Conception et mise en oeuvre d'une approche bioinformatique dédiée à l'identification des ICE, IME et éléments composites dans les génomes de Firmicutes

Julie Lao

Ecole Doctorale SIReNA (Sciences et Ingénierie des Ressources Naturelles)

Thèse

Présentée et soutenue publiquement pour l'obtention du titre de

DOCTEUR DE l'UNIVERSITE DE LORRAINE

Mention : « Ecotoxicologie, Biodiversité, Ecosystèmes »

par **Julie LAO**

Conception et mise en œuvre d'une approche bioinformatique dédiée à l'identification des ICE, IME et éléments composites dans les génomes de Firmicutes

22 février 2021

Membres du jury :

Directrices de thèse :

Mme Nathalie LEBLOND-BOURGET	**Professeur, Université de Lorraine, Vandœuvre-Lès-Nancy, Directrice**
Mme Hélène CHIAPELLO	**Ingénieure de recherche, INRAE, Jouy-en-Josas, Co-directrice**

Rapporteurs :

Mme Christine CITTI	**Directrice de recherche, INRAE, Toulouse**
M. David VALLENET	**Directeur de recherche, CEA, Evry**

Examinateurs :

Mme Marie-Dominique DEVIGNES	**Chargée de recherche, LORIA, Vandœuvre-Lès-Nancy**
M. Vincent BURRUS	**Professeur, Université de Sherbrooke, Québec**
Mme Nathalie LEBLOND-BOURGET	**Professeur, Université de Lorraine, Vandœuvre-Lès-Nancy**
Mme Hélène CHIAPELLO	**Ingénieure de recherche, INRAE, Jouy-en-Josas**

Membres invités :

M. Gérard GUÉDON	**Maître de conférence, Université de Lorraine, Vandœuvre-Lès-Nancy**
M. Thomas LACROIX	**Ingénieur d'étude, INRAE, Jouy-en-Josas**

Laboratoire Dynamique des Génomes et Adaptation Microbienne – UMR 1128
INRAE – Université de Lorraine, Faculté des Sciences et Technologies
Boulevard des Aiguillettes, 54506 Vandœuvre-lès-Nancy

Laboratoire Mathématiques et Informatique Appliquées du Génome à l'Environnement – UR 1404
INRAE – Domaine de Vilvert, 78350 Jouy-en-Josas

Remerciements

Je voudrais d'abord remercier les membres du jury pour avoir accepté d'évaluer mon travail.

Je remercie Bertrand et Sophie de m'avoir accueilli au sein de leur laboratoire et de m'avoir permis de réaliser cette thèse dans des superbes conditions. Je vous remercie aussi pour votre bienveillance.

Je remercie plus particulièrement mes directrices Nathalie et Hélène de m'avoir encadré et fait confiance pour ce projet. Je vous remercie pour votre bienveillance et surtout d'avoir été très patiente avec moi. Votre soutien dans les moments difficiles et tout le long de la thèse m'ont été indispensables. Je n'aurais pas pu aller jusqu'au bout de cette thèse sans vos encouragements.

Merci Gérard pour toutes nos discussions passionnées, parfois tardives, sur les ICE ainsi que d'autres sujets. Je me souviens particulièrement de tes anecdotes très intéressantes sur la série Doctor Who. Je te remercie d'avoir été disponible et de m'avoir éclairé dès que j'avais besoin d'explications. Ton soutien et ta gentillesse m'ont beaucoup touché.

Merci Thomas d'avoir rejoint en cours de route ce projet. Sans ton aide, ma thèse n'aurait sûrement pas pris cette tournure ! C'était très enrichissant de développer le programme avec toi. Je te remercie aussi de m'avoir motivée lorsque mon moral était bas.

Je te remercie tout particulièrement Charles de m'avoir consacré de ton temps lors de la fin de ta thèse et lors de ton post-doc. Grâce à toi j'ai pu m'approprier plus facilement ce sujet très complexe qu'est la recherche automatique des ICE !

Je remercie mes collègues et ex-collègues de mon laboratoire d'accueil DynAMic. Merci pour les pauses cafés et les pauses déjeuner. Merci Razak de m'avoir fait découvrir R. Kelly. Merci Virginie de m'avoir accueilli chez toi lorsque je devais passer en coupe-vent à Nancy ainsi que le soutien moral. Merci Stéphane pour toutes nos discussions matinales sur tout et n'importe quoi.

