résistance à l'érythromycine (Cheng et al., 2000). L'étude des résistances aux métaux lourds ont aussi permis de mettre en évidence certains ICE. Ainsi, l'ICE R391 de *Providencia rettgeri* confère à son hôte la résistance au mercure (Böltner and Osborn, 2004). Enfin certains ICE, comme ICE*St1* et ICE*St3* de *S. thermophilus* portent des systèmes de restriction-modification (R-M) (Burrus et al., 2001; Bellanger et al., 2009). Ces systèmes codent une endonucléase de restriction qui coupe l'ADN et une enzyme de méthylation qui va modifier l'ADN ce qui le protège de la coupure par l'endonucléase. Ces systèmes R-M peuvent conférer un avantage à l'hôte en lui conférant un moyen de défense contre les bactériophages.

Les modules d'adaptation retrouvés sur les ICE et les IME sont très variés et ne s'arrêtent pas aux résistances aux antibiotiques et métaux lourds. L'acquisition de certains ICE ou IME va parfois engendrer un changement notable, voire drastique, du mode vie des organismes qui les reçoivent. Par exemple, certains ICE vont permettent à leur hôte d'exploiter certaines sources de carbone. C'est le cas d'ICE*clc* qui permet à *Pseudomonas* B13 d'utiliser des xénobiotiques, les chlorocatéchols, comme seule source de carbone (Ravatn et al., 1998) ou encore de l'ICE CTn*Scr94* de *Salmonella* (Hochhut et al., 1997) et Tn*5276* de *Lactococcus*

lactis (Rauch and De Vos, 1992) qui permet à leurs l'hôte de fermenter le saccharose. De plus, Tn*5276* code également la synthèse d'une bactériocine, c'est-à-dire d'un peptide ayant des propriétés antibactériennes orientées vers des espèces phylogénétiquement proches de la souche productrice. D'autres ICE sont des ilots de pathogénicité comme PAPI-1 de *Pseudomonas aeruginosa* (Carter et al., 2010) ou encore PAISt de *Streptomyces turgidiscabies* (Kers et al., 2005). Au contraire, certains ICE permettent à leur hôte de rentrer en symbiose avec d'autres organismes. Ainsi, le transfert d'ICE*Ml*Sym^{R7A} à des bactéries du sol leur confère, à lui seul, la capacité à entrer à symbiose avec les racines du lotier corniculé et à fixer l'azote atmosphérique (Sullivan et al., 1995).

4.2. La prévalence des ICE et des IME

Le transfert conjugatif n'a été démontré que pour un nombre relativement limité d'ICE différant par leurs gènes de transfert et pour seulement une quinzaine de types d'IME différents à ce jour (Bellanger et al., 2014; Carraro et al., 2017b; Delavat et al., 2017). Cependant, quelques études récentes de génomes suggèrent que les ICE et les IME seraient extrêmement répandus dans les génomes bactériens et pourraient être les principaux responsables des transferts conjugatifs (pour revue, voir Bellanger et al., 2014). Ainsi, seulement 4 recherches exhaustives d'ICE, et une seule d'IME, ont été réalisées en dehors de ce travail dans un nombre significatif de génomes d'un taxon bactérien donné. La première recherche d'ICE sur 22 génomes complets ou incomplets de firmicutes, réalisée en 2002 par comparaison avec les rares modules de conjugaison et d'intégration d'ICE et de plasmides de firmicutes disponibles à l'époque, a permis de détecter, dans 6 des 22 génomes, 17 ICE putatifs (dont 7 dans *Clostridiodes difficile* 630 et 6 dans *Enterococcus faecalis* V583) (Burrus et al., 2002a). La seule recherche exhaustive d'ICE et d'IME, qui est aussi l'unique étude spécifique aux génomes de streptocoques, réalisée en 2008, a permis de détecter seize ICE, onze IME et vingt-cinq éléments en dérivant dans huit génomes de *Streptococcus agalactiae* (Brochet et al., 2008). La troisième étude, réalisée en 2011, sur 11 souches de *C. difficile*, a permis de mettre en évidence 30 ICE (Brouwer et al., 2011). Enfin, la quatrième recherche exhaustive d'ICE à partir de profils HMM de protéines codées par des élément conjugatifs d'actinobactéries (essentiellement des éléments se transférant en double brin) dans 275 génomes d'actinobactéries a permis de détecter 144 AICE et 17 ICE se transférant sous forme simple brin (Ghinet et al., 2011). De plus, la recherche de gènes de conjugaison sur

1124 génomes (Guglielmini et al., 2011) (dont 72 actinobactéries) à l'aide de profil HMM établis à partir d'une base de données provenant essentiellement de plasmides conjugatifs de protéobactéries (sans aucun élément se transférant en double brin) a permis l'identification de 335 modules chromosomiques de conjugaison. Cependant, aucun des éléments détectés par l'étude de Ghinet at al. (2011) ne l'a été par l'étude de Guglielmini et al. (2011) et inversement. Ce manque de recoupement dans les résultats de ces deux études souligne la difficulté d'analyse d'un très large panel de génomes de groupes éloignés, notamment en raison de la diversité importante des ICE. Ceci suggère également que des ICE ont échappé à ces analyses. Par ailleurs, la recherche de gènes de conjugaison sur les 1124 génomes suggère que les IME seraient nettement plus répandus que les ICE qui, eux-mêmes seraient nettement plus répandus que les plasmides conjugatifs (Guglielmini et al., 2011).

Malgré leur abondance dans les génomes bactériens, les ICE ou IME demeurent mal connus car ils sont difficiles à détecter et pour les IME, difficiles à distinguer des autres éléments mobiles du fait du très petit nombre (ou de l'absence) de gènes codant des protéines impliquées dans la conjugaison (Bellanger et al., 2014). De ce fait, bien que de très nombreux ilots génomiques aient été identifiés et annotés, extrêmement peu sont annotés comme ICE ou IME dans les génomes actuellement disponibles dans les bases de données.

4.3. Les éléments composites

Les ICE, IME et ilots génomiques présentent souvent une structure composite (Bellanger et al., 2014). Ces éléments composites sont en fait des assemblages d'éléments plus petits qui peuvent avoir subi des réorganisations. Ils peuvent ainsi porter d'autres éléments mobiles, non seulement des éléments codant des transposases à DDE comme des IS (pouvant provoquer des délétions) ou des transposons, mais aussi des ICE ou des IME. Ainsi, divers éléments de la famille Tn*5252*, famille répandue chez les streptocoques, portent des ICE de type Tn*916* intégrés de manière non-spécifique (Mingoia et al., 2011). Ils forment ainsi des éléments composites, constitués de deux éléments, l'un inséré dans un autre. Des ICE peuvent également héberger des IME intégrés de façon site-spécifique (figure 11). Ainsi, les ICE ICE*2096-RD.2* de *Streptococcus pyogenes* (Beres and Musser, 2007), ICE*Sp2905* de *S. pyogenes* (Brenciani et al., 2011) et Tn*6103* de *C. difficile* (Brouwer et al., 2011) hébergent respectivement 1, 2 et 3 IME. Dans certains cas, un ICE peut héberger un autre ICE qui héberge lui-même des IME. Ainsi, un ICE de *Streptococcus lutetiensis* apparenté à Tn*5252*,

ICE*Sluvan* porte un ICE de type Tn*1549* qui héberge lui-même un IME putatif (Bjørkeng et al., 2013; Bellanger et al., 2014).

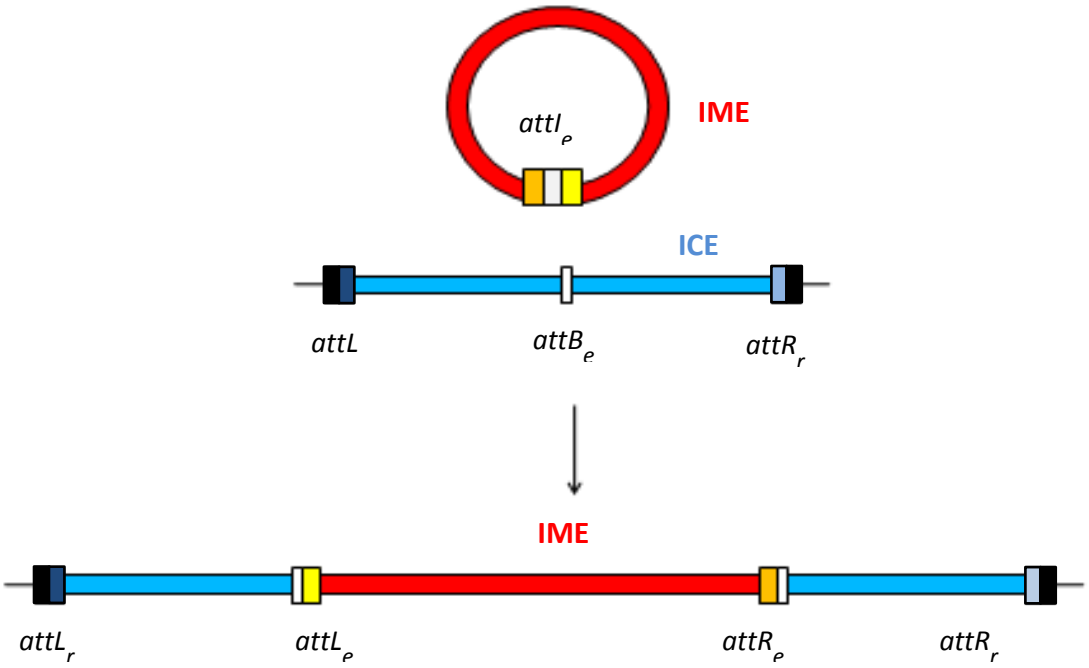

Figure 11 : Formation d'un élément composite par intégration site-spécifique d'un élément entrant (IME) dans un élément résident (ICE). *L'intégrase codée par l'IME catalyse la recombinaison site-spécifique entre les séquences identiques des sites* attI$_e$ *et* attB$_e$ *portées par l'ICE et l'IME conduisant à leur coïntégration et à la formation de sites* attL$_e$ *et* attR$_e$. *Les éléments entrant (IME) et résident (ICE) sont respectivement représentés en trait épais rouge et bleu. Le chromosome de l'hôte est représenté en trait fin noir. Les sites att de l'élément entrant sont marqués par un « e » et ceux de l'élément résident par un« r ». Les rectangles blancs représentent une séquence identique entre les sites* attI *et* attB. *Les rectangles noirs représentent une séquence identique entre les sites* attL$_r$ *et* attR$_r$. *Les rectangles bleus et jaune, clairs et foncés, représentes les bras des sites* attL$_r$, attR$_r$, attL$_e$ *et* attR$_e$ *séquences avec lesquelles l'intégrase se lie avec grande affinité.*

De plus, les éléments qui s'intègrent de façon site spécifique peuvent, s'intégrer non seulement dans le site *attB* mais également dans le site *attL* ou *attR* flanquant un élément résident de même spécificité d'intégration. Ceci conduit à la formation d'un élément composite formé de deux éléments intégrés en tandem (Bellanger et al., 2014). L'élément composite produit par cette accrétion est alors formé des 2 éléments intégratifs séparés par un site *attI* chimérique et peut avoir 2 structures différentes (figure 12). Trois possibilités d'excision sont envisageables à partir de ces structures, soit l'excision du premier élément seulement, soit celle du second seulement, soit celle de l'élément composite dans son ensemble.

La formation d'éléments composites, constitués d'éléments intégrés en tandem ou intégrés les uns dans les autres, peut constituer la première étape d'une mobilisation en *cis* d'éléments génétiques mobiles. En effet, lorsque de tels éléments s'excisent, et qu'au moins un des deux éléments porte une *oriT*, il est envisageable que le transfert de l'un, entraine simultanément le transfert de l'autre. L'élément composite pourrait ensuite s'intégrer dans le génome de la bactérie réceptrice. Des cas de mobilisation en *cis* de CIME (cis-Mobilizable Element) par des ICE ont déjà était décrits dans la littérature (Bellanger et al., 2011; Zhang and Loria, 2017). Ainsi, l'étude du transfert par conjugaison à partir d'une souche portant ICE_*515_tRNA*^{Lys} de *S. agalactiae*, intégré en accrétion avec un CIME, CIME_*Nem_tRNA*^{Lys}, a révélé non seulement le transfert de l'ICE seul mais aussi celui de la forme composite

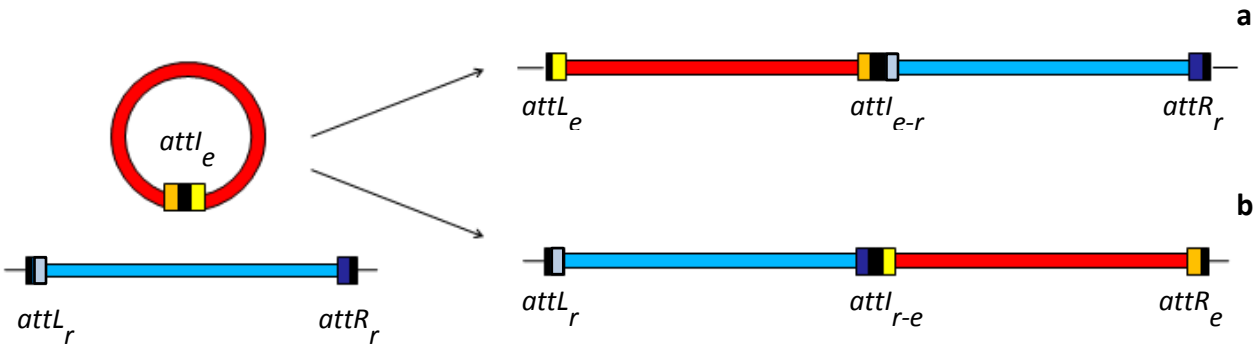

Figure 12 : Formation d'un élément composite par accrétion d'un élément entrant avec un élément résident. *L'intégrase codée par l'élément entrant catalyse la recombinaison site-spécifique entre les séquences identiques des sites* attI$_e$ *et* attL$_r$ *(cas a) ou* attR$_r$ *(cas b) situées aux extrémités de l'élément résident conduisant à la formation d'un élément composite séparé par un site* attI$_{e-r}$*. Les éléments entrant et résident sont représentés respectivement en trait épais rouge et bleu. Le chromosome de l'hôte est représenté en trait fin noir. Les sites* att *de l'élément entrant sont marqués par un « e » et ceux de l'élément résident par un« r ». Les rectangles noirs représentent une séquence identique entre les sites* attL$_e$*,* attR$_e$*,* attL$_r$ *et* attR$_r$*.*

ICE_515_tRNALys-CIME_Nem_tRNALys dans son ensemble (Puymège et al., 2013).

5. Les approches bio-informatiques pour l'identification et la localisation d'ICE et d'IME dans les génomes de procaryotes

5.1. La détection d'ilots génomiques et ses limites

De par leur nature modulaire, leur extrême diversité et étant souvent composés de différents éléments génétiques mobiles, les ilots génomiques sont compliqués à détecter dans les génomes bactériens. En effet, choisir seulement quelques critères de détection

comme le biais d'utilisation des codons ne permettra pas de détecter les ilots génomiques ne répondant pas à ces critères. Par exemple, la différence de pourcentage en G+C ne permet de détecter les ilots génomiques que si le pourcentage en G+C de l'élément est différent de celui du génome de la bactérie. De plus, ce pourcentage varie en fonction des différents éléments génétiques et souvent au sein même des éléments. Ainsi le pourcentage en G+C d'ICE*St1* de *S. thermophilus* varie de 26 % à 42 % (Burrus et al., 2002b). Il est également difficile d'utiliser plusieurs critères de détection comme le pourcentage en G+C et la présence de protéine de transfert ou de virulence : en effet, de par leur diversité, seulement quelques éléments répondront à tous les critères de détection à la fois.

Une des difficultés rencontrées lors de la détection d'ilots génomiques et d'éléments génétiques mobiles est le manque d'une base de données de référence pour comparer les résultats obtenus afin d'estimer l'efficacité de la méthode de détection. Dans le cadre des ICE, il existe la base de données ICEBerg (Bi et al., 2012). Cependant cette base de données n'est pas tenue à jour et contient un grand nombre d'erreurs relatives à la nature des éléments ou à leur classification (Bellanger et al., 2014). En effet, des éléments répertoriés comme ICE ne portent ni module d'intégration ni module de conjugaison. Par ailleurs, cette base, censée n'inclure que des ICE, comporte d'autres éléments génétiques mobiles comme des prophages, référencés dans la base comme ICE.

5.2. Les différentes méthodes de détection

Au moment où ce travail a commencé, aucune méthode dédiée spécifiquement à la détection d'ICE ou d'IME n'était disponible. Cependant de nombreuses méthodes de détection d'ilots génomiques ont été développées. Ainsi, à ce jour, plus d'une trentaine de méthodes sont à la disposition de la communauté scientifique.

5.2.1. Les méthodes d'analyse d'un génome

Diverses méthodes sont basées sur les différences de composition des séquences acquises par transfert horizontal par rapport à celle des gènes d'ossature transmis de façon verticale. Elles s'appuient sur l'idée que les pressions de sélection et de mutations s'appliquant à une espèce résultent en une composition en nucléotides propre à celle-ci. Ainsi les séquences acquises par transfert horizontal pourraient avoir une composition en nucléotides, ou combinaison de nucléotides dans des positions adjacentes, différente du reste du génome ce

qui permettrait de les détecter (Lee et al., 2013) (figure 13 – méthodes basées sur un génomes). Ces méthodes utilisent différents critères tels que le pourcentage en G+C, les biais dans l'usage des codons ou des acides aminés, ou encore la fréquence d'utilisation d'oligonucléotides. Elles peuvent être classées en trois catégories : les méthodes basées sur l'analyse de la composition des gènes, les méthodes basées sur l'analyse des séquences d'ADN et les méthodes basées sur la structure des ilots génomiques (Lu and Leong, 2016).

5.2.1.1. Les méthodes basées sur l'analyse de la composition des gènes

Les méthodes basées sur l'analyse de la composition des gènes sont pensées pour détecter les gènes acquis par transfert horizontal, mais elles ne sont pas très efficaces pour détecter les ilots génomiques ou éléments génétiques mobiles. Alors qu'elles pourraient être utilisées pour détecter, par exemple, des IS, elles ne conviendraient pas, utilisées seules, pour la détection d'ICE et d'IME puisque ces derniers sont composés de différents modules pouvant avoir eu des origines ou des évolutions différentes. De plus, bien que ces méthodes soient généralement faciles à implémenter, elles reposent sur une bonne annotation des gènes et peuvent conduire à la détection de nombreux faux positifs. En effet certains gènes du génome peuvent avoir une composition en nucléotique différente du reste du génome, comme les gènes fortement exprimés tels que les gènes de protéines ribosomiques ou au contraire peu exprimés.

Par exemple, SIGI-HMM (Waack et al., 2006) ou IslandPath-DINUC (Hsiao et al., 2003) (figure 13 – méthodes basées sur la composition des gènes) s'appuient sur un seul critère : le biais d'utilisation des codons pour la première et la composition en dinucléotides des gènes pour la seconde.

5.2.1.2. Les méthodes basées sur l'analyse de la composition des séquences d'ADN.

Les méthodes basées sur l'analyse de la composition des séquences d'ADN ont été pensées pour répondre à une demande croissante de ce genre d'analyse dû au progrès du séquençage haut débit fournissant des séquences d'ADN non assemblées et non annotées. Elles ne diffèrent pas beaucoup des méthodes basées sur la composition des gènes et elles ont les mêmes avantages et inconvénients. La différence réside dans le fait qu'elles ne nécessitent pas une annotation des génomes mais analysent directement la séquence de

l'ADN. Certaines méthodes traitent des sous-ensembles de génomes de taille réglable ou non (méthodes fenêtrées). Ainsi, la méthode fenêtrée AlienHunter (Vernikos and Parkhill, 2006) base sa recherche sur la détection d'utilisation de motifs/oligonucléotides atypiques au sein des séquences. D'autres méthodes traitent les génomes, ou séquences données, dans leur ensemble (méthode non fenêtrées). Ainsi, la méthode non fenêtrée GC-Profile (Gao and Zhang, 2006) utilise les variations en G+C tout au long du génome afin de déterminer les séquences horizontalement acquises (figure 13-méthodes basées sur la composition de l'ADN).

Bien que ces méthodes aient, par le passé, permis de détecter des ilots génomiques putatifs (Elhai et al., 2012), elles ne semblent pas être les mieux adaptées pour la détection d'ICE et d'IME. En effet, les IME sont généralement de petits éléments de quelques kilobases (kb) difficilement détectables par ces méthodes. De plus, les ICE ont généralement des pourcentages en G+C variables en fonction des différents modules, ce qui peut conduire à une mauvaise délimitation de ces éléments. Ces caractéristiques sont notamment retrouvées pour les éléments de famille ICE*St3*/Tn*916*/ICE*Bs1* (Burrus et al., 2002b). Ainsi, le module de conjugaison d'ICE*St1* a un pourcentage en G+C d'environ 42%, tandis que celui de son module d'intégration est de 34% et que des régions du module d'adaptation descendent jusqu'à 26%. De plus, ces méthodes sont généralement capables de détecter les éléments ayant un faible pourcentage en G+C par rapport au reste du génome (Lu and Leong, 2016), ce qui ne convient guère aux génomes de streptocoques ayant déjà un pourcentage en G+C relativement bas.

5.2.1.3. Les méthodes basées sur la structure des îlots génomiques

En plus des critères cités précédemment, des méthodes s'appuient sur les connaissances acquises sur les ilots génomiques. Ainsi, IslandPath utilise non seulement le biais de pourcentage en G+C et celui d'utilisation en dinucléotide, mais aussi la présence de gènes de mobilité tels que les gènes codant des intégrases (Arvey et al., 2009). D'autres méthodes comme RVM (Vernikos and Parkhill, 2008), ou GIDetector (Che et al., 2010) s'appuient sur le « machine learning ». L'approche consiste à fournir au programme un ensemble d'ilots génomiques déjà connus afin que celui-ci détermine des paramètres tels que le site d'insertion, la présence d'intégrases, la densité en gènes, la taille de l'ilot ou le pourcentage

en G+C, afin de détecter automatiquement de nouveaux ilots dans les génomes traités (figure 13-méthodes basées sur la structure des ilots génomiques).

Ces méthodes sont généralement plus robustes que les méthodes précédemment décrites, mais elles dépendent grandement de la qualité de la base fournie au programme et donc du degré de connaissance déjà acquises sur le type d'éléments recherchés. De plus, certaines de ces méthodes ne font que pointer des régions du génome en laissant à l'utilisateur le soin de déterminer si ces régions sont, ou non, des ilots génomiques. Elles reposent donc également sur le degré d'expertise de l'utilisateur.

5.2.2. Les méthodes basées sur la comparaison de plusieurs génomes

Les méthodes basées sur la comparaison de plusieurs génomes utilisent le principe de la synthénie. Elles reposent sur l'alignement de plusieurs génomes de séquences proches avec BLAST ou MAUVE afin de détecter des ilots génomiques putatifs (figure 13 – Méthodes basées sur plusieurs génomes). Ces méthodes sont utilisées par IslandPick (Langille et al., 2008) ou MobilomeFinder (Ou et al., 2007) et sont généralement plus efficaces que toutes les méthodes décrites précédemment. Cependant, ces méthodes ont leurs limites car il est nécessaire d'avoir des génomes proches à disposition pour faire ces comparaisons (Langille et al., 2010). De plus, il est parfois difficile de faire la différence entre gènes acquis par transfert horizontal et perte de gènes (Ravenhall et al., 2015). Enfin, certains réarrangements chromosomiques peuvent rendre difficile l'interprétation des alignements de séquences (Darling et al., 2004).

5.2.3. Les méthodes combinatoires.

Différents programmes tels qu'Islandviewer (Dhillon et al., 2013), EGID (Che et al., 2011) ou PIPs (Soares et al., 2012) utilisent une combinaison de méthodes précédemment décrites (figure 13 – Méthodes combinatoires). Par exemple, Islandviewer combine les programmes SIGI-HMM, IslandPath et IslandPick apportant en plus un système d'annotation des gènes de l'ilot putatif afin de permettre à l'utilisateur de déterminer plus facilement si la région détectée correspond réellement à un ilot. Cette étape, bien que n'influençant pas le taux de faux négatifs, permet de réduire le taux de faux positifs. PIPs, quant à lui, utilise différents paramètres tels que le pourcentage en G+C, les biais dans l'usage de codons, la présence de

protéines hypothétiques et de facteurs de virulence afin de définir si un ilot putatif est un ilot de pathogénicité.

5.3. Les méthodes spécialisées

Les méthodes décrites précédemment permettent de détecter des ilots génomiques. Cependant, elles présentent, à des degrés divers, des problèmes de faux positifs, de faux négatifs ou de délimitation des ilots. De plus, aucune de ces méthodes n'est spécialisée dans la recherche d'éléments chromosomiques capables de se transférer par conjugaison que ce soit des ICE ou des IME. La seule méthode automatisée et systématique se rapprochant le plus d'une recherche d'ICE et d'IME est la méthode employée par Guglielmini et al 2011. En effet, cette approche consiste à rechercher des modules de conjugaison à l'aide de profils HMM (Hidden Markov Model) créés à partir de différentes protéines de ce module provenant d'éléments conjugatifs et mobilisables connus, principalement dans ce cas provenant de plasmides de protéobactéries. Lors de la recherche, ces profils servent de référence et permettent de localiser des protéines apparentées dans les génomes analysés. Bien que cette méthode ne démontre pas la présence d'ICE ou d'IME, puisque ne recherchant pas les modules d'intégration ou les limites de ces éléments, elle reste cependant intéressante. En effet, elle permet de détecter les modules de conjugaison ou gènes de mobilisation portés par les chromosomes qui pourraient provenir, dans certains cas, de plasmides accidentellement intégrés mais proviendraient, probablement dans l'immense majorité des cas, d'éléments intégratifs conjugatifs ou mobilisables.

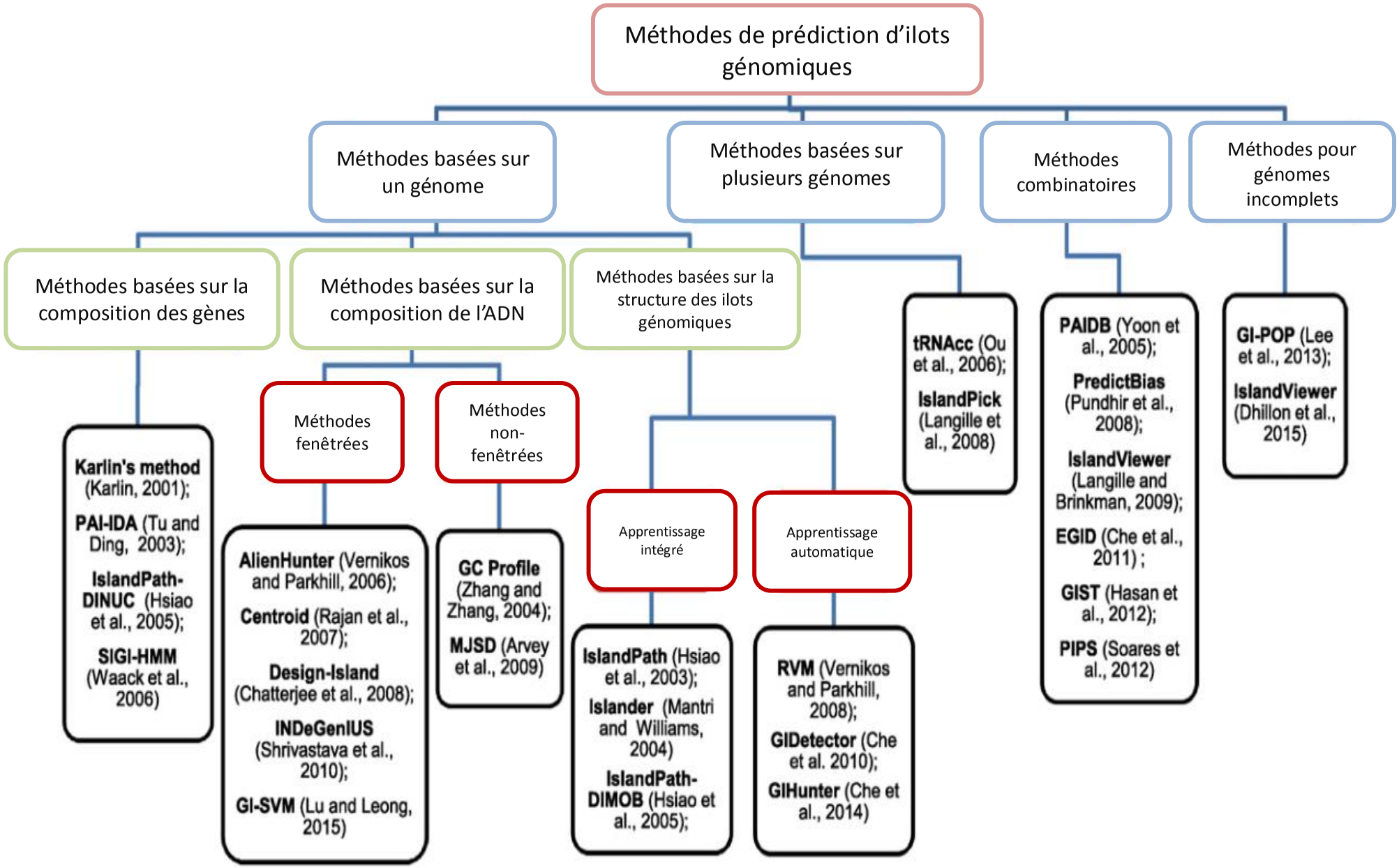

Figure 13 : Vue d'ensemble de différentes méthodes de détection d'ilots génomiques hiérarchisées en fonction des approches utilisées (d'après Lu and Leong, 2016).

RESULTATS

Bien que les ICE et les IME portent des fonctions intervenant dans l'adaptation des organismes et qu'ils soient probablement très répandus au sein des génomes bactériens, peu d'éléments sont actuellement décrits dans la littérature et il n'existe, à l'heure actuelle, aucun outil permettant de les détecter dans ces génomes. Ainsi, un premier objectif de ma thèse était de mettre au point une méthode efficace pour détecter et délimiter les éléments de type ICE et IME dans les génomes de streptocoques. Un second objectif était d'étudier la prévalence et la diversité des éléments détectés, de les classer en familles et d'étudier les gènes d'adaptations qu'ils portent, ce qui permettra de mieux comprendre l'impact que peuvent avoir ces éléments sur leur hôte. Enfin, une fois les données sur ces éléments recueillies et stockées dans une base de données, le dernier objectif consistait à automatiser et améliorer la méthode mise au point à partir de ces données en vue de la rendre plus facilement utilisable pour, à terme, élargir son utilisation, à l'ensemble des firmicutes.

1. Méthodologie ICEFinder : mise au point manuelle de la méthode initiale

La méthode utilisée est fondée sur la détection de protéines signatures dont la combinaison est caractéristique des ICE et IME. Celle-ci a évolué et a dû être mise au point manuellement au fur et à mesure que nos connaissances sur les ICE et les IME augmentaient, ce qui a conduit en particulier à inclure de nouvelles familles de relaxases putatives. L'amélioration progressive des connaissances sur les ICE et IME et la prise en compte des problèmes rencontrés nous a amené à affiner la méthode utilisée, ce qui par souci de clarté ne sera pas détaillé ici. Par ailleurs, la méthode mise au point ne permet pas de détecter les IME ne codant aucune relaxase alors même qu'un élément de ce type était déjà connu chez *S. agalactiae*. En effet, la seule protéine signature de ce type d'élément est l'intégrase (transposase à DDE) qui est malheureusement apparentée à celle des IS de famille IS*1595*, famille d'IS assez répandue chez les streptocoques.

1ère étape : Création d'une base de données de protéines dites « signatures ».

La base de données ICEFinder sert de référence dans la recherche d'ICE et d'IME. Elle contient un ensemble de protéines dites « signatures » :

- l'intégrase (protéine toujours présente chez les ICE et IME),
- la relaxase (toujours présente chez les ICE et présente dans la plupart des IME de firmicutes connus),
- la protéine de couplage (toujours présente chez les ICE et jamais détectée chez les IME avant cette étude)
- et la protéine VirB4 (toujours présente chez les ICE et jamais chez les IME).

A sa création en janvier 2011, cette base de données avait été constituée par Gérard Guédon et Sophie Payot à partir des protéines signatures d'ICE et IME de Firmicutes publiés ainsi que de quelques ICE et IME non publiés mais identifiés lors d'analyses préliminaires de génomes de streptocoques. Notamment, figuraient dans cette base de données, en plus des protéines d'ICE et IME publiés, des protéines signatures provenant d'ICE et d'IME des premiers génomes disponibles de *S. thermophilus* (CNRZ 1066, LMG 18311, LMD9), ainsi que des protéines issues d'ICE apparentés à ICE*St3* (ICE modèle du laboratoire DynAMic) comme ceux de *S. parasanguinis* ATCC 15912 et de *S. dysgalactiae* ATCC 1239, en limitant cependant les redondances entre protéines très proches. La base de données initiale comprenait au total 50 intégrases à tyrosine, 13 intégrases à sérine, 11 transposases à DDE (appartenant à trois familles éloignées, Tn*GBS1*/Tn*GBS2*, IS*1595*, IS*30*), 50 relaxases (principalement des relaxases MobP et MobT d'ICE), 37 protéines de couplage (de familles VirD4 et TcpA) et 26 protéines VirB4. Après une recherche préliminaire de protéines homologues dans un premier ensemble de génomes de streptocoques, le nombre de transposases à DDE utilisées a été réduit aux seules transposases de familleTn*GBS1/*Tn*GBS2*. En effet, les transposases de familles IS*1595* et IS*30* se sont avérées difficilement utilisables en raison du fait que, chez les streptocoques, la quasi-totalité n'étaient associées à aucune autre protéine signature, suggérant que ces protéines étaient des transposases d'IS et non des intégrases d'ICE ou d'IME. Pour IS*30*, ceci était d'autant plus délicat que la transposase peut être produite après décalage du cadre de lecture et qu'en conséquence les gènes codant la transposase d'IS*30* étaient mal annotés (codons de terminaison ou d'initiation sans réalité car la protéine est

produite par décalage du cadre de lecture, et pris à des positions variables). Ceci rendait leur détection ainsi que la distinction entre gènes et pseudogènes extrêmement difficile.

2eme étape : Recherche de protéines signatures.

La seconde étape est la recherche exhaustive des protéines signatures par homologie de séquence dans les génomes de streptocoques complètement séquencés et assemblés en décembre 2013. L'algorithme utilisé est implémenté dans le logiciel ngKlast (Nguyen and Lavenier, 2009) et il réalise une recherche accélérée, très proche de celle réalisée par BlastP. Les paramètres utilisés à ce stade étaient les paramètres par défaut, à l'exception des filtres de faible complexité qui ont été désactivés. En effet, le filtre n'améliore pas la qualité de la recherche du fait de l'absence de grandes régions de faible complexité et était même susceptible de diminuer la sensibilité du test dans certains cas (faux négatifs éventuels pour les protéines homologues présentant des similarités faibles et de petites régions de faible complexité).

La recherche des protéines signatures par homologie nécessite l'extraction de toutes les séquences codantes (CDS) du génome et impose donc que le génome étudié soit annoté. La qualité de l'annotation est cruciale pour ce genre d'analyse. Au cours de ma thèse, 124 génomes de streptocoques provenant de la base de données du NCBI ont été analysés (Voir Annexe 1). Les annotations initiales des génomes réalisées par les auteurs ont été conservées. Mais bien que le NCBI ait émis des consignes et conseils pour l'annotation des génomes avant soumission (https://www.ncbi.nlm.nih.gov/books/NBK174280/), une hétérogénéité dans l'annotation des génomes bactériens a été observée allant du génome très bien annoté (éléments génétiques mobiles compris, voire plus rarement annotation des introns) au génome sans ARN ribosomiques 5s/16s/23s, ni ARNt annotés bien que ces derniers fassent partie des standards d'annotations explicitement exigés. Ce dernier point nous a posé problème dans la mesure où plusieurs éléments sont intégrés dans des gènes d'ARNt qui n'avaient pas été annotés. De plus, les protéines non annotées ne peuvent pas avoir leurs séquences « extraites ». Ceci est, par exemple, le cas lorsqu'un gène pourtant valide est annoté comme pseudogène, qu'un gène est trop petit pour être considéré comme codant par certains programmes d'annotations (cas rencontré pour l'excisionase des modules d'intégration d'ICE et d'IME) ou qu'il est interrompu par un intron ou par l'insertion

d'un IME, ce qui a souvent été le cas pour les protéines de couplage des ICE de la superfamille Tn*5252*. De plus, selon l'algorithme utilisé et son paramétrage, les gènes peuvent être plus ou moins bien annotés. Ainsi, un problème fréquemment rencontré a été une mauvaise identification des codons d'initiation des gènes analysés, non seulement des codons UUG, fréquents dans les génomes de firmicutes et pas toujours pris en compte dans les génomes mal annotés, mais aussi de codons d'initiation rares, jamais pris en compte par l'annotation automatique (par exemple CUG qui est le codon d'initiation des gènes de relaxase des ICE de type Tn*916*). Par ailleurs, une fraction notable des « gènes » annotés et traduits sont en fait des pseudogènes mal annotés. Dans le cas de cette analyse, ceci a posé d'autant plus de problèmes que les gènes de maintien et transfert des éléments mobiles ne sont pas utiles voire nuisibles à l'hôte ; en conséquence, les mutations les inactivant après l'arrivée et l'intégration de l'élément ne sont pas éliminées par la sélection naturelle. En conséquence, une fraction élevée des « séquences protéiques » homologues détectées par BlastP correspondent en fait à des traductions de pseudogènes, pour lesquelles les protéines sont généralement tronquées de leur partie C terminale.

3eme étape : Application des filtres.

Une fois les protéines détectées par BlastP, une série de filtres doit être appliquée afin d'épurer les résultats des faux positifs. Ils ont été établis en considérant les données retrouvées dans la littérature puis ont été affinés après quelques tests sur un petit lot initial de génomes. Les filtres ont pour objectif de permettre la détection des protéines signatures les plus éloignées possibles tout en rejetant un maximum de protéines non apparentées et de faux positifs correspondant à des protéines apparentées mais ne présentant les fonctions recherchées. Les faux-positifs détectés par BlastP incluaient, par exemple i) des protéines de régulation possédant un domaine HTH de liaison à l'ADN homologue à celui des relaxases des ICE de famille Tn*916* et ICE*St3*, ii) des recombinases à sérine et tyrosine impliquées dans les inversions de segments d'ADN (invertases) ou dans la résolution de dimères (résolvases), iii) la protéine FtsK (ADN translocase impliquée dans la division cellulaire) ; iv) les produits de « traduction » de pseudogènes tronqués et v) des protéines de fusion. Les filtres utilisés incluent un critère de taille, un pourcentage de séquence couverte par l'alignement dans la protéine requête et la protéine détectée (taux de couverture), un pourcentage d'identité de l'alignement et une « e-value » maximale de l'alignement (Tableau 1).

Tableau 1 : Filtres utilisés lors de la mise au point manuelle de la méthode ICEFinder

	E-value	Taux de couverture	Pourcentage d'identité	Longueur du « Hit »
Protéine de couplage	>1,00^{E}-05	<25%	<25%	>180 ;<700 ;>1000 ;<1200
Relaxase	>1,00^{E}-04	<25%	<25%	>180
Intégrase à tyrosine	>1,00^{E}-04	<25%	<25%	>320
Intégrase à sérine	>1,00^{E}-04	<25%	<25%	>320
Transposase à DDE	>1,00^{E}-04	<25%	<25%	>320
VirB4	>1,00^{E}-05	<25%	<25%	>500

Sont indiquées pour chacune des protéines étiquettes recherchées, les valeurs seuils à partir desquelles les protéines détectées sont rejetées

Des protéines apparentées et de longueur similaire à des protéines signatures d'ICE et d'IME mais n'intervenant pas dans la conjugaison ne peuvent cependant pas être exclues par les filtres et ont nécessité une élimination manuelle. C'est le cas de XerS, la recombinase site-spécifique à tyrosine impliquée dans la résolution en monomère des dimères de chromosomes bactériens lors de la division cellulaire des streptocoques (Le Bourgeois et al., 2007).

Après application des filtres et validation manuelle des protéines signatures, la base de données est enrichie avec les nouvelles protéines signatures d'ICE et d'IME. Après épuration des protéines redondantes (un représentant a été choisi aléatoirement pour toutes les protéines ayant plus de 90% d'identité entres elles sur au moins 90% de leurs longueurs), une recherche *de novo* est réalisée jusqu'à ce qu'il n'y ait plus aucune nouvelle protéine détectée.

4ème étape : Co-localisation des protéines

Une fois les protéines détectées, la position dans le génome des gènes les codant a été visualisée à l'aide du logiciel Artemis (Rutherford et al., 2000). Cette étape a pour but de vérifier si les gènes sont co-localisés dans le génome, ce qui suggère l'appartenance de ces gènes à un seul élément. Le plus grand ICE ou IME connu de Firmicutes, PAI_UW3114 faisant environ 200 kb (Laverde Gomez et al., 2011), nous avons considéré que pour appartenir à un même élément les gènes devaient être compris dans une région d'au maximum 300 kb. La visualisation par Artemis permet aussi éventuellement d'estimer, à partir du nombre de gènes codant chaque classe de protéines signatures, si la région analysée contient un seul

élément, des éléments intégrés les uns dans les autres, plusieurs éléments en accrétion ou plusieurs éléments intégrés dans des sites proches les uns des autres. Cependant ces hypothèses ne pourront être confirmées qu'à l'issue de l'étape suivante de délimitation des éléments.

5ème étape : Délimitation des éléments

Une délimitation a été entreprise pour tous les éléments putatifs présentant un gène d'intégrase et au moins un gène codant une relaxase ou une protéine de couplage localisés dans les 300 kb avoisinant l'intégrase. Elles s'appuient sur les observations suivantes tirées de la littérature (Bellanger et al., 2014):

i) Les gènes impliqués dans le maintien et dans le transfert de l'élément sont regroupés à l'une des extrémités de l'élément génétique.
ii) L'intégrase est généralement localisée à une des extrémités de l'élément.
iii) L'intégration de la plupart des éléments provoquent la formation de courtes répétitions directes à leur extrémité (de 2 pb à 53 pb).
iv) Les DR incluent le plus souvent l'extrémité 3' du gène d'insertion ciblé.

La délimitation des éléments nécessite l'utilisation de plusieurs approches qui sont utilisées successivement lorsque l'approche précédente n'a pas abouti.

- Tout d'abord, si l'intégrase à tyrosine de l'élément est étroitement apparentée à celle de Tn*916*, les limites de l'élément sont recherchée par BlastN en utilisant comme séquence requête celle de Tn*916*.
- Si l'élément code une intégrase à tyrosine non apparentée à celle de Tn*916* ou une intégrase à sérine, des gènes tels que ceux codant des ARNt ou des protéines ribosomiques, connus pour être des sites d'intégration fréquents d'ICE ou d'IME sont recherchés à proximité du gène de l'intégrase.
- Si un tel gène est trouvé, son extrémité la plus proche de l'intégrase est extraite et cette séquence (environ 15-60 pb selon la longueur du DR attendue) est recherchée par BlastN à l'autre extrémité de l'élément putatif (sur une longueur de 300 kb). Si cette séquence est retrouvée répétée sur le même brin que la première, ces répétitions directes seront considérées comme les limites potentielles de l'élément. Si la région encadrée par ces répétitions contient un gène codant au moins une

intégrase, une relaxase, une protéine de couplage et une protéine VirB4, elle est considérée comme une région codant un ICE putatif. Si la région encadrée par ces répétitions contient un gène codant au moins une intégrase, une relaxase et éventuellement une protéine de couplage mais aucun gène ou pseudogène de VirB4, elle est considérée comme une région codant un IME putatif.

- Si aucun gène cible d'insertion connu n'est détecté à proximité du gène codant l'intégrase, il est alors possible que le gène d'insertion soit localisé à l'autre extrémité de l'élément, à une distance inconnue. Un tel gène est alors recherché dans les 300 kb de part et d'autre du gène de l'intégrase. Si un tel gène est trouvé, le processus décrit ci-dessus est répété.
- Dans les cas où aucun site d'insertion potentiel n'est détecté, la séquence située en aval du gène codant l'intégrase (jusqu'à contenir l'extrémité du gène situé en aval de l'intégrase) est extraite et une recherche de séquences répétées est réalisée à l'autre extrémité de l'élément.
- Si la recherche de séquences répétées n'a pas abouti, les éléments sont délimités par une étude de la synthénie par BlastN : la région contenant potentiellement un ICE ou un IME est comparée à celle de génomes proches phylogénétiquement, idéalement de la même espèce. Le but est de trouver un génome ne contenant aucun élément intégré dans cette région afin de pouvoir délimiter très précisément la région insérée. Une fois la région insérée identifiée (idéalement par son absence dans plusieurs génomes proches), une recherche de répétitions directes à ses extrémités par BlastN est effectuée. Cette recherche a pour but d'une part d'identifier précisément le site d'intégration et d'autre part de déterminer si la région insérée constitue bien un ICE ou un IME et non pas un îlot génomique plus grand.
- Dans certains cas, les génomes disponibles dans les bases de données publiques ne contiennent pas de génomes suffisamment proches pour réaliser une comparaison de la synthénie. Une analyse de « contexte » est alors réalisée : la région avoisinant l'intégrase est analysée et comparée à d'autres régions contenant les intégrases les plus proches phylogénétiquement contenues dans les bases de données publiques. Cette approche a pour but de trouver des similarités entre les régions analysées afin d'identifier un site potentiel d'intégration pour ces intégrases. Si un tel site est identifié, une recherche de DR est réalisée afin de délimiter l'ICE ou l'IME putatif.

6ème étape : Recherche de gène(s) manquant(s).

Lorsqu'un élément possédait tous les gènes nécessaires à son transfert conjugatif et son intégration sauf un, deux approches alternatives ont été utilisées pour rechercher des gènes qui n'auraient pas été annotés : i) le gène manquant peut être recherché manuellement par tBlastN avec une « query » issue de l'élément connu le plus proche, ii) Une comparaison par BlastX contre les protéines de la banque NCBI peut être réalisée à partir de la région qui aurait dû coder la protéine manquante. Par ailleurs, de nombreux éléments codaient une intégrase à tyrosine et une protéine de couplage mais ne portaient aucun gène, ou pseudogène, de relaxase ni de virB4. Dans ces cas, une analyse des domaines des protéines codées par les gènes situés entre les répétitions directes a permis de mettre en évidence la présence quasi-systématique de gènes codant des protéines apparentées à des protéines initiatrices de réplication par cercle roulant. Ces protéines pouvant alors constituer de potentielles relaxases comme cela a déjà été décrit pour les relaxases de la famille MOB_T apparentées, elles aussi, à des protéines initiatrices de la réplication par cercle roulant. Les relaxases putatives identifiées de cette manière ont été rajoutées à la base de données selon la méthode décrite (étape 3) et une recherche *de novo* a été effectuée.

7ème étape : Dénomination des éléments.

Les éléments délimités et codant les 4 protéines signatures (intégrase, relaxase, CP et VirB4) ont été considérés comme ICE. Ceux codant au moins 2 protéines signatures complètes et des pseudogènes des 2 autres, ou encore 3 protéines signatures complètes mais ayant une de leurs extrémités manquante sont considérés comme des dérivés d'ICE appelés dICE. Les éléments plus dégradés ne sont pas pris en compte dans ce travail.

Les éléments délimités et contenant une intégrase et une relaxase (phylogénétiquement éloignée de celles retrouvées chez les ICE) et éventuellement une protéine de couplage éloignée de celles des ICE sont considérés comme IME. Comme aucun des IME, dont la fonctionnalité a été démontrée précédemment, ne codent de protéine de couplage et que diverses comparaisons d'éléments proches suggéraient des pertes, acquisitions ou remplacements récents de ce type de gènes, les éléments contenant un pseudogène de protéine de couplage éloigné de ceux portés par les ICE sont aussi considérés comme IME.

Les éléments ne codant qu'une protéine signature ou codant une intégrase et une relaxase mais ne possédant pas l'une de leurs extrémités ne sont pas pris en compte.

Les éléments ont été nommés en faisant apparaître dans le nom, de façon successive, 3 caractéristiques séparées par des tirets bas :

- le type de l'élément ICE, dICE ou IME,
- des informations sur la souche dans laquelle l'élément a été identifié (première lettre du nom du genre, suivi des deux premières lettres du nom d'espèce et numéros ou lettres caractéristiques de la souche).
- si l'élément code une intégrase site-spécifique, le site dans lequel l'élément est intégré a été ajouté (par exemple, le nom ICE_*Sga43143_tRNAlys*, caractérise un ICE retrouvé intégré dans un gène d'ARN de transfert lysine de la souche *Streptococcus gallolyticus* UCN43143). Dans le cas où l'élément ne code pas d'intégrase site-spécifique (comme l'intégrase de l'ICE Tn*916*), le site d'intégration est remplacé par le nom de la famille de l'ICE (par exemple, ICE_*SsuSC84_Tn916*). Pour les éléments éloignés de Tn*916* et dont le site d'intégration n'a pas pu être identifié, le site d'intégration est remplacé par « ND ».
- Enfin, les rares éléments retrouvés intégrés dans un site probablement secondaire et non dans leur site primaire se sont vus ajouter un astérisque après leur site d'insertion primaire. Par exemple 'IME_*Sparas15912_rpmG**' est intégré dans un site intergénique localisé dans un autre IME IME_*Sparas15912_rpsI* (site secondaire). Il code une intégrase étroitement apparentée à des intégrases catalysant une intégration dans le gène *rpmG* codant la protéine ribosomique L33 (site primaire).
- Dans les quelques cas où l'élément était déjà délimité correctement avant ce travail, et dénommé, la correspondance entre son nom d'origine et le nom donné au cours de ce travail a été conservée et est accessible dans les fichiers supplémentaires des articles publiés sur les ICE et les IME.

8ème étape : Classement des éléments.

Les séquences des protéines signatures de chacun des éléments ont été extraites et comparées entre elles afin de les classer. Les protéines ont été classées en superfamilles en fonction de leur contenu en domaines. Ces superfamilles ont été subdivisées en familles en

fonction de leur parenté avec celle d'éléments connus et/ou entre elles (une famille inclut les protéines partageant plus de 40 % d'identité entre elles). Ces classements ont ensuite servi de base pour classer les ICE et les IME en familles et superfamilles. Le classement des ICE utilisent les données concernant les protéines signatures de leur module de conjugaison (relaxase, protéine de couplage et VirB4). Celui des IME prend en compte uniquement leur relaxase.

2. Etude de la prévalence et de la diversité des ICE au sein des génomes de streptocoques.

Les résultats présentés de façon succincte dans cette partie ont fait l'objet de l'article N°1 (inclus en fin de partie) publié en 2016 dans Frontiers In Microbiology sous le titre :

New Insights into the Classification and Integration Specificity of *Streptococcus* Integrative Conjugative Elements through Extensive Genome Exploration

Bien que des données récentes suggéraient que les ICE soient nombreux dans les génomes bactériens, seulement quelques rares recherches systématiques de protéines signatures d'ICE sans délimitation systématique des éléments, dans un nombre appréciable de génomes bactériens appartenant à un nombre appréciable d'espèce avaient été réalisées avant ce travail (Brouwer et al., 2011; Burrus et al., 2002a; Ghinet et al., 2011). Parmi ces études, une seulement a été réalisée sur un grand nombre de génome bactérien (plus de 1000) (Guglielmini et al., 2011). Une seule recherche systématique de tous les ICE et IME avec délimitation avait été réalisée sur 8 souches de *S. agalactie* (Brochet et al., 2008). Ainsi des données précises sur la prévalence et la diversité des éléments retrouvés chez les streptocoques manquaient.

De plus, il n'existe à l'heure actuelle aucune règle permettant de classer les ICE en familles. Certains éléments, par exemple ceux de la famille ICE*St3,* sont considérés de la même famille lorsque les gènes de leur module de conjugaison sont apparentés même si leurs modules d'intégration et leurs spécificités d'intégration sont différents (Carraro et al., 2011). D'autres classifications se basent uniquement sur la parenté des gènes du module d'intégration, qui ne comporte parfois qu'un seul gène, comme dans le cas du regroupement des ICE de famille Tn*GBS* (Guérillot et al., 2013). D'autres encore se basent sur les 2 critères (famille SXT, famille Tn*916*) (Burrus et al., 2006 ; Roberts and Mullany, 2011). Dans le cas de la famille Tn*916*, les ICE partagent un très fort degré d'identité sur l'ensemble des modules de conjugaison, intégration et régulation (supérieur à 95% d'identité en nucléotide).

Lors de cette étude, une recherche systématique d'ICE a été réalisée sur tous les génomes complets de streptocoques (124 génomes) disponibles dans la base de données du NCBI en décembre 2013. Cette analyse réalisée avec la méthodologie ICEFinder a permis d'identifier 105 ICE et 26 dICE démontrant la forte prévalence des ICE en général au sein des génomes

de streptocoques. Ces éléments ont été identifiés dans la moitié des génomes étudiés (63 sur 124).

Les ICE identifiés portent des modules de recombinaison comportant un gène codant : soit une intégrase à tyrosine, soit une intégrase à sérine, soit une transposase à DDE. Dans certains cas, ils portent des modules de recombinaison comportant un triplet de gènes adjacents codant des intégrases à sérine phylogénétiquement éloignées les unes des autres. Des analyses phylogénétiques suggèrent que cette structure en triplet découlerait de duplications successives dans un module de recombinaison. Par ailleurs, la délimitation précise des 131 ICE et dICE identifiés a permis de déterminer précisément les sites d'intégration de ces éléments. Cette étude met en évidence la grande diversité de sites d'intégration des ICE. Au total, 17 sites d'intégration spécifiques différents ont été mis en évidence dont 8 n'avaient encore jamais été décrits. Il s'agit de *traG* et *mutT* ciblés par des intégrases à sérine et de *ftsK*, *guaA*, *lysS*, *rpmG*, *rpsI* et *ebfC* ciblés par des intégrases à tyrosine. Dans le cas des intégrases à sérine, les sites spécifiques ciblés se situent à l'intérieur du gène et par conséquence, l'intégration de l'élément conduit potentiellement à son inactivation. Dans le cas des intégrases à tyrosine, elles ciblent l'une des extrémités du gène cible (extrémité 3' des ARN de transfert ou l'extrémité 3' ou 5' de gènes codant des protéines de ménage) ce qui ne conduit jamais à l'interruption du gène ciblé. De plus, au cours de cette étude nous avons pu démontrer que les intégrases site-spécifiques phylogénétiquement proches catalysent généralement une intégration dans le même site et ainsi corréler la famille de l'intégrase avec le site d'intégration ciblé par celle-ci.

Afin de mieux caractériser les éléments retrouvés chez les streptocoques, nous les avons classés en famille en fonction des 3 gènes portés par leur module de conjugaison que nous avons recherché : le gène codant pour la relaxase, le gène codant pour la protéine de couplage et le gène codant pour la protéine VirB4. Nous sommes parvenus à classer ces 131 éléments en 3 superfamilles (superfamille Tn*916*/ICE*St3*, superfamille Tn*5252*/Tn*1549*/Tn*GBS2*/VanG et superfamille Tn*GBS1*) et en 8 familles (familles Tn*916*, ICE*St3*, Tn*5252*, Tn*1549*, Tn*GBS2*, VanG et Tn*GBS1*) que nous avons nommées à partir des plus anciens éléments décrits appartenant à ces familles. Par ailleurs, ce classement permet de confirmer la forte prévalence de certaines familles d'éléments (notamment Tn*916*, Tn*GBS1* et Tn*GBS2*) dans les génomes de streptocoques. Il met également en évidence des

différences de diversité entre les familles, certaines familles comme la famille ICE*St3* présente une forte diversité au sein de leur module de conjugaison, tandis que d'autres familles présentent une faible diversité. Ainsi, le module de conjugaison de la famille Tn*916* est extrêmement conservé au sein de tous les éléments de la famille. De plus, cette étude montre que ces ICE de streptocoques possèdent uniquement des T4SS de famille FA (superfamille Tn*916*), FATA (superfamille Tn*5252* ou n'appartenant pas une famille définie (superfamille Tn*GBS1*). Notre analyse réalisée en utilisant le serveur CONJscan-T4SSscan indiquait également qu'ils ne possédaient que des relaxases appartenant aux familles MOB_P (superfamille *Tn5252* et Tn*GBS1*) et MOB_T. Cependant, selon une étude extrêmement récente (Ramachandran et al., 2017), la relaxase de Tn*GBS1* (WP_000383377, non identifiée comme appartenant à un ICE dans Ramachandran et al., 2017) et ses apparentées appartiendraient à une nouvelle famille de relaxases, MOB_L, qui présenterait des similarités notables avec les familles Mob_P, Mob_Q et Mob_V. On retrouve également, comme cela l'a été décrit pour les plasmides (Garcillán-Barcia et al., 2009), une concordance entre les familles de relaxases, les protéines de couplage et les T4SS qui leurs sont associés. En effet, les éléments codant une relaxase appartenant à la famille MOB_T sont associés, à une protéine de couplage de type TcpA et à des T4SS de la famille FA, alors que ceux codant une relaxase appartenant à la famille MOB_P sont associés à des protéines de couplage de type VirD4 et à des T4SS de la famille FATA. Enfin, ceux codant une relaxase appartenant à la famille MOB_L sont associés à des protéines de couplage de type VirD4 et des T4SS n'appartenant pas à une famille définie.