Je remercie aussi mes collègues et ex-collègues du laboratoire MaIAGE. Merci Arnaud pour tous tes conseils, de m'avoir rassuré lorsque je doutais de moi et surtout pour toutes nos discussions et débats passionnés. Merci Cyprien, Sandra, Mahendra pour tout, il y a tellement de choses à dire que je ne sais pas quoi mettre ! Merci Anne-Laure et Ba de m'avoir consacré de votre temps à chaque fois que j'avais des questions techniques (ou non). Je remercie mes voisins « du premier étage » pour tous les moments conviviaux : Cédric, David, Samantha, Quentin, Valentin. Merci Jean-François et tout particulièrement de m'avoir aidé si rapidement quand j'étais bloquée en deuxième année.

Merci Florence, ma meilleure amie, tu as toujours été là pour moi et en particulier lorsque je n'osais pas demander ! Merci Julien, Amaury, Athénaïs et Slim de m'avoir écouté et conseillé lorsque j'en avais besoin. Et merci Jaysen, Hélène, Charlotte, Sam, Ibrahim et Rafaele d'avoir été présents lors de cette aventure.

Et enfin, je remercie mes parents, Cindy, Arnaud et Elise d'avoir toujours été là pour me soutenir et de m'avoir encouragé lorsque j'en avais vraiment besoin.

Sommaire

Liste des tableaux

Liste des figures

Liste des abréviations

- aa : **a**cide **a**minés
- *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)
- *attR* : site d'attachement droit (**R**ight **att**achment site)
- CDS : séquence codante (**C**o**d**ing **S**equence)
- CP : protéine de couplage (**C**oupling **P**rotein)
- DR : répétition directe (**D**irect **R**epeat)
- HMM : modèle de Markov caché (**H**idden **M**arkov **M**odel)
- ICE : élément intégratif conjugatif (**I**ntegrative **C**onjugative **E**lement)
- IME : élément intégratif mobilisable (**I**ntegrative **M**obilizable **E**lement)
- MPF : pore de conjugaison (**M**ating **P**ore **F**ormation)
- nt : **n**ucléo**t**ides
- *oriT* : **ori**gine de **T**ransfert
- pb : **p**aire de **b**ases
- SP : protéine signature (**S**ignature **P**rotein)
- T4SS : système de sécrétion de type IV (**T**ype **4** **S**ecretion **S**ystem

Préambule

Dans ce travail, nous nous sommes intéressés à l'étude des éléments conjugatifs intégrés dans le génome de bactéries du phylum des Firmicutes. Les Firmicutes comportent plus de 250 genres bactériens parmi lesquels les *Streptococcus*, les *Staphylococcus*, les *Lactobacillus*, les *Enterococcus*, les *Bacillus*, les *Listeria* ou encore les *Clostridium*. Ce sont des bactéries très étudiées à l'INRAE en raison de leur présence dans un grand nombre d'écosystèmes environnementaux, alimentaires ou du microbiote des animaux et de l'Homme. Ces bactéries, souvent moins étudiées que les entérobactéries modèles comme *Escherichia coli* ou *Salmonella enterica*, jouent cependant un rôle clé dans l'alimentation (plusieurs lactobacilles sont des probiotiques), dans la santé humaine et des animaux d'élevage. Ainsi, les Firmicutes constituent une grande partie des bactéries du microbiote intestinal humain et d'animaux et plusieurs de ces espèces incluent des souches commensales ou pathogènes. Les Firmicutes sont également présents dans l'environnement (par exemple, les *Bacillus* sont présentes dans le sol, les sédiments, l'eau de mer, les intestins d'animaux et les selles des animaux à sang chaud dont les humains). Les génomes des Firmicutes présentent souvent une grande diversité de taille et de composition probablement due en partie à la présence massive d'éléments génétiques mobiles. On soupçonne aussi ces éléments génétiques mobiles d'être massivement impliqués dans la dissémination, et par conséquent l'acquisition, de résistances aux antibiotiques, un problème majeur de santé publique auquel nous sommes confrontés dans les pays développés.

Introduction

1. Les éléments génétiques mobiles dans les génomes bactériens

Les bactéries sont capables d'acquérir de l'ADN étranger provenant de diverses sources : de leur milieu environnant ou encore d'autres bactéries. Ces dernières années, le développement des méthodes de séquençage et l'explosion des génomes disponibles permettent d'affirmer que les transferts horizontaux de gènes constituent la force évolutive majeure des génomes bactériens (Frost *et al.*, 2005; Treangen and Rocha, 2011).

1.1 Les mécanismes de transferts horizontaux bactériens

L'ADN bactérien, présent dans le cytoplasme des cellules est isolé du milieu extérieur par une ou plusieurs membranes selon que les bactéries soient monoderme ou diderme. Ainsi, les échanges d'ADN nécessitent des mécanismes actifs pour passer ces obstacles. Il existe trois mécanismes bien documentés de transferts horizontaux de gènes : la transformation (Lorenz and Wackernagel, 1994), la transduction et la conjugaison.

La transformation est un mécanisme actif par lequel de l'ADN libre du milieu, provenant généralement de cellules mortes, est absorbé dans le cytoplasme et intégré dans le génome d'une bactérie receveuse. L'acquisition de l'ADN par la cellule réceptrice a lieu lorsqu'elle est dans un état physiologique spécifique que l'on appelle état de compétence. Ce mécanisme pourrait servir à des fins nutritionnelles, mais le fait que certaines bactéries soient très sélectives sur le type d'ADN qu'elles laissent entrer suggère que la transformation sert également à favoriser la recombinaison et donc l'acquisition de gènes entre parents proches (Redfield *et al.*, 1997; Szöllosi *et al.*, 2006; Mell and Redfield, 2014).

La transduction est un mécanisme de transfert d'ADN d'une cellule à l'autre qui se produit grâce aux un bactériophages (i.e. virus de bactéries). Dans le cas de la transduction généralisée, médiée par les phages lytiques, l'ADN fragmenté du génome bactérien peut être véhiculé par erreur à l'intérieur de particules phagiques et injecté à une autre cellule bactérienne, à la place du génome phagique. Certaines espèces de bactéries, essentiellement des alpha-protéobactéries, ont détourné ce mécanisme à leur avantage en recrutant des gènes de bactériophages pour favoriser les échanges génétiques (Lang and Beatty, 2007).

La conjugaison est un mécanisme de transmission unidirectionnel de l'ADN d'une cellule donneuse à une cellule receveuse. Les gènes responsables de la conjugaison sont portés par des éléments génétiques mobiles, appelés éléments conjugatifs, et sont nécessaires pour assurer leur propre transmission. Dans certains cas, la conjugaison n'assure pas exclusivement le transfert de l'élément conjugatif mais peut également favoriser les transferts d'autres éléments non autonomes ou des gènes chromosomiques. Le mécanisme de conjugaison présente la particularité de nécessiter un contact physique entre la cellule donneuse et la receveuse : sans contact, les échanges de gènes sont impossibles.

Dans le cytoplasme, l'ADN étranger peut avoir plusieurs destins : soit il est détruit par les systèmes de dégradation de l'ADN présents dans le cytoplasme de l'hôte (enzymes de restriction, DNAse, etc.), soit il persiste sous forme d'entités réplicatives autonomes comme les plasmides ou est intégré dans le chromosome de l'hôte. L'ADN étranger peut être intégré de manière site-spécifique si l'ADN exogène est un élément mobile tel un phage ou un ICE (Integrative Conjugative Element) ou encore par recombinaison homologue.

1.2 Les éléments génétique mobiles

Les éléments génétiques mobiles sont des éléments capables de changer de position soit à l'intérieur d'un génome (transposons), soit par des mécanismes de transfert horizontal entre deux cellules (ex : bactériophages, plasmides). Les éléments génétiques mobiles se transfèrent donc d'une cellule à une autre et doivent ensuite se maintenir dans le génome de leur hôte après transfert soit sous forme plasmidique ou intégrée au chromosome. Ils sont ainsi souvent porteurs de gènes d'adaptation qui peuvent augmenter la "fitness" de leur hôte à un environnement donné.