Certaines familles d'ICE de streptocoques présentent une forte diversité d'association entre leurs modules de conjugaison et leurs modules d'intégrations. Ainsi le module de conjugaison des ICE de la famille Tn*5252* est associé à des modules d'intégration variés comportant soit une, ou des, intégrase(s) à sérine, soit une intégrase à tyrosine, soit une transposase à DDE. A l'inverse, le module de conjugaison des ICE de la famille ICE*St3* est toujours associé avec des modules d'intégration comportant une intégrase à tyrosine de spécificités d'intégration variées. Ceci conforte l'idée qu'un mécanisme majeur d'évolution des ICE est l'échange de modules (Toussaint and Merlin, 2002). En effet, des ICE codant des modules de conjugaison apparentés, voire même très proches, peuvent être associés à des modules d'intégration ciblant des sites différents ou pouvant même coder des intégrases de

superfamilles différentes. Ainsi, un élément codant un T4SS FATA et une relaxase MOB_P peut être retrouvé associé à un module d'intégration codant une intégrase à tyrosine visant l'extrémité 3' du gène *rplL* ou l'extrémité 5' du gène *rbgA* mais aussi à un module d'intégration codant une intégrase à sérine visant le gène *rumA* (Voir figure 7 de l'article N°1). Dans le cas d'ICE*_SagILRI005_rplL*, une comparaison de séquences nous a permis de mettre en avant cet échange de module. Cet ICE comporte un module de conjugaison apparenté au module de conjugaison d'ICE*_SgaUCN34_TnGBS2*, un ICE de la famille Tn*GBS2* codant une transposase à DDE, tandis que sa partie gauche est proche d'un ICE de la famille Tn*5252*, ICE*_Sag018883_rplL*, codant une intégrase à tyrosine (Figure 10 de l'article N°1). Il est probable que cette structure composite résulte d'une intégration d'un ICE de la famille Tn*GBS2* (codant une transposase à DDE) dans le module de conjugaison d'un ICE de la famille Tn*5252*.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 06 January 2016
doi: 10.3389/fmicb.2015.01483

New Insights into the Classification and Integration Specificity of *Streptococcus* Integrative Conjugative Elements through Extensive Genome Exploration

*Chloé Ambroset[1,2†], Charles Coluzzi[1,2†], Gérard Guédon[1,2], Marie-Dominique Devignes[3,4], Valentin Loux[5], Thomas Lacroix[5], Sophie Payot[1,2] and Nathalie Leblond-Bourget[1,2]**

[1] DynAMic, Faculté des Sciences et Technologies, Université de Lorraine, UMR 1128, Vandœuvre-lès-Nancy, France, [2] DynAMic, Institut National de la Recherche Agronomique, UMR 1128, Vandœuvre-lès-Nancy, France, [3] Laboratoire Lorrain de Recherche en Informatique et ses Applications, Faculté des Sciences et Technologies, Université de Lorraine, UMR 7503, Vandœuvre-lès-Nancy, France, [4] CNRS, Laboratoire Lorrain de Recherche en Informatique et ses Applications, UMR 7503, Vandśuvre-lès-Nancy, France, [5] UR 1404 Mathématiques et Informatique Appliquées du Génome à l'Environnement, Institut National de la Recherche Agronomique, Jouy-en-Josas, France

OPEN ACCESS

Edited by:
John R. Battista,
Louisiana State University and A & M College, USA

Reviewed by:
Awdhesh Kalia,
University of Texas MD Anderson Cancer Center, USA
William John Kelly,
AgResearch Ltd., New Zealand

***Correspondence:**
Nathalie Leblond-Bourget
nathalie.leblond@univ-lorraine.fr

[†]These authors have contributed equally to this work.

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: 27 July 2015
Accepted: 08 December 2015
Published: 06 January 2016

Citation:
Ambroset C, Coluzzi C, Guédon G, Devignes M-D, Loux V, Lacroix T, Payot S and Leblond-Bourget N (2016) New Insights into the Classification and Integration Specificity of Streptococcus Integrative Conjugative Elements through Extensive Genome Exploration. Front. Microbiol. 6:1483. doi: 10.3389/fmicb.2015.01483

Recent genome analyses suggest that integrative and conjugative elements (ICEs) are widespread in bacterial genomes and therefore play an essential role in horizontal transfer. However, only a few of these elements are precisely characterized and correctly delineated within sequenced bacterial genomes. Even though previous analysis showed the presence of ICEs in some species of *Streptococci*, the global prevalence and diversity of ICEs was not analyzed in this genus. In this study, we searched for ICEs in the completely sequenced genomes of 124 strains belonging to 27 streptococcal species. These exhaustive analyses revealed 105 putative ICEs and 26 slightly decayed elements whose limits were assessed and whose insertion site was identified. These ICEs were grouped in seven distinct unrelated or distantly related families, according to their conjugation modules. Integration of these streptococcal ICEs is catalyzed either by a site-specific tyrosine integrase, a low-specificity tyrosine integrase, a site-specific single serine integrase, a triplet of site-specific serine integrases or a DDE transposase. Analysis of their integration site led to the detection of 18 target-genes for streptococcal ICE insertion including eight that had not been identified previously (*ftsK*, *guaA*, *lysS*, *mutT*, *rpmG*, *rpsI*, *traG,* and *ebfC*). It also suggests that all specificities have evolved to minimize the impact of the insertion on the host. This overall analysis of streptococcal ICEs emphasizes their prevalence and diversity and demonstrates that exchanges or acquisitions of conjugation and recombination modules are frequent.

Keywords: integrative and conjugative elements, T4SS, integrase, integration site, *Streptococcus*

INTRODUCTION

Streptococci are Gram positive bacteria belonging to the phylum of Firmicutes. This genus comprises 110 recognized species ([1]July 24, 2015). Almost all streptococci are commensal or pathogen of humans and/or animals. Numerous streptococci, such as *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Streptococcus mutans* or *Streptococcus agalactiae,* are responsible for a wide

[1]http://www.bacterio.net/streptococcus.html

variety of diseases worldwide, ranging from mild to invasive infections that have a severe impact on human and animal health and lead to significant morbidity and mortality (Mitchell, 2003; Kohler, 2007). Streptococci are also ubiquitously present as commensal inhabitants of the gastro-intestinal tracts of healthy adults and/or newborns. *Streptococcus salivarius*, in particular, is one of the first colonizers of the human oral cavity (Park et al., 2005; Nakajima et al., 2013) and is also a dominant part of the early life human intestinal microbiota (Arrieta et al., 2014). At last, two species deriving from commensal streptococci, *S. thermophilus* and *S. macedonicus* are used as starters in dairy industry to transform milk in yogurt and/or cheese (Franciosi et al., 2009).

During the last 20 years, it has become increasingly apparent that horizontal gene transfer (HGT) of genomic islands plays a key role in bacterial evolution and adaptation (Hacker and Kaper, 2000; Hacker and Carniel, 2001; Dobrindt et al., 2004; Juhas et al., 2009). In essence, genomic islands are chromosomal segments acquired by HGT that carry gene sets enhancing the fitness of their hosts. Recent data suggest that numerous genomic islands correspond to non-canonical classes of mobile genetic elements (MGEs) that can transfer by conjugation or are non-mobile elements deriving from such MGEs (Bellanger et al., 2014). Among them, the integrative and conjugative elements (ICEs) are mobile elements integrated in bacterial chromosomes or plasmids which encode their own excision, their transfer by conjugation, and their integration (Burrus et al., 2002b; Bellanger et al., 2014).

Integration of ICEs is catalyzed by three phylogenetically and structurally unrelated families of enzymes: tyrosine integrases, serine integrases, and DDE transposases (Wozniak and Waldor, 2010; Bellanger et al., 2014). Both tyrosine and serine integrases usually catalyze site-specific recombination between small (2–60 bp) similar or identical sequences included in the *attI* site of the circular form of the ICE and the *attB* site of the bacterial genome. This leads to the formation of *attL* and *attR* sites flanking the integrated elements; as a consequence, the integrated ICE is flanked by direct repeats (DR). Usually, tyrosine integrases catalyze ICE integration in a large array of specific sites, including the 3′ end of tRNA genes and the 3′ or 5′ end of genes encoding various housekeeping proteins (Bellanger et al., 2014). One exception is the tyrosine integrase of Tn*916* that shows low integration specificity; as a consequence, Tn*916* and its relatives are not flanked by DRs. Knowledge of the integration specificity of serine integrases from ICE is scarce. The third family of enzymes, DDE transposases, catalyzes transposition of DNA segments. Binding of the enzyme to terminal inverted repeats (IRs) at the extremities of elements enables strand cleavage required for the transposition reaction. DDE transposases have a low specificity of integration and catalyze the duplication of the target sequence (2–13 bp). Up to now, only one subfamily of DDE transposases was described for streptococcal ICEs. These DDE integrases catalyze the integration 15–16 bp upstream of the −35 box of the promoter region of various genes (Brochet et al., 2009; Guérillot et al., 2013).

The first step of the conjugative transfer is the excision of the ICE as a circular form that is ensured by the same enzyme as for integration. In general, tyrosine integrases need additional co-factors to carry out the reverse excision reaction (Groth and Calos, 2004); these are encoded by the element. So far, most ICEs (including all ICEs from Firmicutes) transfer as single-strand DNA. The transfer of the excised ICE would be similar to the transfer of conjugative plasmids that is well-known in Gram negative bacteria (Smillie et al., 2010; Low et al., 2014). The circularized ICE DNA is taken over by a relaxosome, a complex that includes a relaxase. A relaxase is a *trans*-esterase, acting as a dimer, that catalyzes a site and strand-specific cleavage at the *nic* site of the origin of transfer (*oriT*) of its cognate ICE. The relaxase, covalently bound to the 5′ end of the single-stranded DNA, is then recognized by the membrane-associated coupling protein (CP) that interacts with a type IV secretion system (T4SS). T4SSs are ATP-powered and multi-protein complexes that span the cellular envelopes of the donor cell in Gram negative bacteria. CP and T4SS translocate the DNA-relaxase complex through membranes and cell walls into the recipient cell. A rolling-circle replication of the element is likely concomitant to its transfer so that the ICE is not lost in the donor cell. Finally, the relaxase achieves the transfer by recircularizing the ICE (for a review see (Bellanger et al., 2014)). Although the conjugative transfer of DNA in Firmicutes is poorly known, it relies on relaxase, CP and T4SS (Goessweiner-Mohr et al., 2013; Guglielmini et al., 2014). Recent analyses suggest that T4SSs of Firmicutes include an homolog or analog of most T4SS membrane-spanning proteins found in inner membranes of Gram negative bacteria including the ATPase VirB4, VirB3, VirB6, VirB8 and the cell-wall degrading enzyme VirB1 (Goessweiner-Mohr et al., 2013; Guglielmini et al., 2014; Leonetti et al., 2015).

Like all other bacterial MGEs (Toussaint and Merlin, 2002), ICEs have a modular structure, i.e., the genes involved in the same biological function (such as conjugation or integration/excision) are physically linked. In addition to the integrase and recombination directionality factor genes, the integration/excision module includes the recombination site *attI*. It is thus generally located at one end of the integrated element. The conjugation module includes genes coding for the T4SS, the CP, the relaxase (and eventually accessory proteins of the relaxase) and *oriT*. The regulation module encodes all the genes involved in the regulation of ICE dissemination and maintenance. In addition to modules dedicated to gene transfer, all ICEs also carry at least one adaptation module that encodes adaptive traits that might be beneficial for bacteria under certain growth or environmental conditions. Adaptation modules are highly variable and include genes involved in antimicrobial resistance, virulence or alternative metabolic pathways (Burrus et al., 2002a,b; van der Meer and Sentchilo, 2003; Schubert et al., 2004; Heather et al., 2008; Croucher et al., 2009; Chuzeville et al., 2012).

Although ICEs have a major impact on gene flow and genome dynamics in bacteria, their prevalence and diversity remain largely underscored (Bellanger et al., 2014).

In this work, we took advantage of the increase of publicly available genomic sequences (124 complete genomes of *Streptococcus* available at the beginning of this work) to

search for ICEs using the combined presence of signature proteins (from conjugation and integration/excision modules). Coding sequences (CDSs) encoding these signature proteins were localized on the chromosomes and a strategy was developed to search for ICEs boundaries and to identify their integration site. This work (i) gives a general overview of the high prevalence and diversity of ICEs within *Streptococcus* species, (ii) identifies their numerous sites of insertion, and (iii) sheds light on their phylogenetic relationships and on their modular evolution.

MATERIALS AND METHODS

Genomes Examined and Database of Reference Proteins

The dataset of the 124 complete chromosomes from *Streptococcus* species available at the beginning of this work was taken from GenBank ([2]last accessed December 2013).

The initial database of reference proteins contains signature proteins from ICEs reported for Firmicutes in the literature at the beginning of this study. It includes protein sequences from 50 tyrosine integrases, 13 serine integrases, two DDE transposases, 50 relaxases, 37 CP, and 26 VirB4 proteins.

Search Strategy

The overall workflow of our search strategy to detect and characterize ICEs in streptococcal chromosomes is depicted in **Figure 1**.

Detection of Signature Sequences in the Genomes of Streptococci

The first step of our workflow consists in the search for signature proteins (tyrosine integrase, serine recombinase or DDE transposase, CP, relaxase, and VirB4 proteins). It was performed by BlastP comparison, using the accelerated BlastP version implemented in the ngKlast software (Nguyen and Lavenier, 2009; with default parameters except for disabled low-complexity filter). The queries were the sequences of all reference ICE proteins and the target was the set of multifasta files (one per genome) representing all translated CDSs of the studied genomes. Expert filters were designed in the ngKlast system to remove hits corresponding to the translation of pseudogenes and to related proteins not involved in transfer or integration of conjugative elements (i.e., recombinases involved in inversion of DNA segments or in resolution of DNA molecule multimers, transcriptional regulators carrying a HTH DNA binding domain, DNA translocase FtsK involved in cell division, some toxins, etc). These filters include a percentage cover threshold ($>25\%$), an E-value threshold ($<1.10^{-04}$ for relaxases and integrases and $<1.10^{-05}$ for CPs and VirB4) and a length threshold (>320 amino acids for integrases, >180 amino acids for relaxases, between 180 and 700 aa and between 1000 and 1200 aa for CPs and >500 amino acids for VirB4). Related proteins with biological function other than conjugation (i.e., XerS) passing through filters were manually removed. An iterative search was performed with the reference protein database enlarged with the newly found proteins and the same parameters until no new hit was found. Thus, only elements lacking significant sequence similarity with all signature proteins of ICE or carried by a plasmid could be missed. Sequence redundancy was eliminated using the BlastClust program with an identity threshold of 90%. The final reference database for ICE signature proteins contains non-redundant sequences from 106 tyrosine integrases, 43 serine integrases, 10 DDE transposases, 72 relaxases, 45 CPs, and 36 VirB4 proteins.

The relative positions of the CDSs corresponding to the detected signature proteins were visualized using Artemis (Rutherford et al., 2000). This step allowed checking if the CDSs co-localize in the genome and can thus be part of the same ICE. The detection of VirB4 guaranties the retrieval of ICEs rather than IMEs (integrative and mobilizable elements) that never encode this protein.

Detection of Insertion Sites and Delimitation of Putative ICEs

In the second step of the workflow (**Figure 1**), CDSs known to be potential insertion sites for ICE encoding site-specific integrases were searched and retained as potential candidates for insertion sites if located close to the CDSs of signature proteins. In their absence, insertion site was examined by comparing synteny with other genomes.

Putative ICEs were delimited by searching DRs on both sides of the putative ICEs by BLASTn analysis with either the 3′ end or the 5′ end of the potential insertion CDS or tRNA genes as a "query." If no such potential insertion CDS was detected, the intergenic sequence downstream from the integrase gene was used as a query. ICEs closely related to Tn*916* were delimited by BLASTn with the ends of Tn*916* from *Enterococcus faecalis* DS16 as queries.

ICE/Decayed ICE (dICE) Counting

All the elements delimited with DRs and containing CDSs for the four complete signature proteins (integrase, relaxase, CP, and VirB4) were considered as ICEs. When some signature CDSs were missing or were incomplete, the complete CDS encoded by the closest ICE was compared to the putative defective one by tBlastN in order to detect possible genome annotation errors (e.g., mis-identification of an authentic gene as a pseudogene most frequently due to the presence of a type II intron within the gene or mis-identification of START codon suggesting truncated genes). Elements that carry one or two defective signature CDSs, or that lack one of its extremity were considered as decayed ICEs (dICEs).

Domain Composition Analysis

The third step of our workflow (**Figure 1**) involved retrieving domain composition of all ICE signature proteins from Uniprot annotations using the BioMart Central Portal[3]. *De novo* CD-search for conserved domains[4] was performed when no data was available through BioMart.

[2]http://www.ncbi.nlm.nih.gov/genome/browse/

[3]http://central.biomart.org/

[4]http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

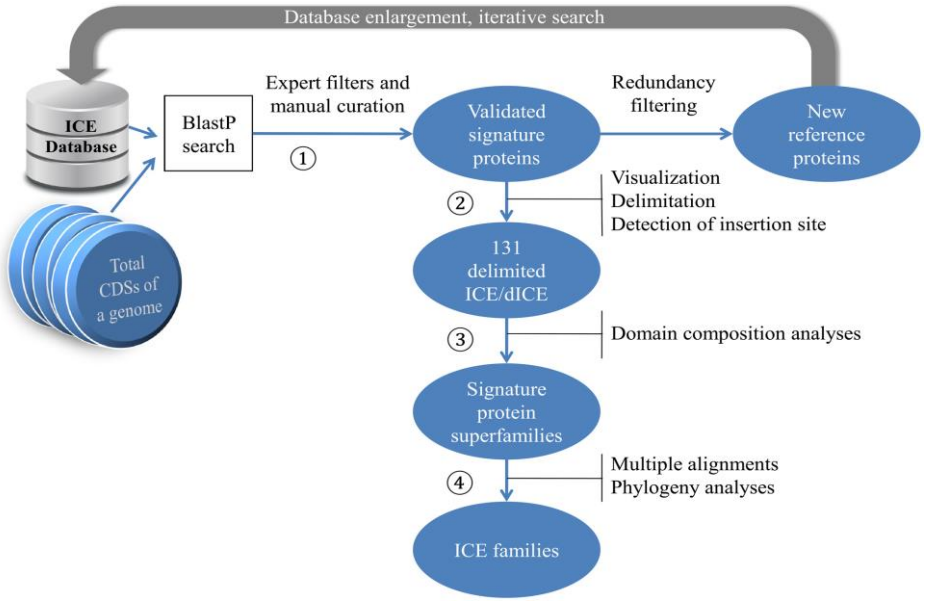

FIGURE 1 | Procedure for identifying candidate ICEs in sequenced genomes. The amino-acid sequences encoded by a chromosome are collected as multifasta files and processed as follows (see Materials and Methods). ① Signature proteins are identified by BlastP search using our reference sequences as query and the set of multifasta files as search database. Resulting hits are filtered and validated. ② The location of genes encoding the validated signature proteins are visualized using Artemis and ICEs are delimited. ③ Domain composition of signature proteins are searched using Biomart and the signature proteins are grouped into classes. ④ Multiple alignments of signature proteins in each class are performed using Clustal Omega and their phylogeny is analyzed using maximum likelihood (ML) methods and BioNJ.

Tree Construction

In this step (Step 4 on **Figure 1**), signature proteins were aligned using Clustal Omega with default parameters (Sievers et al., 2011). Trees of ICE signature proteins were built with MEGA (Tamura et al., 2013) using both maximum likelihood (ML; tree shown) based on JTT with Freqs (+F) model (partial deletion of gaps and missing data (80% cutoff), Gamma distributed with Invariant sites G+I (five categories)) and BioNJ methods with the Poisson model (Gouy et al., 2010). Branch support of the groupings was estimated using bootstrap (100 replicates for the ML method and 1,000 replicates for BioNJ).

ICE Annotation and Comparative Analysis

The comparative analysis of the conserved CDSs within a given ICE family was performed only for those that were non- or mis-characterized and displayed a significant number of ICEs. Functional annotation of ICEs was performed using Agmial (Bryson et al., 2006). Protein product, gene name and EC_number were assigned using similarity with Uniprot databank.

Data mining of the orthologs and the conserved syntenies was performed using Insyght. Sequence alignments were carried out at the protein level (using BLASTp) to achieve the pairwise comparisons of all the CDSs of ICEs belonging to the same family. Two genes were considered orthologous if they gave rise to a bi-directional best hit (BDBH) of the corresponding ICE genomic regions and if the sequence alignment included more than 50% of the total proteins with an e-value less than 0.01. Two CDSs for which the E-value of the sequence alignment was less than 0.01 were considered homologous but were not analyzed unless they belong to a synteny. Syntenies were computed with a dynamic programming algorithm that determines the highest scoring paths amongst the chains of colinear homologs. The scores and penalties used were as follows: minimum synteny score: 8; ortholog BBDH: 4; homolog non BDBH: 2; mismatch: −3; gap creation: −4; gap extension: −2. This setting allows the insertion of small gaps within the conserved synteny. The "Orthologs table" view in Insyght was used to identify the conserved and idiosyncratic loci within ICEs.

RESULTS

Prevalence of ICEs and dICEs within Streptococcal Chromosomal Genomes

A total of 105 ICEs and 26 dICEs were identified among the 124 streptococcal genomes analyzed in this work (Supplementary Table S1). About half (63/124) of the examined strains contain at least one element and among those strains 61% (39/63) harbor several ICEs or dICEs. ICE denomination indicates whether the element is an ICE or a dICE, followed by letters and numbers allowing species and strain identification.

When ICE/dICE encodes a site-specific integrase, its denomination also specifies the name of the target-gene. Otherwise, it indicates the integrase family (Tn*916* or DDE). For elements already well-characterized and named, the correspondence between names is indicated in Supplementary Table S1.

Some streptococcal species show relatively few ICEs and dICEs even if a significant number of genomes were analyzed (**Table 1**). In particular, in the *salivarius* group, only one element was found in the six analyzed genomes of *S. thermophilus* analyzed and no element in the three genomes of *S. salivarius*. By contrast, in the *anginosus* and *bovis* groups, there is an average of more than two ICEs or dICEs per genome, with extreme situations such as 13 and 11 elements found in the three genomes studied in *S. anginosus* and *S. gallolyticus,* respectively. The occurrence of ICEs and dICEs per species or strain can vary within a group. For example in the *pyogenic* group, there were only six elements found in the 19 *S. pyogenes* genomes analyzed but as many as 10 elements in the five *S. dysgalactiae* genomes analyzed.

Diversity of ICE and dICE Relaxases, and VirB4 in Streptococci

A total of 121 relaxases, encoded by ICEs or dICEs, were detected. They can be classified in three distinct classes on the basis of their domains (**Table 2**). The 'Rel-I' regroups 52 relaxases that contain a C-terminal catalytic "Rep_trans" domain (PF02486) associated with an N-terminal Helix-Turn-Helix (PF01381) DNA binding domain. According to the CONJscan-T4SSscan server ([5]Guglielmini et al., 2011), these 52 relaxases belong to the MOB_T family that is related to initiators of rolling-circle replication of some plasmids and prophages (Guglielmini et al., 2014). The 'Rel-II' class regroups 62 relaxases sharing a common N terminal "relaxase" PF03432 catalytic domain and belonging to the MOB_P family. Among them, 20 relaxases also carry a C terminal "Lantibiotic streptin immunity" PF11083 domain of unknown function. The 'Rel-III' class of relaxases contains seven proteins with no identified domains according to CD-search. These proteins are classified

[5]http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::CONJscan-T4SSscan

TABLE 1 | Prevalence of ICEs and dICEs within streptococcal species.

Group of species	*Streptococcus* species or strains	Number of strains	Total number of ICEs/dICEs per species	Minimum number of ICEs/dICEs per genome	Maximal number of ICEs/dICEs per genome	Average number of ICEs/dICEs per genome
anginosus	*S. anginosus*	3	13	2	7	4.3
anginosus	*S. intermedius*	3	7	1	3	2.3
anginosus	*S. constellatus*	3	6	0	4	2.0
bovis	*S. gallolyticus*	3	11	3	5	3.7
bovis	*S. pasteurianus*	1	3	3	3	NA
bovis	*S. infantarius*	1	2	2	2	NA
bovis	*S. lutetiensis*	1	2	2	2	NA
bovis	*S. macedonicus*	1	2	2	2	NA
mitis	*S. pneumoniae*	28	18	0	4	0.6
mitis	*S. oligofermentans*	1	2	2	2	NA
mitis	*S. mitis*	1	1	1	1	NA
mitis	*S. oralis*	1	1	1	1	NA
mitis	*S. pseudopneumoniae*	1	1	1	1	NA
mutans	*S. mutans*	4	1	1	1	0.3
pyogenic	*S. agalactiae*	8	13	0	7	1.6
pyogenic	*S. dysgalactiae*	5	10	0	5	2.0
pyogenic	*S. pyogenes*	19	6	0	3	0.3
pyogenic	*S. equi*	4	4	0	2	1.0
pyogenic	*S. parauberis*	1	2	2	2	NA
pyogenic	*S. iniae*	1	0	0	0	NA
pyogenic	*S. uberis*	1	0	0	0	NA
sanguinis	*S. parasanguinis*	2	3	1	2	1.5
sanguinis	*S. gordonii*	1	0	0	0	NA
sanguinis	*S. sanguinis*	1	0	0	0	NA
suis	*S. suis*	18	21	0	5	1.2
salivarius	*S. thermophilus*	6	1	0	1	0.2
salivarius	*S. salivarius*	3	0	0	0	0.0
ND	S. sp I-G2	1	1	1	1	NA
ND	S. sp I-P16	1	0	0	0	NA

NA, not applicable.

TABLE 2 | Characterization of the Conj$_{Tn916}$, Conj$_{TnGBS2}$, and Conj$_{TnGBS1}$ superfamilies of conjugation modules.

	Relaxases			Coupling proteins			VirB4			ICE/dICE superfamilies of conjugation modules		
Class	**Pfam ID**	**Domain name**	**Conjscan**	**Class**	**Pfam ID**	**Domain name**	**Conjscan**	**Class[a]**	**Pfam ID**	**Domain Name**	**Conjscan**	**Name**	**ICEs**	**dICEs**
Rel-I	PF02486	Rep trans	MOB$_T$	CP-I	PF01580	FtsK_SpoIIIE	TcpA	VirB4-Ia	PF12846	AAA_10	VirB4	Conj$_{Tn916}$	44	9
	PF01381	DNA binding						or Ib						
Rel-II	PF03432	Relaxase	MOB$_P$	CP-IIa[b]	PF02534	T4SS-DNA_transfer	T4cp1/T4cp2	VirB4-Ic	PF12846	AAA_10	VirB4	Conj$_{Tn5252}$	54	15
	PF11083 (optional)	Lantibiotic streptin immunity			PF12696	TraG-D_C								
Rel-III	None	None	MOB$_P$	CP-IIb[c]	PF02534	T4SS-DNA_transfer	T4cp1/T4cp2	VirB4-Id	PF12846	AAA_10	VirB4	Conj$_{TnGBS1}$	7	2
					PF12696	TraG-D_C								

[a] *The four VirB4 subclasses are distinguished on the basis of their phylogenetic relatedness.*
[b] *598–688 aa.*
[c] *1045 aa.*

by the CONJscan-T4SSscan server in the MOB$_P$ family of relaxases.

The 126 CPs encoded by ICEs or dICEs of streptococci are divided into two classes. The CP-I class groups 50 CPs sharing a unique central FtsK_SpoIIIE catalytic domain (PF01580; **Table 2**). According to the CONJscan-T4SSscan server, they belong to a particular class of CPs named TcpA, unrelated to all CPs of Gram-negative bacteria (Guglielmini et al., 2014). The 'CP-II' class contains 76 CPs with a central catalytic TraG/TraD domain (PF02534) and an additional C-terminal 'TraM recognition site of TraD and TraG' PF12696 domain. According to the CONJscan-T4SSscan server, these proteins belong to the main family of CPs named VirD4 (Guglielmini et al., 2014). Among them, 67 CPs, constituting the 'CP-IIa' class, are 598 aa to 688 aa-long. The nine remaining CPs are composed of about 1045 aa and are representatives of the 'CP-IIb' class (**Table 2**).

All 124 VirB4 proteins from ICEs and dICEs show a unique C-terminal PF12846 'AAA-10' catalytic domain. Reconstruction of their phylogenetic relatedness (data not shown) suggested that they can be grouped in four classes designated 'VirB4-Ia' (30 proteins), 'VirB4-Ib' (20 proteins), 'VirB4-Ic' (65 proteins), and 'VirB4-Id' (nine proteins; **Table 2**). These proteins are all identified as VirB4 using the CONJscan-T4SSscan server.

Classification of Conjugative Modules in Superfamilies and Families

Table 2 summarizes the co-occurrence of the different classes of signature proteins within the conjugation modules of ICEs and dICEs. This allowed classification of the conjugation modules into three distinct superfamilies, named according to the first characterized element of each superfamily: Conj$_{Tn916}$, Conj$_{Tn5252}$, and Conj$_{TnGBS1}$.

The Conj$_{Tn916}$ Superfamily of Conjugation Modules

The Conj$_{Tn916}$ superfamily groups together 44 ICEs and nine dICEs. All encode a 'Rel-I' relaxase associated with a 'CP-I' and a 'VirB4-Ia' or '-Ib' VirB4 protein.

The phylogenetic tree of relaxases (**Figure 2A**) of the Conj$_{Tn916}$ superfamily indicates three well-supported groups: relaxases related to the one of Tn*916*, to that of ICE*_SmuA159_tRNAleu* and to that of ICE*St3*. However, the latter two groups are not supported by the phylogenetic trees of CPs and VirB4 (**Figures 2B,C**). Therefore, two families of conjugative modules can be distinguished in the Conj$_{Tn916}$ superfamily: the Conj$_{Tn916}$ family and the Conj$_{ICESt3}$ family.

The Conj$_{Tn916}$ family is well-supported by phylogenetic trees of the relaxase, CP and VirB4 protein sequences (**Figures 2A–C**). It clusters 27 ICEs (and five dICEs) whose signature proteins are almost identical to those of the well-described Tn*916* element originally isolated from the Firmicute *E. faecalis* (Franke and Clewell, 1981). It also includes the more distant ICE*_SmiB6_guaA* whose signature proteins shared only 67, 66, and 80% identity with the *E. faecalis* Tn*916* relaxase, CP and VirB4 protein, respectively. All the 12 genes of the conjugation module of the prototype Tn*916* from *E. faecalis* DS16 were found with a similar organization in the vast majority of ICEs

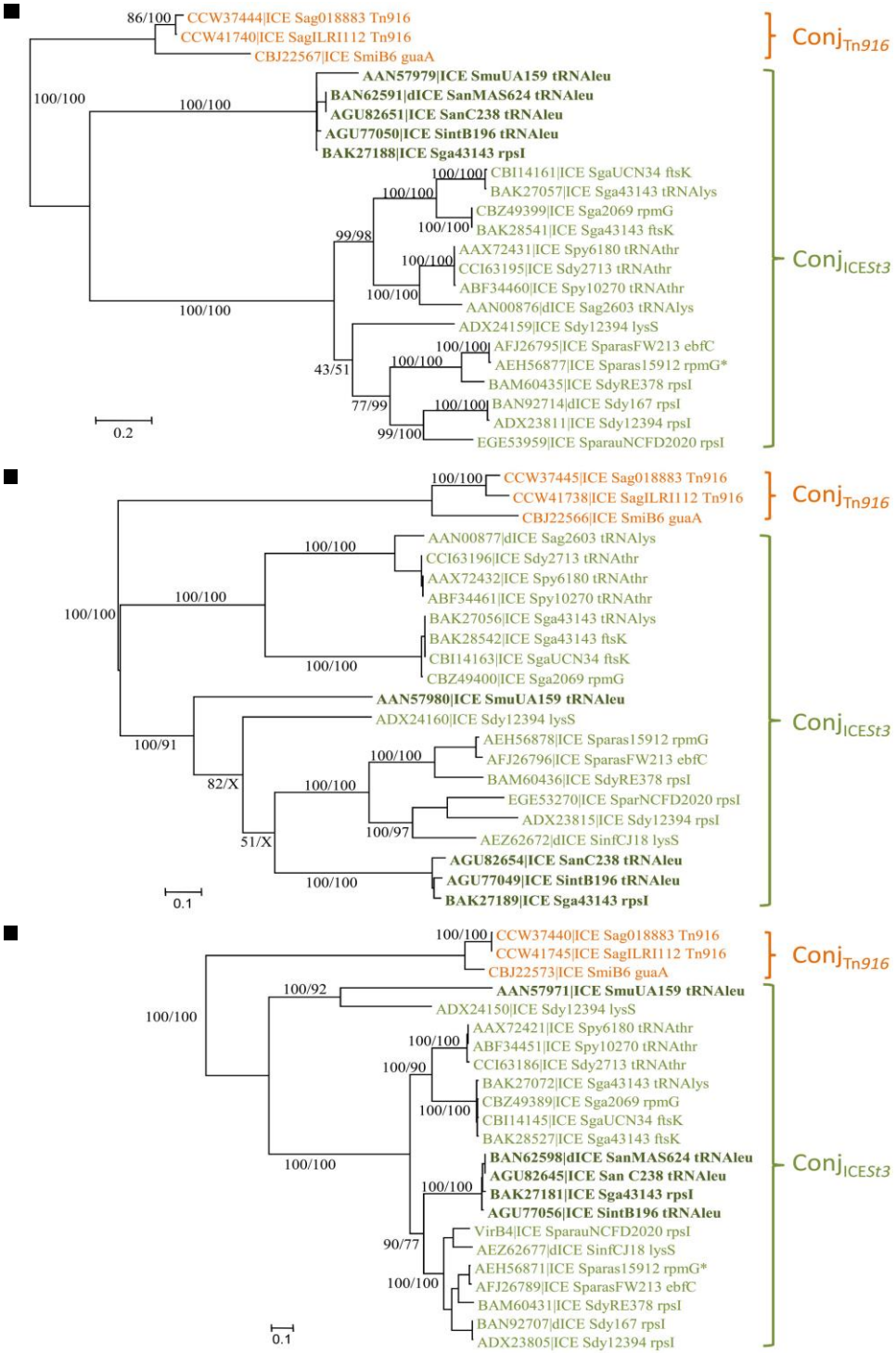

FIGURE 2 | Continued

FIGURE 2 | Continued

Phylogenetic analysis of ICEs and dICE belonging to the Conj$_{Tn916}$ superfamily. (A) Relaxases; **(B)** coupling proteins; **(C)** VirB4 proteins. Bootstrap supports are given as followed: ML/BioNJ. X marks the nodes that are not validated with BioNJ. ICE names are colored according to the family of conjugation module they encode: orange = Conj$_{Tn916}$, green = Conj$_{ICESt3}$ (including elements encoding relaxases related to the one of ICE_*SmuA159_tRNAleu* in bold dark green). Because of their very close phylogenetic relationships, relaxases **(A)**, CP **(B)**, and VirB4 **(C)** of only three ICEs representative of the Conj$_{Tn916}$ family are shown. Refer to Supplementary Table S1 for ICE/dICE and strain details.

of the Conj$_{Tn916}$ family (data not shown) thus confirming their relationships. Others only differ by one or few deletions, pseudogenizations or insertions.

The Conj$_{ICESt3}$ family gathers together 21 elements whose signature proteins are much more variable in sequence than those of the Conj$_{Tn916}$ family. The sequence of relaxases, CPs and VirB4 proteins of the most distantly related elements displayed around 20, 30, and 60% of identity, respectively.

For a better characterization of the Conj$_{ICESt3}$ family, a search for conserved CDSs in all members of this family was undertaken. Orthologous CDSs were identified on the basis of the identity of their product using Insyght. The integration module of all elements of the Conj$_{ICESt3}$ family is composed of both a tyrosine integrase and an excisionase (**Figure 3A**). In addition to the relaxase, CP, VirB4 CDSs, tyrosine integrase and excisionase CDSs, they share seven other CDSs including CDSs involved in the T4SS formation (VirB1, VirB3, VirB6, and VirB8) and a CDS encoding a regulation protein (HTH regulator).

The Conj$_{Tn5252}$ Superfamily of Conjugation Modules

The Conj$_{Tn5252}$ superfamily gathers together 54 ICEs and 15 dICEs encoding a 'Rel-II' relaxase associated with a 'CP-IIa' CP

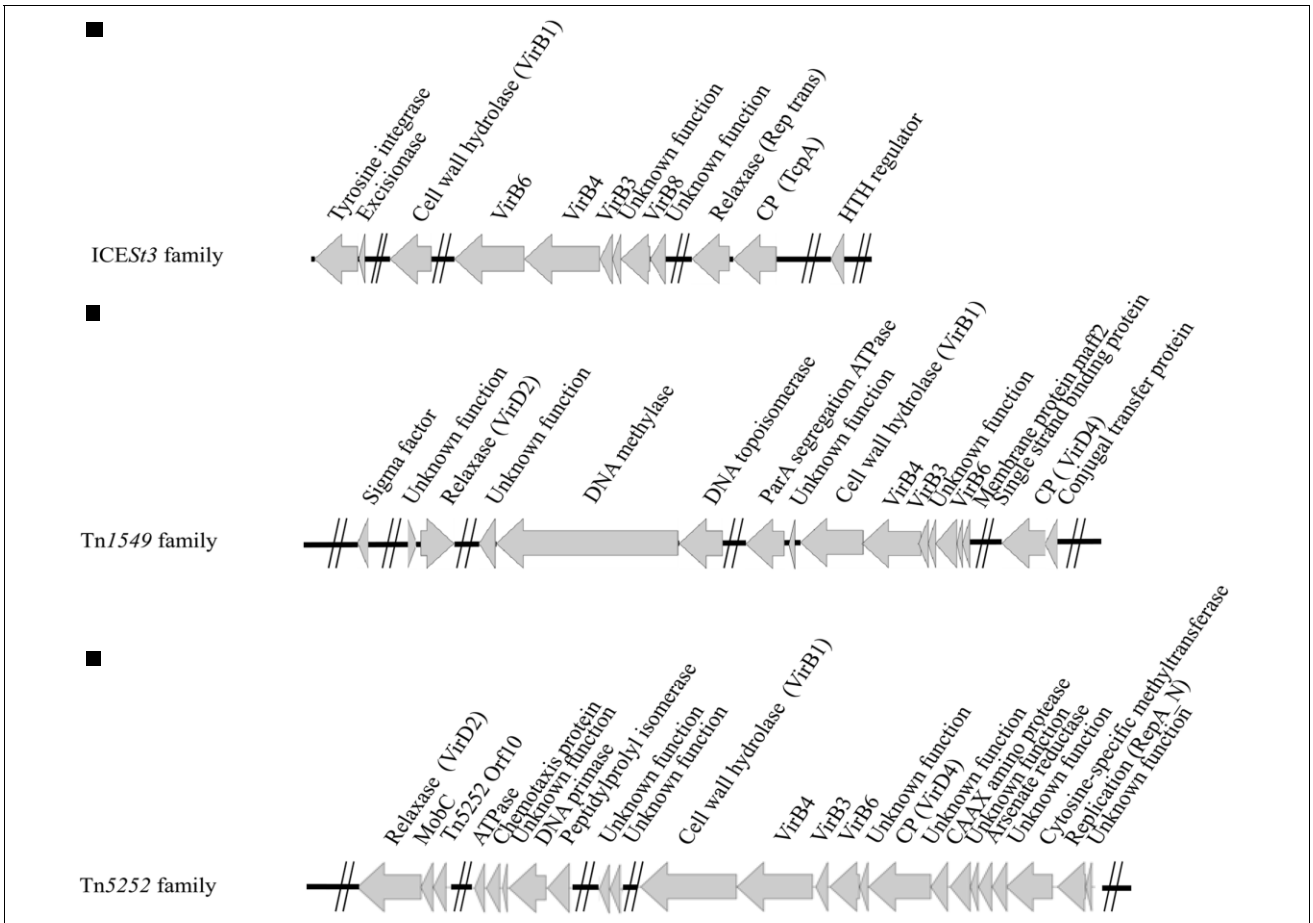

FIGURE 3 | Identification of conserved CDSs within the Conj$_{ICESt3}$, Conj$_{Tn1549}$ and Conj$_{Tn5252}$ families of ICEs. (A) Conj$_{ICESt3}$ family; **(B)** Conj$_{Tn1549}$ family; **(C)** Conj$_{Tn5252}$ family. Genes encoding conserved proteins in all ICEs of a given family are indicated by arrows with below their putative function (or known homologs).

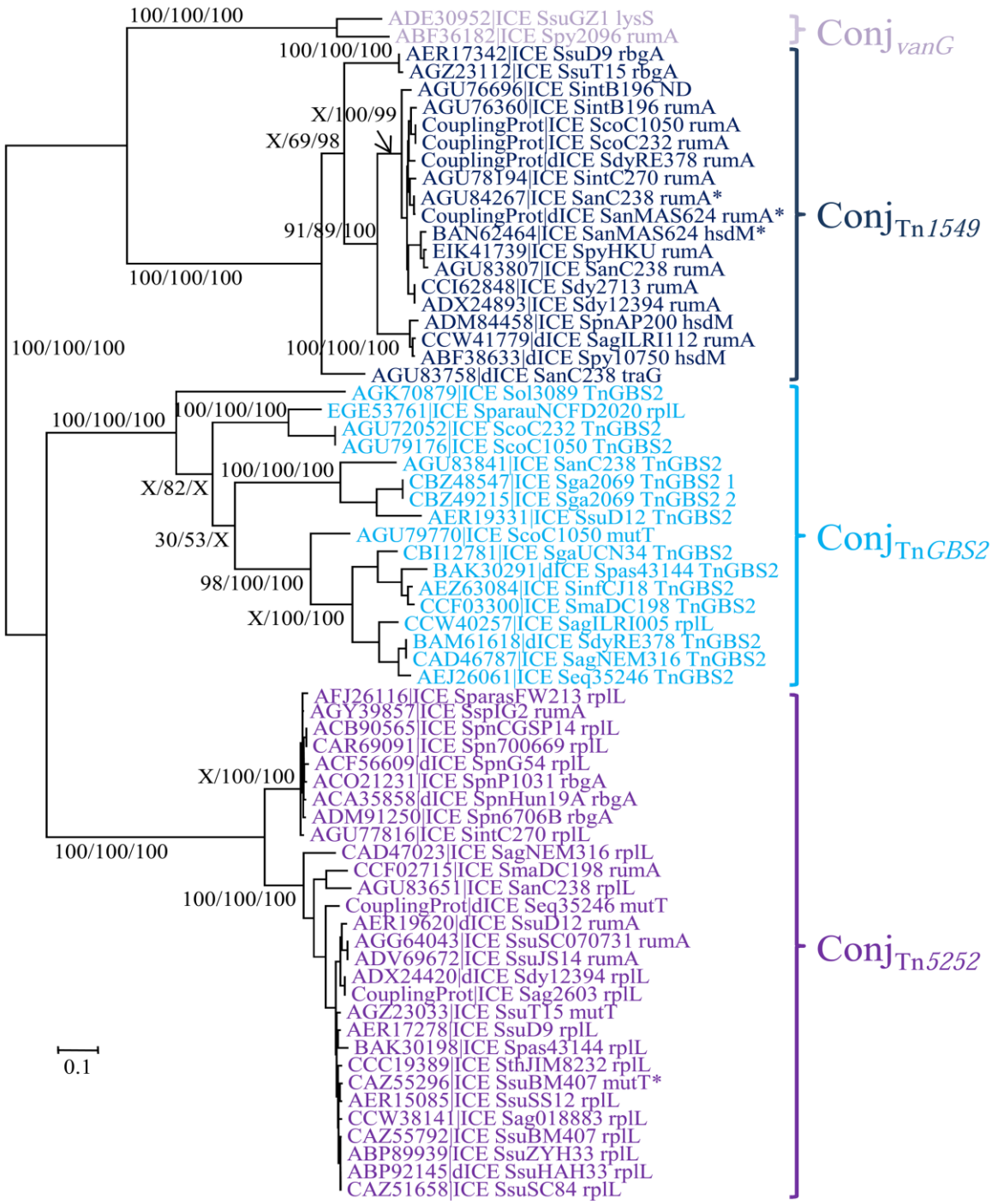

FIGURE 4 | Phylogenetic analysis of ICEs and dICEs belonging to the Conj$_{Tn5252}$ superfamily. The trees of CPs are shown in this figure. ML boostrap values for relaxases/CPs/VirB4 (in this order) are given. X marks the nodes that are not validated with other proteins. Families are also supported by BioNJ analysis (data not shown). In mauve are ICEs/dICEs belonging to the Conj$_{vanG}$ family, in dark blue those belonging to the Conj$_{Tn1549}$ family, in light-blue those of the Conj$_{TnGBS2}$ family and in purple that of the Conj$_{Tn5252}$ family. Elements marked with an asterisk are not integrated in their primary sites but in secondary ones. Refer to **Table 1** for ICE/dICE and strain details.

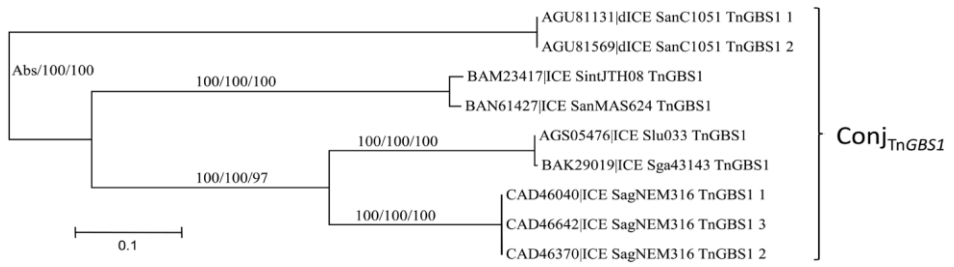

FIGURE 5 | Phylogenetic analysis of ICEs and dICEs belonging to the Conj$_{TnGBS1}$ superfamily. The trees of CPs are shown in this figure. ML boostrap values for relaxases/CPs/VirB4 is given in this order. Families are also supported by BioNJ analysis (data not shown). Abs means that one sequence was missing. Refer to Supplementary Table S1 for ICE/dICE and strain details.

and a 'VirB4-IIc' VirB4 (**Table 2**). The phylogenetic trees obtained independently for 'Rel-II' relaxases, 'CP-IIa' CPs and 'VirB4-Ic' proteins (**Figure 4**) are congruent and therefore only the CP one is shown. These data are consistent with the splitting of the Conj$_{Tn5252}$ superfamily into four distinct families: Conj$_{vanG}$, Conj$_{Tn1549}$, Conj$_{TnGBS2}$, Conj$_{Tn5252}$ that cluster 2, 21, 17, and 29 elements, respectively.

As expected from Guérillot et al. (2013), sequence comparison of the 15 ICEs encoding a Conj$_{TnGBS2}$ module identified 14 CDSs shared by all these Tn*GBS2*-related elements (data not shown).

Using Insyght, CDSs comparison of the 14 ICEs of the Conj$_{Tn1549}$ family identified 14 CDSs shared by all of them, in addition to the relaxase, CP and VirB4 CDSs (**Figure 3B**). One of these CDSs encodes a sigma factor that may be involved in the regulation of ICE transfer; three others being probably involved in the T4SS formation (VirB1, VirB3, VirB6) and one in maintenance of the ICE after excision (segregation ATPase). Many of these ICEs also encode a protein carrying a repA_N domain that could be involved in maintenance of excised ICEs.

As for ICEs of the Conj$_{Tn1549}$, the 23 ICEs of the Conj$_{Tn5252}$ family do not share the same integration module. However, they display 24 conserved CDSs. In addition to the relaxase, and the CP VirB4 CDSs, they share three other CDSs probably involved in the T4SS formation (VirB1, VirB3, VirB6), and one encoding a replication initiator (repA_N; **Figure 3C**).

The Conj$_{TnGBS1}$ Superfamily of Conjugation Module

The Conj$_{TnGBS1}$ superfamily of conjugation modules is the least represented in streptococcal genomes with only seven ICEs and two dICEs exhibiting such a module. It is characterized by the presence of a 'Rel-III' relaxase co-occurring with a 'CP-IIb' CP and a 'VirB4-Id' protein. This superfamily was not identified by Guglielmini et al. (2014).

As for the Conj$_{Tn5252}$ superfamily, whatever the signature protein used, all phylogenetic trees are congruent and therefore only the CP one is shown (**Figure 5**). Sequence comparison of the signature proteins of the most divergent ICEs of this superfamily indicated that they shared more than 56% of identity. Thus, within streptococcal genomes, the Conj$_{TnGBS1}$ superfamily is represented by the unique Conj$_{TnGBS1}$ family. As expected from Guérillot et al. (2013), systematic comparisons of the protein sequence of the members of the Conj$_{TnGBS1}$ family with Tn*GBS1* CDSs confirmed that they are related to Tn*GBS1* (data not shown).

Prevalence of the Different Families of Conjugation Modules within Streptococcal Species

In summary, seven families of conjugation modules belonging to three superfamilies were identified in streptococcal ICEs/dICEs. Most ICEs of the Conj$_{Tn916}$ family are found in *S. suis* and *S. pneumoniae*, while *S. dysgalactiae* and *S. gallolyticus* mainly harbor ICEs of the Conj$_{ICESt3}$ family. ICEs of the ConjTn$_{5252}$ family are frequently found in *S. suis*, and ICEs of the Conj$_{TnGBS1}$ family are only found in *S. agalactiae*, *S. anginosus*, *S. intermedius*, and *S. lutetiensis* genomes.

Diversity of Integration Modules and Integration Sites of Streptococcal ICEs and dICEs

Three unrelated families of integrases (tyrosine integrase, serine integrase, and DDE transposase) are encoded by streptococcal ICEs/dICEs. Most of these integrase genes are located at one extremity of the ICE and adjacent to the integration site.

Prevalence and Integration Site of ICEs/dICEs Encoding a Tyr Integrase

Tyrosine integrases are the most prevalent integrases detected as they are found in 53% of the elements. In total, 73 tyrosine integrases were identified (66 for ICEs and 7 for dICEs). Most of them are ∼400 aa long, except for five that are composed of 502 aa. Despite a variable degree of identity between these proteins, all tyrosine recombinases share a "phage-integrase" PF00589 domain in their N-terminal region. Most of them carry an additional N-terminal binding domain being either: (i) a pfam02920 "DNA binding domain characteristic of Tn*916* integrase" or (ii) a PF14659 "Phage integrase, N-terminal SAM-like domain" (iii) and/or more rarely a PF14657 "AP2-like DNA-binding integrase domain."

The phylogenetic tree of the tyrosine integrases reveals 11 well-supported groups (**Figure 6**). The overall finding is that almost all tyrosine integrases are grouped according to their insertion loci. Exceptions are the two non-grouped tyrosine

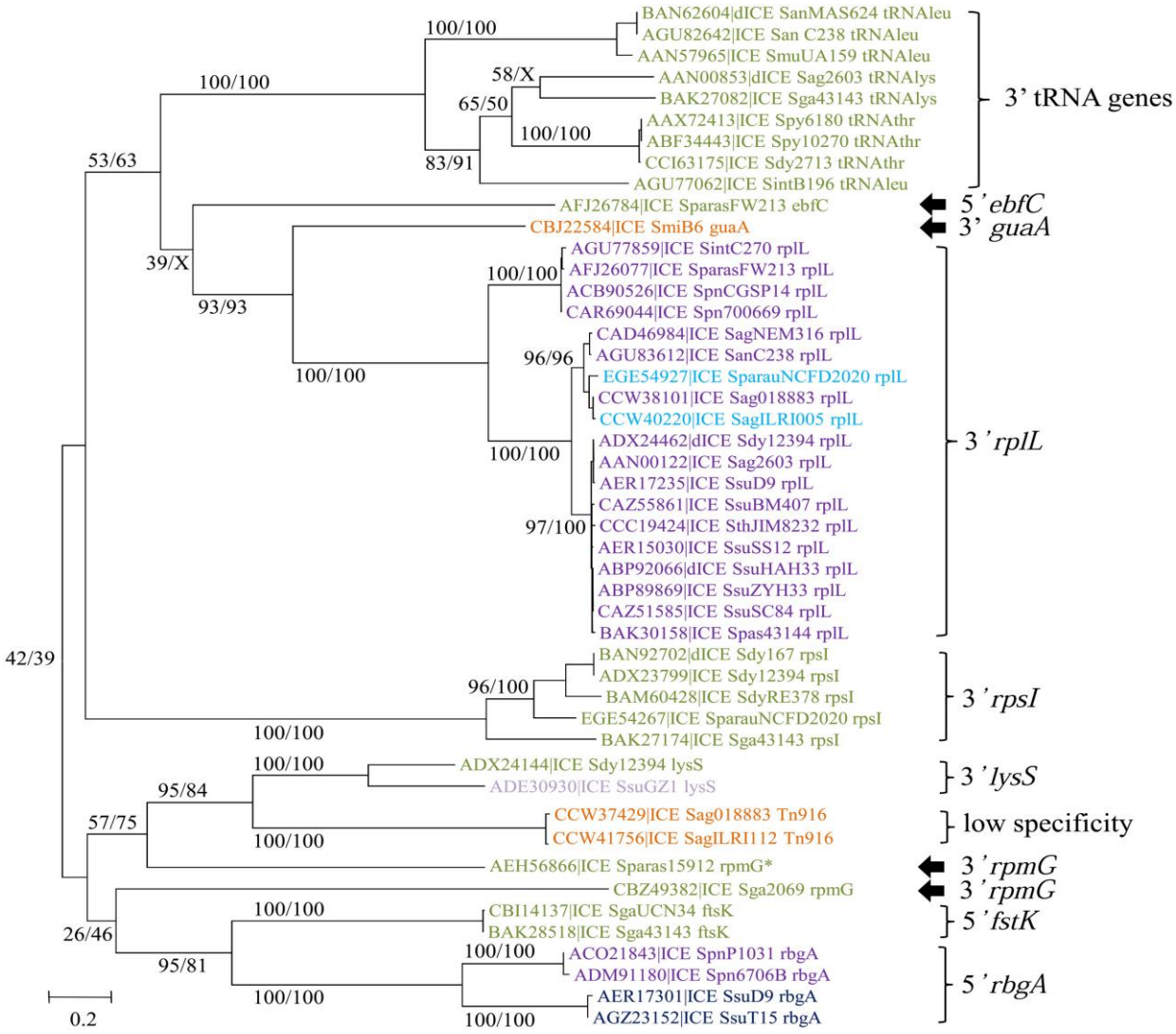

FIGURE 6 | Phylogenetic trees of tyrosine integrases from streptococcal ICEs/dICEs. Bootstrap values are given as followed: ML/BioNJ. X marks the nodes that are not validated with BioNJ. ICE names are colored according to the family of conjugation module they encode: orange = $Conj_{Tn916}$, green = $Conj_{ICESt3}$, mauve = $Conj_{vanG}$, dark-blue = $Conj_{Tn1549}$, purple = $Conj_{Tn5252}$, and in light-blue those of the $Conj_{TnGBS2}$. Brackets gather together closely related integrases sharing the same specificity of integration. Integration specificity of the ICEs/dICEs is indicated. Elements marked with an asterisk are not integrated in their primary sites but in secondary ones. Refer to Supplementary Table S1 for ICE/dICE and strain details.

integrases catalyzing integration in the 3′ end of the *rpmG* gene that display two distinct orientations and locations relative to the *rpmG* gene (**Figure 6**) and share only 21% of identity. Interestingly, all tyrosine recombinases catalyzing integration in a tRNA gene are grouped together. Half of the elements encoding a tyrosine integrase are inserted in the 3′ end of either tRNA CDSs (9/73) or a well-conserved housekeeping genes (30/73) such as *rplL* (L7/L12 ribosomal protein), *rpsI* (S9 ribosomal protein), *lysS* (lysyl-tRNA synthetase), *rpmG* (L33 ribosomal protein), or *guaA* (GMP synthase). In rare cases (7/73), the integration sites are found in the 5′ end of genes such as *ftsK* (DNA translocase involved in cell division), *rbgA* (ribosomal biogenesis GTPase) and *ebfC* (nucleoid associated protein). The remaining tyrosine integrases (27/73) catalyze low-specificity integration as previously shown for the very closely related integrase of Tn*916* (Scott et al., 1994).

All the ICEs/dICEs encoding a site-specific tyrosine integrase are flanked by DRs whose size ranges from 12 to 54 bp. DRs in *rplL*, *guaA* or in genes encoding tRNAthr only contain the exact 3′ end of the target genes. By contrast, DRs of elements integrated in

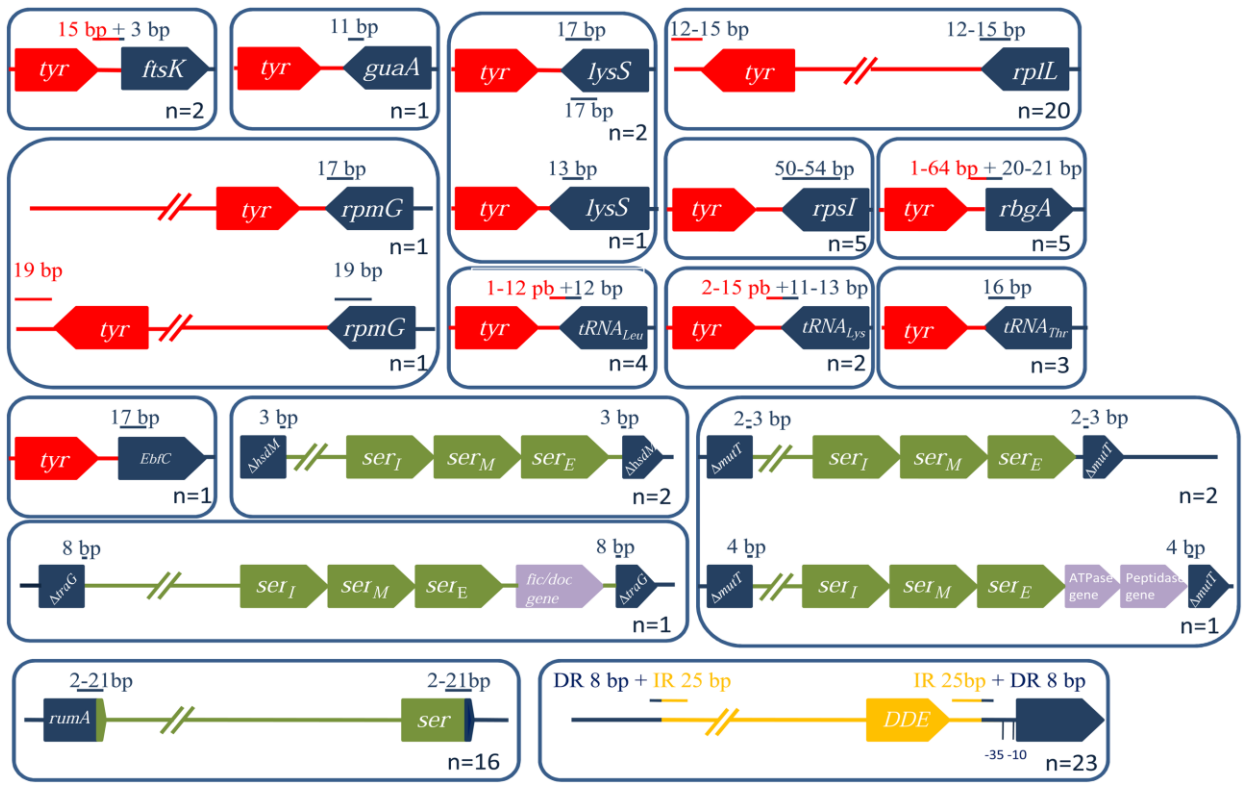

FIGURE 7 | Characterization of ICE/dICE integration loci and their position relative to the integrase CDSs. The genes within (or next to) which an ICE is inserted are in blue. Tyrosine integrases are in red, serine recombinases in green and DDE transposases in yellow. The sizes (in bp) of the DR (or of IRs when specified) are indicated in red when the sequence is inside of the conjugative element (in blue, outside). Numbers represent the number of ICEs integrated in a given target gene.

the 5′ end of *ftsK* and *rbgA*, and in the 3′ end of genes of tRNAleu and tRNAlys, overlap the flanking intergenic regions (**Figure 7**). Three elements are integrated in the *lysS* gene and are flanked by DRs with similar sequences. However, these DR sequences have distinct locations within *lysS* resulting from a slight difference in the length of this gene: one corresponds to the last 17 bp of the *lysS* gene while the others contain 13 bp and are more internal (**Figure 7**).

Prevalence and Integration Sites of ICEs and dICEs Encoding a Serine Integrase

A total of 42 serine integrases were identified: 32 from ICEs and 10 from dICEs. They all displayed an N-terminal catalytic and dimerization 'Resolvase' domain (PF00239) that contains a conserved serine residue. This domain is always associated with a 'Recombinase' PF07508 domain. All of them but 10 also contain a pfam13408 'Zinc ribbon recombinase' domain that is likely to play a DNA-binding role.

Streptococcal conjugative elements encode either a single serine integrase (14 ICEs and four dICEs) or a triplet of serine integrase genes (6 ICEs and 3 dICEs). The phylogenetic tree of the serine integrases is compatible with the existence of two groups of integrases (**Figure 8**). One of the groups clustered all the single serine integrases that target integration in *rumA* [23S rRNA (uracil-5-)methyltransferase]. ICEs/dICEs encoding such serine integrase are flanked by 2–21 bp-DRs localized at one end in the *rumA* gene and at the other end in the serine integrase gene (**Figure 7**). Integration leads to a reciprocal exchange of the 3′ part of the *rumA* and serine integrase genes leading to a modification of sequence and length of the C-terminal end of the corresponding proteins. It should be noticed that the replacement changes the translation frame of both proteins.

The second group of serine integrases is composed of the 24 integrases that are organized in triplets within the elements (**Figure 8**). Interestingly, these serine integrases are clustered according to their position within the triplets: external (E), middle (M), or internal (I) with respect to the ICE/dICE extremity (**Figure 8**). This suggests that all these modules derive from an ancestral module that already encoded three serine recombinases and that the presence of triplet results from two successive duplications. Triplets of serine integrases catalyze site-specific integration within several genes thus leading to their disruption: *mutT* (Nudix hydrolase), *traG* (CP of another ICE, dICE, or ICE remnant) and *hsdM* (methyltransferase subunit of

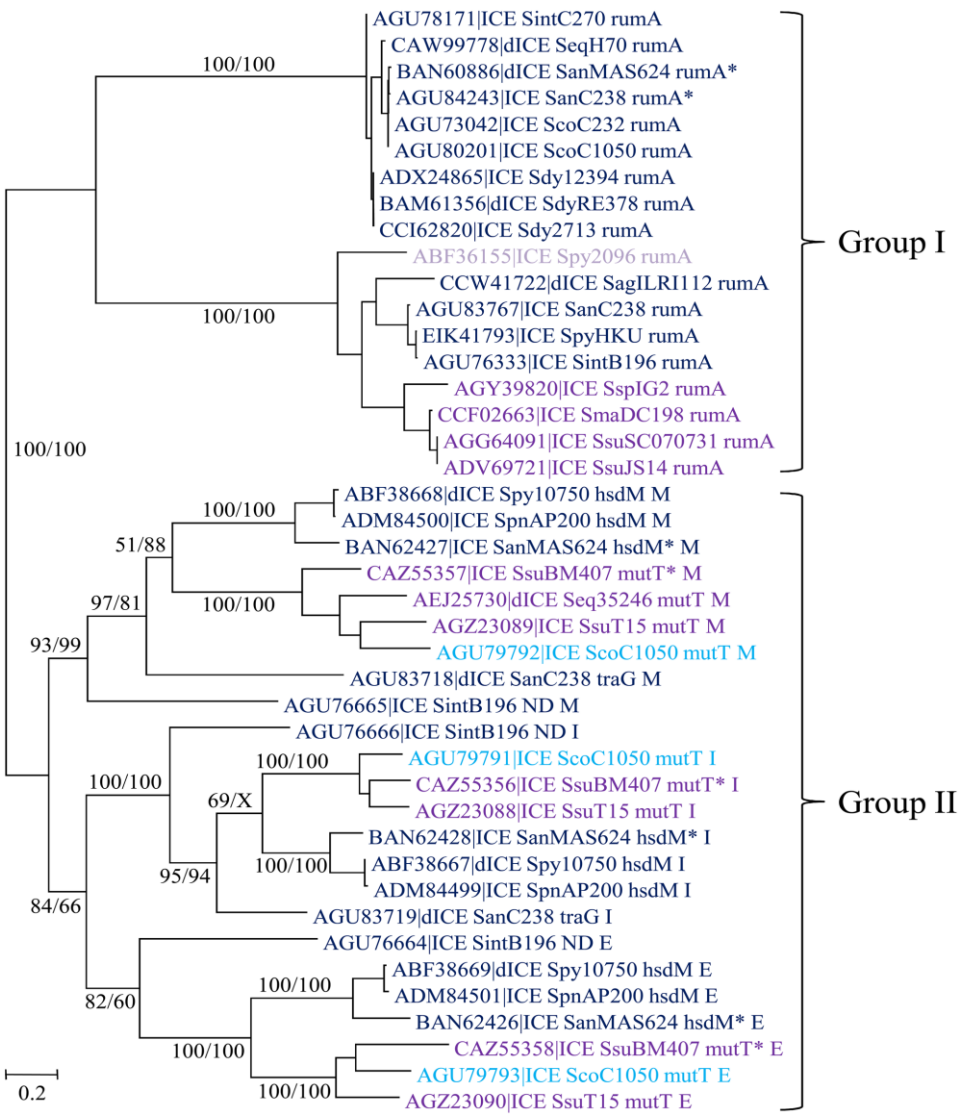

FIGURE 8 | Phylogenetic tree of serine integrases from streptococcal ICE/dICE. Bootstrap values are given as followed: ML/BioNJ. X marks the nodes that are not validated with MP. ICE name are colored according to the family of conjugation module they encode (dark-blue = $Conj_{Tn1549}$, purple = $Conj_{Tn5252}$, mauve = $Conj_{vanG}$), and in light-blue those of the $Conj_{TnGBS2}$. Brackets gather together closely related integrases sharing the same specificity of integration. Genes in which the ICEs/dICEs are integrated are indicated. Refer to Supplementary Table S1 for ICE/dICE and strain details. Elements marked with an asterisk are not integrated in their primary sites but in secondary ones.

type I restriction-modification systems). Insertion of ICEs/dICEs encoding triplets of serine integrases leads to very small 2–8 bp DRs (**Figure 7**). In some cases, one or several CDSs separated the triplets of serine integrases from the target gene (**Figure 7**). All serine integrases have the same orientation with respect to the target gene.

In several cases (element names marked with an asterisk in **Figure 8**), the comparison of the observed integration sites with the ones of elements with closely related integrases strongly suggests that these elements are not integrated in their primary sites but in secondary ones. ICE*_SanMAS624_hsdM** disrupts a gene encoding a protein carrying the domain PF0267 ("Adenine nucleotide alpha hydrolase"). ICE*_SsuBM407_mutT** disrupts a gene encoding a luciferase-like protein. dICE*_SanMAS624_rumA** and ICE*_SanC238_rumA** disrupts *sstT*, a gene encoding a serine/threonine transporter. The relaxase, CP and VirB4 of these last two elements are very closely related (**Figure 4**) as well as their integrases (**Figure 8**). This suggests that they were both inherited from the last common ancestor of their hosts.

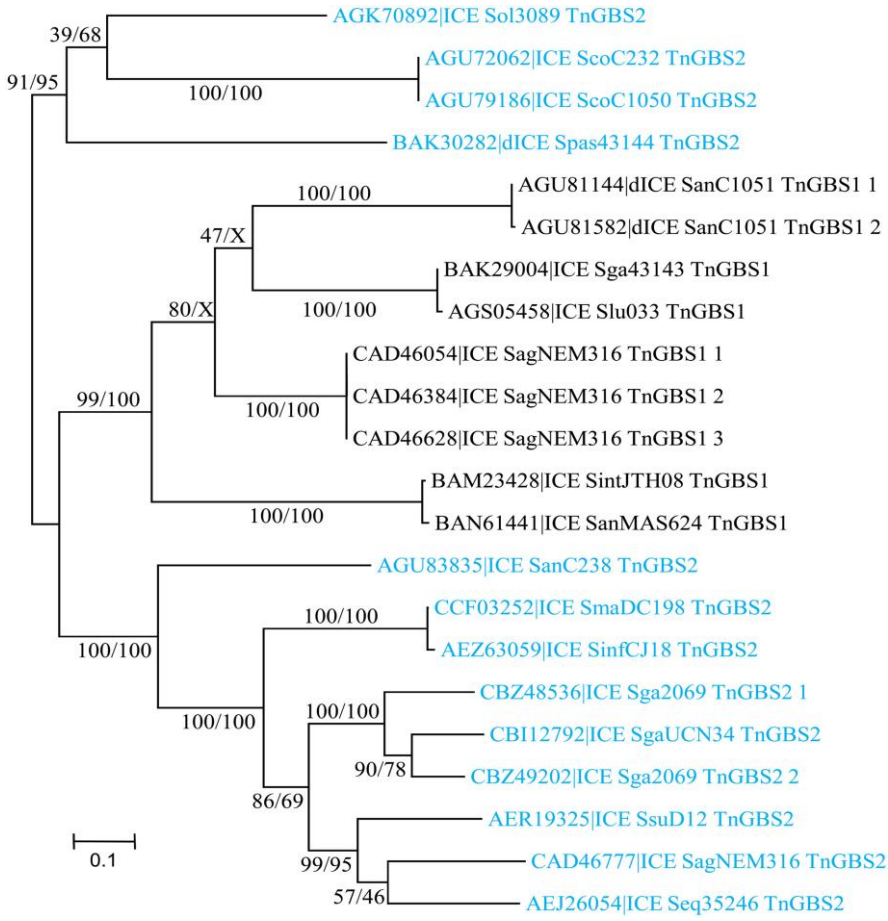

FIGURE 9 | Phylogenetic tree of DDE transposases from streptococcal ICE/dICE. Bootstrap values are given as followed: ML/BioNJ. ICE name are colored according to the family of the conjugation module they encode (black = $Conj_{TnGBS1}$, light-blue = $Conj_{TnGBS2}$. Refer to **Table 1** for ICE/dICE and strain details.

Prevalence and Integration Sites of ICEs/dICEs Encoding a DDE Transposase

A total of 23 DDE transposases related to those encoded by Tn*GBS1* and Tn*GBS2* elements (Guérillot et al., 2013) were detected. All show an "Uncharacterized protein family" PF06782 conserved domain with unknown function. Alignment of the amino acid sequences of the most divergent DDE transposases revealed that they share 42% identity showing that all these transposases belong to the same family. However, DDE transposases encoded by ICEs carrying a $Conj_{Tn\,GBS1}$ module are clustered together and are distinct from those encoded by ICEs with a $CONJ_{Tn\,GBS2}$ module (**Figure 9**).

Analysis of the junction sequence of ICEs encoding DDE transposases shows an 8-bp DR sequence that results from the duplication of the target sequence. Comparison of the insertion sites between all ICEs and dICEs of this group did not reveal any significant sequence similarity among the duplicated sequences but they are all located 15–16 pb upstream from −35 boxes of sigma A promoters as previously reported (Brochet et al., 2009).

Diversity and Evolution of ICEs/dICEs

The determination of the limits of each element allows the comparison of their size. If one excludes the $Conj_{vanG}$ ICE family (only two elements), the $Conj_{Tn916}$ and $Conj_{Tn\,GBS1}$ families are the most homogeneous in size with elements from 18 to 26 kb and from 40 to 53 kb, respectively. By contrast, the size of the $Conj_{ICESt3}$ elements is much more variable: most of them are 19- to 37-kb long and one exceeds 60 kb. The size of the $Conj_{Tn1549}$ elements can double (from 36 to 72 kb) and that of $Conj_{Tn\,GBS2}$ elements shows a very large disparity (from 25 to 82 kb). When ICEs/dICEs are present within *Streptococcus* genomes, they contribute to 1–13% of chromosomal DNA.

Analysis of streptococcal ICEs and dICES also allowed determining their typical combination of conjugation and integration/excision modules (**Table 3**). Almost all conjugation modules of the $Conj_{Tn916}$ family are associated with a tyrosine

integrase identical, or almost identical, to the one of Tn*916*, that is known to have a low specificity of integration (Scott et al., 1994). The only exception is ICE*_SmiB6_guaA* that encodes a conjugation module of the Conj$_{Tn916}$ family but is site-specifically integrated in the 3′ end of *guaA*. Elements with a Conj$_{ICESt3}$ module are associated with site-specific tyrosine integrases catalyzing integration in eight distinct integration sites (3′ end of three types of tRNA encoding genes, *rpsI*, *rpmG,* and *lysS* as well as the 5′ end of *ftsK* or *ebfC*). The conjugation modules of Conj$_{Tn5252}$ superfamily can be associated with a tyrosine recombinase, a serine site specific recombinase or a DDE transposase. Thus, ICEs and dICEs carrying a Conj$_{Tn1549}$ module are associated with: (i) a single serine integrase catalyzing the insertion in *rumA*; (ii) a triplet of serine integrases catalyzing the insertion in *hsdM* or *traG* or (iii) a tyrosine integrase catalyzing the insertion in the 5′ end of *rbgA*. Those carrying a Conj$_{Tn5252}$ module can encode: (i) a tyrosine integrase catalyzing the integration in the 3′ end of *rplL* or the 5′ end of *rbgA*; (ii) a single serine integrase catalyzing the integration in *rumA*; (iii) or a triplet of serine integrases catalyzing the insertion in *mutT*. The two ICE (ICE*_SsuGZ1_lysS* and *ICE_Spy2096_rumA*) displaying a Conj$_{vanG}$ conjugation module are associated with a tyrosine or a serine integrase, respectively. All the Conj$_{TnGBS1}$ and 14 of the 17 Conj$_{TnGBS2}$ conjugation modules are associated with a DDE transposase. However, three ICEs carrying a Conj$_{TnGBS2}$ module are not associated with a DDE transposase: ICE*_Sco1050_mutT* encodes a triplet of serine recombinases and both ICE*_SparauNCFD2020_rplL* and ICE*_SagILRI005_rplL* encode a tyrosine integrase. Sequence comparison suggests that ICE*_SagILRI005_rplL* is composite (**Figure 10**). Its Conj$_{TnGBS2}$ conjugation module is closely related to the one of ICE*_SgaUCN34_TnGBS2*. However, its left part is closely related to the left part of ICE*_Sag018883_rplL*, an ICE carrying a conjugation module belonging to CONJ$_{Tn5252}$ family. It includes not only the recombination module but also a lactose utilization module and a pseudogene of relaxase typical of the Conj$_{Tn5252}$ family. The right end of ICE*_SagILRI005_rplL* carries a gene encoding a repA_N domain closely related to a gene located in the right of ICE*_Sga018883_rplL* and additionally a pseudogen, whose product also carry a repA_N domain.

DISCUSSION

Detection of ICEs and dICEs in Streptococcal Genomes

Burrus et al. (2002a), a precursor analysis of 24 genomes from various Firmicutes, led to the identification of 17 putative ICEs and suggested that these elements are widespread at least in this division of bacteria. Over the last decade, with the revolution of sequencing technology, the number of fully sequenced bacterial genomes greatly increased. By the same time, efforts were made to improve *in silico* analysis of the data sets and in particular those allowing the detection of genomic islands including ICEs. Almost all searches of ICEs in bacterial genomes were only performed with a strain-centric or an ICE family centric point of view. However, few recent studies also reported extensive characterization of these conjugative elements. The first extensive study identified 335 chomosomal conjugative modules by scanning 1124 genomes of prokaryotes for conjugative genes (using HMM profiles of conjugative proteins of essentially proteobacterial plasmids; Guglielmini et al., 2011). Soon after, Ghinet et al. (2011) reported the characterization of 161 ICEs within 275 genomes of Actinobacteria but did not identify their limits. More recently, Puymège et al. (2015) searched for genetic elements integrated in the tRNAlys $_{CTT}$ gene in 303 genomes of *S. agalactiae,* leading to the identification and delimitation of 108 putative ICEs or derivatives. It should be noticed that in 2012, Bi et al. (2012) developed a web database[6], compiling information on ICEs from both Gram$^+$ and Gram$^-$ bacteria. However, if this database has the merit to list a large number and a great diversity of elements, it was not updated since November 2012 and some studies describing novel streptococcal ICEs published before this date escaped to the attention of the authors (for example, Brochet et al., 2008, reporting 10 novel ICEs). ICEberg have limitations considering ICEs from *Streptococcus* since (i) about half of ICE/dICE boundaries are incorrectly delimitated in ICEberg and (ii) the insertion site of many of them, although published, is not registered in this database. More widely, information on numerous elements from Firmicutes is inconsistent or wrong. Thus, while the authors indicated that a family should include only elements that carry both related integration and conjugation modules, they also included in the Tn*916* family ICEs that carry Tn*916*-unrelated integration modules (such as Tn*5397* or ICE*Lm1*), Tn*916*-unrelated conjugation modules (such as Tn*1549*) or elements completely unrelated to Tn*916* but carrying conjugation modules related to the one of Tn*1549* (such as CTn*2* or CTn*5*). Furthermore, although ICE*Lm1* and Tn*5801* carry almost

[6]http://db-mml.sjtu.edu.cn/ICEberg/index.php

TABLE 3 | Various combinations of conjugation and integration modules in ICEs and dICEs.

Conjugation module		Number of elements	Integrases or transposases			
Superfamily	**Family**		**Tyrosine**	**Serine (Single)**	**Serine (Triplet)**	**DDE**
Conj$_{Tn916}$	Conj$_{Tn916}$[1]	32	28	0	0	0
	Conj$_{ICESt3}$	21	21	0	0	0
Conj$_{Tn5252}$	Conj$_{Tn5252}$[2]	29	19	4	3	0
	Conj$_{Tn1549}$[3]	21	2	13	5	0
	Conj$_{TnGBS2}$	17	2	0	1	14
	Conj$_{vanG}$	2	1	1	0	0
Conj$_{TnGBS1}$	Conj$_{TnGBS1}$	9	0	0	0	9

[1]*Four Conj$_{Tn916}$ elements are deprived of an integrase CDS (absent or as a pseudogene).*
[2]*Three Conj$_{Tn5252}$ elements are deprived of an integrase CDS (absent or as a pseudogene).*
[3]*One Conj$_{Tn1549}$ element exhibits an integrase pseudogene.*

identical integration and conjugation modules (Burrus, plasmid 2002, cited in ICEberg for ICE*Lm1*), they were included in different families, Tn*916* and Tn*5801*, respectively. Even more problematically, the Tn*1207.3* family and 10750-RD.1 family only contain elements unrelated to ICEs, prophages for the first one and highly decayed derivates of integrative mobilizable elements for the second one. In general, these failures (and many others not mentioned here) make this database unreliable and very difficult to use for ICEs from *Streptococci* and other Firmicutes.

Here, we present the results of ICE detection within 124 complete streptococcal genomes. Our search is based on an iterative search for genes encoding signature proteins from ICEs. The co-occurrence of an integrase and three proteins of the conjugation module guaranties the retrieval of ICEs or dICEs. When one or two signature CDSs appeared to be a pseudogene or to be absent, an analysis of the whole element was undertaken to confirm its nature.

This work led to the identification and characterization in 63 *Streptococcus* genomes of 131 ICEs/dICEs whose extremities were precisely mapped on the genome. Elements that were already precisely identified are marked by a reference in Supplementary Table S1.

Distribution of ICEs and dICEs in the Different Streptococcal Species

Among the 27 streptococcal species analyzed, all, except 5 (*S. iniae*, *S. gordonii*, *S. uberis, S. sanguinis,* for which only one complete genome is available and *S. salivarius* for which three genome sequences exist), contain ICEs/dICEs showing the ubiquity of these elements in *Streptococcus*. *S. suis* appears as the species containing the highest number of ICEs/dICEs since 61% of the strains carry at least one ICE/dICE. However, the high prevalence of ICEs/dICEs in this species might be due to strain sampling since many strains carrying an ICE are related. This is also the case for *S. pneumoniae* genomes of which 40% (11/28) encode at least one ICE. Among them, seven are derived from clinical isolates known to be resistant to one or multiple antibiotics. All of them carry at least one ICE/dICE suggesting a possible correlation between the presence of ICEs and resistance to antibiotics. Indeed among the families detected, Tn*916* (Roberts and Mullany, 2011), Tn*5252* (Korona-Glowniak et al., 2015), and Tn*1549* (Garnier et al., 2000) are known vectors of antibiotic resistance genes.

Definition of Different Families of Elements on the Basis of their Conjugative Module

These ICEs were classified into seven distinct families belonging to three superfamilies on the basis of their conjugation modules: (i) $Conj_{Tn916}$ and $Conj_{ICESt3}$ belonging to the $Conj_{Tn916}$ superfamily, (ii) $Conj_{vanG}$, $Conj_{Tn5252}$, $Conj_{Tn1549}$, and $Conj_{TnGBS2}$ belonging to the $Conj_{Tn5252}$ superfamily and the $Conj_{TnGBS1}$ family. The $Conj_{Tn916}$ and the $Conj_{Tn5252}$ superfamilies of conjugation modules belong respectively to the MPF_{FA} and to the MPF_{FATA} classes of T4SS involved in bacterial conjugation as defined by Guglielmini et al. (2014). No match was found for the $Conj_{TnGBS1}$ family.

Modular Evolution

The phylogenetic trees obtained for relaxase, CP and VirB4 encoded by elements belonging to the $Conj_{Tn916}$ superfamily (**Figure 2**), the $Conj_{Tn5252}$ superfamily (**Figure 4**), and the $Conj_{TnGBS1}$ family (**Figure 5**) are highly similar, suggesting that genes exchanges or replacements within the conjugation modules of these families have not occurred or are rare. However, incongruences were found for relaxases, CPs and VirB4 proteins of some elements belonging to ICE*St3* family, suggesting that some gene exchanges or replacements have occurred within the conjugation modules belonging to this family.

Unrelated or very distantly related integrases were found to be encoded by at least some of the ICEs belonging to the same family (except for Tn*GBS1* family) and frequently by closely or very related elements (for example in the Tn*5252* family). Furthermore, related site-specific integrases were found in unrelated or distantly related ICEs. Such incongruences are due to multiple exchanges of integration/excision and/or conjugation modules between and/or within ICE families. For most cases, the data do not allow to determine what precisely happened. However, it was previously reported that the last common ancestor of Tn*GBS2* family acquired its DDE transposase from an insertion sequence and that the last common ancestor of Tn*GBS1* family acquired its DDE transposase from an ICE belonging to Tn*GBS2* family (Guérillot et al., 2013). Here, the comparison of phylogenetic trees obtained for $CONJ_{Tn5252}$ superfamily, serine integrases and tyrosine integrases clearly shows three independent replacements of the DDE transposases by unrelated integrases. One of these ICEs, *ICE_SagILRI005_rplL,* probably results from: (i) the integration of an ICE belonging to Tn*GBS2* family (encoding a low specificity/site-preferential DDE transposase and a RepA_N protein) into an ICE belonging to Tn*5252* family (encoding a tyrosine integrase specific of *rplL* and another RepA_N protein), and (ii) the loss of the conjugation and replication modules of the Tn*5252*-related element and the deletion of the DDE transposase gene of the Tn*GBS2*-related ICE.

Integration Specificity of ICEs in Streptococcal Genomes: Impact on Host Fitness and on the Evolution of Elements

In this work, efforts were made to identify the boundaries of the ICEs and therefore to identify the insertion site of each of them. This analysis of integration/excision modules and site specificity is the first one carried for a large array of ICEs encoding their transfer as single-stranded DNA. The ICEs and dICEs of *Streptococci* carry diverse integration/excision modules (75 encoding a tyrosine recombinase, 20 encoding a unique serine recombinase, nine encoding three serine recombinases, and 23 a DDE transposase) and have a large array of integration specificity (low or preferential integration, 18 different site-specific integrations). This work led to the detection of eight new

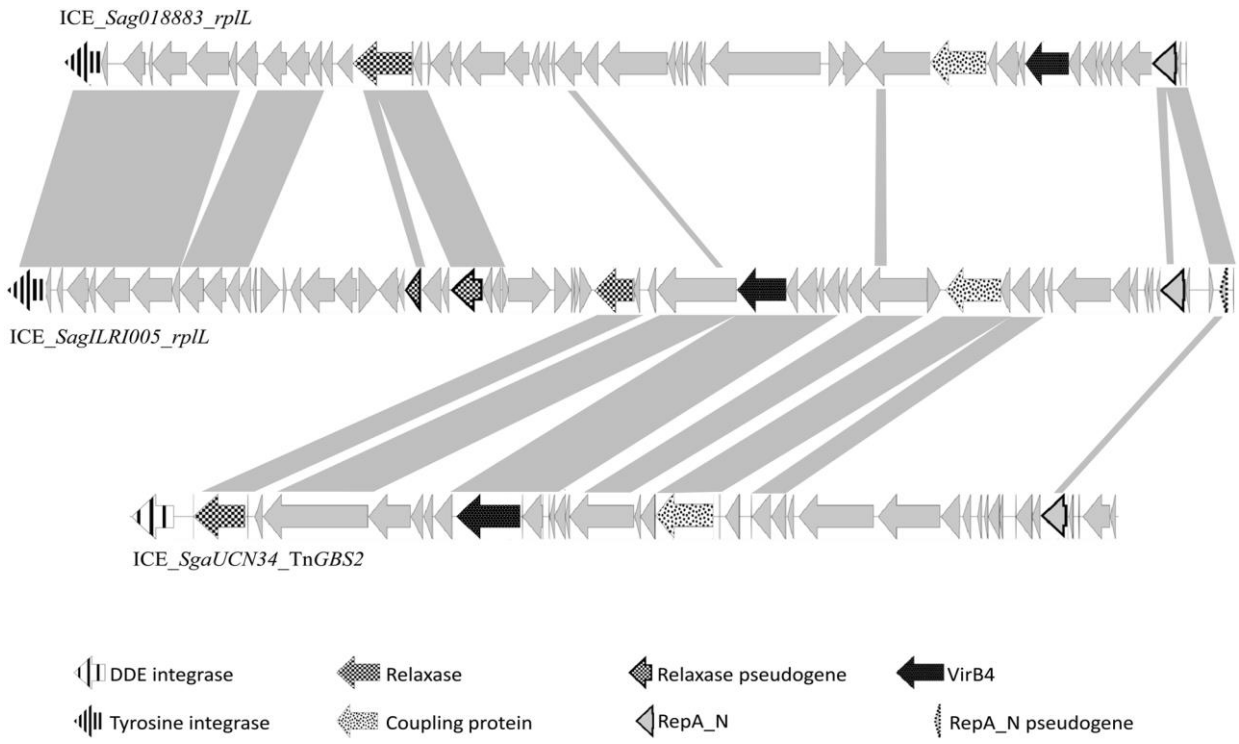

FIGURE 10 | Genetic organization of ICE_*SagILRI005_rplL* and comparison against ICE_*SgaUCN34_TnGBS2* and ICE_ *Sag018883_rplL*. For the comparison, gray shading between the genetic elements represents regions with >50% amino acid sequence identity. The arrows represent the individual ORFs. Putative functions of the conserved genes are those deduced from the functional annotation of the ICE using Agmial.

target-genes for streptococcal ICE insertion that have not been identified previously (*ftsK*, *guaA*, *lysS*, *mutT*, *rpmG*, *rpsI*, *traG*, and *ebfC*).

It should be noticed that, among the 131 ICEs/dICEs identified, only nine (restricted to one family) were found to be integrated into the 3′ end of genes encoding tRNAs. This contrasts with the results of the analysis of actinobacterial ICEs (most of these ICEs carry conjugation modules unrelated to the ones of streptococcal ICEs and transfer as double-stranded DNA). Among the 144 actinobacterial ICEs analyzed, 100 were found integrated in the 3′ end of a tRNA gene (Ghinet et al., 2011).

In most streptococcal ICEs, as for almost all other known ICEs, the *attI* site is located in the vicinity of the integrase gene. However, it should also be noticed that for all streptococcal ICEs/dICEs integrated in *rumA*, the *attR* site is located within the serine integrase gene and consequently the integrase gene carried by the excised ICE has a different length and C terminus. *att* sites are found within the genes of their cognate integrase in very few integrative elements, such as the prophage Mx8 from *Myxococcus xanthus* for which the phage attachment site, *attP*, is located within the tyrosine integrase gene (Magrini et al., 1999). Site-specific integration of Mx8 leads to the replacement of the 112-residue C-terminal sequence by a 13-residue C terminus. This modified integrase is less active that the integrase encoded by the excised element. Therefore, it seems probable that the differences between the integrases encoded by the ICEs integrated in *rumA* and the integrases encoded by the excised elements can lead to differences in the function of the two forms of the integrase.

Besides mechanistic constraints leading to integration in conserved palindromic sequences for many tyrosine integrases (Williams, 2002), one would expect that selection criteria for ICE integration in evolution would be (i) to have the least effect on host fitness and, (ii) since many have a large host range (Bellanger et al., 2014), to allow integration into a wide range of strains and species. Numerous ICEs encoding a tyrosine integrase were found to be site-specifically integrated in the conserved 3′ end of essential conserved genes that are isolated or are the last gene of an operon. The insertion does not modify the gene product (tRNA, ribosomal proteins) or leads to very little modification of the 3′ end of the protein (lysyl-tRNA synthetase). Hence, such integrations would have no effect on host fitness and can occur in a large array of species.

All (except one) the streptococcal ICEs belonging to the Tn*916* family detected in this study encode a tyrosine integrase identical or almost identical to the Tn*916* integrase.

Analyses of a high number of insertion sites in various hosts showed that Tn*916*, despite having a low specificity of integration, still has a preference for AT rich regions (consensus TTTTTnnnnnnAAAAA; Hosking et al., 1998; Cookson et al., 2011). Furthermore, the analyses of insertion sites after conjugal transfer to *Butyrivibrio proteoclasticus* B316^T, whose genome has a similar GC percent as the ones of *Streptococci* (39%), showed that only 34% of the 123 analyzed insertions disrupt annotated ORFs even if 90% of this genome is made of ORFs (Cookson et al., 2011). This may be due to lower GC ratio (34.7%) of intergenic regions. Therefore, the AT-rich region preference of Tn*916* probably leads to a null or low impact of most Tn*916* insertion events on host fitness. Since MGEs have generally a lower G+C percent than their host genome (Rocha and Danchin, 2002), this preference could also explain the frequent presence of Tn*916* or Tn*916*-related elements in plasmids or Tn*5252*-related elements (Clewell and Gawron-Burke, 1986; Ayoubi et al., 1991; Ding et al., 2009; Mingoia et al., 2011; Chancey et al., 2015). This putative preference for MGEs would also lead to a lower impact on the fitness of the bacterial host. It was previously shown that a Tn*916* element carried by a Tn*5252*-related element can be transferred alone (Santoro et al., 2010) or as a part of the Tn*5252* element (Ayoubi et al., 1991). Therefore, besides a low impact on host fitness, this A+T rich preference could increase the transfer ability of Tn*916* (either autonomously or by mobilization *in cis*).

Some ICEs integrate in the conserved 5′ end of the first gene of an operon or of an isolated gene that encodes an essential protein (the DNA translocase FtsK that coordinates cell division and chromosome segregation; the nucleoid-associated protein EbfC; the ribosome assembly GTPase RbgA). Importantly, the insertion does not modify the N-terminus of the protein encoded by the target gene. Moreover, for three ICEs integrated in *rbgA* and two ICEs integrated in *ftsK*, the integration does not change the 15–64 bp sequence located upstream from the START codon. However, in the two other ICEs integrated in *rbgA* and in the one integrated in *ebfC*, the sequence upstream from the gene, including its promoter, is completely different, suggesting that the expression of the gene is impacted by the integration of the element. This situation is reminiscent of the integration of the putative satellite prophage SpyCI from *S. pyogenes*. In stationary phase, SpyCI is integrated into the 5′ end of the DNA mismatch repair gene *mutL*, disrupting its expression and that of three other genes located downstream (Nguyen and McShan, 2014). During early exponential growth, SpyCI excises from the bacterial chromosome and replicates as an episome, thus allowing the expression of *mutL* and of downstream genes. Concerning the *ebfC* gene, it is known that in the spirochaete *Borrelia burgdorferi*, it is highly expressed in rapidly growing bacteria but mRNA is undetectable in stationary phase (Jutras et al., 2012). Thus, ICE integration in the promoter of this gene in *S. parasanguinis* FW213 may not alter host fitness at all: expression of *ebfC* gene would not be required in the stationary phase when the element is integrated and excision of the element in the exponential phase restores the expression of this gene.

Tn*GBS1* and Tn*GBS2* are two known elements from *S. agalactiae* encoding DDE transposases that integrate in various intergenic regions located 15 or 16 bp upstream from the 35 box of sigma A promoters (Brochet et al., 2009; Guérillot et al., 2013). In this study, all elements belonging to Tn*GBS1* and Tn*GBS2* families (except three Tn*GBS2* that encode serine or tyrosine recombinases) are also integrated in such location. Insertion into intergenic regions is expected to minimize the effects caused by the transposon insertion on host fitness. However, such insertions may interfere with the transcription level of the downstream gene. However, it should be noticed that the insertion of Tn*GBS2* does not seem to affect significantly the transcription level of the gene located downstream from the preferred insertion site.

All insertions of the 29 ICEs encoding serine integrase disrupt genes encoding proteins. A large majority are site-specific. Most of these genes are widespread but are not essential for the strain (*rumA*, *hsdM*, *mutT*). The *rumA* gene encodes a widespread rRNA methyltransferase. In *Escherichia coli*, the deletion of this gene has little effect on growth or on the fidelity of translation, but alters the sensitivity of the ribosomes to fusidic acid and capreomycin (Persaud et al., 2010). The *hdsM* gene encodes the methyltransferase subunit of type I restriction-modification systems (Murray, 2000). The *mutT* gene encodes a Nudix hydrolase that removes oxidized nucleotide precursors so that they cannot be incorporated in DNA during replication (Lu et al., 2001). One element, dICE*_SanC238_traG* is site-specifically integrated into the *traG* gene that encodes the CP of an ICE remnant (not detailed in this report because this remnant is too much decayed) belonging to Tn*1549* family.

The consequences of the integration/excision balance of ICE encoding serine recombinases on the expression of the target genes encoding proteins have never been studied. However, several examples can be cited for prophages or prophage-related elements (Stragier et al., 1989; Kunkel et al., 1990; Rabinovich et al., 2012). Thus, the DNA uptake competence system of the intracellular bacterial pathogen *Listeria monocytogenes* serovar 1/2 was considered non-functional because the competence master activator gene, *comK*, is disrupted by the insertion of the temperate prophage A118 encoding a serine recombinase (Rabinovich et al., 2012). However, the prophage excises not only during the activation of lytic phase but also during intracellular growth, primarily within phagosomes of macrophages, without any production of progeny virions, thus allowing expression of the *comK* gene (Rabinovich et al., 2012). In the same way, ICEs integrated within specific genes and disrupting them may excise when these genes are useful for the host cell (*rumA*, *hsdM*, *mutT*) or for the host ICE (*traG*) to reduce the impact on host fitness and guarantee their maintenance in the cell. Four elements encoding serine integrases are integrated in secondary sites within protein-encoding genes. As for primary integration sites, if the elements are still able to excise, expression of the target gene might not be impacted. The integration of CTn*5*, an ICE belonging to the Tn*1549* family and encoding a serine integrase occurs in an adhesin gene in *Clostridium difficile* 630 (Sebaihia et al., 2006). However, comparison of this genome with that of the derived strain 630Δerm showed that CTn*5* has excised from its original location and has inserted in *rumA* (CD3393) of 630Δerm (van Eijk et al., 2015). This suggests that an ICE integrated in a secondary site is able to excise and reintegrate in its primary site and conversely.

Globally, the impact of the integration upstream from promoters, in the 5′ end of CDSs or within CDSs could be reduced if the ICE excises when the targeted genes are expressed. It was initially thought that ICEs do not replicate autonomously in the cell, although conjugative transfer can be seen as an intercellular replication (Burrus et al., 2002b). Therefore, although the excision could be advantageous for the host, if the cell divides when the ICE is excised, the ICE would be lost in one of the daughter cell. Nevertheless, several recent studies showed or strongly suggested that various single-strand DNA transferring ICEs are capable of extrachromosomal replication in both Gram-positive and Gram-negative bacteria (Carraro et al., 2015 and references therein). In particular, replication was found to be involved in maintenance of Tn*GBS1* and Tn*GBS2* in transconjugants before their integration (Guérillot et al., 2013). These elements encode a protein that carries a repA_N domain and is related to the protein controlling the θ replication of various plasmids and a protein related to ParA, a protein involved in maintenance of some plasmids (Guérillot et al., 2013). We found genes encoding a protein with a RepA_N domain in all ICE belonging to the Tn*5252* family and numerous ICEs belonging to Tn*1549* family. We also found ParA segregation ATPase CDSs in all ICEs belonging to Tn*1549* family and *vanG* family suggesting that all these elements can replicate as episomes. Evidence of intracellular extrachromosomal replication was also recently obtained for ICE*St3* (Carraro et al., 2011) and another element belonging to ICE*St3* family, RD2 (i.e., ICE_*Spy6180_tRNAthr*) of *S. pyogenes* (Sitkiewicz et al., 2011) although these elements do not carry any replication module. Moreover, extrachromosomal replication of ICE*Bs1* from *Bacillus subtilis*, an element belonging to the Tn*916* superfamily (Burrus et al., 2002a) was found to be involved in the stability of the element. This intracellular replication is initiated from the ICE*Bs1 oriT* and required the ICE*Bs1*-encoded relaxase (Lee et al., 2010). At last, all ICEs belonging to the Tn*916* superfamily encode a peculiar relaxase (MOB_T) related to rolling circle replication initiators involved in maintenance of various plasmids (Guglielmini et al., 2014), suggesting that all these elements are also able to replicate as episomes.

CONCLUSION

This study greatly enriches our understanding of the classification and integration sites of ICEs/dICEs in streptococci genomes. In the future, it will be updated and further extended to take into account newly sequenced genomes and to confirm all the trends proposed here. An automated bioinformatics procedure will be developed to keep pace with the constantly growing number of available genomes. Extension to other species of Firmicutes and to the search for IMEs is also envisaged.

AUTHOR CONTRIBUTIONS

GG and SP conceived the reference database of signature proteins. GG, SP, NL-B, and M-DD contributed to the conception of the work. CA, CC, GG, NL-B, M-DD, SP, VL, and TL were involved in the acquisition and/or the analysis of the data. NL-B, GG, CC, and SP contribute to the drafting of the manuscript. NL-B and CC elaborated the figures and tables. All authors criticized and finally approved this final version.

ACKNOWLEDGMENTS

CC is recipient of a scholarship of the Ministère de l'Enseignement Supérieur et de la Recherche. This work received financial support from the Région Lorraine and Université de Lorraine.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.01483

REFERENCES

Arrieta, M. C., Stiemsma, L. T., Amenyogbe, N., Brown, E. M., and Finlay, B. (2014). The intestinal microbiome in early life: health and disease. *Front. Immunol.* 5:427. doi: 10.3389/fimmu.2014.00427

Ayoubi, P., Kilic, A. O., and Vijayakumar, M. N. (1991). Tn5253, the pneumococcal omega (cat tet) BM6001 element, is a composite structure of two conjugative transposons, Tn5251 and Tn5252. *J. Bacteriol.* 173, 1617–1622.

Bellanger, X., Payot, S., Leblond-Bourget, N., and Guedon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol. Rev.* 38, 720–760. doi: 10.1111/1574-6976.12058

Bi, D., Xu, Z., Harrison, E. M., Tai, C., Wei, Y., He, X., et al. (2012). ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Res.* 40, D621–D626. doi: 10.1093/nar/gkr846

Brochet, M., Couve, E., Glaser, P., Guedon, G., and Payot, S. (2008). Integrative conjugative elements and related elements are major contributors to the genome diversity of *Streptococcus agalactiae*. *J. Bacteriol.* 190, 6913–6917. doi: 10.1128/JB.00824-08

Brochet, M., Da Cunha, V., Couve, E., Rusniok, C., Trieu-Cuot, P., and Glaser, P. (2009). Atypical association of DDE transposition with conjugation specifies a new family of mobile elements. *Mol. Microbiol.* 71, 948–959. doi: 10.1111/j.1365-2958.2008.06579.x

Bryson, K., Loux, V., Bossy, R., Nicolas, P., Chaillou, S., Van De Guchte, M., et al. (2006). AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. *Nucleic Acids Res.* 34, 3533–3545. doi: 10.1093/nar/gkl471

Burrus, V., Pavlovic, G., Decaris, B., and Guedon, G. (2002a). The ICESt1 element of *Streptococcus thermophilus* belongs to a large family of integrative and conjugative elements that exchange modules and change their specificity of integration. *Plasmid* 48, 77–97. doi: 10.1016/S0147-619X(02)00102-6

Burrus, V., Pavlovic, G., Decaris, B., and Guédon, G. (2002b). Conjugative transposons: the tip of the iceberg. *Mol. Microbiol.* 46, 601–610. doi: 10.1046/j.1365-2958.2002.03191.x

Carraro, N., Libante, V., Morel, C., Decaris, B., Charron-Bourgoin, F., Leblond, P., et al. (2011). Differential regulation of two closely related integrative and conjugative elements from *Streptococcus thermophilus*. *BMC Microbiol.* 11:238. doi: 10.1186/1471-2180-11-238

Carraro, N., Poulin, D., and Burrus, V. (2015). Replication and active partition of Integrative and Conjugative Elements (ICEs) of the SXT/R391 family: the line between ICEs and conjugative plasmids is getting thinner. *PLoS Genet.* 11:e1005298. doi: 10.1371/journal.pgen.1005298

Chancey, S. T., Agrawal, S., Schroeder, M. R., Farley, M. M., Tettelin, H., and Stephens, D. S. (2015). Composite mobile genetic elements disseminating macrolide resistance in *Streptococcus pneumoniae*. *Front. Microbiol.* 6:26. doi: 10.3389/fmicb.2015.00026

Chuzeville, S., Puymege, A., Madec, J. Y., Haenni, M., and Payot, S. (2012). Characterization of a new CAMP factor carried by an integrative and conjugative element in *Streptococcus agalactiae* and spreading in Streptococci. *PLoS ONE* 7:e48918. doi: 10.1371/journal.pone.0048918

Clewell, D. B., and Gawron-Burke, C. (1986). Conjugative transposons and the dissemination of antibiotic resistance in streptococci. *Annu. Rev. Microbiol.* 40, 635–659. doi: 10.1146/annurev.mi.40.100186.003223

Cookson, A. L., Noel, S., Hussein, H., Perry, R., Sang, C., Moon, C. D., et al. (2011). Transposition of Tn916 in the four replicons of the *Butyrivibrio proteoclasticus* B316(T) genome. *FEMS Microbiol. Lett.* 316, 144–151. doi: 10.1111/j.1574-6968.2010.02204.x

Croucher, N. J., Walker, D., Romero, P., Lennard, N., Paterson, G. K., Bason, N. C., et al. (2009). Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae* Spain23F ST81. *J. Bacteriol.* 191, 1480–1489. doi: 10.1128/JB.01343-08

Ding, F., Tang, P., Hsu, M. H., Cui, P., Hu, S., Yu, J., et al. (2009). Genome evolution driven by host adaptations results in a more virulent and antimicrobial-resistant *Streptococcus pneumoniae* serotype 14. *BMC Genomics* 10:158. doi: 10.1186/1471-2164-10-158

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.* 2, 414–424. doi: 10.1038/nrmicro884

Franciosi, E., Settanni, L., Cavazza, A., and Poznanski, E. (2009). Biodiversity and technological potential of wild lactic acid bacteria from raw cows' milk. *Int. Dairy J.* 19, 3–11. doi: 10.1016/j.idairyj.2008.07.008

Franke, A. E., and Clewell, D. B. (1981). Evidence for a chromosome-borne resistance transposon (Tn916) in *Streptococcus faecalis* that is capable of "conjugal" transfer in the absence of a conjugative plasmid. *J. Bacteriol.* 145, 494–502.

Garnier, F., Taourit, S., Glaser, P., Courvalin, P., and Galimand, M. (2000). Characterization of transposon Tn1549, conferring VanB-type resistance in *Enterococcus* spp. *Microbiology* 146(Pt 6), 1481–1489. doi: 10.1099/00221287-146-6-1481

Ghinet, M. G., Bordeleau, E., Beaudin, J., Brzezinski, R., Roy, S., and Burrus, V. (2011). Uncovering the prevalence and diversity of integrating conjugative elements in actinobacteria. *PLoS ONE* 6:e27846. doi: 10.1371/journal.pone.0027846

Goessweiner-Mohr, N., Arends, K., Keller, W., and Grohmann, E. (2013). Conjugative type IV secretion systems in Gram-positive bacteria. *Plasmid* 70, 289–302. doi: 10.1016/j.plasmid.2013.09.005

Gouy, M., Guindon, S. P., and Gascuel, O. (2010). SeaView Version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27, 221–224. doi: 10.1093/molbev/msp259

Groth, A. C., and Calos, M. P. (2004). Phage integrases: biology and applications. *J. Mol. Biol.* 335, 667–678. doi: 10.1016/j.jmb.2003.09.082

Guérillot, R., Da Cunha, V., Sauvage, E., Bouchier, C., and Glaser, P. (2013). Modular evolution of TnGBSs, a new family of integrative and conjugative elements associating insertion sequence transposition, plasmid replication, and conjugation for their spreading. *J. Bacteriol.* 195, 1979–1990. doi: 10.1128/JB.01745-12

Guglielmini, J., Neron, B., Abby, S. S., Garcillan-Barcia, M. P., De La Cruz, F., and Rocha, E. P. (2014). Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.* 42, 5715–5727. doi: 10.1093/nar/gku194

Guglielmini, J., Quintais, L., Garcillan-Barcia, M. P., De La Cruz, F., and Rocha, E. P. (2011). The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* 7:e1002222. doi: 10.1371/journal.pgen.1002222

Hacker, J., and Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.* 2, 376–381. doi: 10.1093/embo-reports/kve097

Hacker, J., and Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* 54, 641–679. doi: 10.1146/annurev.micro.54.1.641

Heather, Z., Holden, M. T., Steward, K. F., Parkhill, J., Song, L., Challis, G. L., et al. (2008). A novel streptococcal integrative conjugative element involved in iron acquisition. *Mol. Microbiol.* 70, 1274–1292. doi: 10.1111/j.1365-2958.2008.06481.x

Hosking, S. L., Deadman, M. E., Moxon, E. R., Peden, J. F., Saunders, N. J., and High, N. J. (1998). An in silico evaluation of Tn916 as a tool for generalized mutagenesis in *Haemophilus influenzae* Rd. *Microbiology* 144(Pt 9), 2525–2530. doi: 10.1099/00221287-144-9-2525

Juhas, M., Van Der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W., and Crook, D. W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol. Rev.* 33, 376–393. doi: 10.1111/j.1574-6976.2008.00136.x

Jutras, B. L., Chenail, A. M., and Stevenson, B. (2012). Changes in bacterial growth rate govern expression of the *Borrelia burgdorferi* OspC and Erp infection-associated surface proteins. *J. Bacteriol.* 195, 757–764. doi: 10.1128/JB.01956-12

Kohler, W. (2007). The present state of species within the genera *Streptococcus* and *Enterococcus*. *Int. J. Med. Microbiol.* 297, 133–150. doi: 10.1016/j.ijmm.2006.11.008

Korona-Glowniak, I., Siwiec, R., and Malm, A. (2015). Resistance determinants and their association with different transposons in the antibiotic-resistant *Streptococcus pneumoniae*. *Biomed. Res. Int.* 2015, 836496. doi: 10.1155/2015/836496

Kunkel, B., Losick, R., and Stragier, P. (1990). The *Bacillus subtilis* gene for the development transcription factor sigma K is generated by excision of a dispensable DNA element containing a sporulation recombinase gene. *Genes Dev.* 4, 525–535. doi: 10.1101/gad.4.4.525

Lee, C. A., Babic, A., and Grossman, A. D. (2010). Autonomous plasmid-like replication of a conjugative transposon. *Mol. Microbiol.* 75, 268–279. doi: 10.1111/j.1365-2958.2009.06985.x

Leonetti, C. T., Hamada, M. A., Laurer, S. J., Broulidakis, M. P., Swerdlow, K. J., Lee, C. A., et al. (2015). Critical components of the conjugation machinery of the Integrative and Conjugative Element ICEBs1 of *Bacillus subtilis*. *J. Bacteriol.* 197, 2558–2567. doi: 10.1128/JB.00142-15

Low, H. H., Gubellini, F., Rivera-Calzada, A., Braun, N., Connery, S., Dujeancourt, A., et al. (2014). Structure of a type IV secretion system. *Nature* 508, 550–553. doi: 10.1038/nature13081

Lu, A. L., Li, X., Gu, Y., Wright, P. M., and Chang, D. Y. (2001). Repair of oxidative DNA damage: mechanisms and functions. *Cell Biochem. Biophys.* 35, 141–170. doi: 10.1385/CBB

Magrini, V., Creighton, C., and Youderian, P. (1999). Site-specific recombination of temperate *Myxococcus xanthus* phage Mx8: genetic elements required for integration. *J. Bacteriol.* 181, 4050–4061.

Mingoia, M., Tili, E., Manso, E., Varaldo, P. E., and Montanari, M. P. (2011). Heterogeneity of Tn5253-like composite elements in clinical *Streptococcus pneumoniae* isolates. *Antimicrob. Agents Chemother.* 55, 1453–1459. doi: 10.1128/AAC.01087-10

Mitchell, T. J. (2003). The pathogenesis of streptococcal infections: from tooth decay to meningitis. *Nat. Rev. Microbiol.* 1, 219–230. doi: 10.1038/nrmicro771

Murray, N. E. (2000). Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). *Microbiol. Mol. Biol. Rev.* 64, 412–434. doi: 10.1128/MMBR.64.2.412-434.2000

Nakajima, T., Nakanishi, S., Mason, C., Montgomery, J., Leggett, P., Matsuda, M., et al. (2013). Population structure and characterization of viridans group streptococci (VGS) isolated from the upper respiratory tract of patients in the community. *Ulster Med. J.* 82, 164–168.

Nguyen, S. V., and McShan, W. M. (2014). Chromosomal islands of *Streptococcus pyogenes* and related streptococci: molecular switches for survival and virulence. *Front. Cell. Infect. Microbiol.* 4:109. doi: 10.3389/fcimb.2014.00109

Nguyen, V. H., and Lavenier, D. (2009). PLAST: parallel local alignment search tool for database comparison. *BMC Bioinformatics* 10:329. doi: 10.1186/1471-2105-10-329

Park, H. K., Shim, S. S., Kim, S. Y., Park, J. H., Park, S. E., Kim, H. J., et al. (2005). Molecular analysis of colonized bacteria in a human newborn infant gut. *J. Microbiol.* 43, 345–353.