1.2.1 Les plasmides

Les plasmides sont des éléments mobiles, extrachromosomiques, capables de se maintenir dans la population hôte. Ils sont vus comme d'importants vecteurs d'adaptation rapide des populations bactériennes aux changements des conditions environnementales (pour revue Heuer and Smalla, 2012; Smalla *et al.*, 2015) et participent à la dissémination des résistances

aux antibiotiques (revues récentes Rozwandowicz *et al.*, 2018; Nang *et al.*, 2019; Mendes Oliveira *et al.*, 2019).

Les plasmides peuvent grossièrement être classés en deux catégories : (1) les plasmides non-conjugatifs et (2) les plasmides conjugatifs et mobilisables qui peuvent transférer par conjugaison. Il existe une très grande diversité de plasmides.

Les plasmides conjugatifs codent toutes les protéines nécessaires pour leur transfert autonome d'une cellule à une autre par conjugaison.

Les plasmides mobilisables codent une partie des protéines nécessaires pour leur transfert par conjugaison. Ainsi, ils ne sont pas autonomes pour leur transfert et vont être mobilisés par d'autres éléments conjugatifs.

Les plasmides non conjugatifs ne codent pas de protéines de transfert par conjugaison. Ils peuvent cependant être transférés s'ils sont intégrés dans un autre élément conjugatif ou mobilisable.

1.2.2 Les îlots génomiques

En dehors des éléments génétiques mobiles traditionnels (prophages, transposons et plasmides), l'analyse de la séquence des génomes bactériens montre que les gènes acquis par transfert horizontal se regroupent dans des régions particulières, mal caractérisées, qui ont été appelées îlots génomiques (Hacker and Kaper, 2000; Juhas *et al.*, 2009). Les îlots génomiques sont définis comme des régions du chromosome présentant des évidences d'un ou de plusieurs transferts horizontaux passés et codant des fonctions pouvant être avantageuses pour l'hôte, telles que la pathogénicité de la souche (Bellanger *et al.*, 2014).

Les évidences suggérant ou démontrant ce transfert sont variées, que ce soit entre différents îlots génomiques ou au sein d'un même îlot. En raison de cette diversité, le concept reste très flou et le contour d'un îlot peut être différent selon la façon de l'identifier. Ainsi, un îlot génomique est le plus souvent présent dans le génome de certaines souches d'une espèce et absent du génome d'autres souches de la même espèce (ou d'espèces très proches), en raison d'une acquisition récente par transfert horizontal. Mais, selon les souches comparées (par

exemple leur degré d'éloignement, ou leur mode de vie), les limites précises de l'îlot pourront être différentes. De même, différentes caractéristiques de la séquence, telles qu'un pourcentage en G+C ou une utilisation des codons différents de celui du reste du génome de l'organisme dans lequel il est présent, sont souvent utilisées pour les identifier. Cependant, de telles caractéristiques peuvent être très variables au sein de la même région (voir Burrus *et al.*, 2002a pour un exemple de variation en pourcentage en G+C au sein de la même région transférée).

La plupart de ces îlots génomiques portent des gènes (ou des pseudogènes) codant des protéines de « mobilité » (par exemple, des protéines de conjugaison, des intégrases, des transposases, des protéines impliquées dans la réplication, etc.), qui pourraient être impliquées dans le transfert d'ADN ou le maintien de l'ADN transféré. Cependant, en général, il est difficile de savoir si ces gènes ont été impliqués dans l'acquisition de l'îlot lui-même ou d'une partie de celui-ci, voire d'un petit élément mobile porté par l'îlot comme par exemple une séquence d'insertion.

Au global, les îlots génomiques peuvent donc être des éléments mobiles codant leur propre transfert et/ou maintien après transfert, en particulier des éléments n'appartenant pas aux catégories classiques d'éléments mobiles (éléments transposables, prophages, plasmides). Ils peuvent aussi correspondre à des éléments immobiles dérivant d'éléments génétiques mobiles par perte de tout ou partie des gènes impliqués dans la mobilité. Ils peuvent aussi correspondre à des combinaisons d'éléments, de même type ou non. Ces combinaisons peuvent aussi bien être des emboîtements (intégration d'un élément dans un autre), des tandems d'éléments (intégrations successives dans le même site) ou des éléments intégrés dans le voisinage les uns des autres, tout en restant séparés par des séquences chromosomiques non transférées.