Persaud, C., Lu, Y., Vila-Sanjurjo, A., Campbell, J. L., Finley, J., and O'connor, M. (2010). Mutagenesis of the modified bases, m(5)U1939 and psi2504, in *Escherichia coli* 23S rRNA. *Biochem. Biophys. Res. Commun.* 392, 223–227. doi: 10.1016/j.bbrc.2010.01.021

Puymège, A., Bertin, S., Guedon, G., and Payot, S. (2015). Analysis of *Streptococcus agalactiae* pan-genome for prevalence, diversity and functionality of integrative and conjugative or mobilizable elements integrated in the tRNA gene. *Mol. Genet. Genomics* 290, 1727–1740. doi: 10.1007/s00438-015-1031-9

Rabinovich, L., Sigal, N., Borovok, I., Nir-Paz, R., and Herskovits, A. A. (2012). Prophage excision activates *Listeria* competence genes that promote phagosomal escape and virulence. *Cell* 150, 792–802. doi: 10.1016/j.cell.2012.06.036

Roberts, A. P., and Mullany, P. (2011). Tn916-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. *FEMS Microbiol. Rev.* 35, 856–871. doi: 10.1111/j.1574-6976.2011.00283.x

Rocha, E. P., and Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet.* 18, 291–294. doi: 10.1016/S0168-9525(02)02690-2

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., et al. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* 16, 944–945. doi: 10.1093/bioinformatics/16.10.944

Santoro, F., Oggioni, M. R., Pozzi, G., and Iannelli, F. (2010). Nucleotide sequence and functional analysis of the tet(M)-carrying conjugative transposon Tn5251 of *Streptococcus pneumoniae*. *FEMS Microbiol. Lett.* 308, 150–158. doi: 10.1111/j.1574-6968.2010.02002.x

Schubert, S., Dufke, S., Sorsa, J., and Heesemann, J. (2004). A novel integrative and conjugative element (ICE) of *Escherichia coli*: the putative progenitor of the *Yersinia* high-pathogenicity island. *Mol. Microbiol.* 51, 837–848. doi: 10.1046/j.1365-2958.2003.03870.x

Scott, J. R., Bringel, F., Marra, D., Van Alstine, G., and Rudy, C. K. (1994). Conjugative transposition of Tn916: preferred targets and evidence for conjugative transfer of a single strand and for a double-stranded circular intermediate. *Mol. Microbiol.* 11, 1099–1108. doi: 10.1111/j.1365-2958.1994.tb00386.x

Sebaihia, M., Wren, B. W., Mullany, P., Fairweather, N. F., Minton, N., Stabler, R., et al. (2006). The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat. Genet.* 38, 779–786. doi: 10.1038/ng1830

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi: 10.1038/msb.2011.75

Sitkiewicz, I., Green, N. M., Guo, N., Mereghetti, L., and Musser, J. M. (2011). Lateral gene transfer of streptococcal ICE element RD2 (region of difference 2) encoding secreted proteins. *BMC Microbiol.* 11:65. doi: 10.1186/1471-2180-11-65

Smillie, C., Garcillan-Barcia, M. P., Francia, M. V., Rocha, E. P., and De La Cruz, F. (2010). Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* 74, 434–452. doi: 10.1128/MMBR.00020-10

Stragier, P., Kunkel, B., Kroos, L., and Losick, R. (1989). Chromosomal rearrangement generating a composite gene for a developmental transcription factor. *Science* 243, 507–512. doi: 10.1126/science.2536191

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Toussaint, A., and Merlin, C. (2002). Mobile elements as a combination of functional modules. *Plasmid* 47, 26–35. doi: 10.1006/plas.2001.1552

van der Meer, J. R., and Sentchilo, V. (2003). Genomic islands and the evolution of catabolic pathways in bacteria. *Curr. Opin. Biotechnol.* 14, 248–254. doi: 10.1016/S0958-1669(03)00058-2

van Eijk, E., Anvar, S. Y., Browne, H. P., Leung, W. Y., Frank, J., Schmitz, A. M., et al. (2015). Complete genome sequence of the *Clostridium difficile* laboratory strain 630Deltaerm reveals differences from strain 630, including translocation of the mobile element CTn5. *BMC Genomics* 16:31. doi: 10.1186/s12864-015-1252-7

Williams, K. P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.* 30, 866–875. doi: 10.1093/nar/30.4.866

Wozniak, R. A., and Waldor, M. K. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat. Rev. Microbiol.* 8, 552–563. doi: 10.1038/nrmicro2382

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

3. Etude de la prévalence et de la diversité des IME au sein des génomes de streptocoques.

Les résultats présentés de façon succincte dans cette partie ont fait l'objet de l'article N°2 (inclus en fin de partie) publié en 2017 dans Frontiers In Microbiology sous le titre :

A Glimpse into the World of Integrative and Mobilizable Elements in Streptococci Reveals an Unexpected Diversity and Novel Families of Mobilization Proteins.

Bien qu'au cours des 15 dernières années les connaissances relatives au transfert, à la prévalence et à la diversité des ICE aient augmenté de façon tout à fait notable, ces mêmes données concernant les IME sont restées des plus incomplètes, voire inexistantes. En effet, à ce jour seulement 3 recherches d'IME dans des génomes bactériens ont été réalisées. D'abord en 2008, Brochet *et al.* ont trouvé 6 IME dans 8 génomes de *S. agalactiae*, tous codant des intégrases à tyrosine et des relaxases de la famille MOB_T. Puis en 2011, une étude portant sur 5 souches de *Clostridioides difficile* par Brouwer *et al.* a révélé la présence de 4 IME codant chacun une intégrase à sérine ou des doublets d'intégrases à sérine ainsi que des relaxases appartenant aux familles MOB_V ou MOB_Q. Enfin, toujours en 2011, une étude réalisée par Guglielmini *et al.* recherchant les relaxases et les modules de conjugaison dans 1124 chromosomes de bactéries et d'archées, a révélé un grand nombre de relaxases non associées à des T4SS pouvant appartenir à des IME. L'étude par Guglielmini *et al.* suggère également que les IME seraient les plus répandus des éléments se transférant par conjugaison (les CIME, éléments dégénérés mobilisables *en cis* par les ICE ou les IME n'étant cependant pas pris en compte car ils ne pouvaient pas être détectés lors de cette analyse). Bien que peu d'études sur les IME aient été réalisées, toutes suggèrent que ces éléments sont très répandus dans les génomes bactériens, ce qui souligne le manque de connaissance que nous avons sur la prévalence et la diversité de ces derniers.

Dans le but de combler le manque d'information concernant les IME, nous avons décidé d'utiliser la méthodologie ICEFinder, pour rechercher systématiquement les IME dans 124 génomes de streptocoques (ceux-là même au sein desquels les ICE ont été recherchés). Cette recherche a conduit à l'identification de 144 IME, soit à un nombre supérieur à celui des ICE, faisant des IME les éléments se transférant par conjugaison les plus répandus dans les génomes de streptocoques. Ils sont de plus présents dans plus de la moitié des génomes

étudiés (77/124). Les IME identifiés ont une taille moyenne de 10 kb (variant de 5 kb à 18 kb à l'exception d'un IME de 53 kb dans lequel est intégré un tandem ICE-IME en accrétion) contre 41 kb pour les ICE démontrant que les IME de streptocoques sont généralement plus petits que les ICE.

La majorité des IME identifiés code une intégrase à tyrosine (128/144) tandis que le reste (16/144) code une intégrase à sérine. Aucune d'entre eux ne code de triplet d'intégrase à sérine ou de transposase à DDE. La délimitation de ces éléments a conduit à l'identification de leur site d'intégration et à la détermination de la spécificité d'intégration de leur intégrase. La majeure partie des éléments codant une intégrase à tyrosine (109/128) sont intégrés dans les extrémités 3' de gènes d'ARN de transfert ($tRNA_{leu}$, $tRNA_{lys}$, $tRNA_{arg}$, $tRNA_{asn}$) ou de protéines ribosomiques (*rplL*, *rpsI*, *rpmE*, *rpmG*) tandis que le reste des éléments est intégré dans l'extrémité 5' des gènes codant les protéines TatD, EbfC et dans l'extrémité 3' du gène codant la GMP synthase. Parmi les 16 IME codant une intégrase à sérine, 2 interrompent le gène gène *rumA* et 12 interrompent des gènes portés par des ICE de la famille Tn*5252*. Les gènes d'ICE interrompus sont le gène *traG* (codant pour la protéine de couplage), le gène *maff2* (codant pour une protéine membranaire) et le gène *snf2* (codant pour une hélicase). Enfin le site d'intégration des 2 derniers éléments codant une intégrase à sérine n'a pas pu être déterminé. Au total, cette étude permet de mettre en évidence 18 spécificités d'intégration dans 15 sites d'intégration différents (certains sites étant localisés à des positions différentes au sein du même gène). Parmi ces sites, seulement 5 étaient déjà connus chez les firmicutes (3'*rpmG*, 3' *rpsI*, 3' tRNAlys, 3' *rumA*, dans *snf2*). Cette diversité de sites d'intégration est équivalente à celle retrouvée pour les ICE pour lesquels 17 spécificités ont été identifiées. Sept des spécificités d'intégration sont communes aux ICE et IME identifiés.

Parmi les 128 IME codant une intégrase à tyrosine, seulement 10 codent des relaxases dites canoniques, appartenant toutes à la famille MOB_V. De plus, tous ces IME codent également une autre relaxase/protéine initiatrice de réplication appartenant à la famille MOB_T/Rep_Trans. Les 118 autres IME possédant une intégrase à tyrosine ne codent pas de relaxases appartenant à l'une des 6 familles canoniques (MOB_V, MOB_P, MOB_F, MOB_C, MOB_Q ou MOB_H) mais codent cependant des protéines apparentées aux initiateurs de réplication par cercle roulant de plasmides ou de virus identifiables par la présence des domaines

PF02486 (45/118) (correspondant à la famille MOB_T), PF01719 (35/118), PHA00330 (21/118), PF01719-PF00910 (15/118) ou PF02407 (2/18). De façon inattendue, parmi ces 118 IME, 85 codent également une protéine de couplage, ce qui n'avait encore jamais décrit chez des IME. Ces protéines de couplage appartiennent à la famille TcpA. Les protéines TcpA sont, chez les ICE, toujours associées à des relaxases non canoniques de la famille MOB_T. Du fait de leur association à des intégrases et à des protéines de couplage, nous avons émis l'hypothèse que les protéines d'IME apparentées à des initiateurs de réplication par cercle roulant pourraient assurer le rôle de relaxase au sein de ces éléments.

Les 16 éléments codant une intégrase à sérine, codent aussi des relaxases reconnues par CONJ-Scan comme appartenant aux familles MOB_Q (12), MOB_C (2), MOB_V (1) et MOB_P (1). Seuls des IME codant des relaxase de la famille MOB_V et MOB_Q avaient jusqu'alors été identifiés au sein des firmicutes. De plus, excepté les IME codant à la fois une relaxase MOB_C et une protéine de couplage appartenant à la famille VirD4, aucun autre ne code de protéine de couplage.

Afin de mieux caractériser les 144 IME, nous avons tenté de les classer en famille comme nous l'avions fait pour les ICE. Pour cela, les relaxases, protéines de couplages et intégrases ont été respectivement classées en groupes : les protéines dont les séquences présentent plus de 40% d'identité sur 40% de leurs longueurs appartenant à un même groupe. Cependant, l'étude a montré qu'une classification se basant sur l'association des gènes de leur module de mobilisation (relaxase et protéine de couplage) n'est pas pertinente pour les IME. D'une part, seulement la moitié des IME code une protéine de couplage et d'autre part aucune concordance dans l'association des 2 gènes du module de mobilisation n'a été constatée. De plus, une classification basée sur les 3 protéines étudiées (intégrases, relaxase, protéine de couplage) serait encore plus problématique. En effet, au total, 39 associations ont en effet été retrouvées entre les gènes codant les 3 types de protéines signatures, ne nous permettant pas de proposer des familles d'IME basées sur ces critères et soulignant la très grande diversité retrouvée au sein de ces éléments (Figure 3 de l'article N°2). Cependant, malgré leur diversité, tous les éléments codant une relaxase apparentée aux initiateurs de réplication et éventuellement une CP présentent une structure compacte. Cette structure compacte est caractérisée par la succession des gènes codant l'intégrase, l'excisionase, la relaxase et éventuellement la protéine de couplage (dans cet ordre).

Malgré l'impossibilité de regrouper les éléments en un petit nombre de familles à partir des gènes du module de mobilisations, les éléments ont été regroupés en fonction de leur relaxase afin de les comparer et de retrouver d'éventuels gènes conservés qui pourraient être nécessaires au maintien ou au transfert de l'élément. Certaines des protéines conservées au sein des IME codant des relaxases MOB_V, MOB_P et MOB_Q possédaient des domaines suggérant qu'elles posséderaient des fonctions impliquées dans le maintien sous forme excisée de ces éléments. Ainsi, nous avons pu identifier des protéines de type « replisome organizer » ou DnaC (protéines impliquées dans l'initiation de la réplication thêta de phage), ou encore des protéines apparentées à ParB, une protéine impliquée dans la partition des chromosomes et des plasmides.

Comme pour les ICE, les comparaisons de séquence des protéines codées par les IME et les analyses phylogénétiques révèlent de nombreux échanges de modules d'intégration entre IME (notamment entre ceux présentant une structure compacte), suggérant que ces éléments évoluent par échanges de modules. Par ailleurs, contrairement aux analyses réalisées sur les modules de conjugaison des ICE, elles révèlent de nombreux échanges de gènes de protéines de couplages entre modules de mobilisation (voir par exemple Figure 5 de l'article N°2). Enfin, la comparaison des IME possédant des relaxases non canoniques révèle également de multiples événements de perte ou acquisition de gènes codant des protéines de couplage de type TcpA. Il est probable que les nombreux échanges, pertes et acquisitions sont responsables de la grande diversité observée des associations intégrase-relaxase-CP.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 20 March 2017
doi: 10.3389/fmicb.2017.00443

A Glimpse into the World of Integrative and Mobilizable Elements in Streptococci Reveals an Unexpected Diversity and Novel Families of Mobilization Proteins

*Charles Coluzzi[1], Gérard Guédon[1], Marie-Dominique Devignes[2], Chloé Ambroset[1], Valentin Loux[3], Thomas Lacroix[3], Sophie Payot[1] and Nathalie Leblond-Bourget[1]**

[1] *UMR1128 DynAMic, Institut National de la Recherche Agronomique, Université de Lorraine, Vandœuvre-lès-Nancy, France,* [2] *UMR7503 Laboratoire Lorrain de Recherche en Informatique et ses Applications, Centre National de la Recherche Scientifique, Université de Lorraine, Vandœuvre-lès-Nancy, France,* [3] *UR1404 Mathématiques et Informatique Appliquées du Génome à l'Environnement, Institut National de la Recherche Agronomique, Université Paris-Saclay, Jouy-en-Josas, France*

OPEN ACCESS

Edited by:
Rakesh Sharma, Institute of Genomics and Integrative Biology (CSIR), India

Reviewed by:
Joshua Peter Ramsay, Curtin University, Australia
Nikolai Ravin, Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences, Russia

***Correspondence:**
Nathalie Leblond-Bourget
nathalie.leblond@univ-lorraine.fr

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *18 January 2017*
Accepted: *03 March 2017*
Published: *20 March 2017*

Citation:
Coluzzi C, Guédon G, Devignes M-D, Ambroset C, Loux V, Lacroix T, Payot S and Leblond-Bourget N (2017) A Glimpse into the World of Integrative and Mobilizable Elements in Streptococci Reveals an Unexpected Diversity and Novel Families of Mobilization Proteins. Front. Microbiol. 8:443. doi: 10.3389/fmicb.2017.00443

Recent analyses of bacterial genomes have shown that integrated elements that transfer by conjugation play an essential role in horizontal gene transfer. Among these elements, the integrative and mobilizable elements (IMEs) are known to encode their own excision and integration machinery, and to carry all the sequences or genes necessary to hijack the mating pore of a conjugative element for their own transfer. However, knowledge of their prevalence and diversity is still severely lacking. In this work, an extensive analysis of 124 genomes from 27 species of *Streptococcus* reveals 144 IMEs. These IMEs encode either tyrosine or serine integrases. The identification of IME boundaries shows that 141 are specifically integrated in 17 target sites. The IME-encoded relaxases belong to nine superfamilies, among which four are previously unknown in any mobilizable or conjugative element. A total of 118 IMEs are found to encode a non-canonical relaxase related to rolling circle replication initiators (belonging to the four novel families or to MobT). Surprisingly, among these, 83 encode a TcpA protein (i.e., a non-canonical coupling protein (CP) that is more closely related to FtsK than VirD4) that was not previously known to be encoded by mobilizable elements. Phylogenetic analyses reveal not only many integration/excision module replacements but also losses, acquisitions or replacements of TcpA genes between IMEs. This glimpse into the still poorly known world of IMEs reveals that mobilizable elements have a very high prevalence. Their diversity is even greater than expected, with most encoding a CP and/or a non-canonical relaxase.

Keywords: mobilizable elements, relaxase, TcpA coupling protein, conjugation, *Streptococcus*

INTRODUCTION

Conjugative elements drive horizontal gene transfer between bacteria, and therefore play a key role in bacterial evolution. These mobile elements encode all factors needed for their autonomous transfer by conjugation. The conjugative transfer of various plasmids from Gram-negative (G-) bacteria, especially proteobacteria, is well understood (Cabezon et al., 2015;

Chandran Darbari and Waksman, 2015; Ilangovan et al., 2015) and proceeds as follows. The plasmid DNA is recognized and processed by the relaxosome, a complex that includes a relaxase protein encoded by the element. Up to now, six superfamilies of relaxases (MobC, MobF, MobH, MobP, MobQ, and MobV) are known to be encoded by conjugative plasmids from proteobacteria (Garcillan-Barcia et al., 2009) and are referred in this paper as canonical relaxases. The relaxase catalyzes a site- and strand-specific cleavage of the origin of transfer (*oriT*) at the *nic* site of its cognate plasmid. The relaxase-tethered DNA is then recruited to the coupling protein (CP) belonging to the VirD4 family. The CP interacts with a multi-protein complex known as a type IV secretion system (T4SS) which spans the cellular envelope of the donor cell. The CP and T4SS subsequently translocate the single-strand DNA-relaxase complex through membranes and cell walls into the recipient cell. The nicking of *oriT* by the relaxase also initiates a rolling-circle replication (RCR) of the plasmid by cellular enzymes so that the donor cell retains the plasmid and the recipient cell acquires the plasmid. In addition to conjugative plasmids, other autonomous elements called integrative and conjugative elements (or ICEs) are found to be integrated in the chromosomes of bacteria. ICEs encode their own excision, transfer by conjugation, and integration (for reviews see Burrus et al., 2002; Bellanger et al., 2014). Apart from the excision and integration steps that are catalyzed by a tyrosine recombinase, a serine recombinase or a DDE transposase, the conjugative transfer of most ICEs is assumed to resemble that of plasmids of G- bacteria, and therefore to involve a relaxase, a CP and a T4SS machinery.

Many other mobile elements, known as mobilizable elements, hijack the conjugative machinery of unrelated conjugative elements (Francia et al., 2004; Garcillan-Barcia et al., 2009; Meyer, 2009). The best known mobilizable elements are the mobilizable plasmids from G- bacteria. While a very few mobilizable plasmids encode a CP, they never encode any other protein belonging to the T4SS. Almost all of them encode a relaxase from one of the six canonical superfamilies, but which is distantly related to those of the conjugative plasmids. These relaxases recognize and cut their cognate *oriT* (Francia et al., 2004; Meyer, 2009). They then recruit the CP and/or T4SS of a conjugative element to mobilize *in trans* the non-autonomous element. In addition to mobilizable plasmids, some integrative elements known as integrative and mobilizable elements (IMEs) also transfer by mobilization. IMEs encode their own excision and integration but carry only some of the sequences or genes necessary for their conjugative transfer (for a review, see Bellanger et al., 2014). Most of the very few IMEs described so far carry their own *oriT* and encode their own relaxase, but none encode a CP or any protein belonging to the T4SS. Other previously described IMEs carry their own *oriT* but do not encode any protein involved in conjugation. To date, very few genomes searches have focused explicitly on IMEs so their prevalence and diversity are essentially unknown (Brochet et al., 2008; Bellanger et al., 2014). However, Guglielmini et al. (2011) performed an extensive search for relaxases and conjugation modules in 1124 archaeal and bacterial genomes, and identified many isolated relaxase genes on chromosomes, which suggests that IMEs are the most prevalent elements that transfer by conjugation.

While conjugation and mobilization mechanisms are well known in G- proteobacteria, they are poorly documented in all other bacterial clades, including the Firmicutes, a major group of Gram-positive (G+) bacteria. The conjugative plasmids and ICEs from firmicutes encode T4SSs belonging to two families, FA and FATA, that have not been found in G- bacteria (Guglielmini et al., 2014). The FATA T4SS was found to be associated with classical CPs (VirB4) and with relaxases belonging to the canonical MobP, MobQ, and MobC superfamilies (Guglielmini et al., 2013; Ambroset et al., 2016). In contrast, most plasmids and ICEs with FA T4SSs encode TcpA CPs instead of VirD4. The TcpA CPs are related to FtsK, the double strand DNA translocase involved in DNA segregation during cell division (Guglielmini et al., 2013; Ambroset et al., 2016). Furthermore, all the plasmids and ICEs that encode FA T4SSs and TcpAs CPs encode non-canonical relaxases. Thus, the pCW3 conjugative plasmid from the Firmicute *Clostridium perfringens*, which encodes a FA T4SS and a TcpA CP, was recently shown to encode a novel type of relaxase related to tyrosine recombinase (Wisniewski et al., 2016). In the same way, all ICEs, which encode a FA T4SS and a TcpA CP, encode a relaxase belonging to the non-canonical MobT superfamily. The MobT relaxases are related to a family of RCR initiators involved in the intracellular replication and maintenance of small plasmids (Guglielmini et al., 2013; Ambroset et al., 2016). MobT encoded by ICE*Bs1* is involved in both conjugative transfer and in replication of the excised ICE [for a review, see (Auchtung et al., 2016)]. It should be mentioned that most families of relaxases involved in the RCR of small plasmids or viruses are clearly distinct from and perhaps unrelated to the superfamilies of relaxases found in conjugative and mobilizable elements. Besides the MobT relaxases encoded by the ICEs with FA T4SS and TcpA CP, the only other exception corresponds to PF01446 RCR initiators of some mobilizable plasmids of firmicutes that are involved in both plasmid replication and mobilization by ICEs encoding TcpA CPs (Naglich and Andrews, 1988; Showsh and Andrews, 1999; Lee et al., 2012).

Streptococci are G+ bacteria belonging to the Firmicutes. Genome and phylogenetic analyses have shown that a large proportion of the streptococcal genomes has experienced horizontal gene transfer (Richards et al., 2014). Previously, our extensive analyses of the genomes of 124 strains belonging to 27 streptococcal species revealed a high prevalence of ICEs (Ambroset et al., 2016), suggesting that a significant fraction of these transfers could be due to those elements. Furthermore, a comprehensive search of IMEs performed on eight available genomes of the firmicute *Streptococcus agalactiae* revealed twelve IMEs (Brochet et al., 2008). Surprisingly, this search also detected nine genomic islands which possess a complete integration/excision module and encode a putative CP belonging to TcpA family, but which do not encode any proteins related to known relaxases. Therefore, these elements could correspond to IMEs encoding a relaxase that is very distantly related or unrelated to known relaxases. A preliminary reanalysis of these nine elements from *S. agalactiae* revealed that they encode

proteins related to RCR initiators from plasmids or viruses, suggesting that these elements are probably IMEs encoding novel types of relaxases related to RCR initiators. In this study, we searched for IMEs in the 124 publicly available complete genomes of *Streptococcus* that were previously used for the ICE search. IMEs were defined by the combined presence of putative integrases and relaxases (related to classical relaxases or to RCR initiators), the eventual presence of putative CPs (VirD4 or TcpA), and the absence of T4SSs. CDSs encoding these signature proteins were localized on the chromosomes and their boundaries and integration site of IMEs were identified. This study (i) gives a general overview of the very high prevalence and diversity of putative IMEs within streptococci, (ii) identifies their numerous specific sites of insertion, (iii) reveals that most IMEs harbor a versatile and compact mobilization module that encodes a non-canonical relaxase related to RCR initiators and generally a non-canonical CP related to FtsK.

MATERIALS AND METHODS

Genomes Examined and Database of Reference Proteins

The dataset of the 124 complete chromosomes from *Streptococcus* species available at the start of the present study was taken from Genbank[1]. This initial database of reference proteins contains signature proteins from ICEs and the few IMEs reported for Firmicutes in the literature. It includes protein sequences from 50 tyrosine integrases, 13 serine integrases, 2 DDE transposases, 50 relaxases, 37 CPs, and 26 VirB4 proteins (a T4SS ATPase). This last protein has never been found in any mobilizable elements and is used here as the main criterion for differentiating between ICEs and IMEs at the detection step.

Workflow

The overall workflow of our search strategy to detect and characterize IMEs in streptococcal chromosomes was described previously (Ambroset et al., 2016). This workflow allows (i) detecting ICEs from the presence of signature CDSs grouped on the genome, (ii) identifying ICE insertion sites and (iii) delineating ICEs. The workflow was adapted to IME detection by modifying the signature CDSs in step (i): a putative IME is detected when no VirB4 CDS is present and when an integrase CDS is found in the vicinity of a relaxase CDS. Steps (ii) and (iii) were conducted in the same way as for ICEs.

When an IME signature CDS was missing or incomplete (pseudogene), the corresponding complete CDS encoded by the closest known IME was taken and compared to the putative defective one by tBlastN in order to detect possible genome annotation errors (e.g., identification of an authentic gene as a pseudogene most frequently due to the presence of a type II intron within the gene or mis-identification of the "start" codon). Because before this work, the known IMEs did not encode a CP, those elements carrying an integrase gene, a relaxase gene, and a CP pseudogene were considered as IMEs. Moreover, our workflow detects elements containing only an integrase and a CP CDS but apparently no relaxase gene or pseudogene. In such cases, an exhaustive manual analysis was performed to search for new relaxase genes, in particular for genes encoding proteins related to RCR initiators. Newly found relaxases were added to the database of reference proteins and step (i) was reiterated on all genomes. In summary, we considered as IMEs all elements delimited by direct repeats and containing at least one CDS for one complete integrase as well as one complete relaxase, but no CDS for VirB4 or other proteins of the T4SS.

Denomination of IMEs

Each IME name indicates by letters and numbers the species and strain of the host bacteria. When an IME encodes a site-specific integrase, its denomination also specifies the name of the target gene. IMEs marked with an asterisk are not integrated in their primary site but in a secondary one as previously observed for ICEs (Ambroset et al., 2016).

Domain Composition Analysis and Tree Construction

The retrieval of the domain composition of all IME signature proteins from Uniprot annotations was done in batch using the BioMart Central Portal[2]. *De novo* conserved domain search (CD-search[3]) and/or PSI-Blast analyses were performed when no data was available through BioMart. The correspondence with Mob families was established using the CONJscan-T4SSscan program[4] (Guglielmini et al., 2011). This tool is no longer accessible but should soon be available on the Pasteur Galaxy server.

The signature proteins were aligned using Clustal omega with default parameters (Sievers et al., 2011). The trees of signature proteins were built with MEGA (Tamura et al., 2013) using (i) maximum likelihood (ML) based on the JTT (Jones–Taylor–Thornton) model including amino acid empirical frequencies (partial deletion of gaps and missing data at 80% cutoff, Gamma distribution in five categories, allowance for invariant sites), and (ii) BioNJ methods with the Poisson model (Gouy et al., 2010). The branch support of the groupings was estimated using bootstrap (100 replicates).

Protein Clustering and Signature Proteins Associations

Protein clustering at 90 and 40% sequence identity was performed using BLASTclust[5] (Alva et al., 2016) with default parameters. Circos[6] was used to visualize the signature protein associations (Krzywinski et al., 2009). The functional annotation of IMEs was performed using Agmial as described previously (Ambroset et al., 2016).

[1]http://www.ncbi.nlm.nih.gov/genome/browse/, last accessed December 2013

[2]http://central.biomart.org/

[3]http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

[4]http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::CONJscan-T4SSscan

[5]https://toolkit.tuebingen.mpg.de/blastclust

[6]http://circos.ca/

RESULTS

Prevalence of IMEs within Streptococcal Chromosomes

The prevalence of IMEs was studied in a large set of species (27 *Streptococcus* species) and strains ($n = 124$). This exhaustive examination led to the identification of 144 IMEs. Their sizes ranged from 5 to 18 kb (Supplementary Table S1) except for IME_*Sparas15912_rpsI* which was 53 kb. The larger size of IME_*Sparas15912_rpsI* has likely resulted from a tandem ICE-IME insertion in the element. Within streptococcal genomes, this study showed that IME sizes (mean = 10 kb) were generally smaller than those of ICEs (mean = 41 kb). This difference is reminiscent of that observed between mobilizable and conjugative plasmids (Smillie et al., 2010).

Ten IMEs were found to be integrated in tandem accretion with other IMEs (IME1-IME2) or with ICEs (ICE-IME or ICE-IME1-IME2) (Supplementary Table S1) and many with decayed elements (data not shown). Nine tandems were found to be integrated in the 3′ end of *rpsI*, *rpmG,* or *rplL*. The last one was integrated in a secondary site inside IME_*Sparas15912_rpsI*. All IMEs in accretion in the same site had a tyrosine integrase but some decayed IMEs in accretion were found to encode serine integrases (data not shown). These integrases had from 21 to 69% sequence identity, showing that there was no particular relatedness between integrases encoded by elements in accretion.

More than half ($n = 78$) of the streptococcal chromosomes contained at least one IME. As seen in Supplementary Table S1, the occurrence of IMEs varied within species and strains. For instance, most of the 20 *S. pyogenes* chromosomes were deprived of IMEs, whereas the 3 *S. anginosus* chromosomes each contained from 4 to 6 IMEs. Interestingly, the genomes that contained the lowest numbers of IMEs (*S. pyogenes* with an average of 0.1 IME per genome) also contain few ICEs (0.1 ICE per genome). The opposite was also true; the genomes from *S. anginosus* were among those containing the highest numbers of IMEs (mean = 4.7) as well as ICEs (mean = 4.3).

Diversity of Integration Modules and Integration Sites of Streptococcal IMEs

Almost all of the integrase genes of streptococcal IMEs were located at one end of the element and were outward facing. These integrases belonged to two unrelated superfamilies: tyrosine integrases and serine integrases.

Diversity of IME Tyrosine Integrases and of Their Integration Sites

Tyrosine integrases were detected in more than 89% of the IMEs ($n = 128/144$). Both phylogenetic analysis and clustering of the tyrosine integrases at 40% sequence identity allowed them to be classified in 10 distinct families (**Figure 1**). In most cases, tyrosine integrases belonging to the same family catalyze integration in the same gene (**Figures 1**, **2**). For instance, all tyrosine integrases allocated to family Tyr_1 target the 3′ end of the tRNAleu gene. Three exceptions to this rule were observed. First, the Tyr_3 family includes tyrosine integrases targeting two distinct insertion genes (*guaA* or *rplL*). Phylogenetic analysis of these integrases suggests an evolution from *guaA* specificity to *rplL* specificity (**Figure 1**). Second, the Tyr_4 family gathers all integrases targeting *rpmE* and one integrase targeting *rpsI*. Phylogenetic analysis of these integrases and the sister Tyr_5 group of integrases suggests an evolution from *rpsI* specificity to *rpmE* specificity. Third, integrases of family Tyr_8 catalyze integration within genes encoding four different tRNAs (tRNAasn, tRNAleu, tRNAlys, and tRNAarg). Although catalyzing integration in distinct tRNA genes, all integrases from family Tyr_8 share the ability to generate short direct repeats (DRs).

In the majority of cases, each insertion gene is specifically targeted by closely related tyrosine integrases grouped in one unique family (**Figures 1**, **2**). Exceptions were *rpsI, rpmG,* and the tRNAleu genes that are targeted by integrases belonging to several families. More specifically, *rpsI* is targeted by tyrosine integrases belonging to three distinct families; the integrases from families Tyr_4 and Tyr_5 lead to the formation of long DRs whereas those from family Tyr_7 all have short DRs (**Figure 2**). Similarly, the integrases targeting the tRNAleu gene belong to two distinct families; the integrases from family Tyr_1 have long DRs, and those from family Tyr_8 have short DRs. Finally, two families of integrases catalyze integration in *rpmG* and lead to two distinct architectures after integration: genes encoding integrases from family Tyr_10 are adjacent to *rpmG,* whereas those encoding integrases from family Tyr_6 are at the opposite end of the IME.

In summary, the analysis of the integration loci of 128 IMEs harboring a tyrosine integrase shows that these integrases specifically target 11 distinct genes, mainly (109/128) at the 3′ end of genes encoding tRNAs or ribosomal proteins. More rarely (19/128), tyrosine integrases from IMEs catalyze integration at the 3′ or 5′ end of other protein-encoding genes (*guaA*, *tatD*, or *ebfC*). This study has therefore extended the known list of possible integration sites for streptococcal IMEs with tyrosine integrases: only four integration sites were previously identified (*oriT* from ICEs belonging to Tn*916* and ICE*St3* families, 3′ *rpsI*, 3′ tRNALys gene and 3′ *rpmG*) (Brochet et al., 2008; Puymege et al., 2015; Lorenzo-Diaz et al., 2016).

Diversity of IME Serine Integrases and of Their Integration Sites

Our collection of IMEs contains 16 IMEs encoding a serine integrase. For two IMEs (IME_*Sol3089_ND* and IME_*SsalCCHSS3_ND*) encoding related integrases belonging to family Ser_2, we were unable to determine the specificity of integration because these two IMEs were integrated in two distinct intergenic regions (**Figure 2**). Two other IMEs with serine integrases were found to be integrated within the bacterial gene *rumA,* which is already known to be a common integration site for ICEs with serine integrases (Ambroset et al., 2016). All other 12 IMEs were found to be integrated within conserved genes from ICEs belonging to the Tn*5252* superfamily: *traG* (encoding a VirD4 CP), *maff2* (encoding a membrane protein), and *snf2* (encoding a helicase) (Supplementary Table S1), thereby leading to gene disruption.

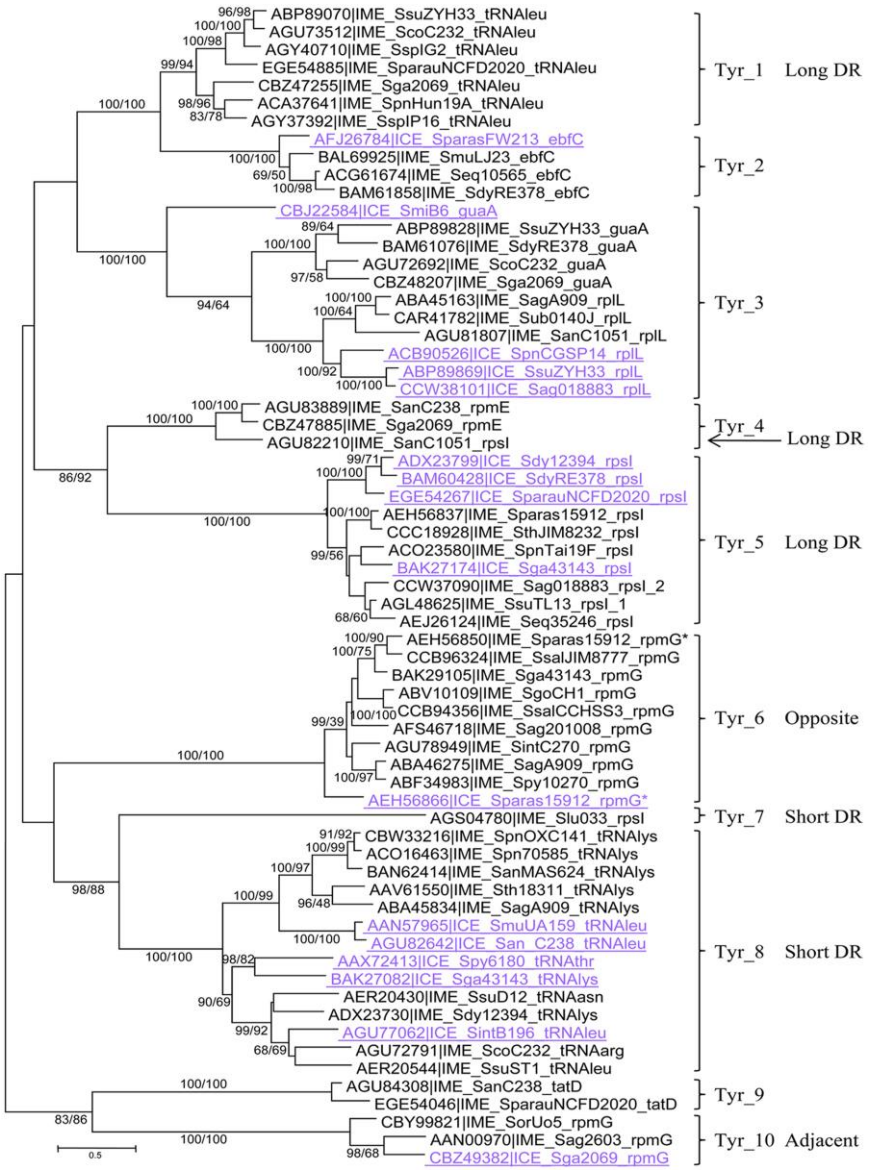

FIGURE 1 | Phylogenetic tree of tyrosine integrases. One representative of each 90% protein identity cluster from integrative and mobilizable elements (IMEs) (in black) and one representative of each 90% protein identity cluster of tyrosine integrases from ICEs targeting the same site as IMEs (in mauve and underlined) are presented in the ML tree. Bootstrap values (BioNJ/ML) are given only when they exceed 50 for both analyses. The target gene is mentioned in the IME/ICE names. Tyrosine integrases sharing more than 40% sequence identity and therefore belonging to the same family are merged with brackets. These families are distinguished with different numbers. The DR length or integrase position is indicated to distinguish tyrosine integrases belonging to different families but targeting the same genes. Refer to Supplementary Table S1 for IME and strain details.

Both phylogenetic analysis and 40% sequence identity clustering of the serine integrases indicate that they may be classified in four distinct families (Ser_1 to 4, **Supplementary Figure S1**). A large majority of the serine integrases sharing the same integration specificity were grouped in the same family. For example, the integrases targeting the *traG*, *snf2,* and *rumA* genes were grouped in families Ser_1, Ser_3, and Ser_4, respectively. Exceptions were the two serine integrases targeting *maff2* that were found in two distinct families: the one belonging to family Ser_1 shows a close relatedness with integrases targeting *traG,* whereas the one belonging to family Ser_3 is related to integrases targeting *snf2*. These two integrases catalyze integration in two different locations within *maff2* and lead to the generation of distinct DRs (**Figure 2**). This variability of integration within the

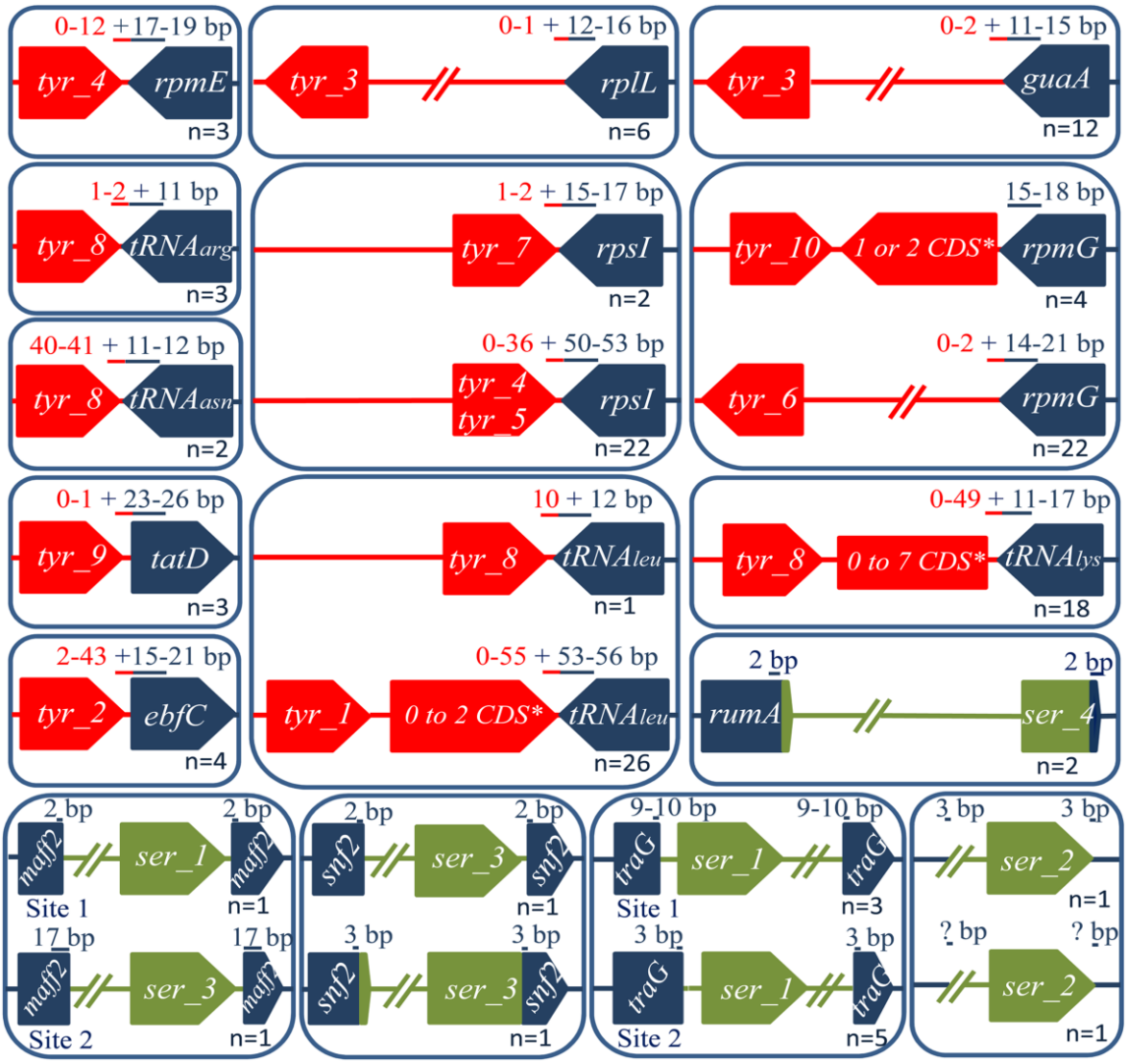

FIGURE 2 | Integrative and mobilizable element integration loci and their position relative to the integrase CDSs. Tyrosine and serine integrase genes are shown in red and green, respectively. The target genes (dark blue) encode, respectively: *ebfC* [nucleoid associated protein], *guaA* [GMP synthase], *maff2* [conserved membrane protein of ICEs belonging to Tn*5252* superfamily], *rpsI* [S9 ribosomal protein], *rplL* [L7/L12 ribosomal protein], *rpmE* [L31 ribosomal protein], *rpmG* [L33 ribosomal protein], *rumA* [23S rRNA (uracil-5-) methyltransferase], *snf2* [helicase of ICEs belonging to Tn*5252* superfamily], *tatD* [DNAse], *traG* [VirD4 CP gene from ICEs belonging to Tn*5252* superfamily]. The DR size (in bp) within the target gene is indicated in blue and the one outside is in red. The number of ICEs integrated in a given site is marked at the bottom of each box.

same gene was also observed for integration inside *traG*. The two IMEs specific to *snf2* are integrated in the same location within *snf2*, but the recombination site of the IME is located within the integrase gene for IME_*Spy2096_SNF2*, whereas it is adjacent to the 3′ end of the integrase gene for IME_*SsuT15_SNF2* (**Figure 2**).

Comparison of the Recombination Modules of IMEs and ICEs

Our collection of IMEs (this work) and ICEs (Ambroset et al., 2016) from streptococci allowed us to compare the recombination modules of these two types of element. Both of them encode serine and tyrosine integrases, and their diversity is similar in IMEs and in ICEs. Indeed, we detected 18 distinct specificities for integrases from streptococcal IMEs (17 targeting a specific site and one with unknown specificity) and at least 17 for the ICEs. However, in contrast to the ICEs, no DDE transposase was found in any of the IMEs.

As shown in **Figure 1** and **Supplementary Figure S1**, several integrase families (such as Tyr_2, Tyr_3, Tyr_5, Tyr_6, Tyr_8, Tyr_10, and Ser_4) grouped both integrases from IMEs and ICEs. For three specificities: *ebfC* (Tyr_2), *rplL* (Tyr_3), and *rpmG*

opposite (Tyr_6), tyrosine integrases from IMEs and ICEs belong to closely related sister groups. In family Tyr_5, the tyrosine integrases from IMEs targeting *rpsI* are mixed with those of the ICEs, and in cluster Ser_4 the serine integrase targeting *rumA* from IME_*SpnAP200_rumA* shares 91–93% identity with those of two ICEs. Altogether, these results suggested that exchange of integration modules between ICEs and IMEs are frequent.

New Relaxase Families Related to RCR Initiator Proteins in IMEs

A total of 154 relaxase genes were detected within IMEs (including three pseudogenes in IMEs harboring two relaxases). Based on their domain composition, the relaxases were classified in nine distinct superfamilies (**Table 1**). Among the four most prevalent relaxase superfamilies, only the Rel_PF02486 superfamily was recognized by the CONJscan-T4SSscan analysis as belonging to a known type of relaxase (MobT). The three others were novel superfamilies of relaxases characterized by the following domains: PF01719 (Rel_PF01719), PHA00330 (Rel_PHA00330) and by the combination of PF001719 and PF00910 (Rel_PF001719-PF00910). The Rel_PF02407 superfamily was the fourth novel superfamily identified in this study. All these new relaxase superfamilies discovered in IMEs harbored domains (PF01719, PHA00330, or PF02407) that were previously found exclusively in RCR initiators from viruses or plasmids (Ebisu et al., 1995; Bachrach et al., 2004; Gibbs et al., 2006; Lorenzo-Diaz et al., 2014). They were assumed to correspond to novel relaxase superfamilies since all these proteins were associated with an integrase and a large majority of these non-canonical relaxases (or all for PF02407) were associated with a CP in our IME collection. Four other relaxase superfamilies described in **Table 1** (Rel_PF03389, Rel_PF01076, Rel_PF13814, and Rel_PF03432) were recognized by the CONJscan-T4SSscan server as the MobQ, MobV, MobC, and MobP superfamilies, respectively.

On the basis of their relaxase content, the IMEs were grouped in nine classes (IME_Class_1 to 9) encoding a unique relaxase, one for each superfamily of relaxase (**Table 2**). An additional class (IME_Class_10) contained 10 IMEs that carry 2 relaxases: one belonging to the Rel_PF02486/MobT and the other to the Rel_PF01076/MobV. Each of the two relaxase superfamilies present in this class also exists as standalone relaxases in IME_Class_1 and IME_Class_6.

Within each superfamily of relaxases, the diversity was estimated by phylogenetic analyses and 40% sequence identity clustering. The most abundant superfamily, Rel_PF02486/MobT (**Supplementary Figure S2**), includes six families (Rel_PF02486_1 to 6), among which the family PF02486_3 was associated with a Rel_PF01076/MobV in IME_Class_10. In contrast, the 35 relaxases from the Rel_PF01719 superfamily are closely related and were therefore clustered in a unique family (**Supplementary Figure S3**). The same was true for the 12 relaxases of the Rel_PF03389/MobQ superfamily (data not shown). The 21 members of the Rel_PHA00330 superfamily were grouped into three families (**Supplementary Figure S4**), and the Rel_PF01719-PF00910 superfamily (15 members) analysis yielded four families (**Supplementary Figure S5**). Finally, the superfamily Rel_PF01076 was clustered in two families, with the Rel_PF01076_1 being always found in IME_Class_10. Overall, the IME relaxases showed a great diversity: they were classified in nine distinct superfamilies subdivided in 20 families (**Table 1**). It should be noted that the diversity of IME relaxases largely exceeds that of ICEs, since ICEs are classified in three superfamilies/eight families according to the same criteria (Ambroset et al., 2016). Only two superfamilies of relaxases corresponding to Rel_PF02486/MobT and Rel_PF03432/MobP are encoded by both IMEs and ICEs. However, our analysis showed that within these superfamilies, the relaxases of IMEs always belong to clearly distinct families from those including ICE relaxases. Moreover, whereas the number of Rel_PF02486/MobT relaxases found for ICEs and IMEs are similar, it can be stressed that Rel_PF03432/MobP relaxases constitute the most abundant superfamily in ICEs (62/105 relaxases) and the least abundant in IMEs (only 1/154 relaxases).

Half of the Streptococcal IMEs Encode a CP

Prior to the present study, none of the previously known or predicted IMEs encode a CP (Bellanger et al., 2014), and the few CPs that are known to be encoded by mobilizable plasmids belong to the VirD4 superfamily (Garcillan-Barcia et al., 2009). Surprisingly, our results show that more than half of the streptococcal IMEs encode a CP (85/144 including 11 pseudogenes of CP) and that almost all these CPs do not belong to the canonical VirD4 family. Indeed, only two IMEs encode a VirD4 CP characterized by a C-terminal VirD4 domain (COG3505 in the NCBI CDD classification). All others (72 proteins excluding pseudogenes) were found to display a unique PF01580 "FtsK-SpoIIIE" catalytic domain and to be more closely related to FtsK (a DNA translocase involved in DNA segregation during cell division) than to the canonical VirD4. According to the CONJscan-T4SSscan analysis, these proteins belong to a particular superfamily of CP named TcpA, found only in Firmicutes (Guglielmini et al., 2013). Reconstruction of their phylogeny and 40% identity clustering allowed their classification in 12 distinct families designated TcpA_1 to TcpA_12 (**Supplementary Figure S6**). The three most abundant families, TcpA_4, TcpA_7, and TcpA_12, contained 11, 11, and 30 proteins, respectively. The nine others were found from only 1 to 6 IMEs.

The two superfamilies (VirD4 and TcpA) of CPs were found in both ICEs (Ambroset et al., 2016) and IMEs with different prevalence. Within *Streptococcus* genomes, VirD4 CPs are encoded by 61/105 ICEs and by only the 2 IMEs belonging to IME_Class_8. In contrast, TcpA CPs are encoded by 44/105 ICEs (belonging to the Tn*916* superfamily) and 83/144 IMEs (belonging to class_1 to 4 and to class_7). Moreover, the diversity of TcpA CPs in IMEs is much larger than in ICEs (12 vs. 3 distinct families) which is reminiscent of relaxases (see above).

TABLE 1 | Relaxase superfamilies based on domain composition.

Superfamily name	Domain(s) ID*	Domain name(s)	Conjscan domain	Number found	Number of clusters
Rel_PF02486/MobT	PF02486	Rep_trans	MobT	55	6
Rel_PF01719	PF01719	Rep_2	No hit	35	1
Rel_PH00330	PHA00330	Not applicable	No hit	21	3
Rel_PF01719-PF00910	PF01719 + PF00910	Rep_2 +RNA_helicase	No hit	15	4
Rel_PF03389/MobQ	PF03389	MobA_MobL	MobQ	12	1
Rel_PF01076/MobV	PF01076	Mob_Pre	MobV	11	1
Rel_PF02407	PF02407	Viral-Rep	No hit	2	1
Rel_PF13814/MobC	PF13814	Replic_Relax	MobC	2	1
Rel_PF03432/MobP	PF03432	Relaxase	MobP	1	1

**All domains are described in the Pfam classification except for domain PHA00330 that was found in the NCBI Conserved Domain classification without any Pfam equivalent.*

TABLE 2 | Diversity of the relaxases and CPs associated with serine and tyrosine integrases.

Integrase type (number)	Relaxase superfamily (number)	CP superfamily (number)	IME class
Tyrosine integrase (128)	Rel_IME_1/MobT(45)	TcpA (24) or none (21)	Class_IME_1
	Rel_IME_2 (35)	TcpA (34*) or none (1)	Class_IME_2
	Rel_IME_3 (21)	TcpA (19*) or none (2)	Class_IME_3
	Rel_IME_4 (15)	TcpA (4) or none (11)	Class_IME_4
	Rel_IME_1/MobT (10*) + Rel_IME_6/MobV(10*)	None	Class_IME_10
	Rel_IME_7 (2)	TcpA (2)	Class_IME_7
Serine integrases (16)	Rel_IME_5/MobQ (12)	None	Class_IME_5
	Rel_IME_8/MobC (2)	VirD4 (2)	Class_IME_8
	Rel_IME_6/MobV (1)	None	Class_IME_6
	Rel_IME_9/MobP (1)	None	Class_IME_9

**Number includes pseudogenes.*

High Diversity of Association of the Different Classes of Signature Genes within IMEs

Our analysis of the co-occurrence of relaxase superfamilies, CPs, and integrases in IMEs reveals some mandatory associations, consistent with the classification defined in section "New Relaxase Families Related to RCR Initiator Proteins in IMEs" (**Table 2**). Tyrosine integrases were found in all IMEs encoding non-canonical relaxases related to RCR initiators, i.e., (i) a Rel_PF02486/MobT relaxase or (ii) relaxases from the four new superfamilies identified in this study. On the other hand, serine integrases were found in all IMEs encoding relaxase from canonical Rel_PF03389/MobQ, Rel_PF13814/MobC, Rel_PF01076/MobV (in the absence of Rel_PF02486/MobT relaxase) and Rel_PF03432/MobP superfamilies. Furthermore, IMEs encoding a single Rel_PF02486/MobT or a relaxase belonging to one of the four new superfamilies also encode either TcpA or no CP. TcpA CPs were never found in IMEs containing one of the canonical relaxases (Rel_PF03389/MobQ, Rel_PF13814/MobC, Rel_PF01076/MobV (alone or with Rel_PF02486/MobT) and Rel_PF03432/MobP). These IMEs contain either no CPs (MobQ, MobV, and MobP) or CPs from the VirD4 superfamily (MobC).

An illustration of all possible ternary Integrase–Relaxase–CP associations in our collection of 144 IMEs is shown in **Figure 3**. For each type of signature protein, the families identified by phylogenetic analyses and clustering at 40% identity are represented by arcs on the circle. The numbers of members in each family are reported in black. The numbers of clusters of sequences sharing more than 90% identity are reported in gray on each arc in order to estimate the sequence diversity in each family of signature protein. Apart from signature protein families with few representatives ($n < 3$) or with low diversity (90% identity clusters <3), this analysis of the co-occurrence of the different families of integrases, CPs, and relaxases (**Figure 3**) reveals no exclusive associations. Altogether, 39 different ternary associations were observed suggesting a high frequency of shuffling between signature proteins.

Organization of Conserved CDSs in IMEs: Predominance of a Common Compact Structure

For a better characterization of the IMEs, a search for conserved CDSs in IMEs encoding the same superfamily of relaxases was undertaken. The various conserved CDS architectures are schematized in **Figure 4**. Apart from IME_Class_10, the IMEs with a tyrosine integrase encode a unique non-canonical relaxase (Rel_PF02486/MobT, Rel_PF01719, Rel_PHA00330, Rel_PF01719-PF00910, and Rel_PF02407). These IMEs display a compact structure composed of successive genes (relaxase, excisionase, and integrase), generally preceded by a TcpA gene (83/118). Interestingly, the same compact structure was

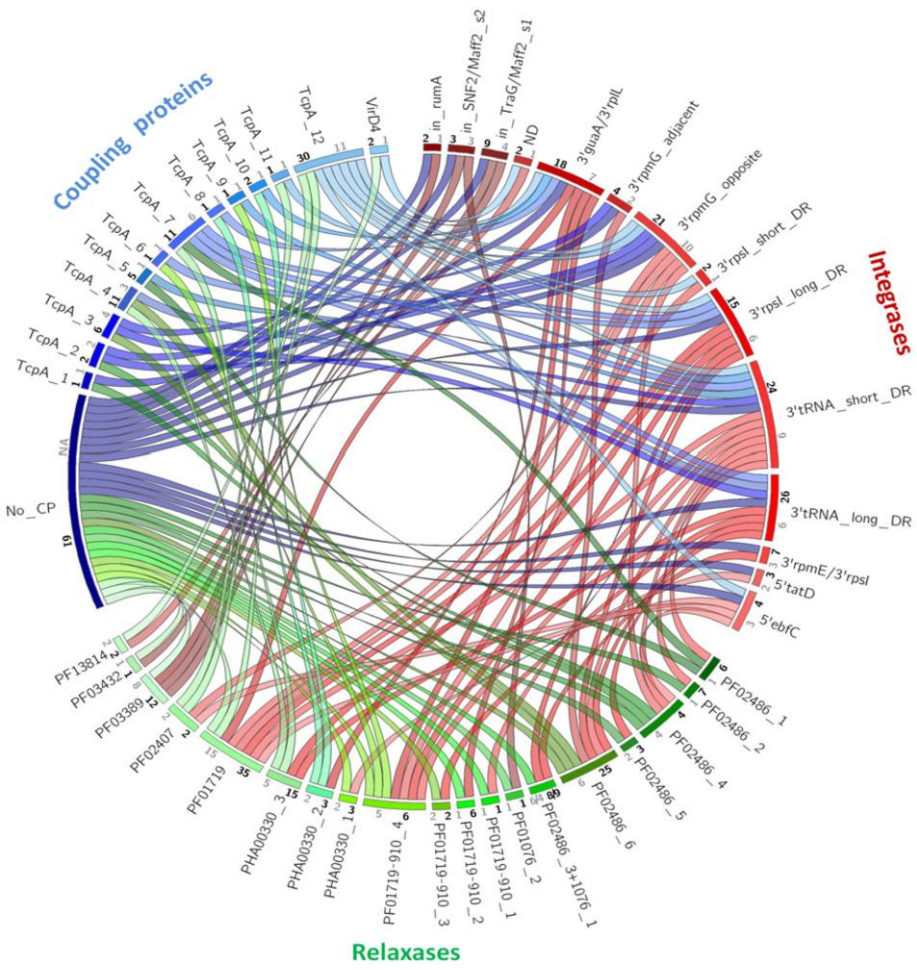

FIGURE 3 | Co-occurrence of relaxases, CPs and integrases among IMEs. Arcs group proteins belonging to the same family according to phylogenetic analysis and 40% sequence identity clustering: red, green, and blue arcs for clustered integrases, relaxases, and CPs, respectively. Ribbons indicate the association between integrases and relaxases in red, relaxases and CPs in green and CPs and integrases in blue. Numbers in black show the number of IMEs belonging to the same family. Numbers in gray show the number of 90% sequence identity clusters within the family.

found in IME_Class_8, even if the encoded proteins are not related (a VirD4 CP instead of a TcpA CP, a canonical Rel_PF13814/MobC relaxase instead of relaxase related to RCR initiators and a serine integrase instead of tyrosine integrase).

All other classes of IMEs encode other types of canonical relaxases and do not encode a CP. In IME_Class_10, the two genes encoding relaxases (a canonical Rel_PF01076/MobV and a non-canonical Rel_PF02486/MobT) are located far upstream of the tyrosine integrase gene in the opposite orientation. As well as the two relaxases and a tyrosine integrase, these IMEs encode a protein with a repA_N domain that is also found in proteins that initiates the theta replication of plasmids and of ICEs belonging to the Tn*5252* superfamily from firmicutes (Weaver et al., 2009; Guerillot et al., 2013; Ambroset et al., 2016). However, repA_N proteins from IMEs are shorter (∼100 amino acids) than those found in plasmids and ICEs (∼340 amino acids), suggesting that they probably serve a different function.

Apart from IME_Class_8, all IMEs encoding a serine integrase and relaxases from canonical superfamilies (Rel_PF03389/MobQ, Rel_PF01076/MobV, and Rel_PF03432/MobP) also encode one or several proteins that could be involved in the maintenance of the excised elements. These proteins include homologs to: (i) 'replisome organizers' or 'DnaC-related' proteins involved in the initiation of theta replication of various phages from Firmicutes such as phi5218, phi4268, or phi9871 (Trotter et al., 2006; Tang et al., 2013; McDonnell et al., 2016), (ii) 'ParB' proteins involved in chromosome and plasmid partitioning (domain TIGR00180) (**Figure 4**). Such proteins are not encoded by any other class of IMEs analyzed in this study.

In summary, whereas the analysis of IME signature proteins shows a remarkable diversity, preventing their classification

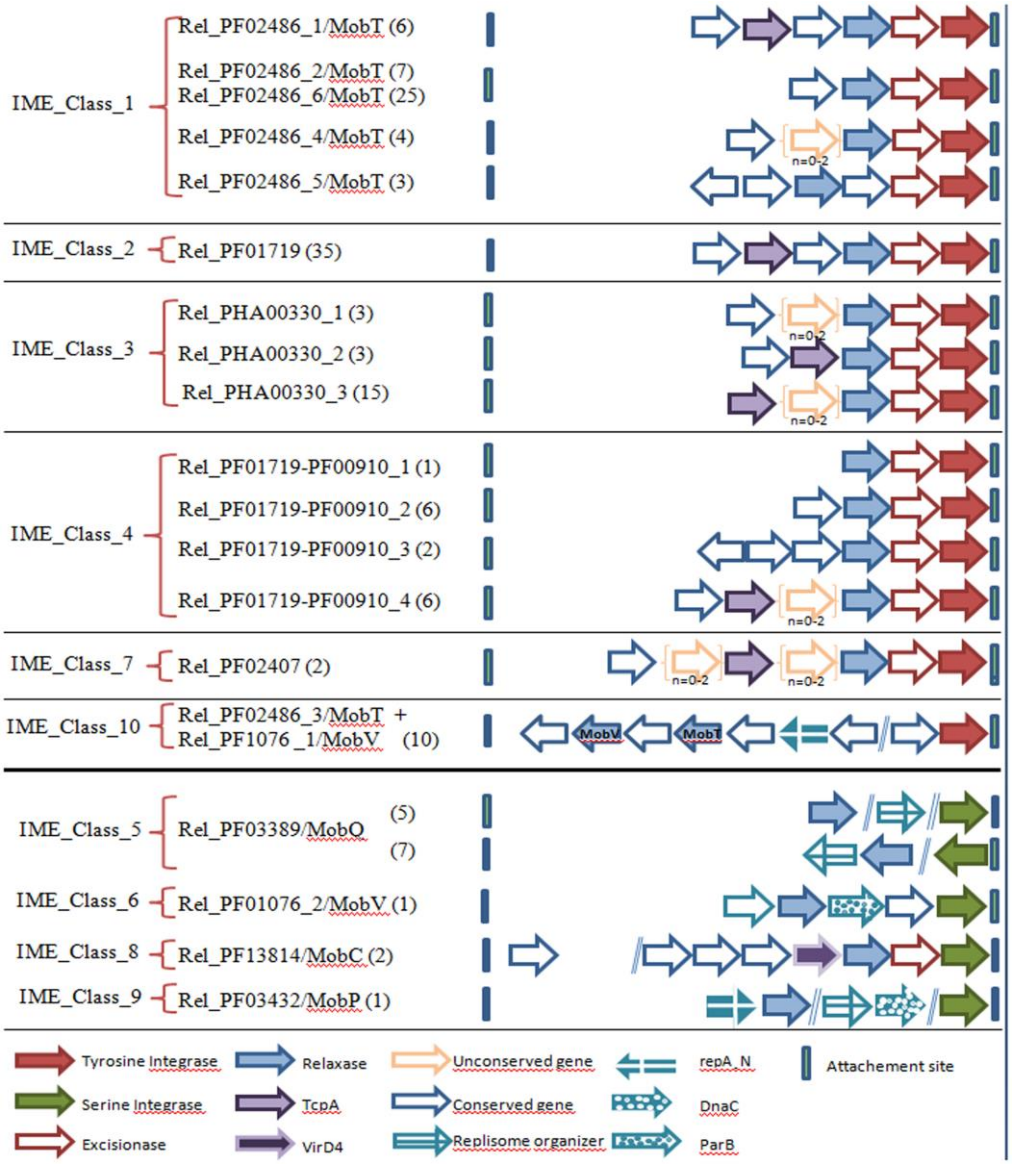

FIGURE 4 | Conserved CDS architectures within IMEs. The classification of IMEs is based on their relaxase content (see also **Table 2**). The genes encoding conserved proteins in all IMEs sharing the same relaxase superfamily are indicated by arrows. The genes between brackets correspond to non-conserved genes located between conserved ones. The numerical values show the numbers of IMEs sharing the same structure.

on the basis of their relationships, the analysis of their CDS organization shows that most of them (IME_Class_1, _2, _3, _4, _7, and _8, representing 120/144 IMEs) harbor a similar compact organization. This conserved compact organization was observed for all the IMEs encoding a single relaxase related to RCR initiators (Rel_PF02486/MobT type and four other new superfamilies discovered in this study) and for all IMEs encoding a CP (TcpA or VirD4).

Modular Evolution of IMEs

Sequence comparison of IMEs suggests that most of the exchanges of modules or CP-encoding genes occur between IMEs with similar structures, especially between elements harboring the compact structure described above. **Figure 5** illustrates two such situations involving IME_Class_2 and IME_Class_7 (**Figure 5A**), and IME_Class_3 and IME_Class1 (**Figure 5B**). In the first example, two members of IME_Class_2 (named here

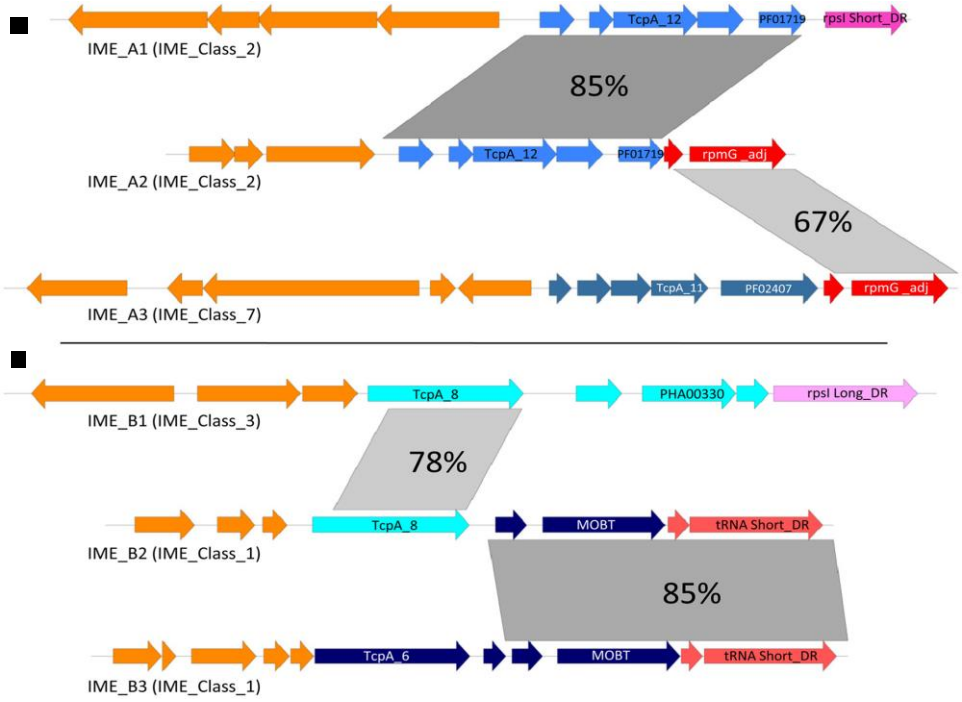

FIGURE 5 | Modular evolution of IMEs. (A) Sequence comparison of IME_*A2* with IME_*A1* and IME_*A3*. **(B)** Sequence comparison of IME_*B2* with IME_*B1* and IME_*B3*. Percentages of nucleic identities are shown in gray. Blue: genes encoding conserved proteins of the mobilization module. Red or pink: genes encoding conserved proteins of integration/excision module. Orange: other genes. IME_*A1* = IME_*SdyRE378_rpsI*, IME_*A2* = IME_*SanC238_rpmG*, IME_*A3* = IME_*Sag201008_rpmG*, IME_*B1* = IME_*Spas43144_rpsI*, IME_*B2* = IME_*SanMAS624_tRNAlys and* IME_*B3* = IME_*SpnOXC141_rpmG*.

IME_*A1* and IME_*A2* for simplicity), exhibit a closely related mobilization module but their integration/excision modules are very different. Moreover, the integration module of IME_*A2* is related to that of IME_*A3* (from IME_Class_7) suggesting a probable exchange of integration modules between IMEs. In the second example, IME_*B2* and IME_*B3* (from IME_Class_1) have related integration and mobilization modules with the exception of their CP. However, the TcpA encoded by IME_*B2* is related to the one encoded by IME_*B1* (from IME_Class3). These data point to an exchange of genes encoding TcpA CPs between IMEs encoding different types of CPs and integrases.

DISCUSSION

Prevalence and Diversity of IMEs

Integrative and mobilizable elements are by far the least known elements that transfer by conjugation. Indeed, until now, very little information on their prevalence has been available. Considering their diversity, very few IMEs with different mobilization and/or recombination modules have been reported (only 15 in 2013, see Bellanger et al., 2014 for a review). Here, we identified all IMEs carried by 124 genomes from 27 species of *Streptococcus* and compared their abundance and diversity to those of ICEs previously identified in the same set of strains (Ambroset et al., 2016). We demonstrated that IMEs have a very high prevalence and are about 40% more abundant than ICEs. We also found that the mobilization modules of IMEs display a larger diversity than the conjugation modules of ICEs.

Only 20% of our collection of streptococcal IMEs encode canonical relaxases (i.e., belong to the Rel_PF03389/MobQ, Rel_PF01076/MobV, Rel_PF13814/MobC, or Rel_PF03432/MobP superfamilies), already found to be encoded by conjugative and mobilizable elements from G- bacteria (Garcillan-Barcia et al., 2009). Among them, none encode any CP (IME_Class_5, _6, and _9) except the two IMEs with a Rel_PF13814/MobC relaxase (IME_Class_8) that encode a canonical VirD4 CP. This is reminiscent of mobilizable plasmids of G- bacteria and firmicutes (Garcillan-Barcia et al., 2009).

Unexpectedly, the large majority (90%, $n = 129/144$) of streptococcal IMEs in our collection encode non-canonical relaxases (Rel_PF02486/MobT, Rel_PF01719, Rel_PHA00330, Rel_PF01719-PF00910, and Rel_PF02407) related to proteins responsible for RCR initiation involved in the maintenance of plasmids from firmicutes or viruses. MobT was previously identified in a few predicted IMEs (Bellanger et al., 2014) and in ICEs belonging to the Tn*916* superfamily (Ambroset et al., 2016). The four other superfamilies of putative relaxases are unrelated to any known relaxase of mobilizable or conjugative elements. It should be emphasized that the MobT "relaxase" of the integrative

and conjugative element ICE*Bs1* from the firmicute *Bacillus subtilis* is involved in not only the initiation of the conjugative transfer of ICE*Bs1* but also the initiation of RCR needed for the maintenance of ICE*Bs1* after excision (Lee et al., 2010). In the same way, another family of "RCR initiators" exhibiting a Rep_1/PF01446 domain is involved not only in the maintenance of three plasmids from firmicutes but also in their mobilization by ICE*Bs1* (Lee et al., 2012). Therefore, the classical distinction between RCR initiators and relaxases could lose its relevance. In IMEs encoding a CP (n = 85/144), RCR initiator-related relaxases are always associated with a non-canonical TcpA CP. The strict association of RCR initiator and TcpA is also observed in conjugative elements encoding a RCR relaxase (i.e., the ICEs belonging to the Tn*916*/ICE*Bs1*/ICE*St3* from firmicutes) (Guglielmini et al., 2013; Ambroset et al., 2016). In these IMEs and ICEs, the *tcpA* gene is located upstream from the RCR "relaxase" gene. Taken together, these findings reveal that non-canonical relaxases related to four types of RCR initiators and non-canonical CPs related to FtsK DNA translocase are involved in the mobilization of most streptococcal IMEs.

FtsK-related proteins were previously found to be encoded by various small plasmids mainly from firmicutes (see for examples, NCBI reference sequences NP_203541, NP_613077, YP_251910, WP_011669127, YP_001967631, YP_006939188) but they were never proposed to be mobilization CPs (see for example Bachrach et al., 2004; Bjorland et al., 2007; Shkoporov et al., 2008). However, according to our analysis using CONJscan-T4SSscan, all these proteins belong to the TcpA family. Interestingly, none of these plasmids encode a canonical relaxase. Rather, they carry a "RCR initiator" gene (located next to the *tcpA*) encoding either one of the domains found in IME relaxases or a PF01446 domain (i.e., the domain found in RCR initiator/relaxases from the small plasmids mobilized by ICE*Bs1*). Therefore, it seems highly probable that these small plasmids are mobilizable. As previously proposed by Lee et al. (2012), it is probable that at least some of the small plasmids from firmicutes encoding a "RCR initiator" and lacking any canonical relaxase gene, CP, or T4SS protein, could be also mobilizable. Finally, it should also be noted that many plasmids devoid of relaxase could also carry an *oriT* related to those of conjugative element and therefore could be mobilizable *in trans,* as recently found for most plasmids from staphylococci (O'Brien et al., 2015; Pollet et al., 2016). Taken together, all these data point to a previous underestimation of the number of mobilizable plasmids and instead support their very high prevalence.

Besides IMEs and ICEs, our analysis of streptococcal genomes reveals many elements that (i) are flanked by DRs and (ii) encode an integrase but (iii) are devoid of CP and of canonical or RCR initiator-related relaxase (data not shown). At least some of these could correspond to IMEs. This hypothesis is supported first by the discovery of a novel type of relaxase related to tyrosine recombinases that is encoded by the pCW3 conjugative plasmid from the firmicute *C. perfringens* (Wisniewski et al., 2016). Second, in proteobacteria, IMEs devoid of relaxase but carrying an *oriT* are found to be mobilizable (Daccord et al., 2010) and one such IME (IME MTn*Sag1*) have previously been described in *S. agalactiae* (Achard and Leclercq, 2007). Thus, the prevalence and diversity of IMEs within analyzed streptococcal genomes could be even greater than that described here.

Modular and Intramodular Evolution

The comparison of phylogenetic analyses of integrases with those of relaxases and CPs reveals many inconsistencies, probably due to multiple replacements of integration/excision or mobilization modules between IMEs. Such replacements were previously observed in streptococcal ICEs (Ambroset et al., 2016). The phylogenetic analysis of integrase families (**Figure 1** and **Supplementary Figure S1**) clearly shows that replacement of integration/excision modules can occur not only between IMEs or between ICEs but also between ICEs and IMEs.

Within the mobilization modules encoding non-canonical relaxases, the comparison of phylogenetic analyses and/or co-occurrence of relaxases and TcpA CPs reveals inconsistencies. Thus, 10 out of 25 IMEs encoding a MobT relaxase that belong the PF02486_6 family do not carry a *tcpA* gene or pseudogene, whereas the 15 others have one. The distribution of these latter suggests that an ancestral IME devoid of TcpA has recently acquired a *tcpA* gene. On the contrary, among IMEs encoding relaxases belonging to the PF01719, PF1719-PF00910, and PHA00330 superfamilies, only 13 do not encode a TcpA and probably lost their *tcpA* by deletion. In some cases, almost identical IME relaxases (see for example the Rel_PF02486_6 cluster in **Supplementary Figure S2**) are associated with TcpA from different families suggesting that intramodular gene replacements occurred. In most cases, the data does not allow us to determine precisely what happened. However, the phylogenetic tree of the PF02486_6 relaxases suggests a replacement of a TcpA_5 CP by a TcpA_6 CP. Interestingly, although two superfamilies of relaxase (Rel_PF02486/MobT and Rel_PF03432/MobP) and the two superfamilies of CPs (VirD4 and TcpA) are shared by ICEs and IMEs from streptococci, there is no evidence of exchange of these genes between ICEs and IMEs.

Various IMEs integrated in *rpsI*, *rpmG*, and *rplL* were found to be integrated in tandem with other IMEs or ICEs encoding either related, distantly related, or unrelated integrases, relaxases, and/or CPs. We have also found many decayed elements or genomic islands in accretion with streptococcal IMEs (data not shown). These accretions result from the integration of an incoming IME or ICE by site-specific recombination in the *attL* or *attR* site of related or unrelated resident element that may not be followed by a deletion (Pavlovic et al., 2004; Bellanger et al., 2011). An accretion between elements targeting the same insertion gene and subsequent deletion of one of the transfer modules and one of the recombination modules is probably responsible for a large part of the replacement of modules or TcpA genes.

Are TcpA CP Needed or not for Mobilization?

In fact, the non-canonical TcpA superfamily of CPs was previously reported to be encoded by conjugation modules from firmicutes, but TcpA has not yet been found in a mobilization module. In conjugative elements, TcpA proteins are associated

with non-canonical relaxases: (i) the relaxase of the conjugative pCW3 plasmid from *C. perfringens* that is related to tyrosine recombinases (Wisniewski et al., 2016) or (ii) the MobT relaxases from the ICEs belonging to ICE*Bs1*/Tn*916*/ICE*St3* superfamily that are related to RCR initiators. In this work, we identified many IMEs encoding a TcpA CP: all of them encode a non-canonical relaxase. Although all IMEs encoding a non-canonical relaxase have a similar organization, their mobilization modules are highly versatile. First, closely related relaxases can be associated or not with a TcpA CP, suggesting that IME-encoded CP might not be needed for mobilization. If so, it can be hypothesized that the non-canonical relaxase might interact with the T4SS of the mobilizing conjugative element, either *via* the CP encoded by the conjugative element or *via* its cognate CP. In this hypothesis, the IME-encoded TcpA might enhance the mobilization efficiency and/or enlarge the mobilization range. Second, closely related relaxases can be associated with different distantly related TcpA CPs, indicating that the IME relaxase can interact with distantly related CPs. We hypothesize that the change of CP might have an impact on the mobilization efficiency and/or range.

IMEs within ICEs, a New Mobilization Mechanism?

The IMEs from streptococci carry diverse recombination modules and have a large array of integration specificity. Almost all serine integrases from IMEs catalyze site-specific integration within genes leading to their disruption. Interestingly, the majority of IME serine integrases (12/16) specifically target several conserved genes from the Tn*5252* superfamily, a group of ICEs that is widespread in streptococci. As previously discussed for ICEs (Ambroset et al., 2016), we would expect that target specificity should be selected to have the least effect on host fitness. Here, the disruption of a conjugation gene would have little or no effect on bacterial host but would be deleterious or lethal for the host ICE. For instance, insertion in *traG*, that encodes a VirD4 CP, would abolish the ICE transfer and therefore the mobilization of the IME by the host ICE. The consequences of the integration/excision balance of ICE or IME encoding serine recombinases have never been studied but are documented for some prophages encoding serine recombinase. For these prophages, excision occurs not only during the activation of lytic phase but also when expression of the host target gene is needed (Kunkel et al., 1990; Rabinovich et al., 2012). By analogy, we can hypothesize that excision of IMEs integrated within specific conjugation genes of ICEs would be caused by the induction of conjugation. After IME excision, the conjugation module would be functional and could be expressed, thus allowing ICE transfer. The IME could use the CP and T4SS of this ICE to transfer and then could integrate in the ICE in the transconjugant. Furthermore, if the ICE that primarily hosts the IME does not transfer or integrate in the recipient cell, the incoming IME could integrate in another resident element (related ICE or decayed ICE as long as it carries the IME integration site). This would explain the presence of such IMEs in many decayed ICEs from streptococci. ICE*Sp2905* from *S. pyogenes,* an ICE integrated in *rumA*, was demonstrated to transfer (Giovanetti et al., 2012) although it carries two IMEs: one integrated in *snf2* and the other in *maff2* (Bellanger et al., 2014). Although we cannot exclude that the disrupted genes are not required for ICE transfer, we can also hypothesize that the transfer could be divided into successive stages including excisions of the ICE and IMEs, independent transfers of the IMEs and of the ICE devoid of IMEs, insertions of the transferred ICE in *rumA* site and insertion of the IMEs within the transferred ICE. Such a mobilization mechanism of an IME integrated in an ICE can also be proposed for IMEs encoding tyrosine integrases that are site-specifically integrated in the putative *oriT* from ICEs belonging to Tn*916* and ICE*St3* families that were found in *S. agalactiae* and in *S. mutans* (Puymege et al., 2015).

Replication of Excised IMEs

Integrative and mobilizable elements and ICEs are integrated in the chromosome and are transmitted to the daughter cell as part of the chromosome. However, an excised IME would be lost in one of the daughter cell during cell division in the absence of replication. Several recent studies indicated that extrachromosomal replication is involved in the maintenance of various ICEs transferring as single-strand DNA (Ramsay et al., 2006; Lee et al., 2012; Guerillot et al., 2013; Carraro et al., 2015), but data suggesting replication of excised IMEs has not been reported so far. All streptococcal IMEs encoding a Rel_PF03432/MobP relaxase and most of those encoding a Rel_PF03389/MobQ carry one or two genes encoding proteins downstream from the relaxase gene that are distantly related to primosome proteins. Such proteins are known to be responsible for the initiation of the theta replication of prophages. Altogether, these data suggest that these IMEs are able to replicate with a theta mechanism after excision. Ten other IMEs (IME_Class_10 family) sharing a similar organization encode two putative relaxases. One belongs to Rel_PF02486/MobT superfamily, i.e., a family including both RCR initiators and relaxases. The other belongs to Rel_PF01076/MobV family. The structure of the region carrying these genes is reminiscent of the one of numerous mobilizable plasmids from firmicutes where the Rel_PF01076/MobV protein was found to be involved in mobilization whereas the Rel_PF02486/MobT protein is involved in RCR (Lorenzo-Diaz et al., 2014). Accordingly, the Rel_PF01076/MobV protein of an IME of the IME_Class_10 family was very recently shown to be a mobilization relaxase (Lorenzo-Diaz et al., 2016). Therefore, it is likely that the Rel_PF01076/MobV protein is the IME relaxase, whereas the Rel_PF02486/MobT protein would be involved in RCR replication of the excised IME. Alternatively, as previously found for the MobT relaxase of ICE*Bs1* (Lee et al., 2012), the Rel_PF02486/MobT protein could ensure both functions. Finally, all IMEs encoding a single Rel_PF01076/MobV relaxase encode a protein related to the ParB proteins that could be involved in the partition of the excised IME. Taken together, these data suggest that almost all streptococcal IMEs encoding canonical relaxases (except the two elements encoding a MobC relaxase) encode proteins involved in the maintenance of their excised

forms. Such proteins were not found for any IME encoding single non-canonical relaxases such as MobT. Since the MobT protein from ICE*Bs1* is both the relaxase and the initiator of RCR involved in the maintenance of excised ICE*Bs1* (Lee et al., 2012), we cannot exclude the possibility that the five superfamilies of non-canonical relaxases might have both functions. Overall, many if not all IMEs and ICEs might exist in two states, namely a main dormant integrated state and an activated excised state which would be maintained by replication.

AUTHOR CONTRIBUTIONS

GG and SP conceived the reference database of signature proteins. CC, NL-B, GG, and M-DD contributed to the conception of the work. CC, GG, NL-B, CA, M-DD, VL, and TL performed the acquisition and analysis of the data. GG, NL-B, CC, and M-DD drafted the manuscript. CC, NL-B, GG, and SP elaborated the figures, tables, and references. All authors criticized and finally approved the final version of the manuscript.

FUNDING

This work was supported by the Région Lorraine and the Université de Lorraine.

ACKNOWLEDGMENT

CC is recipient of a scholarship of the Ministère de l'Enseignement Supérieur et de la Recherche.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.00443/full#supplementary-material

FIGURE S1 | Phylogenetic tree of serine integrases. One representative of each 90% protein identity cluster from IMEs (in black) and one representative of each 90% protein identity cluster of serine integrases from ICEs targeting the same site as IMEs (in mauve and underlined) are presented in the ML tree. Bootstrap values (BioNJ/ML) are given only when they exceed 50 for both analyses. The target gene is mentioned in the IME/ICE names. Serine integrases sharing more than 40% identity and therefore belonging to the same family are merged with brackets. These families are distinguished with different numbers. Refer to Supplementary Table S1 for IME and strain details.

FIGURE S2 | Phylogenetic tree of Rel_PF02486/MobT relaxases. All the Rel_PF02486/MobT relaxases from IMEs (in black) and only one representative of each 90% protein identity cluster of MobT relaxases from ICEs (in mauve and underlined) are presented in the ML tree. Bootstrap values (BioNJ/ML) are given only when they exceed 50 for both analyses. Relaxases sharing more than 40% sequence identity and therefore belonging to the same family are merged with brackets. These families are distinguished with a number preceded by the Pfam identifier of the characteristic domain of this superfamily. The TcpA family associated with each relaxase is indicated.

FIGURE S3 | Phylogenetic tree of Rel_PF01719 relaxases. All the Rel_PF01719 relaxases are presented in the ML tree. Bootstrap values (BioNJ/ML) are given only when they exceed 50 for both analyses. All these relaxases share more than 40% sequence identity and therefore belong to a unique family. The TcpA family associated with each relaxase is indicated.

FIGURE S4 | Phylogenetic tree of Rel_PHA00330 relaxases. All the Rel_PHA00330 relaxases are presented in the ML tree. Bootstrap values (BioNJ/ML) are given only when they exceed 50 for both analyses. The relaxases sharing more than 40% sequence identity and therefore belonging to the same family are merged with brackets. These families are distinguished with a number preceded by the identifier of the characteristic domain of this superfamily. The TcpA family associated with each relaxase is indicated.

FIGURE S5 | Phylogenetic tree of Rel_PF01719-PF00910 relaxases. All the Rel_PF01719-PF00910 relaxases are presented in the ML tree. Bootstrap values (BioNJ/ML) are given only when they exceed 50 for both analyses. The relaxases sharing more than 40% sequence identity and therefore belonging to the same family are merged with brackets. These families are distinguished with a number preceded by the pfam identifier of the characteristic domains of this superfamily. The TcpA family associated with each relaxase is indicated.

FIGURE S6 | Phylogenetic tree of TcpA proteins. All the TcpA CPs from IMEs (in black) and one of each 90% protein identity cluster of TcpA from ICEs (in mauve and underlined) are presented in the BioNJ tree. Bootstrap values are given only when they exceed 50. The TcpA CPs sharing more than 40% sequence identity and therefore belonging to the same family are merged with brackets. These families are distinguished with different numbers. The relaxases families associated with each TcpA family are indicated.

REFERENCES

Achard, A., and Leclercq, R. (2007). Characterization of a small mobilizable transposon, MTnSag1, in *Streptococcus agalactiae*. *J. Bacteriol.* 189, 4328–4331. doi: 10.1128/JB.00213-07

Alva, V., Nam, S. Z., Soding, J., and Lupas, A. N. (2016). The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res.* 44, W410–W415. doi: 10.1093/nar/gkw348

Ambroset, C., Coluzzi, C., Guedon, G., Devignes, M. D., Loux, V., Lacroix, T., et al. (2016). New insights into the classification and integration specificity of *Streptococcus* integrative conjugative elements through extensive genome exploration. *Front. Microbiol.* 6:1483. doi: 10.3389/fmicb.2015.01483

Auchtung, J. M., Aleksanyan, N., Bulku, A., and Berkmen, M. B. (2016). Biology of ICEBs1, an integrative and conjugative element in *Bacillus subtilis*. *Plasmid* 86, 14–25. doi: 10.1016/j.plasmid.2016.07.001

Bachrach, G., Haake, S. K., Glick, A., Hazan, R., Naor, R., Andersen, R. N., et al. (2004). Characterization of the novel *Fusobacterium nucleatum* plasmid pKH9 and evidence of an addiction system. *Appl. Environ. Microbiol.* 70, 6957–6962. doi: 10.1128/AEM.70.12.6957-6962.2004

Bellanger, X., Morel, C., Gonot, F., Puymege, A., Decaris, B., and Guedon, G. (2011). Site-specific accretion of an integrative conjugative element together with a related genomic island leads to cis mobilization and gene capture. *Mol. Microbiol.* 81, 912–925. doi: 10.1111/j.1365-2958.2011.07737.x

Bellanger, X., Payot, S., Leblond-Bourget, N., and Guedon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol. Rev.* 38, 720–760. doi: 10.1111/1574-6976.12058

Bjorland, J., Bratlie, M. S., and Steinum, T. (2007). The smr gene resides on a novel plasmid pSP187 identified in a *Staphylococcus pasteuri* isolate recovered from unpasteurized milk. *Plasmid* 57, 145–155. doi: 10.1016/j.plasmid.2006.08.004

Brochet, M., Couve, E., Glaser, P., Guedon, G., and Payot, S. (2008). Integrative conjugative elements and related elements are major contributors to the genome diversity of *Streptococcus agalactiae*. *J. Bacteriol.* 190, 6913–6917. doi: 10.1128/JB.00824-08

Burrus, V., Pavlovic, G., Decaris, B., and Guedon, G. (2002). Conjugative transposons: the tip of the iceberg. *Mol. Microbiol.* 46, 601–610.

Cabezon, E., Ripoll-Rozada, J., Pena, A., De La Cruz, F., and Arechaga, I. (2015). Towards an integrated model of bacterial conjugation. *FEMS Microbiol. Rev.* 39, 81–95. doi: 10.1111/1574-6976.12085

Carraro, N., Poulin, D., and Burrus, V. (2015). Replication and active partition of integrative and conjugative elements (ICEs) of the SXT/R391 family: the line between ICEs and conjugative plasmids is getting thinner. *PLoS Genet.* 11:e1005298. doi: 10.1371/journal.pgen.1005298

Chandran Darbari, V., and Waksman, G. (2015). Structural biology of bacterial type IV secretion systems. *Annu. Rev. Biochem.* 84, 603–629. doi: 10.1146/annurev-biochem-062911-102821

Daccord, A., Ceccarelli, D., and Burrus, V. (2010). Integrating conjugative elements of the SXT/R391 family trigger the excision and drive the mobilization of a new class of *Vibrio* genomic islands. *Mol. Microbiol.* 78, 576–588. doi: 10.1111/j.1365-2958.2010.07364.x

Ebisu, S., Murahashi, Y., Takagi, H., Kadowaki, K., Yamaguchi, K., Yamagata, H., et al. (1995). Nucleotide sequence and replication properties of the *Bacillus borstelensis* cryptic plasmid pHT926. *Appl. Environ. Microbiol.* 61, 3154–3157.

Francia, M. V., Varsaki, A., Garcillan-Barcia, M. P., Latorre, A., Drainas, C., and De La Cruz, F. (2004). A classification scheme for mobilization regions of bacterial plasmids. *FEMS Microbiol. Rev.* 28, 79–100. doi: 10.1016/j.femsre.2003.09.001

Garcillan-Barcia, M. P., Francia, M. V., and De La Cruz, F. (2009). The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.* 33, 657–687.

Gibbs, M. J., Smeianov, V. V., Steele, J. L., Upcroft, P., and Efimov, B. A. (2006). Two families of rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. *Mol. Biol. Evol.* 23, 1097–1100. doi: 10.1093/molbev/msj122

Giovanetti, E., Brenciani, A., Tiberi, E., Bacciaglia, A., and Varaldo, P. E. (2012). ICESp2905, the erm(TR)-tet(O) element of *Streptococcus pyogenes*, is formed by two independent integrative and conjugative elements. *Antimicrob. Agents Chemother.* 56, 591–594. doi: 10.1128/AAC.05352-11

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27, 221–224. doi: 10.1093/molbev/msp259

Guerillot, R., Da Cunha, V., Sauvage, E., Bouchier, C., and Glaser, P. (2013). Modular evolution of TnGBSs, a new family of integrative and conjugative elements associating insertion sequence transposition, plasmid replication, and conjugation for their spreading. *J. Bacteriol.* 195, 1979–1990. doi: 10.1128/JB.01745-12

Guglielmini, J., De La Cruz, F., and Rocha, E. P. (2013). Evolution of conjugation and type IV secretion systems. *Mol. Biol. Evol.* 30, 315–331. doi: 10.1093/molbev/mss221

Guglielmini, J., Neron, B., Abby, S. S., Garcillan-Barcia, M. P., De La Cruz, F., and Rocha, E. P. (2014). Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.* 42, 5715–5727. doi: 10.1093/nar/gku194

Guglielmini, J., Quintais, L., Garcillan-Barcia, M. P., De La Cruz, F., and Rocha, E. P. (2011). The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* 7:e1002222. doi: 10.1371/journal.pgen.1002222

Ilangovan, A., Connery, S., and Waksman, G. (2015). Structural biology of the gram-negative bacterial conjugation systems. *Trends Microbiol.* 23, 301–310. doi: 10.1016/j.tim.2015.02.012

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Kunkel, B., Losick, R., and Stragier, P. (1990). The *Bacillus subtilis* gene for the development transcription factor sigma K is generated by excision of a dispensable DNA element containing a sporulation recombinase gene. *Genes Dev.* 4, 525–535.

Lee, C. A., Babic, A., and Grossman, A. D. (2010). Autonomous plasmid-like replication of a conjugative transposon. *Mol. Microbiol.* 75, 268–279. doi: 10.1111/j.1365-2958.2009.06985.x

Lee, C. A., Thomas, J., and Grossman, A. D. (2012). The *Bacillus subtilis* conjugative transposon ICEBs1 mobilizes plasmids lacking dedicated mobilization functions. *J. Bacteriol.* 194, 3165–3172. doi: 10.1128/JB.00301-12

Lorenzo-Diaz, F., Fernandez-Lopez, C., Douarre, P. E., Baez-Ortega, A., Flores, C., Glaser, P., et al. (2016). Streptococcal group B integrative and mobilizable element IMESag-rpsI encodes a functional relaxase involved in its transfer. *Open Biol.* 6:160084. doi: 10.1098/rsob.160084

Lorenzo-Diaz, F., Fernandez-Lopez, C., Garcillan-Barcia, M. P., and Espinosa, M. (2014). Bringing them together: plasmid pMV158 rolling circle replication and conjugation under an evolutionary perspective. *Plasmid* 74, 15–31. doi: 10.1016/j.plasmid.2014.05.004

McDonnell, B., Mahony, J., Neve, H., Hanemaaijer, L., Noben, J. P., Kouwen, T., et al. (2016). Identification and analysis of a novel group of bacteriophages infecting the lactic acid bacterium *Streptococcus thermophilus*. *Appl. Environ. Microbiol.* 82, 5153–5165. doi: 10.1128/AEM.00835-16

Meyer, R. (2009). Replication and conjugative mobilization of broad host-range IncQ plasmids. *Plasmid* 62, 57–70. doi: 10.1016/j.plasmid.2009.05.001

Naglich, J. G., and Andrews, R. E. Jr. (1988). Tn916-dependent conjugal transfer of PC194 and PUB110 from *Bacillus subtilis* into *Bacillus thuringiensis* subsp. israelensis. *Plasmid* 20, 113–126.

O'Brien, F. G., Yui Eto, K., Murphy, R. J., Fairhurst, H. M., Coombs, G. W., Grubb, W. B., et al. (2015). Origin-of-transfer sequences facilitate mobilisation of non-conjugative antimicrobial-resistance plasmids in *Staphylococcus aureus*. *Nucleic Acids Res.* 43, 7971–7983. doi: 10.1093/nar/gkv755

Pavlovic, G., Burrus, V., Gintz, B., Decaris, B., and Guedon, G. (2004). Evolution of genomic islands by deletion and tandem accretion by site-specific recombination: ICESt1-related elements from *Streptococcus thermophilus*. *Microbiology* 150, 759–774. doi: 10.1099/mic.0.26883-0

Pollet, R. M., Ingle, J. D., Hymes, J. P., Eakes, T. C., Eto, K. Y., Kwong, S. M., et al. (2016). Processing of nonconjugative resistance plasmids by conjugation nicking enzyme of Staphylococci. *J. Bacteriol.* 198, 888–897. doi: 10.1128/JB.00832-15

Puymege, A., Bertin, S., Guedon, G., and Payot, S. (2015). Analysis of *Streptococcus agalactiae* pan-genome for prevalence, diversity and functionality of integrative and conjugative or mobilizable elements integrated in the tRNA(Lys CTT) gene. *Mol. Genet. Genomics* 290, 1727–1740. doi: 10.1007/s00438-015-1031-9

Rabinovich, L., Sigal, N., Borovok, I., Nir-Paz, R., and Herskovits, A. A. (2012). Prophage excision activates *Listeria* competence genes that promote phagosomal escape and virulence. *Cell* 150, 792–802. doi: 10.1016/j.cell.2012.06.036

Ramsay, J. P., Sullivan, J. T., Stuart, G. S., Lamont, I. L., and Ronson, C. W. (2006). Excision and transfer of the *Mesorhizobium loti* R7A symbiosis island requires an integrase IntS, a novel recombination directionality factor RdfS, and a putative relaxase RlxS. *Mol. Microbiol.* 62, 723–734. doi: 10.1111/j.1365-2958.2006.05396.x

Richards, V. P., Palmer, S. R., Pavinski Bitar, P. D., Qin, X., Weinstock, G. M., Highlander, S. K., et al. (2014). Phylogenomics and the dynamic genome evolution of the genus *Streptococcus*. *Genome Biol. Evol.* 6, 741–753. doi: 10.1093/gbe/evu048

Shkoporov, A. N., Efimov, B. A., Khokhlova, E. V., Steele, J. L., Kafarskaia, L. I., and Smeianov, V. V. (2008). Characterization of plasmids from human infant *Bifidobacterium* strains: sequence analysis and construction of *E. coli-Bifidobacterium* shuttle vectors. *Plasmid* 60, 136–148. doi: 10.1016/j.plasmid.2008.06.005

Showsh, S. A., and Andrews, R. E. Jr. (1999). Analysis of the requirement for a pUB110 mob region during Tn916-dependent mobilization. *Plasmid* 41, 179–186. doi: 10.1006/plas.1999.1398

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi: 10.1038/msb.2011.75

Smillie, C., Garcillan-Barcia, M. P., Francia, M. V., Rocha, E. P., and De La Cruz, F. (2010). Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* 74, 434–452. doi: 10.1128/MMBR.00020-10

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tang, F., Bossers, A., Harders, F., Lu, C., and Smith, H. (2013). Comparative genomic analysis of twelve *Streptococcus suis* (pro)phages. *Genomics* 101, 336–344. doi: 10.1016/j.ygeno.2013.04.005

Trotter, M., Mcauliffe, O., Callanan, M., Edwards, R., Fitzgerald, G. F., Coffey, A., et al. (2006). Genome analysis of the obligately lytic bacteriophage 4268 of *Lactococcus lactis* provides insight into its adaptable nature. *Gene* 366, 189–199. doi: 10.1016/j.gene.2005.09.022

Weaver, K. E., Kwong, S. M., Firth, N., and Francia, M. V. (2009). The RepA_N replicons of gram-positive bacteria: a family of broadly distributed but narrow host range plasmids. *Plasmid* 61, 94–109. doi: 10.1016/j.plasmid.2008.11.004

Wisniewski, J. A., Traore, D. A., Bannam, T. L., Lyras, D., Whisstock, J. C., and Rood, J. I. (2016). TcpM: a novel relaxase that mediates transfer of large conjugative plasmids from *Clostridium perfringens*. *Mol. Microbiol.* 99, 884–896. doi: 10.1111/mmi.13270

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

4. Adaptation et automatisation de la méthode.

Une fois l'analyse des 124 génomes effectuée, la méthode a évolué en prenant en compte le savoir acquis pour que celle-ci soit plus adaptée aux ICE et aux IME de streptocoques. Le développement de la méthodologie ICEFinder et son utilisation pour l'analyse des 124 génomes de streptocoques a été très couteuse en temps. C'est pourquoi l'un des objectifs de cette thèse était d'automatiser un maximum d'étapes de l'analyse en utilisant le langage informatique Python : un langage de programmation orienté objet. J'ai profité de cette étape d'automatisation pour modifier quelques procédures. Ainsi, par souci de compatibilité et d'autonomie de l'utilisateur, la version semi-automatisée d'ICEFinder n'utilise plus le logiciel NgKlast (qui n'est plus commercialisé), mais le logiciel gratuit BLAST. Un des scripts Python que j'ai écrit est présenté en annexe. Il contient plusieurs fonctions, chacune étant chargée de réaliser une ou plusieurs étapes de la méthodologie ICEFinder.

4.1. Automatisation de la recherche par BLAST.

La première étape à avoir été automatisée est l'extraction des CDS de chaque génome et l'harmonisation de leur nom. Chaque génome étant annoté différemment, il est en effet important de renommer tous les CDS de la même façon. Ainsi une fonction crée pour chaque génome un fichier au format FASTA contenant toutes les séquences protéiques annotées dans le génome. Le nom de chaque CDS est harmonisé pour comporter le nom du génome, le nom de la CDS ainsi que sa position dans le génome (e.g. >**CP002215**_cdsid_**ADX23568.1**|**11067..12353)**.

La seconde étape automatisée est la recherche des protéines signatures. Une fonction est chargée d'effectuer une recherche de similarité par BLAST entre chaque séquence protéique extraite du génome (« hits » potentiels) et les séquences protéiques des protéines signatures contenues dans la base de données (« queries »). La base de données ICEFinder utilisée est la base de données enrichie après analyse des 124 génomes de streptocoques. Elle contient actuellement 106 intégrases à tyrosine, 43 intégrases à sérine, 10 transposases à DDE de famille Tn*GBS1*/Tn*GBS2*, 72 relaxases, 45 protéines de couplage et 36 VirB4. Les résultats bruts sont ensuite stockés en fonction de la catégorie de protéines signatures testées (intégrase, relaxase, protéines de couplage et VirB4) dans un dossier différent pour chaque génome.

La troisième fonction développée permet d'appliquer les filtres et de ne conserver que la meilleure « query » par « hit ». Les filtres utilisés par cette fonction diffèrent un peu des filtres utilisés lors de l'analyse initiale (Tableau 1). En particulier, j'ai modifié le filtre correspondant aux CP parce que certaines CP d'IME et d'ICE étaient exclues par les filtres de tailles. Ainsi, les filtres de tailles des CP excluent désormais seulement les protéines dont la taille est inférieure à 180 acides aminés (Tableau 2). Cependant, cette modification a pour conséquence la non-élimination des protéines FtsK (translocase d'ADN impliquée dans la division cellulaire). L'élimination de ces protéines reste pour l'instant manuelle mais il est envisagé de faire une recherche active de la protéine FtsK, avec la protéine FtsK en « querie » afin de pouvoir la différencier des potentielles protéines de couplages d'ICE et d'IME.

	E-value	Taux de couverture	Pourcentage d'identité	Longueur du « Hit »
Protéine de couplage	>1,00^{E}-05	<25%	<25%	<180
Relaxase	>1,00^{E}-04	<25%	<25%	<180
Intégrase à tyrosine	>1,00^{E}-04	<25%	<25%	<320
Intégrase à sérine	>1,00^{E}-04	<25%	<25%	<320
Transposase à DDE	>1,00^{E}-04	<25%	<25%	<320
VirB4	>1,00^{E}-05	<25%	<25%	<500

Tableau 2 : Filtres utilisés par la méthode ICEFinder semi-automatisée. *Sont indiquées pour chacune des protéines étiquettes recherchées, les valeurs seuils à partir desquelles les protéines détectées sont rejetées*

La troisième fonction applique les filtres correspondant à chaque type de protéines, par exemple si une intégrase contenue dans la base de données détecte une protéine dans un génome, les filtres correspondant à la catégorie « intégrase » seront appliqués (filtres de taille, « e-value », pourcentage d'identité et de couverture) (tableau 2). De plus, lorsque plusieurs protéines franchissent les filtres, cette fonction ne conserve que la protéine ayant la meilleure correspondance avec la protéine identifiée dans le génome et ce, en fonction de l'e-value de l'alignement. Comme il est très fréquent que la base de données contienne plusieurs protéines « query » capables de détecter une protéine « hit » codée par le génome analysé, ceci permettra de connaître la « query » ressemblant le plus à la protéine « hit » détectée.

La 4ème fonction développée est chargée de regrouper dans un tableau au format « .csv » tous les gènes codant des protéines signatures identifiés dans un génome et de trier ces gènes en fonction de leur position dans le génome afin de déterminer facilement si ces derniers sont co-localisés à une distance compatible avec celle attendue s'ils sont codés par un ICE ou un IME. De plus, la fonction récupère les informations contenues dans la base de données pour la meilleure « query » associée à chaque « hit » et permet ainsi de savoir à quelle famille la protéine « query » appartient, par quel élément elle est codée et, pour les intégrases de savoir quel site d'intégration elles ciblent probablement. Ceci permet de prédire et déterminer rapidement les sites d'intégration des éléments détectés lorsque leur intégrase est proche d'une intégrase de spécificité connue. Ceci permet également de déterminer rapidement à quelle famille appartiennent les éléments lorsqu'ils sont proches d'un élément connu.

L'étape de recherche de gènes manquants, quant à elle, ne pouvant pas être automatisée de par sa complexité, nécessite toujours une intervention manuelle et est dépendante de l'expertise de l'utilisateur. Elle peut rester cependant nécessaire en cas d'annotation des génomes de qualité insuffisante ou pour identifier de nouveaux éléments qui coderaient des protéines signatures trop différentes de celles incluses dans la base de données.

Bien que l'étape de délimitation des éléments ne soit pas automatisée, les connaissances acquises au cours de l'analyse des 124 génomes de streptocoques nous ont permis d'affiner notre méthode en fonction du type d'intégrase et du site d'intégration (Figure 14) et à proposer la procédure suivante :

Lorsque l'élément à délimiter code une intégrase à tyrosine, différentes approches vont être utilisées :

- Les éléments du type Tn*916* étant extrêmement proches entre eux et codant des intégrases ne catalysant pas une intégration site spécifique, les extrémités internes de Tn*916* (environ 50 pb) sont utilisées en « queries » afin de délimiter par BlastN ce type d'élément.

- Si l'élément à délimiter code une intégrase générant des DR longues conservées (plus de 12 pb), comme c'est le cas pour les intégrases ciblant les ARNt ou *rplL*, alors la séquence nucléotidique de l'extrémité ciblée (5' ou 3' en fonction du gène) du gène cible est utilisée en « query » afin de rechercher par BlastN les DR de part et d'autre de l'élément.

- Si l'élément code une intégrase générant des DR longues légèrement dégénérées (dû à la redondance du code génétique, la séquence protéique est, quant à elle, conservée), comme pour les intégrases ciblant *rpmG* ou *rpsI*, alors une recherche de DR par BlastP utilisant la séquence protéique de l'extrémité du gène ciblé en « query » est réalisée.

- Si l'élément code une intégrase générant des DR courtes (inférieure à 12 pb), comme certaines intégrases ciblant *guaA*, il est alors difficile de faire une simple recherche de DR compte tenu de leur taille. Dans ce cas, une délimitation par synthénie est réalisée.

- Enfin, si l'intégrase détectée a une spécificité inconnue, alors une analyse manuelle du contexte est réalisée afin d'identifier le site d'intégration ciblé. Si un site d'intégration est trouvé et qu'une délimitation par recherche de DR est possible alors une recherche de DR est réalisée sinon, l'élément est délimité par synthénie.

Si l'élément à délimiter code une intégrase à sérine deux approches sont envisagées :

- Toutes les intégrases à sérine identifiées lors de l'analyse des 124 génomes s'intègrent à l'intérieur des gènes qu'elles ciblent, fragmentant le gène cible en deux parties localisées de part et d'autre de l'élément intégré. De ce fait, la délimitation des éléments codant une intégrase à sérine apparentée à une intégrase ciblant un site déjà connu se fait en réalisant un BlastN du gène ciblé, tronqué, contre un gène cible complet issu de la banque NCBI. Cette démarche nous permet de déterminer l'endroit exact de la troncature qui correspond au site d'insertion.

- Si l'intégrase codée par l'élément à délimiter n'a pas de spécificité connue alors une analyse par BlastX de la zone entourant l'élément contre la banque du NCBI est réalisée afin d'identifier un gène fragmenté en deux. Si un tel gène est identifié, une analyse de contexte est réalisée afin de déterminer si ce gène peut être le site d'intégration de l'élément. Si oui, l'élément est délimité selon la méthode décrite précédemment. Cependant, si aucun gène tronqué n'est identifié et qu'aucune spécificité d'intégration n'a été déterminée, alors l'élément est délimité par synthénie.

Enfin les transposases à DDE des ICE de la famille Tn*GBS1*/Tn*GBS2* n'ayant pas de spécificité stricte, ces éléments sont délimités par synthénie. De plus, ces limites peuvent être

confirmées par la présence de répétitions inversées conservées situées aux extrémités internes des éléments et la duplication du site cible de 8 pb.

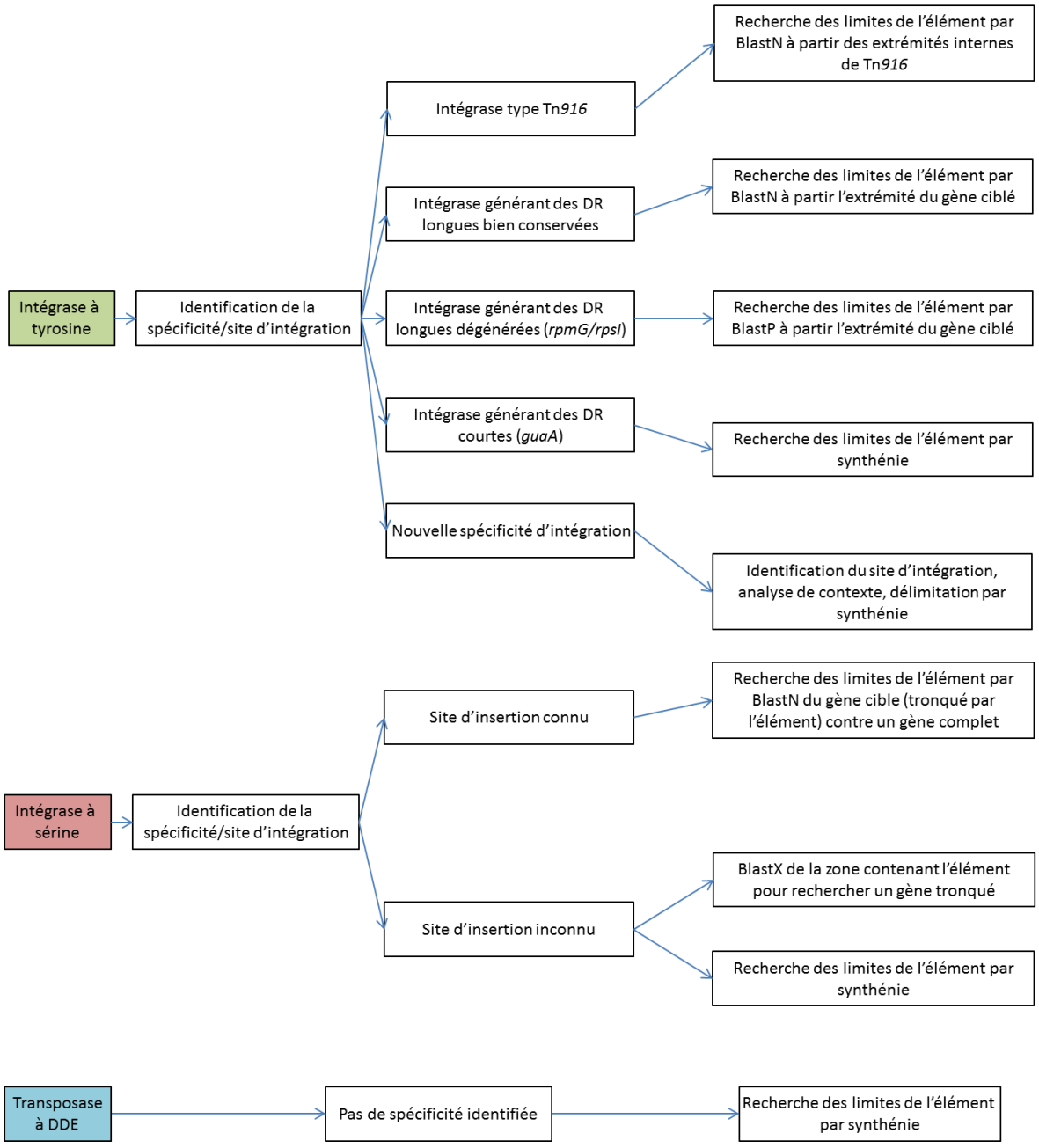

Figure 14 : Arbre de décision des différentes approches utilisées pour délimiter les ICE et les IME rencontrés chez les streptocoques en fonction du type d'intégrase et du site d'intégration. *En vert le chemin décisionnel pour un intégrase à tyrosine, en rouge pour une intégrase à sérine, en bleu pour une transposase à DDE.*

4.2. Elargissement de l'analyse à d'autres groupes bactériens.

Après avoir analysé tous les génomes de streptocoques disponibles dans les bases de données au début de l'étude, la suite logique serait d'élargir l'étude à des groupes

phylogénétiques plus larges comme l'ensemble des bactéries lactiques ou d'étudier plus largement le phylum des firmicutes. Ceci permettrait probablement d'identifier de nouvelles familles d'ICE ou IME, permettant ainsi de mieux appréhender leur diversité et leur impact dans le monde bactérien. Cependant, il semble probable qu'une partie des éléments hébergés présenteront des séquences éloignées des séquences des protéines signatures actuellement présentes dans notre base de données. Dans ce cadre, les logiciels comme BLAST ou NgKlast utilisent des algorithmes moyennement performants pour détecter des séquences phylogénétiquement éloignées. Il est donc préférable d'opter pour des programmes utilisant des modèles de Markov cachés plus efficaces pour détecter des séquences éloignées (Söding, 2005). De ce fait, la méthodologie ICEFinder a été modifiée afin d'intégrer le programme HMMer. Ce programme utilise, non pas les séquences protéiques présentent dans la base de données ICEFinder, mais des profils HMM (Hidden Markov Model) préalablement créés par l'utilisateur à partir de ces séquences. Un profil HMM est un consensus, c'est-à-dire l'expression probabiliste condensée d'un ensemble de motifs, basée sur une chaîne de Markov cachée. Les profils HMM sont principalement utilisés pour modéliser des familles de protéines dans les bases de données dédiées (Sonnhammer et al., 1997). Afin de créer ces profils HMM pour chacune des protéines signatures, (i) les protéines partageant plus de 40% d'identité sur au moins 40% de leur longueur ont été regroupées, (ii) un alignement multiple pour chaque groupe a ensuite été réalisé, et (iii) les profils ont été créés à partir de ces alignements à l'aide du programme HMMer. Les groupes comportant moins de 10 protéines ont été enrichis avec des protéines proches retrouvées dans les bases publiques (NCBI). Au total, 24 profils d'intégrase à tyrosine ont été créés, 11 d'intégrase à sérine, 2 profils de transposase à DDE, 33 profils de relaxases, 18 profils de protéines de couplage et 8 profils de protéines VirB4.