1.2.3 Les ICE et les IME

Depuis plus de vingt ans, l'étude des bactéries et de leur(s) chromosome(s) a révélé qu'en dehors des transposons, prophages et plasmides, existent d'autres types d'éléments génétiques mobiles répandus comme les éléments intégratifs conjugatifs (ICE) et les éléments intégratifs mobilisables (IME). L'étude du transfert par conjugaison et des îlots génomiques

révèle que de nombreux éléments chromosomiques sont capables de se transférer par conjugaison (Bellanger *et al.*, 2014). Parmi eux, les éléments intégratifs conjugatifs, ou ICE codent leur propre excision, leur transfert par conjugaison d'une bactérie donatrice vers une bactérie réceptrice et leur intégration.

Comme les ICE, les IME codent toutes les fonctions nécessaires à leur excision et intégration. Contrairement aux ICE, ils ne sont pas autonomes pour leur transfert mais détournent la machinerie de conjugaison d'un autre élément conjugatif (plasmide conjugatif ou bien un ICE) résidant dans la même cellule pour promouvoir leur propre transfert.

Ces éléments possèdent des caractéristiques semblables à celles des transposons, prophages et plasmides. Ainsi, les premiers ICE identifiés ont été appelés transposons conjugatifs à cause de leur faible spécificité d'insertion semblable à ceux des transposons ou plasmides conjugatifs intégratifs du fait de l'existence d'une forme circulaire (Burrus *et al.*, 2002b)

2. Structure et caractéristiques des ICE et des IME

Compte tenu de leur biologie, les ICE et les IME vont posséder des gènes responsables et/ou nécessaires à leur transfert par conjugaison, des gènes responsables de leur excision et de leur intégration et généralement des gènes pouvant conférer un avantage adaptatif.

La suite du document concerne uniquement les éléments conjugatifs intégrés de Firmicutes, qui comprennent de nombreuses bactéries de microbiotes humains et d'animaux et des espèces pathogènes (*Enterococcus*, *Streptococcus*, *Staphylococcus*, *Mycoplasma*...).

2.1 Les modules fonctionnels des éléments

Tous les éléments génétiques mobiles sont constitués d'un seul module ou d'une combinaison de modules (Toussaint and Merlin, 2002; Burrus *et al.*, 2002b; Bellanger *et al.*, 2014). 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 (voir Figure 1).

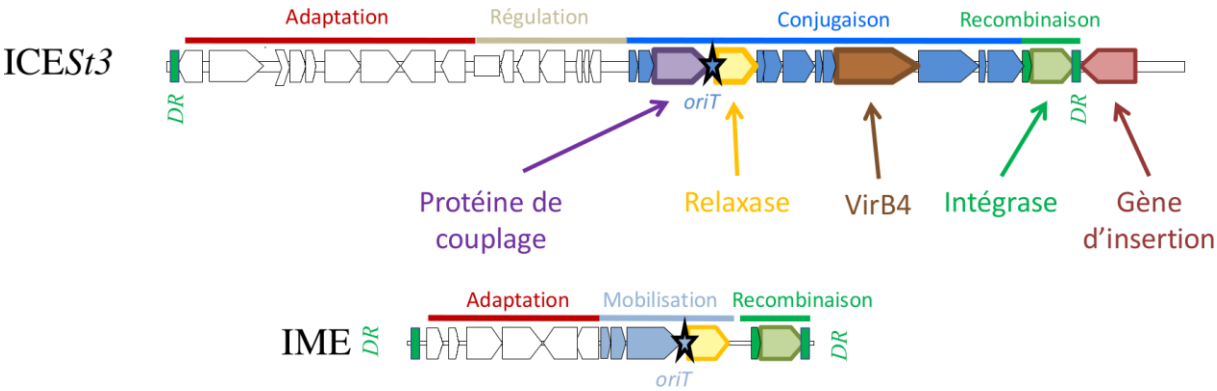

Figure 1 : Représentation schématique de la structure modulaire des ICE et des IME. Les modules fonctionnels d'un ICE de type ICE*St3* et d'un IME sont schématisés. Les gènes intervenant pour une même fonction biologique sont regroupés physiquement sous forme de module. Chez les ICE, il y a quatre modules principaux : le module de recombinaison (indiqué par le trait vert), le module de conjugaison (indiqué par le trait bleu), le module de régulation (indiqué par le trait gris) et un ou plusieurs modules d'adaptation (indiqué par le trait rouge). Les IME possèdent un module de recombinaison, un module de mobilisation (indiqué par le trait bleu clair) et un ou plusieurs modules d'adaptation. Le module de régulation peut être absent. Les rectangles verts représentent les répétitions directes (DR) qui délimitent l'ICE ICE*St3* et l'IME dans ce schéma. Les gènes codants pour des protéines indispensables pour le transfert et l'intégration des éléments sont indiquées par des flèches. Enfin, le gène cible pour l'insertion de l'ICE est représenté en rouge.

Tous les ICE portent ainsi :

- un module de recombinaison qui peut être plus ou moins apparenté à ceux de prophages (bactériophages intégrés) ou de séquences d'insertion[1],
- un module de conjugaison qui peut être plus ou moins apparenté à ceux de plasmides conjugatifs,
- un module de régulation qui contient souvent des protéines apparentées à celles de prophages,
- et un ou plusieurs modules d'adaptation d'origine diversifiée souvent apparentés à ceux de plasmides, prophages ou de transposons.

Chez les IME, le module de mobilisation remplace le module de conjugaison et les modules de régulation et d'adaptation peuvent être absents. Le module de mobilisation des IME peut être plus ou moins apparenté à ceux des plasmides mobilisables.

Les modules de recombinaison, de transfert et de régulation présentent des gènes et des séquences requises à des moments précis du mécanisme de transfert. Dans cette introduction, les mécanismes de régulation du transfert ne seront pas évoqués.

Actuellement, les données concernant la mécanistique du transfert des ICE et des IME de Firmicutes restent très parcellaires. Néanmoins, la mécanistique du transfert des plasmides conjugatifs est beaucoup mieux connue et, par analogie, permet d'imaginer le scénario suivant (voir Figure 2 ci-dessous).

[1] Élément transposable simple ne possédant que les gènes nécessaires à sa transposition.