Les scripts ont donc été modifiés pour intégrer HMMer à l'analyse. Ainsi, la seconde fonction a été modifiée afin de remplacer l'analyse BLAST de chaque génome par une analyse HMMEr en utilisant tous les profils disponibles dans la base de données ICEFinder. Les analyses réalisées par HMMer ne permettant pas d'obtenir un pourcentage d'identité entre un profil et la protéine trouvée par celui-ci, cela a contraint à éliminer dans la 3ème fonction le pourcentage d'identité de la liste des filtres utilisés.

Des analyses préliminaires ont été réalisées afin de tester cette méthode sur les 124 génomes de streptocoques déjà analysés et sur différents génomes de firmicutes phylogénétiquement proches des streptocoques. Les premiers résultats sont encourageants car d'une part cette méthode nous permet de détecter tout les protéines détectées lors de l'analyse des génomes de streptocoques et d'autre part elle permet la détection d'ICE et d'IME dans des génomes autres que ceux des streptocoques. Cependant, ils suggèrent que pour certaines classes de protéines les filtres devront être modifiés pour élargir la recherche à des éléments éloignés de ceux de streptocoques. Par exemple, les filtres de taille des protéines de couplage excluaient des protéines de couplage d'ICE retrouvés chez des ICE de superfamille Tn*GBS1* de *Lactobacillus casei.* De plus, les informations concernant les sites d'intégration ciblés par chaque famille d'intégrases s'avèrent être, dans certains cas, inutilisables du fait de la trop grande distance phylogénétique entre les intégrases contenues dans la base de données et les intégrases retrouvées dans les génomes analysés, alors que les gènes ciblés soient de même type.

5. Caractérisation des gènes d'adaptation véhiculés par les ICE et les IME de streptocoques.

Notre capacité à borner les ICE et les IME permet d'avoir accès à l'ensemble des gènes transportés par ces élements : non seulement ceux nécessaires au transfert de l'élément mais également tous les gènes dits « d'adaptation ». Sachant que de nombreux ICE décrits dans la littérature sont porteurs de gènes de résistances aux antibiotiques, dont la dissémination est un fléau pour la santé publique, il nous a semblé important d'étudier la nature des gènes d'adaptation codés par les ICE et IME de streptocoques.

Pour ce faire, deux approches ont été envisagées. La première consistait à déduire la fonction des gènes d'adaptation en utilisant les informations (annotation, analyse de domaines…) les caractérisant. La seconde consistait en l'utilisation de bases de données expertes, pour rechercher des fonctions précises, comme les gènes de résistances à des antibiotiques. Bien que la première stratégie ait l'avantage d'explorer toutes les fonctions *sans a priori*, c'est la seconde approche qui a été choisie. En effet, après analyse des gènes d'adaptation d'un petit set d'éléments, nous avons constaté que la majorité d'entre eux (plus d'un gène sur deux) étaient annotée comme codant des « protéines hypothétiques » et

que la plupart de ces protéines hypothétiques ne comportaient pas de domaines identifiés ou portaient des domaines indiquant des fonctions biochimiques ne permettant pas de prédire une fonction biologique tels que « ABC transporteur » ou « domaine de fixation à l'ADN ». Ce manque d'information rendait l'approche sans *a priori* inopérante.

5.1. Base de données expertes utilisées.

Un préalable à l'analyse des gènes d'adaptation via l'utilisation de bases de données expertes était : (i) d'identifier quels gènes d'adaptation nous étions susceptibles de retrouver parmi les ICE et les IME de streptocoques et (ii) lorsque des bases de données étaient disponibles, de choisir les plus fiables, complètes et faciles d'emploi. Après une recherche bibliographique, nous avons décidé de concentrer notre recherche sur 4 catégories de gènes d'adaptation pour lesquels des bases de données expertes sont disponibles : (i) les résistances aux antibiotiques, (ii) les résistances aux ions de métaux lourds et transporteurs d'ions métalliques, (iii) les systèmes de restriction-modification (R-M) qui confèrent des une résistance aux bactériophages, et (iv) les bactériocines. Pour cela, nous avons eu recours à 4 bases de données spécialisées.

5.1.1. CARD : The Comprehensive Antibiotic Resistance Database

CARD est une base de données de résistances aux antibiotiques (Jia et al., 2017). Décrite comme la base de données de référence pour ce type de fonction par le journal « Nucleic Acid Research » (Galperin et al., 2017), elle s'appuie sur près de 2400 publications et contient plus de 2300 séquences de référence. Cette base de données est mise à jour tous les mois et est épurée manuellement. De plus, CARD est accompagné d'un outil de détection adapté, appelé RGI (Resistance Gene Identifier), ce qui en facilite l'emploi. Cette base de données est accessible sur le site dédié : https://card.mcmaster.ca/

5.1.2. BacMet : Antibacterial Biocide and Metal Resistance Genes Database

BacMet est une base de données de résistances aux ions de métaux lourds et biocides bactériens (Pal et al., 2014). Elle contient aussi des gènes codant des transporteurs d'ions métalliques. Elle fait également partie des bases de données de référence recensées par le journal « Nucleic Acid Research » (Galperin et al., 2017). Comme elle a été mise à jour pour la dernière fois le 18 janvier 2014 (http://bacmet.biomedicine.gu.se/), il est probable qu'elle

ne soit pas complète. Elle contient cependant plus de 40 000 gènes prédits informatiquement, plus de 700 gènes confirmés expérimentalement et a été manuellement épurée. Cette base de données est disponible sur le site dédié : http://bacmet.biomedicine.gu.se/

5.1.3. REBASE : The Restriction Enzyme Database

Décrite comme base de données de référence pour les systèmes R-M (Galperin et al., 2017), REBASE contient plus de 20 000 endonucléases de restriction et méthyl-transférases (Roberts et al., 2015). Ces enzymes peuvent appartenir soit aux systèmes R-M de types I, II, IIC et III ou aux endonucléase de type 4 (composé d'une seule enzyme coupant l'ADN méthylé). REBASE est mise à jour régulièrement (plus d'une fois par mois) et est épurée manuellement. Cette base de données est disponible sur le site dédié : http://rebase.neb.com/rebase/rebase.html

5.1.4. BAGEL : automated bacteriocin mining

BAGEL est une base de données de bactériocines (de Jong et al., 2006). C'est la seule base de données utilisée au cours de nos travaux ne faisant pas partie des bases de référence recensées comme « gold standards » par le journal « Nucleic Acid Research », cependant c'est aussi la seule base de données dédiée aux bactériocines disponible. Cette base de données regroupe 3 bases de données différentes : une base de données comportant les petites bactériocines modifiées, connues en tant que lantibiotiques (bactériocine de Classe I), une base de données comportant les petites bactériocines non modifiées (bactériocines de Classe II) et enfin une base de données comportant les bactériocines de taille supérieure à 10 kDa (bactériocines de Classe III). Cependant, les bases de données de BAGEL n'ont pas été mise à jour depuis le 21 janvier 2013 (http://bagel.molgenrug.nl/index.php/bacteriocin-database) et pourraient donc présenter des carences. Enfin, un outil dédié à la recherche de bactériocines BAGEL3 (van Heel et al., 2013) utilisant ces bases de données est disponible sur le site http://bagel.molgenrug.nl/index.php/bagel3 (dernière mise à jour 17 octobre 2013).

5.2. Méthode de détection utilisée

Afin de rechercher les gènes d'adaptation dans les ICE et les IME, l'ensemble des CDS portées par ces éléments ont été extraites. Lorsqu'un outil de détection dédié à une base de

données était disponible, cet outil a été utilisé avec les paramètres par défaut. C'était le cas des bases de données CARD et BAGEL. Lorsqu'aucun outil de détection, facilement utilisable, n'était disponible, l'ensemble des séquences contenues dans les bases de données ont été téléchargées et une recherche par homologie de séquence avec BLASTP a été réalisée. C'était le cas des bases de données BacMet et REBASE. Un seuil de 80% d'identité sur 80% de la longueur a été appliqué aux « hits » détectés afin d'éviter la détection de faux positifs. De ce fait, il est possible que des gènes codant les fonctions recherchés mais dont les séquences sont trop éloignées des séquences présentes dans les bases de données n'aient pas été détecté.

5.3. Gènes d'adaptation détectés chez les ICE

En avant-propos, il est important de noter que lorsqu'un élément (ICE ou IME) portait un gène d'adaptation et était intégré dans un autre élément (ICE ou IME), le gène d'adaptation a été comptabilisé à la fois pour l'élément intégré et l'élément hôte. En effet, le transfert de l'élément intégré seul est bien entendu possible, mais le transfert de l'élément composite est également envisageable.

a. Résistances aux antibiotiques

Sur l'ensemble des 105 ICE et 26 dICE analysés, 59 éléments portent une ou plusieurs résistances aux antibiotiques (Figure 15). L'essentiel de ces ICE codant une résistance à un antibiotique (48/59) appartiennent aux familles Tn*916* et Tn*5252*. La très grande majorité de ces résistances conférent la résistance à la tétracycline (gène *tetO*) ou à l'érythromycine (gène *ermB*). Les gènes conférant la résistance à la tétracycline sont généralement portés par des éléments de la famille Tn*916* (tous les éléments de cette famille la portent) et les gènes conférant la résistance à l'érythromycine par les éléments de la famille Tn*5252.* Quatre gènes conférant une résistance au chloramphénicol ont aussi été détectés, tous portés par des ICE de la famille Tn*5252.* Sept ICE appartenant à la famille Tn*1549* codent également des résistances aux antibiotiques. Parmi eux, 3 codent une résistance à la tétracycline procurée par un gène différent (gène *tet32*) de celui généralement porté par les éléments de la famille Tn*916* (gène *tetO*). Trois codent une résistance à l'érythromycine procurée par un gène différent (gène *ermA*) de celui généralement porté par les éléments de

la famille Tn*5252* (gène *ermB*). Enfin, un ICE portait un gène conférant la résistance au lincosamide, seul gène conférant cette résistance détectée au cours de cette étude.

Les 2 ICE appartenant à la famille *VanG* (famille ainsi nommée par référence au premier ICE de cette famille qui portait le gène de résistance à la vancomycine *vanG*) portent également des gènes de résistance à des antibiotiques. Cependant ces gènes ne conférent pas la résistance à la vancomycine, mais à la tétracycline (gène *tet32*) pour ICE_SsuGZ1_lysS, et à la tigécycline (gène *mepA*) pour ICE_*Spy2096_rumA*.

Enfin, un seul ICE de la famille ICE*St3* (ICE_*SgaUCN34_ftsK*) et un seul ICE de la famille Tn*GBS2* (ICE_*ScoC1050_mutT*) portent des gènes résistances à des antibiotiques. Dans le cas d'ICE_*SgaUCN34_ftsK*, cette résistance à la tétracycline est due à l'intégration d'un ICE de la famille Tn*916.* Dans le cas d'ICE_*ScoC1050_mutT*, le gène détecté confère potentiellement la résistance à la polymyxine (gène *pmrE*), un antibiotique utilisé généralement contre les bactéries Gram- (https://www.vidal.fr/substances/6857/polymyxine_b/). De plus, cet ICE est le seul ICE que nous avons identifié codant à la fois un module de conjugaison appartenant à la famille Tn*GBS2* et un module d'intégration comportant 3 intégrases à sérine. Cette association atypique pour les éléments de la famille Tn*GBS2* résulte vraisemblablement d'un réarrangement entre deux éléments, dont un de famille Tn*GBS2* et l'autre de famille Tn*5252*. Ce réarrangement pourrait être à l'origine de la présence de cette résistance à la polymyxine dans un élément de la famille Tn*GBS2.*

a. Système de restriction-modification

Initialement, la base de données utilisée pour notre recherche de système R-M contenait : les gènes impliqués dans les systèmes de restriction-modification de type I, II, IIG, III et IV ; les gènes « S » contrôlant la spécificité de ces systèmes ainsi que les gènes « C » impliqués dans la régulation des systèmes R-M. Cependant, l'utilisation de l'ensemble de cette base montre que les gènes « S » et « C » détectent de nombreux faux positifs sans aucun rapport avec les systèmes RM (par exemple les excisionases d'ICE) et ont donc été écartés de l'étude.

Au total, sur les 105 ICE et 26 dICE analysés, 59 éléments portent, au moins, une enzyme pouvant appartenir à un système R-M. Ces éléments codent dans la plupart des cas, soit des méthyl-transférase de type II isolées (qui pourraient ou non appartenir à un système R-M

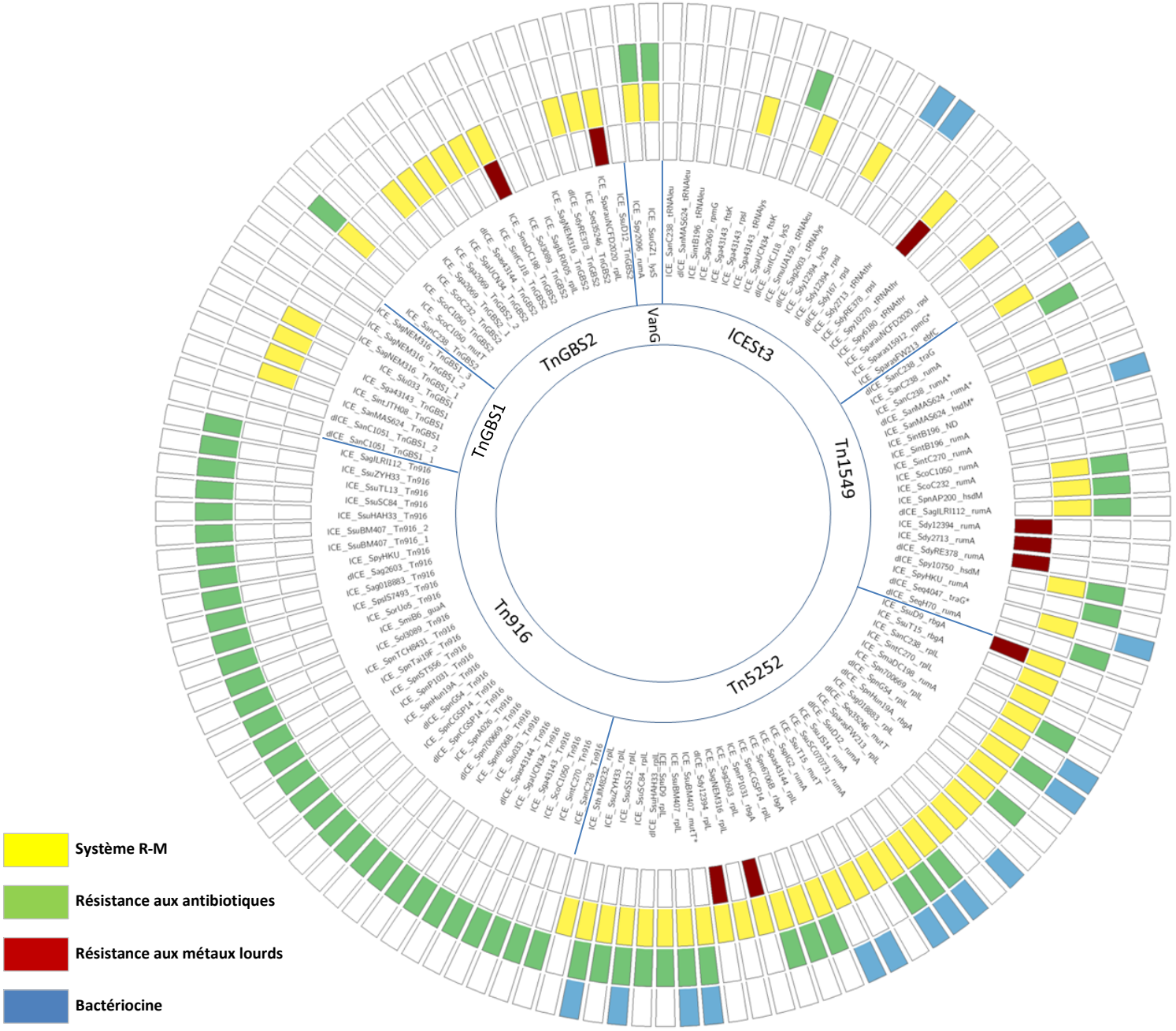

Figure 15 : Fonctions d'adaptation portées par les ICE de streptocoques. *Les ICE sont classés en famille autour du cercle. Un carré de couleur est indiqué lorsqu'au moins un gène impliqué dans la fonction d'adaptation a été détecté. En jaune : système de restriction-modification ou méthyltransférase isolée ; en vert : résistance aux antibiotiques ; en rouge : résistance aux métaux lourds (contient également des transporteurs d'ions) ; en bleu : synthèse de bactériocine.*

dont l'endonucléase de restriction aurait échappé à la détection), soit des méthyl-transférase de type II associées à des enzymes de restriction de type II.

Parmi ces 59 ICE, on retrouve sans exception tous les éléments (29) de la famille Tn*5252* qui codent une, ou plusieurs, méthyl-transférase isolées de type II. Ceci suggère que ces enzymes pourraient ne pas être appartenir à un système R-M mais plutôt intervenir dans le cycle de vie des éléments de cette famille.

De même, 10 éléments de la famille Tn*GBS2* portent des gènes pouvant être impliqués dans des systèmes R-M. Sept d'entre eux semblent coder des systèmes de restriction de type I, tandis que 3 ne semblent coder que des méthyl-transférases de type II. De plus, 9 ICE de la famille Tn*1549* codent des enzymes impliquées dans des systèmes R-M de type II, 6 codant des systèmes complets et 3 ne codant apparemment que la méthyl-transférase. Par ailleurs, quatre éléments de la famille Tn*GBS1* codent des méthyl-transférases de type II dont une est associée à une enzyme de restriction de type II. Enfin, les 2 éléments de la famille *VanG,* ICE_*SsuGZ1_lysS* et ICE_ *Spy2096_rumA,* codent respectivement une méthyl-transférase de type IV et 2 méthyl-transférase de type II.

Bien que ICE*St3*, prototype de la famille du même nom, soit connu pour coder un système R-M (Burrus et al., 2001; Johnson and Grossman, 2015), seulement 5 des 53 ICE/dICE de la famille ICE*St3* détectés dans les 124 génomes de streptocoques portent de tels gènes. Quatre ICE semblent coder des systèmes R-M de type II complets, tandis qu'un ICE semble coder une méthyl-transférase de type III isolée. Enfin, parmi les 59 éléments portant des gènes pouvant être impliqués dans des systèmes R-M, on ne retrouve aucun élément de la famille Tn*916*.

b. Synthèse de bactériocines et résistance aux métaux lourds.

Au cours de notre analyse du contenu en gènes d'adaptation des ICE, nous avons aussi détectés, en plus petit nombre, des gènes impliqués dans la synthèse de bactériocines ainsi que des gènes pouvant conférer une résistance aux métaux lourds.

Ainsi, 17 ICE appartenant à 3 familles portent des gènes pouvant être impliqués dans la synthèse de bactériocines. La plupart de ces ICE appartiennent à la famille Tn*5252* (12/17) et, dans une moindre mesure, aux familles ICE*St3* (3/17) et Tn*1549* (2/17). Tous les gènes détectés codent des bactériocines de classe I (lantibiotiques) pouvant être de 4 types

différents (connus pour être synthétisés par les streptocoques) apparentées à la macédovicine, thermophiline, salivaricine ou nisine sans que la présence d'aucune des 4 ne puissent être corrélée spécifiquement à une famille d'éléments.

Enfin parmi les 131 ICE/dICE analysés, seulement 9 ICE portent des gènes pouvant coder une résistance aux métaux lourds ou un transport d'ions métalliques. Parmi ces 9 ICE, 8 sont de la superfamille Tn*5252* (4 de la famille Tn*1549*, 2 de la famille Tn*1549* et 2 de la famille Tn*GBS2*) et un de la famille ICE*St3.* Les 4 éléments de la famille Tn*1549* codent tous un transporteur du fer (Fe^{3+}). Ces éléments étant retrouvés intégrés dans des souches pouvant être pathogènes comme *S. dysgalactiae* et *S. suis.* Le fer libre étant en quantité limité dans les hôtes de ces bactéries, il semblerait plus probable que ces transporteurs ne confèrent non pas une résistance au fer, mais permette au contraire à la bactérie d'importer du fer dans la cellule. Parmi les 5 autres gènes conférant une résistance aux métaux lourds détectés, 4 confèraient une résistance aux ions Cd^{++} (portés par 2 ICE de la famille Tn*5252*, un ICE de la famille Tn*GBS2* et un ICE de la famille ICE*St3*). Enfin le dernier gène détecté, porté par un élément de la famille Tn*GBS2*, confèrerait une résistance à l'arséniate.

5.4. Gènes d'adaptation détectés chez les IME

Bien que l'étude des gènes d'adaptation portés par les IME ne soit pas entièrement finalisée, quelques tendances générales semblent ressortir. Tout d'abord, sur les 144 éléments étudiés, seulement 36 codent une des fonctions d'adaptation recherchées. Seulement deux IME codent des résistances aux antibiotiques : IME_*ScoC232_maff2* porte un gène de résistance à la tétracycline (gène *tet32*) et IME_*Slu033_rpsI* un gène de résistance à l'érythromycine (gène *ermB*). De même, seulement 18 IME codent des gènes impliqués dans des systèmes R-M, la plupart de ces systèmes sont des systèmes de type IIG, pour lesquels une seule enzyme est requise pour la méthylation et la restriction.

Par ailleurs, un plus grand nombre d'IME que d'ICE (14 contre 9) portent des gènes de résistance aux métaux lourds. La grande majorité de ces gènes (10/14) codent des systèmes d'efflux spécifiques des ions Cd^{++}, tandis que 2 d'entre eux sont impliqués dans la résistance à l'arséniate. Enfin, deux IME portent des gènes codant un transporteur du fer (Fe^{3+}). Ces IME étant intégrés dans la bactérie pathogène *S. pneumoniae*, il se peut que les protéines

codées par ces gènes ne confèrent pas une résistance au fer, mais permettent au contraire à la bactérie d'importer du fer.

Enfin, seulement 2 IME portant des gènes codant pour des bactériocines ont été détectés. Il s'agit de bactériocines de classe I apparentées à la salivaricine et à la mutacine.

Bien que cette faible prévalence en gènes d'adaptation au sein des IME puisse être due à la plus petite taille des IME par rapport à celle des ICE, ce résultat pourrait aussi être dû au fait que seule une fraction faible des gènes d'adaptation correspondent à ceux recherchés. Globalement, l'impact que les IME peuvent avoir sur le mode de vie des organismes dans lesquels ils sont intégrés reste donc à caractériserCeci sera fait en poursuivant l'identification des fonctions adaptatives portées par ces élements.

DISCUSSION

1. ICE et IME : des éléments méconnus

En 2002, Burrus et al. ont réalisé la première recherche d'ICE dans les 22 génomes de firmicutes disponibles à l'époque. Au cours de cette étude, 17 ICE ont été identifiés, suggérant que ces éléments sont répandus dans ce groupe bactérien. Depuis, les progrès réalisés dans les méthodes de séquençage ont permis d'augmenter considérablement le nombre de génomes disponibles pour les scientifiques. Cependant, les recherches d'ICE réalisées à partir de ces nouveaux génomes se sont presque toutes concentrées sur un nombre restreint de génomes de la même espèce (voire un seul génome) et/ou sur une famille particulière d'ICE. Ainsi, deux recherches exhaustives d'ICE sur 8 génomes de *S. agalactiae* (Brochet et al., 2008) et 11 génomes de *C. difficile* (Brouwer et al., 2011) ont respectivement révélé 12 et 30 ICE. Plus récemment en 2015, Puymège et al. ont montré que des ICE apparentés sont intégrés dans l'extrémité 3' des gènes codant un $ARNt^{lys}$ dans 88 des 303 génomes de *S. agalactiae* analysés. Avant cette thèse, seules deux études avaient tenté d'apprécier la diversité et la prévalence des ICE dans un large set de génomes. La première en 2011, réalisée par Ghinet et al. s'intéressant essentiellement aux AICE, des ICE d'actinobactéries se transférant par conjugaison sous la forme double brin, a mis en évidence 144 AICE et 17 ICE dans 275 génomes d'actinobactéries. La seconde, réalisée en 2011 par Guglielmini et al et recherchant non pas des ICE, mais des gènes de conjugaison dans 1124 chromosomes bactériens, a identifié 335 modules chromosomiques de conjugaison correspondant potentiellement à 335 ICE. Ainsi, bien que la plupart des études réalisées aient permis d'enrichir les connaissances sur une famille d'ICE ou sur une espèce bactérienne, extrêmement peu permettent d'avoir une vue d'ensemble de la prévalence et de la diversité des ICE au sein d'un groupe bactérien.

Parmi les éléments se transférant par conjugaison, les IME sont, de loin, les moins bien connus. En 2014, seul le transfert conjugatif de moins de 15 IME, différant par leur module de mobilisation ou d'intégration, était décrit dans la littérature (Bellanger et al., 2014). Au début de cette thèse, les recherches se réduisaient, en tout et pour tout, à une seule recherche systématique d'IME, et ce uniquement sur 8 génomes de *S. agalactiae* (Brochet et al., 2008). Depuis, seules deux recherches d'une famille d'IME sur un set important de souches ont été réalisées. Toutefois, ces deux études se sont concentrées, comme la plupart

des études portant sur les ICE, sur une seule espèce bactérienne (*S. agalactiae*). Elles ont permis de montrer la forte prévalence de 2 familles d'IME dans cette espèce : l'une intégrée dans l'extrémité 3' du gène *rpsI* (112 sur 204) (Lorenzo-Diaz et al., 2016) et l'autre dans l'extrémité 3' d'un gène codant un $ARNt^{lys}$ (69 sur 303) (Puymège et al., 2015). Ainsi, bien plus encore que pour les ICE, les données concernant la prévalence et la diversité des IME au sein des génomes bactériens sont des plus fragmentaires voire inexistantes. Il est à noter cependant que lors de l'étude réalisée par Guglielmini et al. en 2011, les auteurs ont détecté un grand nombre de relaxases isolées dans les chromosomes analysés. Ils suggèrent que ces gènes de relaxases, éloignées de gènes de T4SS, puissent correspondre à des IME, faisant ainsi des IME les éléments se transférant par conjugaison les plus répandus.

Les efforts réalisés par la communauté scientifique pour mettre en place des outils bio-informatiques permettant d'analyser les génomes a conduit à la création d'outils de recherche de certaines classes d'éléments mobiles comme les prophages (PHAST) (Zhou et al., 2011) ou les IS (ISFinder) (Siguier et al., 2006). Cependant, il n'existe actuellement aucun outil ou méthode dédié à la détection des ICE et des IME. Ceci est lié au fait que ces éléments sont à la fois mal connus et très divers, rendant ainsi leur détection difficile. De plus, l'absence d'une base de données de référence d'ICE et d'IME, fiable et à jour, rend le développement d'un outil de détection plus difficile pour les équipes non-expertes. En effet, même si ces équipes tentent de développer un outil permettant de détecter les ICE et les IME, elles n'auront aucun moyen de valider leurs résultats sans base de données de référence. Une tentative de création de base de données dédiée aux ICE (ICEBerg) a tout de même été réalisée par Bi et al. en 2012. Cette base de données regroupe des ICE de bactéries Gram+ et Gram- et comporte un grand nombre d'éléments. Cependant, elle n'a pas été mise à jour depuis Novembre 2012 et n'est ni exhaustive, ni fiable. En effet, la base de données ICEBerg n'inclut pas certains ICE décrits dans la littérature avant sa publication, comme les 10 ICE de *S. agalactiae* décrits par Brochet et al. en 2008. De nombreux sites d'intégration inclus dans ICEBerg sont faux ou manquants et ce même lorsqu'ils étaient décrits de façon exacte et précise dans les publications originelles. Notre vérification des limites des ICE de streptocoques inclus dans ICEBerg indique qu'environ la moitié des limites de ces ICE sont fausses. De la même manière, la classification proposée par cette base de données est incohérente voire franchement déroutante. On y trouve, par exemple, une

famille nommée Tn*916* qui comporte des éléments (tels qu'ICE*Lm1*) ayant des modules d'intégration non apparentés à celui de Tn916 et d'autres éléments (tels que l'ICE Tn*1549*) ayant des modules de conjugaison non apparentés à celui de Tn*916*. En conséquence, ICE*Lm1* et Tn*1549* sont classés par ICEBerg dans la même famille bien que ces élements n'aient aucun module apparenté. De même, des éléments décrits comme appartenant à la famille Tn*916* tel que CTn5 ou CTn2 ne portent aucun module apparenté à Tn*916*, mais uniquement des modules de conjugaison apparentés à ceux de Tn*1549*. Une autre erreur, encore plus problématique pour une base de référence d'ICE, est la présence de familles de prétendus « ICE » constituées exclusivement d'éléments non apparentés aux ICE. C'est le cas de la famille Tn*1207.3* composée uniquement de prophages. Ainsi, le manque de fiabilité d'ICEBerg, du moins concernant les ICE de streptocoques et de firmicutes pour lesquels nous avons une expertise, exclut toute utilisation comme base de données de référence.

L'un des objectifs de ce travail était de combler le manque de connaissance concernant la prévalence et la diversité des ICE et des IME au sein des streptocoques. Pour cela, les 124 génomes de streptocoques disponibles dans la base de données publique du NCBI au début de l'étude ont été analysés grâce à une méthode mise au point lors de ce travail, la méthode ICEFinder. Puis, à l'aide des connaissances acquises lors de cette analyse, la méthode devait être améliorée/affinée et automatisée le plus possible afin de pouvoir, à terme, fournir à la communauté scientifique un outil de détection d'ICE et d'IME (dédié dans un premier temps aux streptocoques).

Afin de réaliser ces objectifs, la méthode utilisée devait nous permettre de trouver le plus grand nombre possible d'éléments tout en évitant au maximum les faux positifs. Dans cet objectif, la base de données utilisée devait être pertinente. Afin d'évaluer la diversité des éléments identifiés et la diversité de leur site d'intégration, ces derniers devaient être délimités précisément et classés en famille. De plus, pour caractériser au mieux ces éléments mal connus, ils ont été comparés entre eux afin de souligner leurs points communs et différences, et leur contenu en gènes a été analysé.

2. Les éléments détectés sont-ils fonctionnels?

A la suite de l'analyse des 124 génomes, 105 ICE, 26 dICE (éléments légèrement dégénérés dérivant d'ICE) et 144 IME putatifs ont été identifiés. Un point important est de bien définir ce que nous avons considérés comme ICE, dICE et IME au cours de ces travaux.

Les éléments délimités et codant les 4 protéines étiquettes complètes (intégrase, relaxase, CP et VirB4) ont été considérés comme ICE. Ont été considérés comme dérivés d'ICE (dICE) les éléments bien délimités et codant au moins 2 protéines étiquettes complètes et des pseudogènes des 2 autres, ainsi que les éléments codant3 protéines étiquettes complètes mais ayant apparemment perdu l'une de leurs extrémités. Cette définition s'appuie notamment sur le fait que ces 4 protéines sont retrouvées dans tous les ICE se transférant sous forme simple brin décrits dans la littérature et que ce sont ces 4 protéines qui sont recherchées par la méthode ICEFinder.

L'un des défauts de cette définition est qu'elle surestime très probablement la prévalence des éléments considérés comme « complets » ou potentiellement « fonctionnels », ici dénommés ICE. Cependant, une comparaison réalisée sur les gènes conservés du module de conjugaison de chacune des familles suggère que le nombre d'éléments concernés par cette surestimation serait faible. Il faut malgré tout prendre en considération le fait que même si ces gènes conservés sont présents, ils peuvent coder des protéines non fonctionnelles du fait de mutations. Un autre défaut de cette définition est qu'elle sous-estime grandement la prévalence des éléments dégradés, dérivant des ICE, ici dénommés dICE. Bien que ces éléments n'aient pas été pris en compte au cours de cette étude, il est important de noter que, en dehors des 26 éléments considérés comme dICE, plus d'une quarantaine d'éléments suffisamment peu dégénérés pour ne laisser aucun doute sur le fait qu'ils dérivent d'ICE n'ont pas été comptabilisés comme dICE du fait des critères choisis. Il faut également être conscient que ces éléments peu dégénérés, qu'ils soient considérés comme dICE ou non, pourraient avoir conservé une certaine capacité à se transférer, à être mobilisés par d'autres éléments, ou à mobiliser en *trans* d'autres éléments comme c'est le cas d'éléments dégénérés d'autres bactéries (Rice and Carias, 1998; Baker et al., 2008; Haskett et al., 2016).

Les éléments délimités par des répétitions directes et contenant une intégrase et une relaxase (phylogénétiquement éloignée de celles retrouvées chez les ICE) et éventuellement

une protéine de couplage éloignée de celles des ICE sont considérés comme IME. Comme pour les ICE, cette définition a pour défaut qu'elle surestime probablement la prévalence des éléments considérés comme fonctionnels. Cependant du fait de leur petite taille et du faible nombre de gènes que les IME codent, il est peu vraisemblable que des gènes nécessaires à leur transfert n'aient pas été identifiés. Néanmoins, la définition choisie pour les IME inclut aussi les IME portant un pseudogène de protéine de couplage, éléments qui auraient pu être considérés comme « dIME ». Cependant, les analyses phylogénétiques et les comparaisons entre IME indiquent que des IME ayant des relaxases proches peuvent porter, ou non, des protéines de couplage et que les acquisitions et pertes de CP au sein des IME sont très fréquentes. Ici, nous avons envisagé que la relaxase de l'IME pouvait interagir avec le T4SS de l'élément « helper » soit *via* la CP codée par l'IME (lorsque celui en codait une), soit directement *via* la protéine de couplage de l'élément « helper ». Dans cette hypothèse, les CP codées pas les IME ne seraient alors pas indispensables à leur mobilisation mais augmenteraient le spectre d'éléments mobilisateurs avec lesquels ils pourraient interagir ou augmenterait l'efficacité de leur transfert. Par ailleurs, contrairement à la définition utilisée pour les ICE, celle des IME pourrait sous-estimer grandement leur prévalence. En effet, cette définition ne prend pas en compte les IME dépourvus de relaxase. Pourtant, un IME ne codant qu'une intégrase et une *oriT*, MTn*Sag1*, a déjà été identifié chez *S. agalactiae* (Achard et al., 2007). Ces éléments trop difficiles à détecter ont été volontairement ignorés lors de ce travail.

Afin d'améliorer la finesse de notre méthode de détection ainsi que notre capacité à distinguer les éléments potentiellement fonctionnels des éléments dégradés, l'inclusion de nouvelles protéines signatures à notre base de données pourrait être envisagée. La méthode utilisée au cours de ces travaux repose sur une base de données experte, créée et mise à jour au laboratoire DynAMic. Cette base de données contient à ce jour 4 types de protéines. Les intégrases, les relaxases, les protéines de couplage et les protéines VirB4. Ces protéines ont été choisies selon différents critères. Les intégrases ont été incorporées à la base de données parce qu'elles jouent un rôle essentiel dans le cycle de vie des ICE comme des IME et assurent la fonction d'intégration des éléments dans un réplicon. Cette fonction est indispensable pour qu'un élément soit qualifié d'élément intégratif. La relaxase a été choisie comme protéine signature car cette protéine joue un rôle clé dans le transfert par

conjugaison des éléments. De plus, au début de cette étude, elle était considérée comme une des seules protéines impliquées dans le transfert par conjugaison retrouvée à la fois chez les ICE et chez les IME. La protéine de couplage et la protéine VirB4 sont les deux ATPase du T4SS. Elles ont été sélectionnées, entre autres, parce qu'elles sont relativement bien conservées au sein du module de conjugaison et qu'elles sont relativement faciles à détecter dans les génomes. Bien que nous ayons jugé cet ensemble de protéines suffisant pour détecter les ICE et les IME dans les génomes étudiés, l'inclusion d'autres protéines pourrait être envisagée et améliorerait la finesse de notre analyse. Par exemple, lors de sa recherche d'ICE dans les génomes d'actinomycètes, l'équipe de Ghinet et al. a utilisé comme protéines signature, non seulement des intégrases et des protéines impliquées dans le transfert par conjugaison (telle que la protéine de couplage VirD4 ou la protéine Tra/TraB (seule protéine requise pour le transfert par conjugaison des AICE (ICE d'actinomycètes), mais également des protéines impliquées dans la réplication sous forme excisée de l'élément (telles que des protéines initiatrices de réplication par cercle roulant ou réplication thêta (protéines Rep et RepA_N respectivement). Bien que l'utilisation de protéines impliquées dans la réplication semble être efficace pour la détection d'AICE (les 144 AICE identifiés dans l'étude de Ghinet et al. possédent tous, au moins, une protéine impliquée dans la réplication de l'élément), seulement certaines familles d'ICE et d'IME de firmicutes codent des protéines impliquées uniquement dans la réplication sous forme excisée. Ainsi, les éléments de famille Tn*GBS2* semblent tous coder une protéine de la superfamille RepA_N, qui est toujours située à l'extrémité opposée à celle de l'intégrase et est requise pour le maintien de l'élément sous forme circulaire (Guérillot et al., 2013). L'utilisation de telles protéines serait donc utile pour discriminer certaines familles ainsi que pour déterminer si un élément est potentiellement fonctionnel, mais n'augmenterait probablement pas la sensibilité de détection des ICE et des IME de firmicutes. De plus, ce type de protéines est retrouvé aussi dans d'autres types d'éléments génétiques mobiles tels que les prophages et pourrait ainsi augmenter notre taux de faux positifs. Cependant d'autres protéines pourraient être ajoutées à la base de données afin d'augmenter notre sensibilité de détection. Ainsi, la recherche des protéines de la partie interne du T4SS présente aussi bien dans les MPF FA et FATA des firmicutes (Guglielmini et al., 2013), pourrait nous permettre d'améliorer notre capacité à déterminer si un ICE est potentiellement fonctionnel ou non.

La prévalence des éléments détectés ainsi que celle des éléments considérés comme fonctionnels sont, sans doute, aussi influencées par la qualité de l'annotation des génomes analysés. Au cours de ces travaux, les 124 génomes utilisés proviennent de la base de données du NCBI. Cette base de données à l'avantage d'être facile d'accès et de contenir un grand nombre de génomes complets. De plus, tous ces génomes doivent atteindre un certain standard d'annotation pour être admis (https://www.ncbi.nlm.nih.gov/books/NBK174280/). Toutefois, comme ces génomes ne sont pas tous annotés avec les mêmes programmes, la qualité de leur annotation est hétérogène, ce qui rend leur analyse plus complexe. Ainsi en fonction de la qualité de l'annotation, un même gène pourra être annoté comme gène fonctionnel, pseudogène ou ne pas être annoté du tout (par exemple, certains programmes d'annotation n'annotent pas les CDS lorsque les gènes contiennent des introns). Ceci aura pour conséquence qu'un même élément peut être considéré comme ICE ou dICE, ou voire même dans certains cas ne pas être pris en compte comme élément (ICE/dICE/IME) par nos critères. Idéalement, dans l'optique d'une analyse automatique des génomes, une ré-annotation systématique et rationnelle de ces derniers serait souhaitable.

3. ICE et IME : des éléments très répandus et très divers

3.1. Prévalence des ICE et des IME au sein des streptocoques

Au cours de ces travaux, 124 génomes provenant de 27 espèces différentes de streptocoques ont été étudiés. Nous sommes parvenus à identifier des ICE/dICE dans 22 espèces ce qui souligne l'omniprésence de ces éléments au sein des génomes de streptocoques. De plus, les 5 espèces dépourvus d'ICE, *S. salivarius*, *S. iniae*, *S. gordonii*, *S. uberis* et *S. sanguinis* n'étaient représentées que par 3 génomes dans les cas de *S. salivarius* et d'un seul pour les 4 autres. Ainsi, nous pouvons estimer qu'un plus grand nombre de représentants de ces espèces nous auraient permis d'identifier des ICE/dICE en leur sein. Une étude récente a d'ailleurs décrit la présence de 13 ICE de la famille ICE*St3* au sein de génomes de *S. salivarius* (Dahmane et al., 2017).

Au cours de ces travaux, la recherche exhaustive d'IME nous a permis d'identifier 144 IME répartis dans plus de la moitié des génomes étudiés (78/124) et souligne l'omniprésence des IME au sein des génomes de streptocoques. De plus, ceci tend à confirmer l'hypothèse émise

par Guglielmini et al. en 2012, suggérant que les IME sont encore plus répandus que les ICE au sein des génomes bactériens. Cet écart entre ICE et IME est sans doute encore plus marqué que ces chiffres ne le laissent envisager. En effet, d'une part, la détection des IME ne reposant sur la détection que de 2 ou 3 protéines, le risque de non-détection d'un IME est plus grand que celui d'un ICE. D'autre part, les critères que nous avons établis pour qu'un élément soit considéré comme ICE/dICE ou IME tend à sous-estimer davantage la prévalence des IME que celle des ICE. Ainsi, une vingtaine d'éléments détectés lors de cette étude qui auraient pu être appelés « dIME » n'ont pas été pris en considération car ils possédaient soit un gène d'intégrase associé à un pseudogène de relaxase ou un gène de relaxase associé à un pseudogène d'intégrase. De plus, la méthode utilisée élimine d'emblée la plupart des pseudogènes suggérant que beaucoup pourraient avoir échappé à l'analyse.

3.2. Distribution des ICE et des IME dans les espèces de streptocoques

Les espèces qui présentent le plus d'ICE semblent être *S. suis* et *S. pneumoniae,* pour lesquelles 61% (8/13) et 40% (11/28) des génomes analysés contiennent au moins un ICE ou dICE. Cependant, il ne peut pas être exclu que cette forte prévalence en ICE dans certaines espèces soit due à un biais d'échantillonnage dans les génomes analysés. En effet, les ICE identifiés au sein des génomes de ces espèces sont souvent étroitement apparentés entre eux. De la même manière, la prévalence des IME détectés au cours de notre analyse varie en fonction des espèces. Ainsi, sur les 20 souches de *S. pyogenes* analysées, seulement 2 contiennent 1 IME tandis que les 3 souches de *S. anginosus* contiennent chacune entre 4 et 6 IME. De plus, les espèces contenant un faible nombre d'IME, telle *S. pyogenes* (avec une moyenne de 0,1 IME par génome), contiennent également un faible nombre d'ICE (avec une moyenne de 0,1 ICE par génome). Inversement les espèces contenant de nombreux IME, telle *S. anginosus* (avec une moyenne de 4,7 IME par génome), comportent aussi de nombreux ICE (avec une moyenne de 4,3 ICE par génome). Pour les ICE comme les IME, nous ne pouvons pas exclure que cette différence de distribution au sein des différentes espèces de streptocoques soit due à un biais d'échantillonnage. En effet, il faut garder à l'esprit que les génomes présents dans la base de données sont essentiellement des génomes d'intérêt pour l'homme (pathogène de l'homme ou d'animaux utiles à l'homme) et que de nombreuses souches proviennent de la même source ou de la même étude (isolats cliniques pour *S. pneumoniae*, *S. anginosus* et *S. pyogenes* ou d'élevages porcins pour *S. suis*). Il aurait

été préférable, pour tirer de meilleures conclusions, de partir d'un set de génomes provenant de sources diversifiées et représentatif des bactéries dans leur environnement.

Néanmoins, il est également possible que la faible prévalence en éléments se transférant par conjugaison (ICE et IME confondus) constatée dans certaines espèces soit due à une différence dans le mode de vie de ces organismes, favorisant l'évolution par transfert d'ICE et IME dans le cas de *S. anginosus* ou au contraire la défavorisant dans le cas *S. pyogenes*.

3.3. Diversité au sein des différentes familles d'ICE

Un classement des ICE basé sur l'analyse de leurs modules de conjugaison conduit à les cataloguer en 7 familles appartenant à 3 superfamilles. La superfamille Tn*916* comporte les familles *Tn916* et ICE*St3*, la superfamille Tn*5252* comporte les familles *vanG*, Tn*5252*, Tn*1549*, et Tn*GBS2* et enfin la superfamille Tn*GBS1* comporte uniquement, pour le moment, la famille Tn*GBS1*. Par ailleurs, les comparaisons et analyses phylogénétiques révèlent une évolution modulaire caractérisée, en particulier, par de nombreux échanges de modules d'intégration entre ICE. Ainsi les ICE de la famille Tn*5252* peuvent porter un module de conjugaison de la famille Tn*5252* en association avec un module d'intégration codant soit une intégrase à tyrosine, une intégrase à sérine ou encore un triplet d'intégrases à sérine.

Les analyses phylogénétiques effectuées à partir des gènes de conjugaison, a permis non seulement de classer les ICE en famille, mais aussi d'estimer la diversité au sein de ces familles. Il s'avère que les modules de conjugaison de certaines familles sont plus divers que d'autres. Ainsi au sein de la superfamille Tn*916*, les éléments de la famille ICE*St3* présentent une forte diversité de leur module de conjugaison tandis que les éléments de la famille Tn*916* ne présente que peu, voire aucune, diversité dans ce module. De la même manière, au sein de la superfamille Tn*5252*, les éléments de la famille Tn*GBS2* présentent, certes une plus faible diversité que les éléments de la famille ICE*St3*, mais ils présentent une plus grande diversité dans leur module de conjugaison que les éléments des familles Tn*1549* et Tn*5252*. De plus, non seulement les éléments de la famille ICE*St3* présentent une plus grande diversité au sein de leur module de conjugaison, mais ces éléments présentent aussi une plus grande diversité de sites d'intégration que les ICE de la famille Tn*5252* ou Tn*1549* (11 sites d'intégration différents contre 5 et 4 respectivement). Cette différence pourrait s'expliquer par une invasion récente du genre streptocoques par des éléments appartenant

à certaines familles. Cette invasion récente aurait pu être facilitée par une forte pression de sélection exercée par l'homme et l'usage intensif d'antibiotiques. Ce constat est particulièrement marquant pour les éléments de la famille Tn*916* codant tous la résistance à la tétracycline (Roberts and Mullany, 2011) et pour lesquels on retrouve une très faible diversité. Dans une moindre mesure, cette hypothèse peut également être envisagée pour les éléments des familles *Tn5252* et Tn*1549* connus aussi pour portés des gènes conférant des résistances aux antibiotiques (Garnier et al., 2000; Korona-Glowniak et al., 2015).

3.4. Diversité des modules de mobilisation des IME

La diversité des modules de mobilisation des IME est beaucoup plus grande que celles des modules de conjugaison d'ICE. Ainsi, si seulement 3 superfamilles de relaxases (MobT, MobP et MobL) sont représentées chez les ICE, les IME ne comptent pas de moins de 9 familles de relaxases putatives.

3.4.1. Diversité des relaxases

Seulement 18% des IME identifiés (26/144) codent une relaxase dite « canonique ». Ces relaxases appartiennent aux superfamilles MOB_Q/PF03389 (12/25), MOB_C/PF13814 (2/25), MOB_V/PF01076 (11/25) ou MOB_P/PF03432 (1/25). En dehors des 2 IME codant les relaxases de la famille MOB_C et également des protéines de couplage canoniques (VirD4), aucun de ces IME ne code de protéine de couplage. Cette association MOB_C/VirD4 a déjà été observée dans des plasmides conjugatifs et mobilisables de protéobactéries γ tels que le plasmide CloDF13 (Garcillán-Barcia et al., 2009).

Une large majorité des IME putatifs identifiés (118 sur 144) codent des protéines apparentées à des protéines initiatrices de réplication par cercle roulant impliquées dans le maintien par réplication de plasmides de firmicutes, ou dans la réplication de génomes viraux. Ces protéines portent les domaines PF02486/MobT, PF01719, PHA00330, PF01719-PF00910, and PF02407. Bien que les relaxases des ICE de la superfamille ICE*Bs1*/Tn*916*/ICE*St3* appartiennent à la famille MOB_T et que divers IME putatifs de *S. agalactiae* possédent des relaxases de cette famille (Bellanger et al., 2014; Douard et al., 2015), les 4 autres familles ne sont apparentées à aucune famille connue de relaxases. Cependant, des études récentes ont démontré que la même protéine peut à la fois être responsable de l'initiation de la réplication par cercle roulant nécessaire au maintien de

l'élément et être la relaxase impliquée dans le transfert par conjugaison. Ainsi, les relaxases appartenant à la famille MOB_T d'ICE*Bs1* de *B. subtilis* et Tn*916* sont impliquées non seulement dans le transfert par conjugaison mais aussi dans l'initiation de la réplication par cercle roulant de la forme excisée de l'élément (Lee et al., 2010; Wright and Grossman, 2016). De même, des protéines initiatrices de la réplication par cercle roulant de plasmides de firmicutes, portant un domaine Rep1/PF01446, sont impliquées non seulement dans le maintien par réplication des éléments mais aussi dans leur mobilisation par ICE*Bs1* en tant que relaxases (Lee et al., 2012). De plus, il semblerait que les relaxases canoniques des ICE R391 (MOB_H) et ICE*Ml*Sym^{R7A} (MOB_F) puissent également être impliquées dans le maintien de la forme excisée de l'ICE (Carraro and Burrus, 2015; Ramsay et al., 2006). L'ensemble de ces données suggère que la distinction classique entre protéines initiatrices de la réplication par cercle roulant et relaxases pourrait ne pas être pertinente.

Par ailleurs, plus de la moitié des IME codant une protéine apparentée à des protéines initiatrices de la réplication par cercle roulant codent aussi une protéine de couplage. Cette CP est toujours apparentée à TcpA et non à VirD4. Cette association RCR/TcpA est aussi retrouvée au sein des ICE de la superfamille Tn*916* (contenant les ICE ICE*Bs1*, ICE*St3*, Tn*916*) qui codent des relaxases de la famille MOB_T (Guglielmini et al., 2013).

Globalement, l'ensemble de ces données suggèrent que ces protéines apparentées à des protéines initiatrices de réplication par cercle roulant (MobT, PF01719, PHA00330, PF01719-PF00910, and PF02407) ont soit une unique fonction de relaxase de mobilisation, soit la double fonction d'initiation de la réplication de la forme excisée de l'IME nécessaire à son maintien et de relaxase de mobilisation.

3.4.2. Diversité intra- et inter-modulaire

Comme pour les ICE, nous avons tenté de classer les IME en différentes familles en se basant sur des gènes du transfert conjugatif. Cependant, tandis que les associations entre les gènes du module de conjugaison des ICE (relaxase, CP, VirB4) sont congruentes dans la quasi-totalité des cas et nous a donc permis de proposer des familles sur cette base, les associations entre les gènes du module de mobilisation des IME (relaxase/CP) révèlent quant à elles de nombreux échanges au sein des IME. De ce fait, il nous a été impossible de proposer des familles pour les IME basées sur leur module de mobilisation. De plus, les

comparaisons et analyses phylogénétiques révèlent également de nombreux échanges de modules d'intégration entre IME. Ainsi dans certains cas, des relaxases pratiquement identiques (comme celle du Cluster PF02486_6, voir figure supplémentaire de l'article 2) sont associées à des protéines de couplage de familles différentes (moins de 40% d'identités en protéine), ces mêmes familles de protéines de couplage pouvant être retrouvées associées avec d'autres familles de relaxases.

Bien que nous ne soyons pas parvenus à classer les IME en familles sur la base de leur module de mobilisation, nous avons regroupé les IME en fonction de la superfamille de relaxase qu'ils codent. Une structure commune aux IME codant une relaxase apparentée aux protéines initiatrice de réplication par cercle roulant est alors ressortie (figure 4 de l'article 2) : i) tous ces éléments codent une intégrase à tyrosine, ii) tous partagent une structure compacte dans laquelle les gènes codant pour l'intégrase, l'excisionase, la relaxase et la protéine de couplage (TcpA) lorsqu'elle est présente, sont regroupés dans le même ordre à l'une des extrémités de l'élément. De plus, tous les échanges, entre module de mobilisation d'IME, et la plupart des échanges de modules d'intégration, ont été constatés entre les IME partageant cette même structure compacte, suggérant que cette structure favorise les échanges entre les IME la partageant.

3.5. Prévalence et diversité des ICE et des IME : questions non résolues

Lors de cette analyse, les ICE et les IME présents sur les plasmides des souches étudiées n'ont volontairement pas été recherchés pour éviter la confusion entre ICE portés par un plasmide et plasmides conjugatifs ainsi qu'entre IME portés par un plasmide et plasmides mobilisables. En effet, les systèmes de conjugaison ou mobilisation des plasmides conjugatifs et mobilisables étant apparentés à celui des ICE et des IME, notre méthode de détection aurait immanquablement détecté les gènes impliqués dans le transfert de ces plasmides, ce qui aurait considérablement compliqué l'analyse. Dans cette perspective, il doit être souligné que certains ICE codant des intégrases ayant une faible spécificité tels que Tn*916* ou Tn*1549* sont fréquemment intégrés dans des plasmides (Clewell and Gawron-Burke, 1986; Garnier et al., 2000). De même, l'IME Tn*4451* de *C. perfringens* codant une intégrase à sérine de faible spécificité, a initialement été identifié comme un transposon porté par un plasmide conjugatif (Abraham and Rood, 1987; Crellin and Rood, 1998). Par ailleurs, de la même

manière que certains ICE et IME s'intègrent spécifiquement dans des gènes conservés portés par des ICE (par exemple les IME intégrés dans le gène *traG* codant la protéine de couplage des ICE de la superfamille Tn*5252*), il est possible que des ICE ou des IME puissent coder des intégrases catalysant des intégrations ciblant spécifiquement les gènes de conjugaison des plasmides conjugatifs. Pour toutes ces raisons, il est probable que la prévalence des ICE et IME présentant une faible spécificité d'intégration ainsi que celle des IME utilisant des gènes de conjugaison comme cible d'intégration aient été quelque peu sous-estimées.

Par ailleurs, en dehors des IME et les ICE, nous avons détecté un grand nombre d'éléments flanqués de DR et codant une intégrase, mais ne codant, ni protéine de couplage, ni relaxase canonique, ni protéine apparentée à des initiateurs de réplications en cercle roulant. Il est possible qu'au moins une partie de ces éléments correspondent à des éléments dérivant d'IME. Ils pourraient aussi correspondre soit à des IME ne codant pas de relaxase, soit à des IME codant des relaxases encore inconnues, ou non recherchées au cours de notre étude. Celles-ci pourraient par exemple appartenir à la nouvelle superfamille de relaxases récemment découverte chez le plasmide conjugatif pCW3 de *C. perfringens* et les plasmides apparentés (Wisniewski et al., 2016). Par ailleurs, aucun IME sans relaxase et ayant une *oriT* apparentée à celle de leurs éléments mobilisateurs n'est actuellement identifié chez les streptocoques et plus généralement chez les firmicutes. Néanmoins, ces éléments semblent être fréquents chez les protéobactéries (Carraro et al., 2016; Daccord et al., 2013; Marcoleta et al., 2016; Mulvey et al., 2006). De plus, comme de nombreux plasmides mobilisables ne codant pas de relaxases et possédant des *oriT* apparentées à celles de leurs plasmides helpers ont été décrits récemment chez les staphylocoques (F. G. O'Brien et al., 2015; Pollet et al., 2016), il semble possible que des IME présentant des caractéristiques similaires puissent aussi exister chez les streptocoques. Des IME de protéobactéries ne codant pas leur propre relaxase, codent cependant leur propre VirB6 (Carraro et al., 2017a), protéine de la partie interne du T4SS qui est aussi présente dans les MPF FA et FATA des firmicutes (Guglielmini et al., 2013). Ceci pourrait être mis à profit en ajoutant les protéines VirB6 des MPF FA et FATA dans la base de données utilisée par ICEfinder pour nous permettre de détecter la présence éventuelle d'IME indétectables jusqu'alors par l'approche utilisée.

Enfin, la méthode utilisée au cours de cette étude prend appui sur les connaissances des ICE et IME identifiés de firmicutes, principalement de streptocoques. Il est tout à fait possible

d'imaginer que des ICE et des IME puissent porter des modules de conjugaison ou mobilisation totalement différents de ceux déjà caractérisés. Ainsi, le plasmide conjugatif pXO16 du firmicute *Bacillus thuringiensis* ne code aucune protéine apparentée à des relaxases ou protéines de T4SS connues (Makart et al., 2015). Notre méthode de détection serait bien sûr incapable de détecter un ICE qui se transférerait par un mécanisme similaire à celui de ce plasmide.

Globalement, il est fort probable que nous ayons sous-estimé la prévalence et/ou la diversité des ICE et des IME au sein des génomes étudiés ainsi que la prévalence des éléments en dérivant.

4. Délimitation des éléments

Au cours de ces travaux, une attention particulière a été portée à la délimitation des ICE et des IME. Notre méthode de délimitation repose principalement sur la détection du site d'intégration de l'élément et sur la recherche des répétitions directes le bornant. Bien que cette approche permette de délimiter avec précision et certitude les éléments détectés, elle comporte 2 défauts majeurs. D'une part, elle repose sur une solide connaissance des éléments et de leurs sites d'intégrations potentiels. Ceci signifie que l'exploration de génomes encore jamais analysés pour leur contenu en ICE et IME exigera une longue période d'apprentissage durant laquelle des informations sur les sites d'intégration des éléments devront être obtenues. D'autre part, bien que cette méthode détecte efficacement les DR longues (>10-12 b), elle devient inefficace lorsque que l'intégration de l'élément ne génère pas de DR ou que celles-ci sont trop courtes et/ou trop dégénérées pour être détectées. Ceci est le cas, par exemple pour les éléments codant une intégrase de type Tn*916* ou pour de nombreux IME ou ICE codant des intégrases à sérine, car elles génèrent généralement des répétitions directes de 2 à 4 pb non détectables. Dans ces cas, la délimitation des éléments s'est faite par génomique comparative/synthénie. Cependant, cette approche présente aussi des limites. En effet, elle nécessite tout d'abord d'avoir à disposition un ou des génome(s) suffisamment proche(s) du génome analysé pour permettre la comparaison. De plus, ceux-ci doivent être dépourvus d'élément apparenté (même très dégénéré comme un CIME) intégré au même site pour que la délimitation de l'élément soit possible. Or, lors de notre étude, pratiquement la moitié des espèces étudiées (13/27 : *S. iniae, S. infantarius, S. mitis, S.*

lutetientis, S. gordonii, S. parasanguinis, S. pasteurianus, S. oligofermentans, S. oralis, S. macedonicus, S. uberis) n'étaient représentées que par 1 seul génome disponible dans les bases de données du NCBI. Une autre limite de cette approche est que, bien que celle-ci permette de délimiter efficacement les ilots génomiques, elle ne permet pas de délimiter spécifiquement les ICE et les IME. En effet, dans les cas où l'ilot génomique est composé d'un seul ICE ou IME, les limites de l'ilot correspondent aux limites de l'élément, cependant lorsque l'ilot est composé de plusieurs éléments en accrétion ou d'élément imbriqués les uns dans les autres, délimiter l'ilot ne permet pas de délimiter tous les éléments.

Pour résumer, la délimitation des ICE et des IME combine génomique comparative et recherche de DR et lorsque l'approche de génomique comparative est employée, elle nécessite d'avoir à disposition plusieurs génomes de la même espèce.

5. Spécificité d'intégration et impacts sur le « fitness » de l'hôte

5.1. Spécificité d'intégration des ICE et impacts sur le fitness

En dehors de l'étude portant sur la prévalence des ICE d'actinobactéries (essentiellement des AICE se transférant sous forme double brin) (Ghinet et al., 2011), nos résultats constituent de très loin le plus important set de données concernant les modules d'intégration et les spécificités d'insertions d'ICE. Parmi les 131 ICE/dICE identifiés, 75 portent des modules d'intégration codant une intégrase à tyrosine, 20 une intégrase à sérine, 9 un triplet d'intégrase à sérine et 23 une transposase à DDE. Bien que lors de leur analyse, Ghinet et al. n'aient pas détecté d'ICE codant de transposase à DDE, les auteurs avaient eux aussi détecté un plus grand nombre d'ICE codant une intégrase à tyrosine qu'à sérine (128 codant une intégrase à tyrosine, contre 19 codant une intégrase à sérine). Au total, les modules d'intégration des 131 ICE et dICEs de streptocoques détectés lors de ce travail présentent 18 spécificités d'intégration différentes. Parmi ces spécificités d'intégration, on retrouve principalement des gènes codant des protéines ribosomiques, des gènes codant des protéines de ménage et à moindre mesure des ARN de transfert (9/131). Ceci contraste avec les résultats obtenus par Ghinet et al., qui montrent qu'environ les 2/3 des AICE identifiés (100 des 144) sont intégrés dans l'extrémité 3' d'ARN de transfert. Une autre différence importante est que les ICE de streptocoques codant une intégrase à sérine

s'intègrent de façon spécifique dans des gènes codant des protéines et les interrompent, alors qu'une partie importante des AICE codant des intégrases à sérine d'actinobactéries s'intègre dans l'extrémité 3' des gènes d'ARNt sans les interrompre.

En dehors des contraintes mécanistiques comme la présence de séquences palindromiques et de séquences répétées dans le site d'intégration visé (Williams, 2002), la sélection devrait favoriser les ICE dont l'insertion a un d'impact minimal sur le fitness de l'hôte. De ce point de vue, de nombreux ICE de streptocoques codant une intégrase à tyrosine s'intègrent de façon spécifique dans l'extrémité 3' de gènes conservés, isolés ou situés à la fin d'un opéron. Lorsque c'est le cas, ces insertions ne modifient quasiment jamais le produit du gène et ne modifieraient probablement pas son expression. Par ailleurs, des ICE capables de s'intégrer dans un large panel d'espèces sans gêner leur hôte devraient être favorisés. Dans cette perspective, les sites visés par les modules d'intégration d'ICE sont généralement bien conservés au sein de nombreuses espèces de streptocoques (ARNt, protéines ribosomiques), leur permettant ainsi de s'intégrer dans un large panel d'espèces, tout en ayant peu ou pas d'impact sur le fitness de l'hôte.

Seulement, six ICE ont été retrouvés dans des sites d'insertions dit « secondaires ». Les insertions dans les sites secondaires semblent se réaliser naturellement à plus petite fréquence que les insertions dans le site « primaire » et seraient plus fréquemment observés lorsque celui-ci est occupé ou absent (Menard and Grossman, 2013). De plus, contrairement aux insertions dans les sites « primaires », celles se produisant dans les sites « secondaires » semblent être défavorables à la fois pour l'hôte et pour l'élément. En effet, l'insertion d'ICE*Bs1* dans un site « secondaire » diminue à la fois le taux d'excision de l'élément et le taux de prolifération de la cellule hôte (Menard and Grossman, 2013). De la même manière le taux d'excision de l'ICE SXT depuis un site « secondaire » est réduit de 3-4 fois par rapport à l'excision depuis un site « primaire » (Burrus and Waldor, 2003). Il semblerait donc y avoir une pression de sélection contre les intégrations des éléments dans des sites secondaires.

Tous les éléments appartenant à la famille Tn*916* identifiés dans cette étude (excepté un) codent une intégrase identique ou quasi identique à l'intégrase du Tn*916* originel. Cette intégrase catalyse une intégration préférentielle dans les régions riches en A-T (Hosking et al., 1998; Cookson et al., 2011). Les éléments codant ce type d'intégrase n'ont donc pas de

site d'intégration « primaire » et peuvent être retrouvés intégrés dans de nombreux sites de séquences différentes. Cependant l'analyse des sites d'insertion après le transfert par conjugaison de Tn*916* vers *Butyrivibrio proteoclasticus* B316^T a montré que seulement un tiers des 123 insertions analysées interrompaient une ORF (Cookson et al., 2011). Ceci est probablement lié au fait que les régions inter-géniques ont un pourcentage en G+C plus faible que le reste du génome, critère favorisant l'intégration de Tn*916* dans ces régions, quelque soit l'endroit précis. Ainsi, dans la majorité des cas, l'insertion de Tn*916* aurait un impact faible sur le fitness de l'hôte. Les éléments mobiles présentent aussi, le plus généralement, un faible pourcentage en G+C, ce qui pourrait expliquer l'insertion fréquente des Tn*916* dans les éléments de la superfamille Tn*5252* ou dans des plasmides. Dans ce cadre on peut imaginer que l'impact de l'insertion d'un Tn*916* sur le fitness de l'hôte soit faible. Par ailleurs un ICE de type Tn*916* porté par ICE de type Tn*5252* peut se transférer seul (Santoro et al., 2010) ou en tant que partie de Tn*5252* (Ayoubi et al., 1991). Ainsi, la préférence pour les régions riches en A+T pourrait favoriser la dissémination de Tn*916* en lui conférant une seconde façon de se transférer (mobilisation en *cis* par les éléments conjugatifs qui le portent).

Au cours de cette étude, tous les éléments appartenant aux familles Tn*GBS1* et Tn*GBS2* codant une transposase à DDE sont intégrés dans des régions inter-géniques situées 15 ou 16 pb en amont de la boite -35 de promoteur sigma A. Ce résultat concorde avec ce qui est décrit pour les ICE Tn*GBS1* et Tn*GBS2* (Brochet et al., 2009; Guérillot et al., 2013). Bien que ces insertions soient inter-géniques, le site ciblé se trouve proche d'un promoteur. Même s'il est possible qu'elle ait un faible effet sur le fitness de l'hôte, on ne peut exclure que l'insertion d'un tel ICE puisse interférer avec l'expression du gène en aval de ce promoteur. Cependant, l'insertion de Tn*GBS2* ne semble pas affecter l'expression du gène en aval lorsque celui-ci s'intègre dans son site d'intégration préférentiel (Guérillot et al., 2013, 2014).

Tous les ICE codant des intégrases à serine sont intégrés de manière site-spécifique, non pas à l'extrémité de gènes, mais à l'intérieur provoquant ainsi l'interruption de la séquence codante. Bien que ceci puisse avoir un effet négatif sur le fitness de l'hôte, l'ensemble des gènes ciblés par ces intégrases à sérine (*rumA*, *hsdM*, *mutT*) sont certes conservés, mais ne

semblent pas être indispensables pour la bactérie. Le gène *rumA* code par exemple une ARNr méthyl-transférase. Chez *E. coli*, la délétion de ce gène n'a pas d'impact sur la croissance de la bactérie, mais modifie cependant la sensibilité du ribosome aux deux antibiotiques que sont l'acide fusidique et la capréomycine (Persaud et al., 2010).

L'impact que peut avoir l'intégration d'ICE codant des intégrases à sérine sur l'expression des gènes dans lesquels ils sont intégrés n'a jamais été étudié. Cependant, cet impact a été étudié pour certains prophages (Stragier et al., 1989; Kunkel et al., 1990; Rabinovich et al., 2012) et plus récemment pour un élément « non-prophage(like) » de *Bacillus cereus* ATCC10987 appelé *gin* (Abe et al., 2017). Bien que le mécanisme de transfert de cet élément soit inconnu, les auteurs émettent l'hypothèse qu'il serait capable de se transférer par conjugaison en présence du plasmide conjugatif pBc10987 présent naturellement dans la souche. Le gène *gerE* code pour un facteur de transcription exprimé au moment de la sporulation chez *Bacillus cereus.* Dans la souche *Bacillus cereus* ATCC10987, ce gène est interrompu par l'intégration site-spécifique d'un élément codant 3 intégrases à sérine appelé *gin.* Au moment de la sporulation, l'élément *gin*, normalement intégré, s'excise du chromosome permettant le réarrangement du gène *gerE* et ainsi l'expression d'une protéine fonctionnelle (Abe et al., 2017). Par analogie, on peut postuler que les ICE intégrés de manière site-spécifique à l'intérieur de gènes ait un comportement similaire ce qui limiterait l'impact de l'intégration sur le fitness de l'hôte.

5.2. Spécificité d'intégration des IME et intégration au sein des ICE

Les IME comme les ICE présentent de nombreuses spécificités d'intégration (18 spécificité pour les IME et au moins 17 pour les ICE). Cependant, aucun d'eux ne code une intégrase appartenant à la superfamille des transposases à DDE. L'absence de détection de telles transposases au sein des IME pourrait résulter, en partie, de la faible diversité des transposases à DDE de notre base de données. Il est ainsi possible que des IME codant à la fois une transposase à DDE et une relaxase toutes deux éloignées de celles présentes dans notre base de données n'aient pas été détecté.

Comme nous l'avions constaté pour les ICE, des IME sont intégrés dans des ICE. Tous ces IME (12) codent une intégrase à sérine et s'intégrent de manière spécifique au sein des gènes du module de conjugaison des ICE la superfamille Tn*5252*. L'intégration de ces IME interrompt

les gènes dans lesquels ils sont intégrés, ce qui pourrait avoir pour conséquence de gêner, voire d'empêcher le transfert des ICE concernés. Bien que l'impact de l'intégration d'ICE ou d'IME codant une intégrase à sérine pour l'hôte (ou ici l'élément hôte) n'ait jamais été étudié, nous pouvons cependant postuler que l'excision de ces IME serait induite quand la conjugaison de l'ICE est elle-même induite. Après excision de l'IME, le gène cible serait alors de nouveau fonctionnel, ce qui permettrait ainsi le transfert par conjugaison de l'ICE. L'IME pourrait alors emprunter/pirater le pore de conjugaison de l'ICE et ainsi être mobilisé en *trans* par l'ICE hôte. Dans la réceptrice, l'IME pourrait s'intégrer soit dans l'ICE nouvellement acquis par conjugaison si le transfert de l'ICE a été couronné de succès, soit dans un ICE ou un dérivé d'ICE résident. L'IME utiliserait donc l'ICE non seulement comme site d'intégration mais aussi comme élément « helper ». Le transfert d'ICE*Sp2905* de *S. pyogenes* comportant 2 IME intégrés dans des gènes conservés localisés à proximité des gènes de conjugaison identifiés (un IME intégré dans *snf2* et un IME intégré dans *maff2)* a été déjà démontré (Giovanetti et al., 2012; Bellanger et al., 2014). Bien que nous ne puissions pas exclure le fait que ces gènes ne soient pas indispensables au transfert par conjugaison de l'ICE, il semble probable que ce transfert se passe en plusieurs étapes successives : i) l'excision des 2 IME et de l'ICE, ii) le transfert indépendant de l'ICE et des 2 IME, iii) l'intégration de l'ICE dans son site d'intégration et des 2 IME dans l'ICE.

PERSPECTIVES

Au début de cette étude (décembre 2013), les 124 génomes complets de streptocoques présents dans la base de données du NCBI ont été extraits afin d'être analysés par la méthode ICEFinder. Aujourd'hui, moins de 4 ans plus tard, on compte déjà plus du double de génomes de streptocoques disponibles en téléchargement (284 au mois d'octobre 2017). Tandis que notre capacité à séquencer les génomes s'est considérablement améliorée ces dernières années, notre capacité à les assembler, les analyser et à les annoter progresse quant à elle plus lentement. Il n'est pas rare d'ailleurs de trouver dans les bases de données du NCBI des génomes comportant essentiellement des protéines annotées « hypothetical protein ». La tendance actuelle qui consiste à priviligier la quantité de génomes au détriment de leur qualité d'analyse peut d'ailleurs devenir gênante à terme, les protéines annotées correctement se retrouvant noyées dans une masse de protéines « hypothétiques » ou de

fonction annotée erronée. La forte proportion de protéines hypothétiques est d'autant plus marquée dans les zones des génomes correspondants à des ilots génomiques ou éléments génétiques mobiles (ces régions étant généralement moins bien connues que le reste du génome). C'est pourquoi, afin d'inverser cette tendance, il est important de mettre au point des outils d'annotation des ilots génomiques et de démocratiser leur utilisation par la communauté scientifique. Pour que l'outil ICEFinder soit mis à disposition de la collectivité, deux étapes sont nécessaires. D'une part, il faudra élargir le spectre de détection des ICE et des IME à un groupe bactérien plus large que les streptocoques. Un premier objectif serait d'élargir la recherche aux éléments présents dans les génomes de bactéries lactiques et de progressivement s'intéresser à ceux de l'ensemble des firmicutes. D'autre part, pour que l'outil soit utilisable par le plus grand nombre de scientifiques, il est nécessaire de rendre l'outil accessible aux utilisateurs n'ayant pas forcément un « background » en informatique. Ce but pourrait être alors atteint au travers le développement d'une interface graphique ou d'un site web dédié à ICEFinder.

1. Automatisation de la méthode et élargissement de la recherche aux firmicutes

L'automatisation des étapes d'extraction des CDS, de recherche de protéines signatures, d'applications des filtres et de co-localisation des protéines a permis de réduire grandement le temps dédié à la recherche des ICE et des IME dans les génomes de streptocoques. Les résultats fournis par les scripts sont désormais triés et épurés. De plus, l'extraction des données depuis la base de données nous permet d'obtenir directement les informations relatives aux familles des gènes détectés. Cependant, malgré l'automatisation de ces étapes, la méthode ICEFinder nécessite toujours des interventions manuelles notamment au moment de la délimitation des éléments, ce qui rend son utilisation difficile pour des utilisateurs non experts. De plus, l'automatisation de certaines étapes comme l'enrichissement de la base de données est difficilement envisageable. Cette étape est pourtant primordiale car l'efficacité de la méthode repose entièrement sur la pertinence de la base de données de référence. En effet, l'incorporation de mauvaises protéines signatures engendrerait inévitablement la détection de faux positifs. Ainsi, la mise à jour de la base de données nécessite l'intervention d'un expert. De la même manière, la recherche de gènes manquants semble difficilement automatisable. En effet, chaque cas étant différent, prévoir

toutes les possibilités est difficile. De plus, la recherche d'éléments dans des genres ou espèces encore jamais étudiés devrait conduire à la détection d'éléments très différents de ceux déjà identifiés. Il serait alors très difficile de faire la différence entre : (i) un gène manquant mais nécessaire au maintien ou au transfert de l'élément, (ii) un gène présent mais non détecté comme gène intervenant dans la conjugaison, et (iii) un gène naturellement absent de l'élément, élément qui serait très différent des éléments de la base de données mais parfaitement fonctionnel. Il en est de même pour différencier une famille d'IME dont la relaxase n'a pas été détectée, d'une famille d'IME ne codant pas de relaxase. Seule l'intervention d'un expert des ICE et des IME permettra de trancher et d'établir de nouvelles règles en vue d'une possible automatisation.

2. Élargissement à d'autres génomes de firmicutes

Compte tenu des données acquises sur les ICE et les IME de streptocoques et de l'enrichissement de la base de données, la méthode ICEFinder est désormais adaptée à la recherche des ICE et des IME dans les génomes de streptocoques. Cependant, rien ne garantit que cette méthode de détection, basée sur BLAST, permette de détecter des éléments éloignés de ceux identifiés dans d'autres genres bactériens. C'est pourquoi, dans l'optique d'augmenter la sensibilité de détection, nous avons amélioré l'approche de détection des protéines signatures en remplaçant les analyses par Blast par des analyses utilisant des profils HMM des protéines signatures connues pour être plus sensibles (Söding, 2005).

Afin de tester l'efficacité de cette nouvelle méthode de détection, 8 génomes de *Lactobacillus casei* ont été analysés d'une part avec la méthode utilisant BLAST et d'autre part celle utilisant des profils HMM. Les deux méthodes ont détecté le même nombre d'éléments (6 ICE et 9 IME). Cependant, les protéines détectées par l'approche basées sur les profils HMM sont détectées par une meilleure e-value que lorsqu'elles l'étaient par l'approche par Blast où certaines protéines étaient à la limite de détection. Bien que dans ce test, les deux approches aient abouti au même nombre d'éléments détectés, l'utilisation de profils HMM rendrait donc notre détection plus sensible. De plus, *L. casei* étant phylogénétiquement assez proche des streptocoques (ordre des lactobacillales), il est possible que les ICE et les IME présent dans les génomes de *L. casei* soient suffisamment

proches phylogénétiquement des ICE et des IME de streptocoques pour permettre aisément leur détection par BLAST (même famille/ même superfamille), alors que des éléments provenant de bactéries plus éloignées ne le seraient pas.

Il est à noter que, lors de ce test sur les génomes de *L. casei*, 2 ICE appartenant à une nouvelle famille de la superfamille Tn*GBS1* ont été identifiés. Cette famille d'ICE code des relaxases appartenant à la famille MOB_L, une nouvelle famille de relaxase récemment décrite par Ramachandran et al. (2017) et retrouvée quasi-exclusivement au sein des firmicutes. Ce résultat est particulièrement encourageant car il indique que notre approche permet de détecter des éléments appartenant à de nouvelles familles. Cependant, ce résultat souligne également à quel point les ICE et les IME sont des éléments méconnus. En effet, si un simple test sur un petit lot de génomes proches des streptocoques détecte une nouvelle famille d'ICE, il est fort probable qu'une recherche sur l'ensemble des firmicutes nous permette de détecter de très nombreuses familles d'éléments encore jamais décrits. Ce dernier point, bien que motivant et excitant, rappelle la difficulté du challenge que représente la mise au point d'une méthode de détection adaptée à l'ensemble des firmicutes. Cette détection serait pourtant d'autant plus utile que les IME et ICE sont probablement les plus fréquents de tous les éléments transférables par conjugaison et qu'ils portent des gènes d'adaptation bactérienne dont certains, comme les gènes de résistance aux antibiotiques et de pathogénicité, sont susceptibles d'avoir un immense impact pour l'humanité.

BIBLIOGRAPHIE

Abajy, M.Y., Kopeć, J., Schiwon, K., Burzynski, M., Döring, M., Bohn, C., Grohmann, E., 2007. A type IV-secretion-like system is required for conjugative DNA transport of broad-host-range plasmid pIP501 in gram-positive bacteria. J. Bacteriol. 189, 2487–2496. doi:10.1128/JB.01491-06

Abe, K., Shimizu, S.-Y., Tsuda, S., Sato, T., 2017. A novel non prophage(-like) gene-intervening element within *gerE* that is reconstituted during sporulation in Bacillus cereus ATCC10987. Sci. Rep. 7, 11426. doi:10.1038/s41598-017-11796-8

Abraham, L.J., Rood, J.I., 1987. Identification of Tn4451 and Tn4452, chloramphenicol resistance transposons from *Clostridium perfringens*. J. Bacteriol. 169, 1579–1584.

Achard, A., Leclercq, R., 2007. Characterization of a small mobilizable transposon, MTn*Sag1*, in *Streptococcus agalactiae*. J. Bacteriol. 189, 4328–4331. doi:10.1128/JB.00213-07

Alvarez-Martinez, C.E., Christie, P.J., 2009. Biological diversity of prokaryotic type IV secretion systems. Microbiol. Mol. Biol. Rev. 73, 775–808. doi:10.1128/MMBR.00023-09

Anisimova, M., Bielawski, J., Dunn, K., Yang, Z., 2007. Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. BMC Evol. Biol. 7, 154. doi:10.1186/1471-2148-7-154

Arends, K., Celik, E.-K., Probst, I., Goessweiner-Mohr, N., Fercher, C., Grumet, L., Soellue, C., Abajy, M.Y., Sakinc, T., Broszat, M., Schiwon, K., Koraimann, G., Keller, W., Grohmann, E., 2013. TraG Encoded by the pIP501 Type IV Secretion System Is a Two-Domain Peptidoglycan-Degrading Enzyme Essential for Conjugative Transfer. J. Bacteriol. 195, 4436–4444. doi:10.1128/JB.02263-12

Arvey, A.J., Azad, R.K., Raval, A., Lawrence, J.G., 2009. Detection of genomic islands via segmental genome heterogeneity. Nucleic Acids Res. 37, 5255–5266. doi:10.1093/nar/gkp576

Auchtung, J.M., Aleksanyan, N., Bulku, A., Berkmen, M.B., 2016. Biology of ICEBs1, an integrative and conjugative element in *Bacillus subtilis*. Plasmid 86, 14–25. doi:10.1016/j.plasmid.2016.07.001

Auchtung, J.M., Lee, C.A., Garrison, K.L., Grossman, A.D., 2007. Identification and characterization of the immunity repressor (ImmR) that controls the mobile genetic element ICE*Bs1* of *Bacillus subtilis*. Mol. Microbiol. 64, 1515–1528. doi:10.1111/j.1365-2958.2007.05748.x

Auchtung, J.M., Lee, C.A., Monson, R.E., Lehman, A.P., Grossman, A.D., 2005. Regulation of a *Bacillus subtilis* mobile genetic element by intercellular signaling and the global DNA damage response. Proc. Natl. Acad. Sci. U. S. A. 102, 12554–12559. doi:10.1073/pnas.0505835102

Ayoubi, P., Kilic, A.O., Vijayakumar, M.N., 1991. Tn*5253*, the pneumococcal omega (cat tet) BM6001 element, is a composite structure of two conjugative transposons, Tn*5251* and Tn*5252*. J. Bacteriol. 173, 1617–1622.

Bai, H., Sun, M., Ghosh, P., Hatfull, G.F., Grindley, N.D.F., Marko, J.F., 2011. Single-molecule analysis reveals the molecular bearing mechanism of DNA strand exchange by a serine recombinase. Proc. Natl. Acad. Sci. 108, 7419–7424. doi:10.1073/pnas.1018436108

Baker, J.R., Pritchard, D.G., 2000. Action pattern and substrate specificity of the hyaluronan lyase from group B streptococci. Biochem. J. 348 Pt 2, 465–471.

Baker, S., Pickard, D., Whitehead, S., Farrar, J., Dougan, G., 2008. Mobilization of the incQ plasmid R300B with a chromosomal conjugation system in *Salmonella enterica* serovar typhi. J. Bacteriol. 190, 4084–4087. doi:10.1128/JB.00065-08

Balter, S.E., Dowell, S.F., 2000. Update on acute otitis media. Curr. Opin. Infect. Dis. 13, 165–170.

Bellanger, X., Morel, C., Decaris, B., Guédon, G., 2007. Derepression of excision of integrative and potentially conjugative elements from *Streptococcus thermophilus* by DNA damage response: implication of a cI-related repressor. J. Bacteriol. 189, 1478–1481. doi:10.1128/JB.01125-06

Bellanger, X., Morel, C., Gonot, F., Puymège, A., Decaris, B., Guédon, G., 2011. Site-specific accretion of an integrative conjugative element together with a related genomic island leads to *cis* mobilization and gene capture. Mol. Microbiol. 81, 912–925. doi:10.1111/j.1365-2958.2011.07737.x

Bellanger, X., Payot, S., Leblond-Bourget, N., Guédon, G., 2014. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. FEMS Microbiol. Rev. 38, 720–760. doi:10.1111/1574-6976.12058

Bellanger, X., Roberts, A.P., Morel, C., Choulet, F., Pavlovic, G., Mullany, P., Decaris, B., Guédon, G., 2009. Conjugative Transfer of the Integrative Conjugative Elements ICESt1 and ICESt3 from Streptococcus thermophilus. J. Bacteriol. 191, 2764–2775. doi:10.1128/JB.01412-08

Beres, S.B., Musser, J.M., 2007. Contribution of Exogenous Genetic Elements to the Group A *Streptococcus* Metagenome. PLoS ONE 2. doi:10.1371/journal.pone.0000800

Bessen, D.E., McShan, W.M., Nguyen, S.V., Shetty, A., Agrawal, S., Tettelin, H., 2015. Molecular Epidemiology and Genomics of Group A *Streptococcus*. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. 33, 393–418. doi:10.1016/j.meegid.2014.10.011

Bhatty, M., Laverde Gomez, J.A., Christie, P.J., 2013. The expanding bacterial type IV secretion lexicon. Res. Microbiol. 164, 620–639. doi:10.1016/j.resmic.2013.03.012

Bi, D., Xu, Z., Harrison, E.M., Tai, C., Wei, Y., He, X., Jia, S., Deng, Z., Rajakumar, K., Ou, H.-Y., 2012. ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. Nucleic Acids Res. 40, D621-D626. doi:10.1093/nar/gkr846

Bjørkeng, E.K., Hjerde, E., Pedersen, T., Sundsfjord, A., Hegstad, K., 2013. ICE*Sluvan*, a 94-Kilobase Mosaic Integrative Conjugative Element Conferring Interspecies Transfer of VanB-Type Glycopeptide Resistance, a Novel Bacitracin Resistance Locus, and a Toxin-Antitoxin Stabilization System. J. Bacteriol. 195, 5381–5390. doi:10.1128/JB.02165-12

Bolotin, A., Quinquis, B., Renault, P., Sorokin, A., Ehrlich, S.D., Kulakauskas, S., Lapidus, A., Goltsman, E., Mazur, M., Pusch, G.D., Fonstein, M., Overbeek, R., Kyprides, N., Purnelle, B., Prozzi, D., Ngui, K., Masuy, D., Hancy, F., Burteau, S., Boutry, M., Delcour, J., Goffeau, A., Hols, P., 2004. Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. Nat. Biotechnol. 22, 1554–1558. doi:10.1038/nbt1034

Böltner, D., Osborn, A.M., 2004. Structural comparison of the integrative and conjugative elements R391, pMERPH, R997, and SXT. Plasmid 51, 12–23.

Bose, B., Auchtung, J.M., Lee, C.A., Grossman, A.D., 2008. A conserved anti-repressor controls horizontal gene transfer by proteolysis. Mol. Microbiol. 70, 570–582. doi:10.1111/j.1365-2958.2008.06414.x

Bose, B., Grossman, A.D., 2011. Regulation of horizontal gene transfer in *Bacillus subtilis* by activation of a conserved site-specific protease. J. Bacteriol. 193, 22–29. doi:10.1128/JB.01143-10

Brenciani, A., Tiberi, E., Bacciaglia, A., Petrelli, D., Varaldo, P.E., Giovanetti, E., 2011. Two distinct genetic elements are responsible for erm(TR)-mediated erythromycin resistance in tetracycline-susceptible and tetracycline-resistant strains of *Streptococcus pyogenes*. Antimicrob. Agents Chemother. 55, 2106–2112. doi:10.1128/AAC.01378-10

Brochet, M., Couvé, E., Glaser, P., Guédon, G., Payot, S., 2008. Integrative conjugative elements and related elements are major contributors to the genome diversity of *Streptococcus agalactiae*. J. Bacteriol. 190, 6913–6917. doi:10.1128/JB.00824-08

Brochet, M., Da Cunha, V., Couvé, E., Rusniok, C., Trieu-Cuot, P., Glaser, P., 2009. Atypical association of DDE transposition with conjugation specifies a new family of mobile elements. Mol. Microbiol. 71, 948–959. doi:10.1111/j.1365-2958.2008.06579.x

Brouwer, M.S.M., Warburton, P.J., Roberts, A.P., Mullany, P., Allan, E., 2011. Genetic organisation, mobility and predicted functions of genes on integrated, mobile genetic elements in sequenced strains of *Clostridium difficile*. PloS One 6, e23014. doi:10.1371/journal.pone.0023014

Burrus, V., Bontemps, C., Decaris, B., Guédon, G., 2001. Characterization of a Novel Type II Restriction-Modification System, *Sth*368I, Encoded by the Integrative Element ICE*St1* of *Streptococcus thermophilus* CNRZ368. Appl. Environ. Microbiol. 67, 1522–1528. doi:10.1128/AEM.67.4.1522-1528.2001

Burrus, V., Pavlovic, G., Decaris, B., Guédon, G., 2002a. Conjugative transposons: the tip of the iceberg. Mol. Microbiol. 46, 601–610.

Burrus, V., Pavlovic, G., Decaris, B., Guédon, G., 2002b. The ICESt1 element of *Streptococcus thermophilus* belongs to a large family of integrative and conjugative elements that exchange modules and change their specificity of integration. Plasmid 48, 77–97.

Burrus, V., Waldor, M.K., 2003. Control of SXT integration and excision. J. Bacteriol. 185, 5045–5054.

Cabezón, E., Ripoll-Rozada, J., Peña, A., de la Cruz, F., Arechaga, I., 2015. Towards an integrated model of bacterial conjugation. FEMS Microbiol. Rev. 39, 81–95. doi:10.1111/1574-6976.12085

Cabezón, E., Sastre, J.I., de la Cruz, F., 1997. Genetic evidence of a coupling role for the TraG protein family in bacterial conjugation. Mol. Gen. Genet. MGG 254, 400–406.

Carapetis, J.R., Steer, A.C., Mulholland, E.K., Weber, M., 2005. The global burden of group A streptococcal diseases. Lancet Infect. Dis. 5, 685–694. doi:10.1016/S1473-3099(05)70267-X

Carballeira, J.D., González-Pérez, B., Moncalián, G., de la Cruz, F., 2014. A high security double lock and key mechanism in HUH relaxases controls oriT-processing for plasmid conjugation. Nucleic Acids Res. 42, 10632–10643. doi:10.1093/nar/gku741

Carraro, N., Burrus, V., 2015. The dualistic nature of integrative and conjugative elements. Mob. Genet. Elem. 5, 98–102. doi:10.1080/2159256X.2015.1102796

Carraro, N., Durand, R., Rivard, N., Anquetil, C., Barrette, C., Humbert, M., Burrus, V., 2017a. Salmonella genomic island 1 (SGI1) reshapes the mating apparatus of IncC conjugative plasmids to promote self-propagation. PLoS Genet. 13, e1006705. doi:10.1371/journal.pgen.1006705

Carraro, N., Rivard, N., Burrus, V., Ceccarelli, D., 2017b. Mobilizable genomic islands, different strategies for the dissemination of multidrug resistance and other adaptive traits. Mob. Genet. Elem. 7, 1–6. doi:10.1080/2159256X.2017.1304193

Carraro, N., Rivard, N., Ceccarelli, D., Colwell, R.R., Burrus, V., 2016. IncA/C Conjugative Plasmids Mobilize a New Family of Multidrug Resistance Islands in Clinical *Vibrio cholerae* Non-O1/Non-O139 Isolates from Haiti. mBio 7. doi:10.1128/mBio.00509-16

Carter, M.Q., Chen, J., Lory, S., 2010. The *Pseudomonas aeruginosa* Pathogenicity Island PAPI-1 Is Transferred via a Novel Type IV Pilus. J. Bacteriol. 192, 3249–3258. doi:10.1128/JB.00041-10

Chandler, M., de la Cruz, F., Dyda, F., Hickman, A.B., Moncalian, G., Ton-Hoang, B., 2013. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nat. Rev. Microbiol. 11, 525–538. doi:10.1038/nrmicro3067

Che, D., Hasan, M.S., Wang, H., Fazekas, J., Huang, J., Liu, Q., 2011. EGID: an ensemble algorithm for improved genomic island detection in genomic sequences. Bioinformation 7, 311–314.

Che, D., Hockenbury, C., Marmelstein, R., Rasheed, K., 2010. Classification of genomic islands using decision trees and their ensemble algorithms. BMC Genomics 11 Suppl 2, S1. doi:10.1186/1471-2164-11-S2-S1

Cheng, Q., Paszkiet, B.J., Shoemaker, N.B., Gardner, J.F., Salyers, A.A., 2000. Integration and excision of a *Bacteroides* conjugative transposon, CTnDOT. J. Bacteriol. 182, 4035–4043.

Christie, G.E., Dokland, T., 2012. Pirates of the Caudovirales. Virology 434, 210–221. doi:10.1016/j.virol.2012.10.028

Christie, P.J., 2016. The Mosaic Type IV Secretion Systems. EcoSal Plus 7. doi:10.1128/ecosalplus.ESP-0020-2015

Claverys, J.-P., Martin, B., Håvarstein, L.S., 2007. Competence-induced fratricide in streptococci. Mol. Microbiol. 64, 1423–1433. doi:10.1111/j.1365-2958.2007.05757.x

Clewell, D.B., Gawron-Burke, C., 1986. Conjugative Transposons and the Dissemination of Antibiotic Resistance in *Streptococci*. Annu. Rev. Microbiol. 40, 635–659. doi:10.1146/annurev.mi.40.100186.003223

Cookson, A.L., Noel, S., Hussein, H., Perry, R., Sang, C., Moon, C.D., Leahy, S.C., Altermann, E., Kelly, W.J., Attwood, G.T., 2011. Transposition of Tn*916* in the four replicons of the *Butyrivibrio proteoclasticus* B316(T) genome. FEMS Microbiol. Lett. 316, 144–151. doi:10.1111/j.1574-6968.2010.02204.x

Crellin, P.K., Rood, J.I, 1997. The resolvase/invertase domain of the site-specific recombinase TnpX is functional and recognizes a target sequence that resembles the junction of the circular form of the Clostridium perfringens transposon Tn4451. J. Bacteriol 178, 5148-5156.

Crellin, P.K., Rood, J.I., 1998. Tn4451 from Clostridium perfringens is a mobilizable transposon that encodes the functional Mob protein, TnpZ. Mol. Microbiol. 27, 631–642.

Daccord, A., Ceccarelli, D., Rodrigue, S., Burrus, V., 2013. Comparative analysis of mobilizable genomic islands. J. Bacteriol. 195, 606–614. doi:10.1128/JB.01985-12

Dahmane, N., Libante, V., Charron-Bourgoin, F., Guédon, E., Guédon, G., Leblond-Bourget, N., Payot, S., 2017. Diversity of Integrative and Conjugative Elements of *Streptococcus salivarius* and Their Intra- and Interspecies Transfer. Appl. Environ. Microbiol. 83. doi:10.1128/AEM.00337-17

Darling, A.C.E., Mau, B., Blattner, F.R., Perna, N.T., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14, 1394–1403. doi:10.1101/gr.2289704

de Jong, A., van Hijum, S.A.F.T., Bijlsma, J.J.E., Kok, J., Kuipers, O.P., 2006. BAGEL: a web-based bacteriocin genome mining tool. Nucleic Acids Res. 34, W273–W279. doi:10.1093/nar/gkl237

de la Cruz, F., Frost, L.S., Meyer, R.J., Zechner, E.L., 2010. Conjugative DNA metabolism in Gram-negative bacteria. FEMS Microbiol. Rev. 34, 18–40. doi:10.1111/j.1574-6976.2009.00195.x

Delavat, F., Miyazaki, R., Carraro, N., Pradervand, N., van der Meer, J.R., 2017. The hidden life of integrative and conjugative elements. FEMS Microbiol. Rev. 41, 512–537. doi:10.1093/femsre/fux008

Delorme, C., Abraham, A.-L., Renault, P., Guédon, E., 2015. Genomics of *Streptococcus salivarius*, a major human commensal. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. 33, 381–392. doi:10.1016/j.meegid.2014.10.001

Dermer, P., Lee, C., Eggert, J., Few, B., 2004. A history of neonatal group B streptococcus with its related morbidity and mortality rates in the United States. J. Pediatr. Nurs. 19, 357–363. doi:10.1016/j.pedn.2004.05.012

DeWitt, T., Grossman, A.D., 2014. The Bifunctional Cell Wall Hydrolase CwlT Is Needed for Conjugation of the Integrative and Conjugative Element ICE*Bs1* in *Bacillus subtilis* and *B. anthracis*. J. Bacteriol. 196, 1588–1596. doi:10.1128/JB.00012-14

Dhillon, B.K., Chiu, T.A., Laird, M.R., Langille, M.G.I., Brinkman, F.S.L., 2013. IslandViewer update: Improved genomic island discovery and visualization. Nucleic Acids Res. 41, W129-132. doi:10.1093/nar/gkt394

Douarre, P.-E., Sauvage, E., Poyart, C., Glaser, P., 2015. Host specificity in the diversity and transfer of lsa resistance genes in group B *Streptococcus*. J. Antimicrob. Chemother. 70, 3205–3213. doi:10.1093/jac/dkv277

Elhai, J., Liu, H., Taton, A., 2012. Detection of horizontal transfer of individual genes by anomalous oligomer frequencies. BMC Genomics 13, 245. doi:10.1186/1471-2164-13-245

Evans, R.P., Macrina, F.L., 1983. Streptococcal R plasmid pIP501: endonuclease site map, resistance determinant location, and construction of novel derivatives. J. Bacteriol. 154, 1347–1355.

Farías, M.E., Espinosa, M., 2000. Conjugal transfer of plasmid pMV158: uncoupling of the pMV158 origin of transfer from the mobilization gene mobM, and modulation of pMV158 transfer in *Escherichia coli* mediated by IncP plasmids. Microbiology 146, 2259–2265. doi:10.1099/00221287-146-9-2259

Farías, M.E., Grohmann, E., Espinosa, M., 1999. Expression of the mobM gene of the streptococcal plasmid pMV158 in *Lactococcus lacti*s subsp. *lactis*. FEMS Microbiol. Lett. 176, 403–410.

Fernández-López, C., Bravo, A., Ruiz-Cruz, S., Solano-Collado, V., Garsin, D.A., Lorenzo-Díaz, F., Espinosa, M., 2014. Mobilizable Rolling-Circle Replicating Plasmids from Gram-Positive Bacteria: A Low-Cost Conjugative Transfer. Microbiol. Spectr. 2. doi:10.1128/microbiolspec.PLAS-0008-2013

Ferretti, J.J., Stevens, D.L., Fischetti, V.A. (Eds.), 2016. *Streptococcus pyogenes* : Basic Biology to Clinical Manifestations. University of Oklahoma Health Sciences Center, Oklahoma City (OK).

Fogg, P.C.M., Colloms, S., Rosser, S., Stark, M., Smith, M.C.M., 2014. New applications for phage integrases. J. Mol. Biol. 426, 2703–2716. doi:10.1016/j.jmb.2014.05.014

Fontaine, L., Boutry, C., de Frahan, M.H., Delplace, B., Fremaux, C., Horvath, P., Boyaval, P., Hols, P., 2010. A novel pheromone quorum-sensing system controls the development of natural competence in *Streptococcus thermophilus* and *Streptococcus salivarius*. J. Bacteriol. 192, 1444–1454. doi:10.1128/JB.01251-09

Francia, M.V., Clewell, D.B., 2002. Transfer origins in the conjugative *Enterococcus faecalis* plasmids pAD1 and pAM373: identification of the pAD1 nic site, a specific relaxase and a possible TraG-like protein. Mol. Microbiol. 45, 375–395.

Francia, M.V., Varsaki, A., Garcillán-Barcia, M.P., Latorre, A., Drainas, C., de la Cruz, F., 2004. A classification scheme for mobilization regions of bacterial plasmids. FEMS Microbiol. Rev. 28, 79–100. doi:10.1016/j.femsre.2003.09.001

Franke, A.E., Clewell, D.B., 1981. Evidence for a chromosome-borne resistance transposon (Tn*916*) in *Streptococcus faecalis* that is capable of “conjugal” transfer in the absence of a conjugative plasmid. J. Bacteriol. 145, 494–502.

Fronzes, R., Christie, P.J., Waksman, G., 2009. The structural biology of type IV secretion systems. Nat. Rev. Microbiol. 7, 703–714. doi:10.1038/nrmicro2218

Frost, L.S., Leplae, R., Summers, A.O., Toussaint, A., 2005. Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. 3, 722–732. doi:10.1038/nrmicro1235

Furuya, N., Komano, T., 1995. Specific binding of the NikA protein to one arm of 17-base-pair inverted repeat sequences within the *oriT* region of plasmid R64. J. Bacteriol. 177, 46–51.

Galperin, M.Y., Fernández-Suárez, X.M., Rigden, D.J., 2017. The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. Nucleic Acids Res. 45, D1–D11. doi:10.1093/nar/gkw1188

Gao, F., Zhang, C.-T., 2006. GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. Nucleic Acids Res. 34, W686-691. doi:10.1093/nar/gkl040

Garcillán-Barcia, M.P., Francia, M.V., de la Cruz, F., 2009. The diversity of conjugative relaxases and its application in plasmid classification. FEMS Microbiol. Rev. 33, 657–687.

Garnier, F., Janapatla, R.P., Charpentier, E., Masson, G., Grélaud, C., Stach, J.F., Denis, F., Ploy, M.-C., 2007. Insertion sequence 1515 in the *ply* gene of a type 1 clinical isolate of *Streptococcus pneumoniae* abolishes pneumolysin expression. J. Clin. Microbiol. 45, 2296–2297. doi:10.1128/JCM.02168-06

Garnier, F., Taourit, S., Glaser, P., Courvalin, P., Galimand, M., 2000. Characterization of transposon Tn*1549*, conferring VanB-type resistance in *Enterococcus* spp. Microbiology146, 1481–1489. doi:10.1099/00221287-146-6-1481

Ghinet, M.G., Bordeleau, E., Beaudin, J., Brzezinski, R., Roy, S., Burrus, V., 2011. Uncovering the prevalence and diversity of integrating conjugative elements in actinobacteria. PloS One 6, e27846. doi:10.1371/journal.pone.0027846

Ghosh, P., Wasil, L.R., Hatfull, G.F., 2006. Control of Phage Bxb1 Excision by a Novel Recombination Directionality Factor. PLOS Biol. 4, e186. doi:10.1371/journal.pbio.0040186

Giovanetti, E., Brenciani, A., Tiberi, E., Bacciaglia, A., Varaldo, P.E., 2012. ICESp2905, the erm(TR)-tet(O) element of *Streptococcus pyogenes*, is formed by two independent integrative and conjugative elements. Antimicrob. Agents Chemother. 56, 591–594. doi:10.1128/AAC.05352-11

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., Rappé, M.S., Short, J.M., Carrington, J.C., Mathur, E.J., 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science 309, 1242–1245. doi:10.1126/science.1114057

Goessweiner-Mohr, N., Arends, K., Keller, W., Grohmann, E., 2014. Conjugation in Gram-Positive Bacteria. Microbiol. Spectr. 2, PLAS-0004-2013. doi:10.1128/microbiolspec.PLAS-0004-2013

Goessweiner-Mohr, N., Arends, K., Keller, W., Grohmann, E., 2013. Conjugative type IV secretion systems in Gram-positive bacteria. Plasmid 70, 289–302. doi:10.1016/j.plasmid.2013.09.005

Gogarten, J.P., Townsend, J.P., 2005. Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3, 679–687. doi:10.1038/nrmicro1204

Gomis-Rüth, F.X., Solà, M., de la Cruz, F., Coll, M., 2004. Coupling factors in macromolecular type-IV secretion machineries. Curr. Pharm. Des. 10, 1551–1565.

Gottschalk, M., Xu, J., Calzas, C., Segura, M., 2010. *Streptococcus suis*: a new emerging or an old neglected zoonotic pathogen? Future Microbiol. 5, 371–391. doi:10.2217/fmb.10.2

Grindley, N.D.F., Whiteson, K.L., Rice, P.A., 2006. Mechanisms of site-specific recombination. Annu. Rev. Biochem. 75, 567–605. doi:10.1146/annurev.biochem.73.011303.073908

Grohmann, E., Muth, G., Espinosa, M., 2003. Conjugative plasmid transfer in gram-positive bacteria. Microbiol. Mol. Biol. Rev. 67, 277–301.

Groth, A.C., Calos, M.P., 2004. Phage Integrases: Biology and Applications. J. Mol. Biol. 335, 667–678. doi:10.1016/j.jmb.2003.09.082

Guérillot, R., Da Cunha, V., Sauvage, E., Bouchier, C., Glaser, P., 2013. Modular evolution of TnGBSs, a new family of integrative and conjugative elements associating insertion sequence transposition, plasmid replication, and conjugation for their spreading. J. Bacteriol. 195, 1979–1990. doi:10.1128/JB.01745-12

Guérillot, R., Siguier, P., Gourbeyre, E., Chandler, M., Glaser, P., 2014. The diversity of prokaryotic DDE transposases of the mutator superfamily, insertion specificity, and association with conjugation machineries. Genome Biol. Evol. 6, 260–272. doi:10.1093/gbe/evu010

Guglielmini, J., de la Cruz, F., Rocha, E.P.C., 2013. Evolution of Conjugation and Type IV Secretion Systems. Mol. Biol. Evol. 30, 315–331. doi:10.1093/molbev/mss221

Guglielmini, J., Néron, B., Abby, S.S., Garcillán-Barcia, M.P., de la Cruz, F., Rocha, E.P.C., 2014. Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. Nucleic Acids Res. 42, 5715–5727. doi:10.1093/nar/gku194

Guglielmini, J., Quintais, L., Garcillán-Barcia, M.P., de la Cruz, F., Rocha, E.P.C., 2011. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. PLoS Genet. 7, e1002222. doi:10.1371/journal.pgen.1002222

Guzmán, L.M., Espinosa, M., 1997. The mobilization protein, MobM, of the streptococcal plasmid pMV158 specifically cleaves supercoiled DNA at the plasmid *oriT*. J. Mol. Biol. 266, 688–702. doi:10.1006/jmbi.1996.0824

Hacker, J., Kaper, J.B., 2000. Pathogenicity islands and the evolution of microbes. Annu. Rev. Microbiol. 54, 641–679. doi:10.1146/annurev.micro.54.1.641

Hamilton, H.L., Dillard, J.P., 2006. Natural transformation of *Neisseria gonorrhoeae*: from DNA donation to homologous recombination. Mol. Microbiol. 59, 376–385. doi:10.1111/j.1365-2958.2005.04964.x

Haskett, T.L., Terpolilli, J.J., Bekuma, A., O'Hara, G.W., Sullivan, J.T., Wang, P., Ronson, C.W., Ramsay, J.P., 2016. Assembly and transfer of tripartite integrative and conjugative genetic elements. Proc. Natl. Acad. Sci. U. S. A. 113, 12268–12273. doi:10.1073/pnas.1613358113

Haustenne, L., Bastin, G., Hols, P., Fontaine, L., 2015. Modeling of the ComRS Signaling Pathway Reveals the Limiting Factors Controlling Competence in *Streptococcus thermophilus*. Front. Microbiol. 6, 1413. doi:10.3389/fmicb.2015.01413

Hayes, C.S., Aoki, S.K., Low, D.A., 2010. Bacterial contact-dependent delivery systems. Annu. Rev. Genet. 44, 71–90. doi:10.1146/annurev.genet.42.110807.091449

Heather, Z., Holden, M.T.G., Steward, K.F., Parkhill, J., Song, L., Challis, G.L., Robinson, C., Davis-Poynter, N., Waller, A.S., 2008. A novel streptococcal integrative conjugative element involved in iron acquisition. Mol. Microbiol. 70, 1274–1292. doi:10.1111/j.1365-2958.2008.06481.x

Hegstad, K., Mikalsen, T., Coque, T.M., Werner, G., Sundsfjord, A., 2010. Mobile genetic elements and their contribution to the emergence of antimicrobial resistant *Enterococcus faecalis* and *Enterococcus faecium*. Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis. 16, 541–554. doi:10.1111/j.1469-0691.2010.03226.x

Hochhut, B., Jahreis, K., Lengeler, J.W., Schmid, K., 1997. CTn*scr94*, a conjugative transposon found in enterobacteria. J. Bacteriol. 179, 2097–2102.

Hosking, S.L., Deadman, M.E., Moxon, E.R., Peden, J.F., Saunders, N.J., High, N.J., 1998. An in silico evaluation of Tn916 as a tool for generalized mutagenesis in *Haemophilus influenzae* Rd. Microbiology. 144, 2525–2530. doi:10.1099/00221287-144-9-2525

Hsiao, W., Wan, I., Jones, S.J., Brinkman, F.S.L., 2003. IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics. 19, 418–420.

Huang, J., Ma, J., Shang, K., Hu, X., Liang, Y., Li, D., Wu, Z., Dai, L., Chen, L., Wang, L., 2016. Evolution and Diversity of the Antimicrobial Resistance Associated Mobilome in *Streptococcus suis*: A Probable Mobile Genetic Elements Reservoir for Other Streptococci. Front. Cell. Infect. Microbiol. 6. doi:10.3389/fcimb.2016.00118

Iannelli, F., Santoro, F., Oggioni, M.R., Pozzi, G., 2014. Nucleotide sequence analysis of integrative conjugative element Tn*5253* of *Streptococcus pneumoniae*. Antimicrob. Agents Chemother. 58, 1235–1239. doi:10.1128/AAC.01764-13

Ilangovan, A., Connery, S., Waksman, G., 2015. Structural biology of the Gram-negative bacterial conjugation systems. Trends Microbiol. 23, 301–310. doi:10.1016/j.tim.2015.02.012

Iyer, L.M., Makarova, K.S., Koonin, E.V., Aravind, L., 2004. Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. Nucleic Acids Res. 32, 5260–5279. doi:10.1093/nar/gkh828

Jia, B., Raphenya, A.R., Alcock, B., Waglechner, N., Guo, P., Tsang, K.K., Lago, B.A., Dave, B.M., Pereira, S., Sharma, A.N., Doshi, S., Courtot, M., Lo, R., Williams, L.E., Frye, J.G., Elsayegh, T., Sardar, D., Westman, E.L., Pawlowski, A.C., Johnson, T.A., Brinkman, F.S.L., Wright, G.D., McArthur, A.G., 2017. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 45, D566–D573. doi:10.1093/nar/gkw1004

Johnson, C.M., Grossman, A.D., 2015. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. Annu. Rev. Genet. 49, 577–601. doi:10.1146/annurev-genet-112414-055018

Juhas, M., van der Meer, J.R., Gaillard, M., Harding, R.M., Hood, D.W., Crook, D.W., 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol. Rev. 33, 376–393. doi:10.1111/j.1574-6976.2008.00136.x

Kers, J.A., Cameron, K.D., Joshi, M.V., Bukhalid, R.A., Morello, J.E., Wach, M.J., Gibson, D.M., Loria, R., 2005. A large, mobile pathogenicity island confers plant pathogenicity on *Streptomyces* species. Mol. Microbiol. 55, 1025–1033. doi:10.1111/j.1365-2958.2004.04461.x

Khaleel, T., Younger, E., McEwan, A.R., Varghese, A.S., Smith, M.C.M., 2011. A phage protein that binds ϕC31 integrase to switch its directionality. Mol. Microbiol. 80, 1450–1463. doi:10.1111/j.1365-2958.2011.07696.x

Köhler, W., 2007. The present state of species within the genera *Streptococcus* and *Enterococcus*. Int. J. Med. Microbiol. IJMM 297, 133–150. doi:10.1016/j.ijmm.2006.11.008

Korona-Glowniak, I., Siwiec, R., Malm, A., 2015. Resistance determinants and their association with different transposons in the antibiotic-resistant *Streptococcus pneumoniae*. BioMed Res. Int. 2015, 836496. doi:10.1155/2015/836496

Krah, E.R., Macrina, F.L., 1991. Identification of a region that influences host range of the streptococcal conjugative plasmid pIP501. Plasmid 25, 64–69.

Kunkel, B., Losick, R., Stragier, P., 1990. The *Bacillus subtilis* gene for the development transcription factor sigma K is generated by excision of a dispensable DNA element containing a sporulation recombinase gene. Genes Dev. 4, 525–535.

Kurenbach, B., Bohn, C., Prabhu, J., Abudukerim, M., Szewzyk, U., Grohmann, E., 2003. Intergeneric transfer of the *Enterococcus faecalis* plasmid pIP501 to *Escherichia coli* and S*treptomyces lividans* and sequence analysis of its tra region. Plasmid 50, 86–93.

Kurenbach, B., Grothe, D., Farías, M.E., Szewzyk, U., Grohmann, E., 2002. The tra region of the conjugative plasmid pIP501 is organized in an operon with the first gene encoding the relaxase. J. Bacteriol. 184, 1801–1805.

Kurenbach, B., Kopeć, J., Mägdefrau, M., Andreas, K., Keller, W., Bohn, C., Abajy, M.Y., Grohmann, E., 2006. The TraA relaxase autoregulates the putative type IV secretion-like system encoded by the broad-host-range *Streptococcus agalactiae* plasmid pIP501. Microbiology. 152, 637–645. doi:10.1099/mic.0.28468-0

Langille, M.G.I., Hsiao, W.W.L., Brinkman, F.S.L., 2010. Detecting genomic islands using bioinformatics approaches. Nat. Rev. Microbiol. 8, 373–382. doi:10.1038/nrmicro2350

Langille, M.G.I., Hsiao, W.W.L., Brinkman, F.S.L., 2008. Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinformatics 9, 329. doi:10.1186/1471-2105-9-329

Lanka, E., Wilkins, B.M., 1995. DNA processing reactions in bacterial conjugation. Annu. Rev. Biochem. 64, 141–169. doi:10.1146/annurev.bi.64.070195.001041

Laverde Gomez, J.A., Hendrickx, A.P.A., Willems, R.J., Top, J., Sava, I., Huebner, J., Witte, W., Werner, G., 2011. Intra- and Interspecies Genomic Transfer of the *Enterococcus faecalis* Pathogenicity Island. PLoS ONE 6. doi:10.1371/journal.pone.0016720

Le Bourgeois, P., Bugarel, M., Campo, N., Daveran-Mingot, M.-L., Labonté, J., Lanfranchi, D., Lautier, T., Pagès, C., Ritzenthaler, P., 2007. The Unconventional Xer Recombination Machinery of *Streptococci*/*Lactococci*. PLoS Genet. 3. doi:10.1371/journal.pgen.0030117

Lederberg, J., 1998. Plasmid (1952-1997). Plasmid 39, 1–9. doi:10.1006/plas.1997.1320

Lederberg, J., Tatum, E.L., 1946. Gene recombination in *Escherichia coli*. Nature 158, 558.