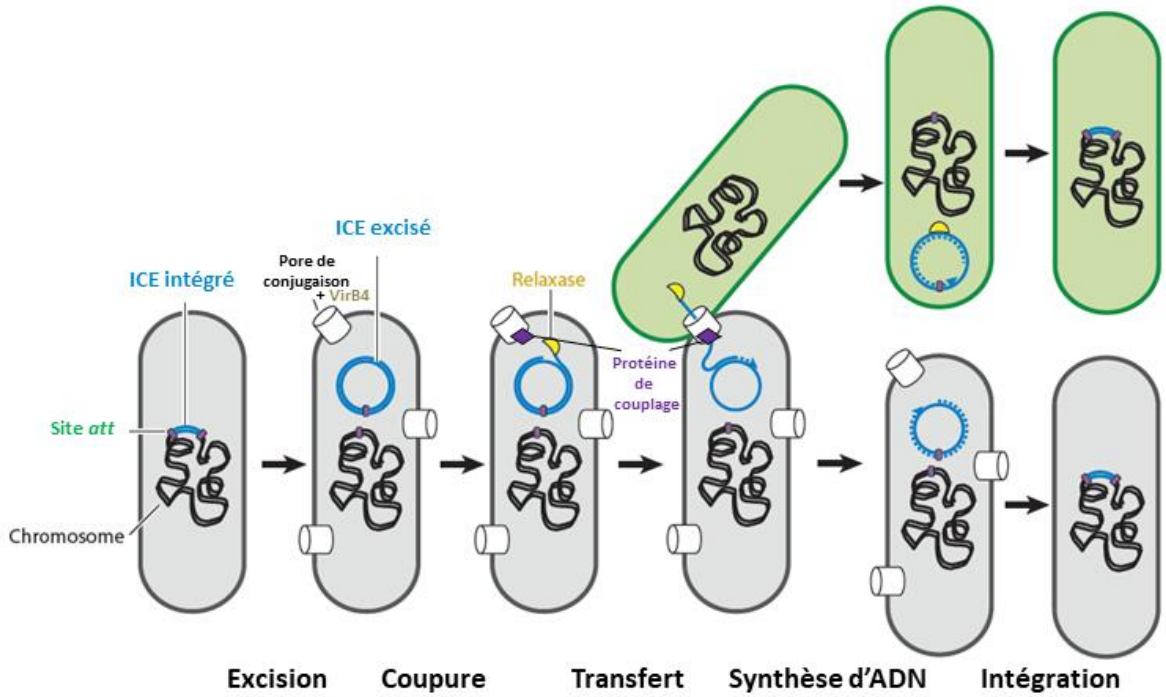

Figure 2 : Schéma du transfert d'un ICE (adapté de Johnson et Grossman, Annu. Rev. Genet., 2015). Des protéines indispensables au transfert ont été ajoutées au schéma : en jaune la relaxase, en violet la protéine de couplage et en gris la protéine VirB4 du pore de conjugaison. Le transfert de l'ICE d'une cellule donneuse (en gris) vers une cellule receveuse (en vert) comporte cinq étapes. (1) L'ICE intégré dans le chromosome de la bactérie hôte (en gris) s'excise sous forme circulaire double brin par recombinaison des sites att (indiqué en vert), cette excision est catalysée par l'intégrase codée par l'ICE lui-même. (2) Un dimère de relaxase codée par l'ICE va réaliser une coupure d'un des brins au niveau du site nick de l'*oriT* et se fixer à l'extrémité 5' du brin de manière covalente. (3) Le complexe ADN-relaxase est ensuite transféré via le pore de conjugaison par un mécanisme de réplication par cercle roulant, la protéine de couplage (en violet) pourrait être la protéine initiant ce transfert, la VirB4 fourni l'énergie nécessaire au transport du complexe. (4) L'ADN simple brin de la cellule donneuse et de la cellule receveuse est ensuite répliqué puis recircularisé grâce à la relaxase pour former un intermédiaire circulaire double brin. (5) L'ICE est ensuite intégré dans les chromosomes grâce à l'intégrase/recombinase.

2.1.1 Module de recombinaison

Avant son transfert, l'ICE est intégré dans le chromosome de la bactérie donneuse (en bleu sur la figure 2). Le module de recombinaison code les enzymes et séquences requises pour l'intégration de l'élément ainsi que pour son excison afin de générer un intermédiaire circulaire.

Trois types de modules de recombinaison non apparentés ont été observés dans les ICE et les IME décrits : des modules codant une intégrase à tyrosine, des modules codant une ou plusieurs intégrases à sérine et des modules codant une transposase à DDE qui comme son nom l'indique possède un motif constitué de trois résidus non contigus, deux aspartates et une glutamate (motif DDE) (Haren *et al.*, 1999).

2.1.1.1 Les modules codant une intégrase à tyrosine

L'intégration et l'excision de la plupart des ICE et des IME connues, dont ceux des Firmicutes, sont catalysées par une intégrase appartenant à la superfamille des recombinases à tyrosine (Bellanger *et al.*, 2014).

Les intégrases d'ICE et d'IME de Firmicutes ont une taille comprise entre 380 aa et 450 aa et sont toutes caractérisées par la présence d'un domaine catalytique à leur extrémité C-terminale. Dans tous les cas où la fonction de l'intégrase à tyrosine a été bien étudiée, la région N-terminale porte un domaine de liaison à l'ADN (Grindley *et al.*, 2006).

Dans leur grande majorité, les intégrases à tyrosine d'ICE et d'IME de Firmicutes catalysent une recombinaison site-spécifique entre deux séquences courtes identiques (ou presque) appartenant au site *attB* du chromosome bactérien et au site *attI* porté par l'ICE :

- Le site *attB* contient généralement la fin ou parfois au début d'un gène conservé dans l'espèce, pouvant être un gène codant des ARN de transfert ou un gène codant des protéines de ménage, souvent des protéines ribosomiques.

- L'insertion par recombinaison entre séquences identiques (ou presque) correspondant à l'extrémité du gène va provoquer l'apparition de répétitions directes (DR) flanquant

l'élément, parfaites ou dégénérées. Ces DR ont une taille comprise entre 8 et 100 nt pour les ICE et IME connus de Firmicutes.