Lee, C.A., Auchtung, J.M., Monson, R.E., Grossman, A.D., 2007. Identification and characterization of int (integrase), xis (excisionase) and chromosomal attachment sites of the integrative and conjugative element ICE*Bs1* of *Bacillus subtilis*. Mol. Microbiol. 66, 1356–1369. doi:10.1111/j.1365-2958.2007.06000.x

Lee, C.A., Babic, A., Grossman, A.D., 2010. Autonomous plasmid-like replication of a conjugative transposon. Mol. Microbiol. 75, 268–279. doi:10.1111/j.1365-2958.2009.06985.x

Lee, C.A., Grossman, A.D., 2007. Identification of the Origin of Transfer (*oriT*) and DNA Relaxase Required for Conjugation of the Integrative and Conjugative Element ICE*Bs1* of *Bacillus subtilis*. J. Bacteriol. 189, 7254–7261. doi:10.1128/JB.00932-07

Lee, C.A., Thomas, J., Grossman, A.D., 2012. The *Bacillus subtilis* Conjugative Transposon ICE*Bs1* Mobilizes Plasmids Lacking Dedicated Mobilization Functions. J. Bacteriol. 194, 3165–3172. doi:10.1128/JB.00301-12

Lee, C.-C., Chen, Y.-P.P., Yao, T.-J., Ma, C.-Y., Lo, W.-C., Lyu, P.-C., Tang, C.Y., 2013. GI-POP: a combinational annotation and genomic island prediction pipeline for ongoing microbial genome projects. Gene 518, 114–123. doi:10.1016/j.gene.2012.11.063

Lefébure, T., Stanhope, M.J., 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. 8, R71. doi:10.1186/gb-2007-8-5-r71

Leonetti, C.T., Hamada, M.A., Laurer, S.J., Broulidakis, M.P., Swerdlow, K.J., Lee, C.A., Grossman, A.D., Berkmen, M.B., 2015. Critical Components of the Conjugation Machinery of the Integrative and Conjugative Element ICE*Bs1* of *Bacillus subtilis*. J. Bacteriol. 197, 2558–2567. doi:10.1128/JB.00142-15

Lesic, B., Carniel, E., 2005. Horizontal transfer of the high-pathogenicity island of *Yersinia pseudotuberculosis*. J. Bacteriol. 187, 3352–3358. doi:10.1128/JB.187.10.3352-3358.2005

Li, J., Busscher, H.J., van der Mei, H.C., Sjollema, J., 2013. Surface enhanced bacterial fluorescence and enumeration of bacterial adhesion. Biofouling 29, 11–19. doi:10.1080/08927014.2012.742074

Lorenz, M.G., Wackernagel, W., 1994. Bacterial gene transfer by natural genetic transformation in the environment. Microbiol. Rev. 58, 563–602.

Lorenzo-Diaz, F., Fernández-Lopez, C., Douarre, P.-E., Baez-Ortega, A., Flores, C., Glaser, P., Espinosa, M., 2016. Streptococcal group B integrative and mobilizable element IME*Sag-rpsI* encodes a functional relaxase involved in its transfer. Open Biol. 6. doi:10.1098/rsob.160084

Lorenzo-Díaz, F., Fernández-López, C., Garcillán-Barcia, M.P., Espinosa, M., 2014. Bringing them together: plasmid pMV158 rolling circle replication and conjugation under an evolutionary perspective. Plasmid 74, 15–31. doi:10.1016/j.plasmid.2014.05.004

Lu, B., Leong, H.W., 2016. Computational methods for predicting genomic islands in microbial genomes. Comput. Struct. Biotechnol. J. 14, 200–206. doi:10.1016/j.csbj.2016.05.001

Magot, M., Carlier, J.P., Popoff, M.R., 1983. Identification of *Clostridium butyricum* and *Clostridium beijerinckii* by gas-liquid chromatography and sugar fermentation: correlation with DNA homologies and electrophoretic patterns. J. Gen. Microbiol. 129, 2837–2845. doi:10.1099/00221287-129-9-2837

Makarova, K.S., Koonin, E.V., 2007. Evolutionary genomics of lactic acid bacteria. J. Bacteriol. 189, 1199–1208. doi:10.1128/JB.01351-06

Makart, L., Gillis, A., Mahillon, J., 2015. pXO16 from *Bacillus thuringiensis* serovar israelensis: Almost 350 kb of terra incognita. Plasmid 80, 8–15. doi:10.1016/j.plasmid.2015.03.002

Marcoleta, A.E., Berríos-Pastén, C., Nuñez, G., Monasterio, O., Lagos, R., 2016. *Klebsiella pneumoniae* Asparagine tDNAs Are Integration Hotspots for Different Genomic Islands Encoding Microcin E492 Production Determinants and Other Putative Virulence Factors Present in Hypervirulent Strains. Front. Microbiol. 7, 849. doi:10.3389/fmicb.2016.00849

Mays, T.D., Smith, C.J., Welch, R.A., Delfini, C., Macrina, F.L., 1982. Novel antibiotic resistance transfer in *Bacteroides*. Antimicrob. Agents Chemother. 21, 110–118.

Mazodier, P., Davies, J., 1991. Gene transfer between distantly related bacteria. Annu. Rev. Genet. 25, 147–171. doi:10.1146/annurev.ge.25.120191.001051

McShan, W.M., Nguyen, S.V., 2016. The Bacteriophages of S*treptococcus pyogenes*, in: Ferretti, J.J., Stevens, D.L., Fischetti, V.A. (Eds.), *Streptococcus Pyogenes* : Basic Biology to Clinical Manifestations. University of Oklahoma Health Sciences Center, Oklahoma City (OK).

Menard, K.L., Grossman, A.D., 2013. Selective pressures to maintain attachment site specificity of integrative and conjugative elements. PLoS Genet. 9, e1003623. doi:10.1371/journal.pgen.1003623

Mingoia, M., Morici, E., Marini, E., Brenciani, A., Giovanetti, E., Varaldo, P.E., 2016. Macrolide resistance gene erm(TR) and erm(TR)-carrying genetic elements in *Streptococcus agalactiae*: characterization of ICE*SagTR7*, a new composite element containing IME*Sp2907*. J. Antimicrob. Chemother. 71, 593–600. doi:10.1093/jac/dkv408

Mingoia, M., Tili, E., Manso, E., Varaldo, P.E., Montanari, M.P., 2011. Heterogeneity of Tn*5253*-like composite elements in clinical *Streptococcus pneumoniae* isolates. Antimicrob. Agents Chemother. 55, 1453–1459. doi:10.1128/AAC.01087-10

Mitchell, T.J., 2003. The pathogenesis of streptococcal infections: from tooth decay to meningitis. Nat. Rev. Microbiol. 1, 219–230. doi:10.1038/nrmicro771

Miyazaki, R., van der Meer, J.R., 2013. A New Large-DNA-Fragment Delivery System Based on Integrase Activity from an Integrative and Conjugative Element. Appl. Environ. Microbiol. 79, 4440–4447. doi:10.1128/AEM.00711-13

Moran, R.A., Holt, K.E., Hall, R.M., 2016. pCERC3 from a commensal ST95 *Escherichia coli*: A ColV virulence-multiresistance plasmid carrying a sul3-associated class 1 integron. Plasmid 84–85, 11–19. doi:10.1016/j.plasmid.2016.02.002

Mullany, P., Williams, R., Langridge, G.C., Turner, D.J., Whalan, R., Clayton, C., Lawley, T., Hussain, H., McCurrie, K., Morden, N., Allan, E., Roberts, A.P., 2012. Behavior and Target Site Selection of Conjugative Transposon Tn*916* in Two Different Strains of Toxigenic Clostridium difficile. Appl. Environ. Microbiol. 78, 2147–2153. doi:10.1128/AEM.06193-11

Mulvey, M.R., Boyd, D.A., Olson, A.B., Doublet, B., Cloeckaert, A., 2006. The genetics of *Salmonella* genomic island 1. Microbes Infect. 8, 1915–1922. doi:10.1016/j.micinf.2005.12.028

Nesmelova, I.V., Hackett, P.B., 2010. DDE transposases: Structural similarity and diversity. Adv. Drug Deliv. Rev. 62, 1187–1195. doi:10.1016/j.addr.2010.06.006

Nguyen, V.H., Lavenier, D., 2009. PLAST: parallel local alignment search tool for database comparison. BMC Bioinformatics 10, 329. doi:10.1186/1471-2105-10-329

Nunes-Düby, S.E., Kwon, H.J., Tirumalai, R.S., Ellenberger, T., Landy, A., 1998. Similarities and differences among 105 members of the Int family of site-specific recombinases. Nucleic Acids Res. 26, 391–406.

Núñez, B., De La Cruz, F., 2001. Two atypical mobilization proteins are involved in plasmid CloDF13 relaxation. Mol. Microbiol. 39, 1088–1099.

O'Brien, F.G., Ramsay, J.P., Monecke, S., Coombs, G.W., Robinson, O.J., Htet, Z., Alshaikh, F. a. M., Grubb, W.B., 2015. *Staphylococcus aureus* plasmids without mobilization genes are mobilized by a novel conjugative plasmid from community isolates. J. Antimicrob. Chemother. 70, 649–652. doi:10.1093/jac/dku454

O'Brien, F.G., Yui Eto, K., Murphy, R.J.T., Fairhurst, H.M., Coombs, G.W., Grubb, W.B., Ramsay, J.P., 2015. Origin-of-transfer sequences facilitate mobilisation of non-conjugative antimicrobial-resistance plasmids in *Staphylococcus aureus*. Nucleic Acids Res. 43, 7971–7983. doi:10.1093/nar/gkv755

O'Brien, K.L., Nohynek, H., World Health Organization Pneumococcal Vaccine Trials Carraige Working Group, 2003. Report from a WHO working group: standard method for detecting upper respiratory carriage of *Streptococcus pneumoniae*. Pediatr. Infect. Dis. J. 22, 133–140. doi:10.1097/01.inf.0000048676.93549.d1

Ochman, H., Lawrence, J.G., Groisman, E.A., 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405, 299–304. doi:10.1038/35012500

Ou, H.-Y., He, X., Harrison, E.M., Kulasekara, B.R., Thani, A.B., Kadioglu, A., Lory, S., Hinton, J.C.D., Barer, M.R., Deng, Z., Rajakumar, K., 2007. MobilomeFINDER: web-based tools for in

silico and experimental discovery of bacterial genomic islands. Nucleic Acids Res. 35, W97–W104. doi:10.1093/nar/gkm380

Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., Larsson, D.G.J., 2014. BacMet: antibacterial biocide and metal resistance genes database. Nucleic Acids Res. 42, D737–D743. doi:10.1093/nar/gkt1252

Partridge, S.R., Hall, R.M., 2003. The IS*1111* family members IS*4321* and IS*5075* have subterminal inverted repeats and target the terminal inverted repeats of Tn*21* family transposons. J. Bacteriol. 185, 6371–6384.

Pavlovic, G., Burrus, V., Gintz, B., Decaris, B., Guédon, G., 2004. Evolution of genomic islands by deletion and tandem accretion by site-specific recombination: ICE*St1*-related elements from *Streptococcus thermophilus*. Microbiology. 150, 759–774. doi:10.1099/mic.0.26883-0

Persaud, C., Lu, Y., Vila-Sanjurjo, A., Campbell, J.L., Finley, J., O'Connor, M., 2010. Mutagenesis of the modified bases, m(5)U1939 and psi2504, in Escherichia coli 23S rRNA. Biochem. Biophys. Res. Commun. 392, 223–227. doi:10.1016/j.bbrc.2010.01.021

Piednoël, M., Gonçalves, I.R., Higuet, D., Bonnivard, E., 2011. Eukaryote DIRS1-like retrotransposons: an overview. BMC Genomics 12, 621. doi:10.1186/1471-2164-12-621

Pollet, R.M., Ingle, J.D., Hymes, J.P., Eakes, T.C., Eto, K.Y., Kwong, S.M., Ramsay, J.P., Firth, N., Redinbo, M.R., 2016. Processing of Nonconjugative Resistance Plasmids by Conjugation Nicking Enzyme of Staphylococci. J. Bacteriol. 198, 888–897. doi:10.1128/JB.00832-15

Porter, C.J., Bantwal, R., Bannam, T.L., Rosado, C.J., Pearce, M.C., Adams, V., Lyras, D., Whisstock, J.C., Rood, J.I., 2012. The conjugation protein TcpC from *Clostridium perfringens* is structurally related to the type IV secretion system protein VirB8 from Gram-negative bacteria. Mol. Microbiol. 83, 275–288. doi:10.1111/j.1365-2958.2011.07930.x

Puymège, A., Bertin, S., Chuzeville, S., Guédon, G., Payot, S., 2013. Conjugative transfer and cis-mobilization of a genomic island by an integrative and conjugative element of *Streptococcus agalactiae*. J. Bacteriol. 195, 1142–1151. doi:10.1128/JB.02199-12

Puymège, A., Bertin, S., Guédon, G., Payot, S., 2015. Analysis of *Streptococcus agalactiae* pan-genome for prevalence, diversity and functionality of integrative and conjugative or mobilizable elements integrated in the tRNA(Lys CTT) gene. Mol. Genet. Genomics MGG 290, 1727–1740. doi:10.1007/s00438-015-1031-9

Rabinovich, L., Sigal, N., Borovok, I., Nir-Paz, R., Herskovits, A.A., 2012. Prophage excision activates Listeria competence genes that promote phagosomal escape and virulence. Cell 150, 792–802. doi:10.1016/j.cell.2012.06.036

Ralph, A.P., Carapetis, J.R., 2013. Group a streptococcal diseases and their global burden. Curr. Top. Microbiol. Immunol. 368, 1–27. doi:10.1007/82_2012_280

Ramachandran, G., Miguel-Arribas, A., Abia, D., Singh, P.K., Crespo, I., Gago-Córdoba, C., Hao, J.A., Luque-Ortega, J.R., Alfonso, C., Wu, L.J., Boer, D.R., Meijer, W.J.J., 2017. Discovery of a new family of relaxases in Firmicutes bacteria. PLOS Genet. 13, e1006586. doi:10.1371/journal.pgen.1006586

Ramsay, J.P., Firth, N., 2017. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. Curr. Opin. Microbiol. 38, 1–9. doi:10.1016/j.mib.2017.03.003

Ramsay, J.P., Kwong, S.M., Murphy, R.J.T., Yui Eto, K., Price, K.J., Nguyen, Q.T., O'Brien, F.G., Grubb, W.B., Coombs, G.W., Firth, N., 2016. An updated view of plasmid conjugation and mobilization in *Staphylococcus*. Mob. Genet. Elem. 6, e1208317. doi:10.1080/2159256X.2016.1208317

Ramsay, J.P., Sullivan, J.T., Stuart, G.S., Lamont, I.L., Ronson, C.W., 2006. Excision and transfer of the Mesorhizobium loti R7A symbiosis island requires an integrase IntS, a novel recombination directionality factor RdfS, and a putative relaxase RlxS. Mol. Microbiol. 62, 723–734. doi:10.1111/j.1365-2958.2006.05396.x

Rashtchian, A., Dubes, G.R., Booth, S.J., 1982. Transferable resistance to cefoxitin in Bacteroides thetaiotaomicron. Antimicrob. Agents Chemother. 22, 701–703.

Rauch, P.J., De Vos, W.M., 1992. Characterization of the novel nisin-sucrose conjugative transposon Tn5276 and its insertion in Lactococcus lactis. J. Bacteriol. 174, 1280–1287.

Ravatn, R., Studer, S., Springael, D., Zehnder, A.J.B., van der Meer, J.R., 1998. Chromosomal Integration, Tandem Amplification, and Deamplification in Pseudomonas putida F1 of a 105-Kilobase Genetic Element Containing the Chlorocatechol Degradative Genes from Pseudomonas sp. Strain B13. J. Bacteriol. 180, 4360–4369.

Ravenhall, M., Škunca, N., Lassalle, F., Dessimoz, C., 2015. Inferring horizontal gene transfer. PLoS Comput. Biol. 11, e1004095. doi:10.1371/journal.pcbi.1004095

Rawlings, D.E., Tietze, E., 2001. Comparative biology of IncQ and IncQ-like plasmids. Microbiol. Mol. Biol. Rev. 65, 481–496. doi:10.1128/MMBR.65.4.481-496.2001

Redzej, A., Ukleja, M., Connery, S., Trokter, M., Felisberto-Rodrigues, C., Cryar, A., Thalassinos, K., Hayward, R.D., Orlova, E.V., Waksman, G., 2017. Structure of a VirD4 coupling protein bound to a VirB type IV secretion machinery. EMBO J. 36, 3080–3095. doi:10.15252/embj.201796629

Reuther, J., Wohlleben, W., Muth, G., 2006. Modular architecture of the conjugative plasmid pSVH1 from *Streptomyces venezuelae*. Plasmid 55, 201–209. doi:10.1016/j.plasmid.2005.11.007

Rice, L.B., Carias, L.L., 1998. Transfer of Tn*5385*, a composite, multiresistance chromosomal element from *Enterococcus faecalis*. J. Bacteriol. 180, 714–721.

Richards, V.P., Palmer, S.R., Pavinski Bitar, P.D., Qin, X., Weinstock, G.M., Highlander, S.K., Town, C.D., Burne, R.A., Stanhope, M.J., 2014. Phylogenomics and the dynamic genome evolution of the genus *Streptococcus*. Genome Biol. Evol. 6, 741–753. doi:10.1093/gbe/evu048

Roberts, A.P., Chandler, M., Courvalin, P., Guédon, G., Mullany, P., Pembroke, T., Rood, J.I., Smith, C.J., Summers, A.O., Tsuda, M., Berg, D.E., 2008. Revised nomenclature for transposable genetic elements. Plasmid 60, 167–173. doi:10.1016/j.plasmid.2008.08.001

Roberts, A.P., Mullany, P., 2011. Tn916-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. FEMS Microbiol. Rev. 35, 856–871. doi:10.1111/j.1574-6976.2011.00283.x

Roberts, A.P., Mullany, P., 2009. A modular master on the move: the Tn916 family of mobile genetic elements. Trends Microbiol. 17, 251–258. doi:10.1016/j.tim.2009.03.002

Roberts, R.J., Vincze, T., Posfai, J., Macelis, D., 2015. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 43, D298-299. doi:10.1093/nar/gku1046

Rocco, J.M., Churchward, G., 2006. The integrase of the conjugative transposon Tn*916* directs strand- and sequence-specific cleavage of the origin of conjugal transfer, *oriT*, by the endonuclease Orf20. J. Bacteriol. 188, 2207–2213. doi:10.1128/JB.188.6.2207-2213.2006

Ruiz-Masó, J.A., MachóN, C., Bordanaba-Ruiseco, L., Espinosa, M., Coll, M., Del Solar, G., 2015. Plasmid Rolling-Circle Replication. Microbiol. Spectr. 3, PLAS-0035-2014. doi:10.1128/microbiolspec.PLAS-0035-2014

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A., Barrell, B., 2000. Artemis: sequence visualization and annotation. Bioinformatics 16, 944–945. doi:10.1093/bioinformatics/16.10.944

Santoro, F., Oggioni, M.R., Pozzi, G., Iannelli, F., 2010. Nucleotide sequence and functional analysis of the tet (M)-carrying conjugative transposon Tn*5251* of *Streptococcus pneumoniae*. FEMS Microbiol. Lett. 308, 150–158. doi:10.1111/j.1574-6968.2010.02002.x

Schaberg, D.R., Clewell, D.B., Glatzer, L., 1982. Conjugative transfer of R-plasmids from *Streptococcus faecalis* to *Staphylococcus aureus*. Antimicrob. Agents Chemother. 22, 204–207.

Showsh, S.A., Andrews, R.E., 1992. Tetracycline enhances Tn*916*-mediated conjugal transfer. Plasmid 28, 213–224.

Siguier, P., Gourbeyre, E., Varani, A., Ton-Hoang, B., Chandler, M., 2015. Everyman's Guide to Bacterial Insertion Sequences. Microbiol. Spectr. 3, MDNA3-0030-2014. doi:10.1128/microbiolspec.MDNA3-0030-2014

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., Chandler, M., 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34, D32-36. doi:10.1093/nar/gkj014

Smillie, C., Garcillán-Barcia, M.P., Francia, M.V., Rocha, E.P.C., de la Cruz, F., 2010. Mobility of plasmids. Microbiol. Mol. Biol. Rev. 74, 434–452. doi:10.1128/MMBR.00020-10

Smith, C.J., Parker, A.C., 1998. The transfer origin for *Bacteroides* mobilizable transposon Tn*4555* is related to a plasmid family from gram-positive bacteria. J. Bacteriol. 180, 435–439.

Smith, M.C.M., 2015. Phage-encoded Serine Integrases and Other Large Serine Recombinases. Microbiol. Spectr. 3. doi:10.1128/microbiolspec.MDNA3-0059-2014

Smyth, D.S., Robinson, D.A., 2009. Integrative and sequence characteristics of a novel genetic element, ICE*6013*, in *Staphylococcus aureus*. J. Bacteriol. 191, 5964–5975. doi:10.1128/JB.00352-09

Soares, S.C., Abreu, V.A.C., Ramos, R.T.J., Cerdeira, L., Silva, A., Baumbach, J., Trost, E., Tauch, A., Hirata, R., Mattos-Guaraldi, A.L., Miyoshi, A., Azevedo, V., 2012. PIPS: pathogenicity island prediction software. PloS One 7, e30848. doi:10.1371/journal.pone.0030848

Söding, J., 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics. 21, 951–960. doi:10.1093/bioinformatics/bti125

Somkuti, G.A., Steinberg, D.H., 2007. Molecular organization of plasmid pER13 in *Streptococcus thermophilus*. Biotechnol. Lett. 29, 1991–1999. doi:10.1007/s10529-007-9542-z

Sonnhammer, E.L., Eddy, S.R., Durbin, R., 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28, 405–420.

Srinivasan, V., Metcalf, B.J., Knipe, K.M., Ouattara, M., McGee, L., Shewmaker, P.L., Glennen, A., Nichols, M., Harris, C., Brimmage, M., Ostrowsky, B., Park, C.J., Schrag, S.J., Frace, M.A., Sammons, S.A., Beall, B., 2014. *vanG* element insertions within a conserved chromosomal site conferring vancomycin resistance to *Streptococcus agalactiae* and *Streptococcus anginosus*. mBio 5, e01386-01314. doi:10.1128/mBio.01386-14

Stark, W.M., 2017. Making serine integrases work for us. Curr. Opin. Microbiol. 38, 130–136. doi:10.1016/j.mib.2017.04.006

Stark, W.M., 2014. The Serine Recombinases. Microbiol. Spectr. 2. doi:10.1128/microbiolspec.MDNA3-0046-2014

Steen, J.A., Bannam, T.L., Teng, W.L., Devenish, R.J., Rood, J.I., 2009. The putative coupling protein TcpA interacts with other pCW3-encoded proteins to form an essential part of the conjugation complex. J. Bacteriol. 191, 2926–2933. doi:10.1128/JB.00032-09

Stragier, P., Kunkel, B., Kroos, L., Losick, R., 1989. Chromosomal rearrangement generating a composite gene for a developmental transcription factor. Science 243, 507–512.

Straume, D., Stamsås, G.A., Håvarstein, L.S., 2015. Natural transformation and genome evolution in *Streptococcus pneumoniae*. Infect. Genet. Evol. 33, 371–380. doi:10.1016/j.meegid.2014.10.020

Sullivan, J.T., Patrick, H.N., Lowther, W.L., Scott, D.B., Ronson, C.W., 1995. Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment. Proc. Natl. Acad. Sci. U. S. A. 92, 8985–8989.

Szpirer, C.Y., Faelen, M., Couturier, M., 2001. Mobilization function of the pBHR1 plasmid, a derivative of the broad-host-range plasmid pBBR1. J. Bacteriol. 183, 2101–2110. doi:10.1128/JB.183.6.2101-2110.2001

te Poele, E.M., Bolhuis, H., Dijkhuizen, L., 2008. Actinomycete integrative and conjugative elements. Antonie Van Leeuwenhoek 94, 127–143. doi:10.1007/s10482-008-9255-x

Teng, W.L., Bannam, T.L., Parsons, J.A., Rood, J.I., 2008. Functional characterization and localization of the TcpH conjugation protein from *Clostridium perfringens*. J. Bacteriol. 190, 5075–5086. doi:10.1128/JB.00386-08

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J.B., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M., 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." Proc. Natl. Acad. Sci. U. S. A. 102, 13950–13955. doi:10.1073/pnas.0506758102

Thoma, L., Muth, G., 2016. Conjugative DNA-transfer in *Streptomyces*, a mycelial organism. Plasmid 87–88, 1–9. doi:10.1016/j.plasmid.2016.09.004

Thomas, J., Lee, C.A., Grossman, A.D., 2013. A conserved helicase processivity factor is needed for conjugation and replication of an integrative and conjugative element. PLoS Genet. 9, e1003198. doi:10.1371/journal.pgen.1003198

Thompson, J.K., Collins, M.A., 1988. Evidence for the conjugal transfer of the broad host range plasmid pIP501 into strains of *Lactobacillus helveticus*. J. Appl. Bacteriol. 65, 309–319.

Toussaint, A., Merlin, C., 2002. Mobile elements as a combination of functional modules. Plasmid 47, 26–35. doi:10.1006/plas.2001.1552

Toussaint, A., Rice, P.A., 2017. Transposable phages, DNA reorganization and transfer. Curr. Opin. Microbiol. 38, 88–94. doi:10.1016/j.mib.2017.04.009

van Heel, A.J., de Jong, A., Montalbán-López, M., Kok, J., Kuipers, O.P., 2013. BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. Nucleic Acids Res. 41, W448–W453. doi:10.1093/nar/gkt391

Vernikos, G.S., Parkhill, J., 2008. Resolving the structural features of genomic islands: a machine learning approach. Genome Res. 18, 331–342. doi:10.1101/gr.7004508

Vernikos, G.S., Parkhill, J., 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics. 22, 2196–2203. doi:10.1093/bioinformatics/btl369

Vijayakumar, M.N., Priebe, S.D., Guild, W.R., 1986. Structure of a conjugative element in *Streptococcus pneumoniae*. J. Bacteriol. 166, 978–984.

Vos, M., Hesselman, M.C., te Beek, T.A., van Passel, M.W.J., Eyre-Walker, A., 2015. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? Trends Microbiol. 23, 598–605. doi:10.1016/j.tim.2015.07.006

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W.F., Surovcik, K., Meinicke, P., Merkl, R., 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics 7, 142. doi:10.1186/1471-2105-7-142

Waksman, G., Orlova, E.V., 2014. Structural organisation of the type IV secretion systems. Curr. Opin. Microbiol. 17, 24–31. doi:10.1016/j.mib.2013.11.001

Waldor, M.K., 2010. Mobilizable genomic islands: going mobile with *oriT* mimicry. Mol. Microbiol. 78, 537–540.

Waldor, M.K., Tschäpe, H., Mekalanos, J.J., 1996. A new type of conjugative transposon encodes resistance to sulfamethoxazole, trimethoprim, and streptomycin in *Vibrio cholerae* O139. J. Bacteriol. 178, 4157–4165.

Walldén, K., Williams, R., Yan, J., Lian, P.W., Wang, L., Thalassinos, K., Orlova, E.V., Waksman, G., 2012. Structure of the VirB4 ATPase, alone and bound to the core complex of a type IV secretion system. Proc. Natl. Acad. Sci. U. S. A. 109, 11348–11353. doi:10.1073/pnas.1201428109

Wang, H., Mullany P. 2000. The large resolvase TndX is required and sufficient for integration and excision of derivatives of the novel conjugative transposon Tn5397. J. Bacteriol. 182, 6577-6583.

Williams, K.P., 2002. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res. 30, 866–875.

Williamson, C.M., Pullinger, G.D., Lax, A.J., 1990. Identification of proteins expressed by the essential virulence region of the *Salmonella* dublin plasmid. Microb. Pathog. 9, 61–66.

Wisniewski, J.A., Traore, D.A., Bannam, T.L., Lyras, D., Whisstock, J.C., Rood, J.I., 2016. TcpM: a novel relaxase that mediates transfer of large conjugative plasmids from *Clostridium perfringen*s. Mol. Microbiol. 99, 884–896. doi:10.1111/mmi.13270

Wright, L.D., Grossman, A.D., 2016. Autonomous Replication of the Conjugative Transposon Tn*916*. J. Bacteriol. 198, 3355–3366. doi:10.1128/JB.00639-16

Wright, L.D., Johnson, C.M., Grossman, A.D., 2015. Identification of a Single Strand Origin of Replication in the Integrative and Conjugative Element ICE*Bs1* of *Bacillus subtilis*. PLoS Genet. 11, e1005556. doi:10.1371/journal.pgen.1005556

Yuan, P., Gupta, K., Duyne, G.D.V., 2008. Tetrameric Structure of a Serine Integrase Catalytic Domain. Structure 16, 1275–1286. doi:10.1016/j.str.2008.04.018

Zadoks, R.N., Middleton, J.R., McDougall, S., Katholm, J., Schukken, Y.H., 2011. Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. J. Mammary Gland Biol. Neoplasia 16, 357–372. doi:10.1007/s10911-011-9236-y

Zhang, Y., Loria, R., 2017. Emergence of Novel Pathogenic *Streptomyces* Species by Site-Specific Accretion and cis-Mobilization of Pathogenicity Islands. Mol. Plant-Microbe Interact. MPMI 30, 72–82. doi:10.1094/MPMI-09-16-0190-R

Zhou, Y., Liang, Y., Lynch, K.H., Dennis, J.J., Wishart, D.S., 2011. PHAST: A Fast Phage Search Tool. Nucleic Acids Res. 39, W347–W352. doi:10.1093/nar/gkr485

Zúñiga, M., Pardo, I., Ferrer, S., 2003. Conjugative plasmid pIP501 undergoes specific deletions after transfer from *Lactococcus lactis* to *Oenococcus oeni*. Arch. Microbiol. 180, 367–373. doi:10.1007/s00203-003-0599-3

ANNEXES

Annexe 1 : Liste des génomes analysés.

Organisme	Taille (Mb)	N° Genbank	Lot de génome
Streptococcus agalactiae 2603V/R	2,2	AE009948.1	Lot 1
Streptococcus agalactiae A909	2,1	CP000114.1	Lot 1
Streptococcus agalactiae NEM316	2,2	AL732656.1	Lot 1
Streptococcus dysgalactiae subsp.*equisimilis* 12394	2,2	CP002215.1	Lot 1
Streptococcus dysgalactiae subsp. *equisimilis* GGS_124	2,1	AP010935.1	Lot 1
Streptococcus equi subsp. equi 4047	2,3	FM204883.1	Lot 1
Streptococcus equi subsp. *zooepidemicus* ATCC 35246	2,2	CP002904	Lot 1
Streptococcus equi subsp. *zooepidemicus* MGCS10565	2,0	CP001129.1	Lot 1
Streptococcus equi subsp. *zooepidemicus* str. H70	2,2	FM204884.1	Lot 1
Streptococcus gallolyticus ATCC 43143	2,4	AP012053	Lot 1
Streptococcus gallolyticus UCN34	2,4	FN597254.1	Lot 1
Streptococcus gordonii str. Challis substr. CH1	2,2	CP000725.1	Lot 1
Streptococcus mitis B6	2,2	FN568063.1	Lot 1
Streptococcus mutans NN2025	2	AP010655.1	Lot 1
Streptococcus mutans UA159	2	AE014133.1	Lot 1
Streptococcus parasanguinis ATCC 15912	2,2	CP002843	Lot 1
Streptococcus parauberis KCTC 11537	2,1	CP002471.1	Lot 1
Streptococcus pasteurianus ATCC 43144	2,1	AP012054.1	Lot 1
Streptococcus pneumoniae 670-6B	2,2	CP002176.1	Lot 1
Streptococcus pneumoniae 70585	2,2	CP000918.1	Lot 1
Streptococcus pneumoniae AP200	2,1	CP002121.1	Lot 1
Streptococcus pneumoniae CGSP14	2,2	CP001033.1	Lot 1
Streptococcus pneumoniae D39	2,1	CP000410.1	Lot 1
Streptococcus pneumoniae G54	2,1	CP001015.1	Lot 1
Streptococcus pneumoniae Hungary19A-6	2,3	CP000936.1	Lot 1
Streptococcus pneumoniae INV104	2,1	FQ312030	Lot 1
Streptococcus pneumoniae INV200	2,1	FQ312029	Lot 1
Streptococcus pneumoniae JJA	2,1	CP000919.1	Lot 1
Streptococcus pneumoniae OXC141	2	FQ312027	Lot 1
Streptococcus pneumoniae P1031	2,1	CP000920.1	Lot 1
Streptococcus pneumoniae R6	2	AE007317.1	Lot 1
Streptococcus pneumoniae Taiwan19F-14	2,1	CP000921.1	Lot 1
Streptococcus pneumoniae TCH8431/19A	2,1	CP001993.1	Lot 1
Streptococcus pneumoniae TIGR4	2,2	AE005672.3	Lot 1
Streptococcus pseudopneumoniae IS7493	2,2	CP002925	Lot 1
Streptococcus pyogenes Alab49	1,8	CP003068.1	Lot 1
Streptococcus pyogenes M1 GAS	1,9	AE004092.1	Lot 1
Streptococcus pyogenes MGAS10270	1,9	CP000260.1	Lot 1
Streptococcus pyogenes MGAS10394	1,9	CP000003.1	Lot 1

Organisme	Taille (Mb)	N° Genbank	Lot de génome
Streptococcus pyogenes MGAS10750	2	CP000262.1	Lot 1
Streptococcus pyogenes MGAS2096	1,9	CP000261.1	Lot 1
Streptococcus pyogenes MGAS315	1,9	AE014074.1	Lot 1
Streptococcus pyogenes MGAS5005	1,8	CP000017.1	Lot 1
Streptococcus pyogenes MGAS6180	1,9	CP000056.1	Lot 1
Streptococcus pyogenes MGAS8232	1,9	AE009949.1	Lot 1
Streptococcus pyogenes MGAS9429	1,8	CP000259.1	Lot 1
Streptococcus pyogenes NZ131	1,8	CP000829.1	Lot 1
Streptococcus pyogenes SSI-1	1,9	BA000034.2	Lot 1
Streptococcus pyogenes str. Manfredo	1,9	AM295007.1	Lot 1
Streptococcus salivarius 57.I	2,2	CP002888	Lot 1
Streptococcus salivarius CCHSS3	2,2	FR873481.1	Lot 1
Streptococcus salivarius JIM8777	2,2	FR873482	Lot 1
Streptococcus sanguinis SK36	2,4	CP000387.1	Lot 1
Streptococcus suis 05ZYH33	2,1	CP000407.1	Lot 1
Streptococcus suis 98HAH33	2,1	CP000408.1	Lot 1
Streptococcus suis A7	2	CP002570	Lot 1
Streptococcus suis BM407	2,2	FM252032.1	Lot 1
Streptococcus suis D12	2,2	CP002644	Lot 1
Streptococcus suis D9	2,2	CP002641	Lot 1
Streptococcus suis GZ1	2	CP000837	Lot 1
Streptococcus suis JS14	2,1	CP002465	Lot 1
Streptococcus suis P1/7	2	AM946016.1	Lot 1
Streptococcus suis SC84	2,1	FM252031.1	Lot 1
Streptococcus suis SS12	2,1	CP002640	Lot 1
Streptococcus suis ST1	2	CP002651	Lot 1
Streptococcus suis ST3	2	CP002633.1	Lot 1
Streptococcus thermophilus CNRZ1066	1,8	CP000024.1	Lot 1
Streptococcus thermophilus JIM 8232	1,9	FR875178.1	Lot 1
Streptococcus thermophilus LMD-9	1,9	CP000419.1	Lot 1
Streptococcus thermophilus LMG 18311	1,8	CP000023.1	Lot 1
Streptococcus thermophilus ND03	1,8	CP002340	Lot 1
Streptococcus uberis 0140J	1,9	AM946015.1	Lot 1
Streptococcus agalactiae 09mas018883	2,1	HF952104.1	Lot 2
Streptococcus agalactiae GD201008-001	2,1	CP003810.1	Lot 2
Streptococcus agalactiae ILRI005	2,1	HF952105.1	Lot 2
Streptococcus agalactiae ILRI112	2	HF952106.1	Lot 2
Streptococcus agalactiae SA20-06	1,8	CP003919.1	Lot 2
Streptococcus anginosus C1051	1,9	CP003860.1	Lot 2
Streptococcus anginosus C238	2,2	CP003861.1	Lot 2
Streptococcus constellatus subsp. *pharyngis* C818	2	CP003840.1	Lot 2
Streptococcus dysgalactiae subsp. *equisimilis* 167	2	AP012976.1	Lot 2

Organisme	Taille (Mb)	N° Genbank	Lot de génome
Streptococcus dysgalactiae subsp. equisimilis AC-2713	2,2	HE858529.1	Lot 2
Streptococcus dysgalactiae subsp. equisimilis RE378	2,2	AP011114.1	Lot 2
Streptococcus gallolyticus subsp. gallolyticus ATCC 2069	2,4	FR824043.1	Lot 2
Streptococcus infantarius subsp. infantarius CJ18	2	CP003295.1	Lot 2
Streptococcus iniae SF1	2,2	CP005941.1	Lot 2
Streptococcus intermedius B196	2	CP003857.1	Lot 2
Streptococcus intermedius C270	2	CP003858.1	Lot 2
Streptococcus intermedius JTH08	2	AP010969.1	Lot 2
Streptococcus lutetiensis 033	2	CP003025.1	Lot 2
Streptococcus macedonicus ACA-DC 198	2,1	HE613569.1	Lot 2
Streptococcus mutans GS-5	2	CP003686.1	Lot 2
Streptococcus mutans LJ23	2	AP012336.1	Lot 2
Streptococcus oligofermentans AS 1.3089	2,1	CP004409.1	Lot 2
Streptococcus oralis Uo5	2	FR720602.1	Lot 2
Streptococcus parasanguinis FW213	2,2	CP003122.1	Lot 2
Streptococcus pneumoniae A026	2,1	CP006844.1	Lot 2
Streptococcus pneumoniae ATCC 700669	2,2	FM211187.1	Lot 2
Streptococcus pneumoniae gamPNI0373	2,1	CP001845.1	Lot 2
Streptococcus pneumoniae PCS8235	2,1	CM001835.1	Lot 2
Streptococcus pneumoniae SPN032672	2,1	FQ312039	Lot 2
Streptococcus pneumoniae SPN033038	2,1	FQ312042	Lot 2
Streptococcus pneumoniae SPN034156	2	FQ312045	Lot 2
Streptococcus pneumoniae SPN034183	2	FQ312043	Lot 2
Streptococcus pneumoniae SPN994038	2	FQ312041	Lot 2
Streptococcus pneumoniae SPN994039	2	FQ312044	Lot 2
Streptococcus pneumoniae SPNA45	2,1	HE983624.1	Lot 2
Streptococcus pneumoniae ST556	2,2	CP003357.1	Lot 2
Streptococcus pyogenes A20	1,9	CP003901.1	Lot 2
Streptococcus pyogenes HSC5	1,8	CP006366.1	Lot 2
Streptococcus pyogenes M1 476	1,8	AP012491.2	Lot 2
Streptococcus pyogenes MGAS15252	1,8	CP003116.1	Lot 2
Streptococcus pyogenes MGAS1882	1,8	CP003121.1	Lot 2
Streptococcus sp. I-G2	2	CP006805.1	Lot 2
Streptococcus sp. I-P16	2	CP006776.1	Lot 2
Streptococcus suis S735	2	CP003736.1	Lot 2
Streptococcus suis SC070731	2,1	CP003922.1	Lot 2
Streptococcus suis T15	2,2	CP006246.1	Lot 2
Streptococcus suis TL13	2	CP003993.1	Lot 2
Streptococcus suis YB51	2	CP006645.1	Lot 2
Streptococcus anginosus subsp. whileyi MAS624	2,1	AP013072.1	Lot 2
Streptococcus constellatus subsp. pharyngis C1050	2	CP003859.1	Lot 2

Annexe 2 : Script utilisé par ICEFinder

```
#!/usr/bin/python
# -*- coding: latin-1 -*-
import os
import sys
import argparse
import operator, csv
from operator import itemgetter
from Bio import Entrez, SeqIO, SeqFeature
from Bio.Alphabet import IUPAC
pathname = os.getcwd()
listdir = str(pathname)+'/Fichier_a_traiter'
listdir_db = str(pathname)+'/DataBase_HMM'
listdir_gb = str(pathname)+'/Genbank'

def DataBase_press():
    Database = os.listdir(listdir_db)
    for i in Database :
        DataBase_links = listdir_db + '/' + i
        profils_list = os.listdir(DataBase_links)
        for j in profils_list:
            k=j.split('.')
            K = k[-1]
            z = j + ".h3i"
            if K == 'hmm' and z not in profils_list:
                HMMpress = 'hmmpress %s/%s/%s' % (listdir_db,i,j)
                print(str(HMMpress))
                os.system(str(HMMpress))
```

```
def genbank_convert():
    fichiers_Genbank = os.listdir(listdir_gb)
    for i in fichiers_Genbank:
        j = os.path.splitext(i)[0]
        j = j.replace(' ', '_')
        genbank_path = '%s/%s'%(listdir_gb,i)
        print(genbank_path)
        input_handle = open(genbank_path, "rU")
        for seq_record in SeqIO.parse(input_handle, "genbank") :
            Fasta_output_temp = '%s/Fichier_a_traiter/%s.faa'%(pathname,j)
            print(Fasta_output_temp)
            Fasta = open(Fasta_output_temp, 'a')
            for feature in seq_record.features:
                x = feature.location.start
                y = feature.location.end
                z = feature.location.strand
                CDS_seq = 'Genbank file is not a Genbank full'
                protid = ''
                geneproduct = ''
                if z == 1:
                    a = "+" + '|' + str(x) + ".." + str(y)
                if z == -1:
                    a = "-" + '|' + str(x) + ".." + str(y)
                if feature.type == "CDS":
                    if 'translation' in feature.qualifiers:
                        CDS_seq = feature.qualifiers['translation'][0]
                    if str(CDS_seq) == 'Genbank file is not a Genbank full':
```

```
                                        nuc_seq =
feature.location.extract(seq_record).seq
                                        CDS_seq = nuc_seq.translate()
                                    if 'product' in feature.qualifiers:
                                        geneproduct = feature.qualifiers['product'][0]
                                        geneproduct = geneproduct.replace(',', '')
                                        geneproduct = geneproduct.replace(' ', '_')
                                    if 'db_xref' in feature.qualifiers:
                                        protid = feature.qualifiers['db_xref'][0]
                                        protid = protid.replace(',', '')
                                        protid = protid.replace(' ', '_')
                                    if 'locus_tag' in feature.qualifiers :
                                        protid = feature.qualifiers['locus_tag'][0]
                                        protid = protid.replace(',', '')
                                        protid = protid.replace(' ', '_')
                                    if 'label' in feature.qualifiers:
                                        protid = feature.qualifiers['label'][0]
                                        protid = protid.replace(',', '')
                                        protid = protid.replace(' ', '_')
                                    if 'protein_id' in feature.qualifiers:
                                        protid = feature.qualifiers['protein_id'][0]
                                        protid = protid.replace(',', '')
                                        protid = protid.replace(' ', '_')
                                    Fasta.write(">" + j + "_cdsid_" + protid + "|" + str(a) +
"\n" + str(CDS_seq) + "\n")
                        Fasta.close()
```

```
def Ouput_gather(input):
    fichiers_a_traiter = os.listdir(input)
    for i in fichiers_a_traiter:
        j = os.path.splitext(i)[0]
        j = j.replace(' ', '_')
        output_links = str(pathname) +'/Fichiers_traites' + '/' + str(j) +
'/HMMscan_Output'
        dir1 = str(pathname) +'/Fichiers_traites' + '/' + str(j) +
'/HMM_results_compiled'
        try:
            os.makedirs(dir1)
        except OSError:
            print('Directory already exists')
            pass
        hmm_output_list = os.listdir(output_links)
        print(hmm_output_list)
        for k in hmm_output_list:
            out_put_newfile_link = dir1 +'/'+ str(j)+'.csv'
            out_put_newfile = open (out_put_newfile_link, 'a')
            outputfiles_link = str(output_links) +'/'+ str(k)
            print(outputfiles_link)
            handle = open(outputfiles_link, 'rU')
            hit_name = ''
            align_lenght_hmm = 0
            align_lenght_hit= 0
            for line in handle:
                if line[0] != '#':
                    raw_results = str(line)
                    split_results = raw_results.split() #black magic to
remove all useless space
```

```
                query_info = split_results[3]
                query_info = query_info.split('|')
                query_ID = query_info[0] #ID du hit - oui du hit !
                query_strand = query_info[1] #strand du hit
                query_pos = query_info[2]
                query_pos = query_pos.split('..')
                query_start = query_pos[0] #start hit - oui du hit !
                query_end = query_pos[1] #end query - oui du hit !
                query_len = split_results[5] #longueur du hit
                evalue = split_results[6]#evalue
                hmm_name = split_results[0]#nom du profil qui trouve
le hit
                hmm_lenght = split_results[2]#longueur du profil, pour
faire le pourcentage de couverture
                hmm_name2 = hmm_name.split('_')
                num_info = len(hmm_name2)
                prot_type = hmm_name2[0]
                prot_info = ''
                for count in range(1,num_info):
                                prot_info = str(hmm_name2[count])+'-
'+str(prot_info)
                if str(hit_name) != str(query_info):
                        align_lenght_hmm = int(split_results[16])-
int(split_results[15])
                        print (int(split_results[16]))
                        print (int(split_results[15]))
                        align_lenght_hit = int(split_results[18])-
int(split_results[17])
                        hit_name = str(query_info)
                        print (str(split_results[9] == split_results[10]))
```

```
                                        if split_results[9] == split_results[10]:
                                                print(str('rtinging'))
                                                line =
(str(query_ID)+','+str(query_strand)+','+str(query_start)+','+str(query_end)+','+str(query_len
)+','+str(evalue)+','+str(prot_type)+','+str(prot_info[:-
1])+','+str(hmm_lenght)+','+str(align_lenght_hit)+','+str(align_lenght_hmm)+'\n')
                                                out_put_newfile.write(str(line))
                                        else :
                                                temp_long_hmm = int(split_results[16])-
int(split_results[15])
                                                temp_long_hit = int(split_results[18])-
int(split_results[17])
                                                align_lenght_hit = int(align_lenght_hit) +
int(temp_long_hit)
                                                align_lenght_hmm = int(align_lenght_hmm) +
int(temp_long_hmm)
                                                hit_name = str(query_info)
                                                print (str(split_results[9] == split_results[10]))
                                                if split_results[9] == split_results[10]:
                                                        print(str('rtingin'))
                                                        line =
(str(query_ID)+','+str(query_strand)+','+str(query_start)+','+str(query_end)+','+str(query_len
)+','+str(evalue)+','+str(prot_type)+','+str(prot_info[:-
1])+','+str(hmm_lenght)+','+str(align_lenght_hit)+','+str(align_lenght_hmm)+'\n')
                                                        out_put_newfile.write(str(line))
        out_put_newfile.close()
```

def hmmfilter(QCover, HCover, input, evalue):

```
        print(str('QCover ='+str(QCover)))
        fichiers_a_traiter = os.listdir(input)
        for i in fichiers_a_traiter:
```

```
		j = os.path.splitext(i)[0]
		j = j.replace(' ', '_')
		HMM_Output_path = '%s/Fichiers_traites/%s/HMM_results_compiled'%(pathname,j)
		HMM_Output = os.listdir(HMM_Output_path)
		for x in HMM_Output:
			a = ''
			b = ''
			y=os.path.splitext(x)[0]
			y=y.replace(' ', '_')
			HMM_Output_csv_path = '%s/Fichiers_traites/%s/HMM_results_compiled/%s'%(pathname,j,x)
			HMM_Output_csv_temp = open(HMM_Output_csv_path,'r')
			HMM_Output_csv = csv.reader(HMM_Output_csv_temp, delimiter=',')
			try :
				HMM_Output_csv_sorted = sorted(HMM_Output_csv, key=lambda p: (float(p[5])), reverse=False)
			except ValueError:
				HMM_Output_csv_sorted = sorted(HMM_Output_csv, key=lambda p: (int(p[5])), reverse=False)
			HMM_Output_csv_sorted.sort(key=operator.itemgetter(0))
			output_temp = '%s/Fichiers_traites/%s/HMM_results_compiled/%s_sorted.csv'%(pathname,j,y)
			output = open(output_temp,'w')
			for line in HMM_Output_csv_sorted:
				a = line[0]
				if a != b:
					if str(line[6]) == 'Couplage':
						if float(line[4]) >= 180 and float(line[5]) <= float(evalue) and float(line[10])/float(line[8])*100 >= int(QCover) and float(line[9])/float(line[4])*100 >= int(HCover):#En gros l'équivalent des filtres = l'alignement à 40% de couverture avec les query et subject, le % d'identité et la taille du hit (query ici)
```

```
                                        newline = str(line)
                                        newline = newline.replace('[','')#Il faut enlever les '[','[' et les "'" parce que le fait de trier le csv rajoute ces caractères et si on les laisse ils vont s'accumuler tout du long
                                        newline = newline.replace(']','')
                                        newline = newline.replace("'",'')
                                        newline = newline.replace(" ","")
                                        output.write(newline+'\n')
                                        b = line[0]
                                elif str(line[6]) == 'Relaxase':
                                        if float(line[4]) >= 180 and float(line[5]) <= float(evalue) and float(line[10])/float(line[8])*100 >= int(QCover) and float(line[9])/float(line[4])*100 >= int(HCover):# l'équivalent des filtres = l'alignement à 40% de couverture avec les query et subject, le % d'identité et la taille du hit (query ici)
                                        newline = str(line)
                                        newline = newline.replace('[','')#Il faut enlever les '[','[' et les "'" parce que le fait de trier le csv rajoute ces caractères et si on les laisse ils vont s'accumuler tout du long
                                        newline = newline.replace(']','')
                                        newline = newline.replace("'",'')
                                        newline = newline.replace(" ","")
                                        output.write(newline+'\n')
                                        b = line[0]
                                elif str(line[6]) == 'VirB4':
                                        if float(line[4]) >= 500 and float(line[5]) <= float(evalue) and float(line[10])/float(line[8])*100 >= int(QCover) and float(line[9])/float(line[4])*100 >= int(HCover):# l'équivalent des filtres = l'alignement à 40% de couverture avec les query et subject, le % d'identité et la taille du hit (query ici)
                                        newline = str(line)
                                        newline = newline.replace('[','')#Il faut enlever les '[','[' et les "'" parce que le fait de trier le csv rajoute ces caractères et si on les laisse ils vont s'accumuler tout du long
                                        newline = newline.replace(']','')
```

```
                                        newline = newline.replace("'",'')
                                        newline = newline.replace(" ","")
                                        output.write(newline+'\n')
                                        b = line[0]
                                else :
                                        if float(line[4]) >= 320 and float(line[5]) <= float(evalue) and float(line[10])/float(line[8])*100 >= int(QCover) and float(line[9])/float(line[4])*100 >= int(HCover):# l'équivalent des filtres = l'alignement à 40% de couverture avec les query et subject, le % d'identité et la taille du hit (query ici)
                                                newline = str(line)
                                                newline = newline.replace('[','')#Il faut enlever les '[','[' et les "'" parce que le fait de trier le csv rajoute ces caractères et si on les laisse ils vont s'accumuler tout du long
                                                newline = newline.replace(']','')
                                                newline = newline.replace("'",'')
                                                newline = newline.replace(" ","")
                                                output.write(newline+'\n')
                                                b = line[0]
                        output.close()
                        input_temp = '%s/Fichiers_traites/%s/HMM_results_compiled/%s_sorted.csv'%(pathname,j,y)
                        input_csv = open(input_temp,'r')
                        HMM_Output_sorting = csv.reader(input_csv, delimiter=',')
                        try :
                                Sorted_result = sorted(HMM_Output_sorting,  key=lambda p: int(p[2]), reverse=False)
                        except ValueError:
                                print ('boloss')
                        output2 = open(input_temp,'w')
                        output2.write('Hit_name,Hit_strand,Hit_start,Hit_end,Hit_length,e-value,protein_type,Query_family,Query_length,Cover_percentage_Query,Cover_percentage_Hit'+'\n')
```

```
                for line in Sorted_result :
                        newline = str(line)
                        newline = newline.replace('[','')#Il faut enlever les '[','[' et les """ parce que le fait de trier le csv rajoute ces caractères et si on les laisse ils vont s'accumuler tout du long
                        newline = newline.replace(']','')
                        newline = newline.replace("""','')
                        newline = newline.strip(' ')
                        newline = newline.split(',')
                        coverhit = (float(newline[9])/float(newline[4]))*100
                        coverquery = (float(newline[10])/float(newline[8]))*100
                        coverhit = round(coverhit,2)
                        coverquery = round(coverquery,2)
                        newline = str(newline[0]+','+newline[1]+','+newline[2]+','+newline[3]+','+newline[4]+','+newline[5]+','+newline[6]+','+newline[7]+','+newline[8]+','+str(coverquery)+','+str(coverhit))
                        output2.write(str(newline)+'\n')
                output2.close()
```

def hmmscan(input):

```
        fichiers_a_traiter = os.listdir(input)
        print (str(fichiers_a_traiter))
        proteins_type = os.listdir(listdir_db)
        for i in fichiers_a_traiter :
                j = os.path.splitext(i)[0]
                j = j.replace(' ', '_')
                dir1 = '%s/Fichiers_traites/%s'%(pathname,j)
                dir2 = '%s/Fichiers_traites/%s/HMMscan_Output'%(pathname,j)
                try:
                        os.makedirs(dir1)
```

```
                except OSError:
                        print('Directory already exists')
                        pass
                try:
                        os.makedirs(dir2)
                except OSError:
                        print('Directory already exists')
                        pass
                for k in proteins_type :
                        DataBase_links = listdir_db + '/' + k
                        profils_list = os.listdir(DataBase_links)
                        for l in profils_list:
                                m=l.split('.')
                                M = m[-1]
                                n = m[0]
                                if M == 'hmm':
                                        HMMscan = 'hmmscan --domtblout
%s/Fichiers_traites/%s/HMMscan_Output/%s_%s.txt %s/%s %s/Fichier_a_traiter/%s' %
(pathname,j,j,n,DataBase_links,l,pathname,i)
                                        print(str(HMMscan))
                                        os.system(str(HMMscan))
        print('Scan done')


def main(args):
        if args.input == None:
                input = listdir


        if args.input != None:
                input = args.input
```

```
	QCover = args.QCover
	HCover = args.HCover
	evalue = args.evalue
	inType = args.inType
	if args.inType == 'prot':
		DataBase_press()
		hmmscan(input)
		Ouput_gather(input)
		hmmfilter(QCover, HCover, input, evalue)
	if args.inType == 'gb':
		DataBase_press()
		genbank_convert()
		hmmscan(input)
		Ouput_gather(input)
		hmmfilter(QCover, HCover, input, evalue)



def run():
	parser = argparse.ArgumentParser(description='ICEFinder')
	parser.add_argument('-i','--input', dest="input", help="Input file(s) must be Genbank (.gb) Full (with proteins sequences or genome dna sequence), default: Genbank folder")
	parser.add_argument('-H','--HitCover', default="40", dest="HCover", help="Cover percentage on hit threshold, low cover percentage threshold can induce false positive (<25), default: 40")
	parser.add_argument('-Q', '--QueryCover',  dest="QCover", default="40", help="Cover percentage on query threshold threshold, low identity percentage threshold can induce false positive (<25), default: 40")
	parser.add_argument('-t', '--inType',  dest="inType", default="gb", help="input type, must be either 'gb' for Genbank (.gb) or 'prot' for fasta/multifasta proteins (.faa)")
	parser.add_argument('-e', '--evalue', default ='10E-5',  dest="evalue", help="set the evalue threshold exemple: 10E-10, default : 10E-5")
```

```
        args = parser.parse_args()
        main(args)

if __name__ == '__main__':
        run()
```