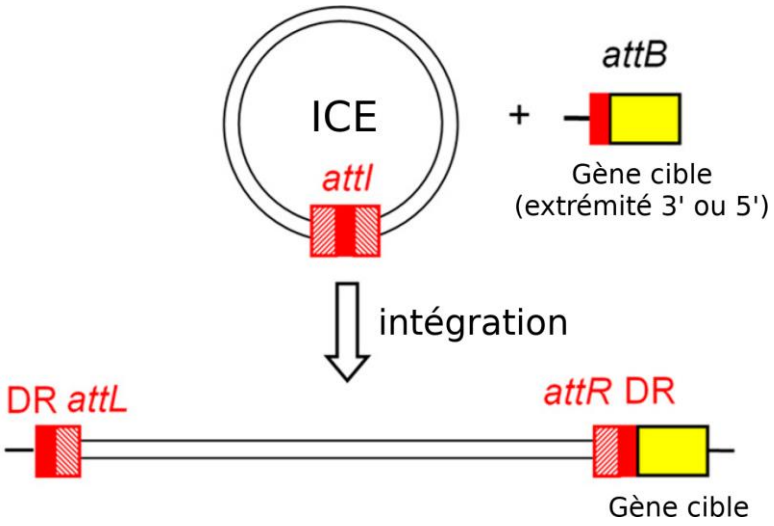

Figure 3 : Représentation schématique de l'intégration d'un ICE codant une intégrase site-spécifique (adapté de Guédon *et al.*, Genes, 2017). L'intégration de l'ICE est médiée par l'intégrase/recombinase de l'élément qui catalyse la recombinaison entre deux séquences identiques courtes (rectangle rouge plein) se trouvant dans le site *attI* (rectangles hachuré) de l'ICE et le site *attB* qui se trouve à une des extrémités 3' ou 5' du gène ciblé (rectangle jaune) par l'ICE sur le chromosome bactérien (représenté par la ligne noire). L'intégration de l'ICE entraîne la formation des sites *attL* et *attR*.

Certaines intégrases à tyrosine d'ICE et d'IME catalysent une intégration préférentielle plutôt que spécifique. Ainsi, l'intégrase de Tn*916*, un ICE répandu chez les Firmicutes, catalyse l'intégration de cet ICE dans divers sites ayant pour point commun d'être riches en bases AT (Roberts and Mullany, 2009), ces sites pouvant ou non appartenir à d'autres éléments conjugatifs (Mingoia *et al.*, 2011; Iannelli *et al.*, 2014).

2.1.1.2 Les modules codant une intégrase à sérine

Le module d'intégration de divers ICE et IME de Firmicutes porte non pas un gène d'intégrase à tyrosine mais un gène d'intégrase à sérine (Wang and Mullany, 2000) ou un triplet (trois gènes adjacents et de même sens) d'intégrases à sérine (Beres and Musser, 2007; Holden *et al.*, 2009; Camilli *et al.*, 2011). La taille de ces protéines est variable mais toujours supérieure à 500 aa. Elles incluent toutes, à leur extrémité C-terminale, trois domaines conservés indispensables à l'activité de l'enzyme. Les intégrases à sérine d'ICE et IME de Firmicutes catalysent le plus souvent une intégration site-spécifique. Contrairement aux intégrases à tyrosine, la cible de l'intégrase est un site interne d'un gène codant une protéine

répandue mais non indispensable à la bactérie. La recombinaison aboutit à la formation de répétitions directes flanquant l'élément qui peuvent être des DR parfaites de 2 nt ou des DR plus longues (jusqu'à une vingtaine de nucléotides) souvent dégénérées. Comme pour les intégrases à tyrosine, certaines intégrases à sérine d'ICE et d'IME de Firmicutes catalysent une intégration peu spécifique (Bellanger *et al.*, 2014).

2.1.1.3 Les modules codant une transposase à DDE

Divers ICE et IME de Firmicutes codent des intégrases appartenant à la superfamille des transposases à DDE. Ils incluent plus particulièrement :

- des ICE des familles Tn*GBS1* et Tn*GBS2* codant des transposases à DDE appartenant à la famille IS*Lre2* dont la taille varie généralement entre 500 et 630 acides aminés (Guérillot *et al.*, 2014; Siguier *et al.*, 2015). Ces ICE sont répandus chez les streptocoques ;
- presque tous les ICE connus de mycoplasmes. Ces ICE codent des transposases à DDE apparentées aux transposases de la famille IS*Lre2* (Frisoni *et al.*, 2013) ;
- divers ICE de staphylocoques dont ICE*6013* qui code des transposases à DDE de la famille IS30 (Smyth and Robinson, 2009).

Les transposases à DDE des ICE Tn*GBS1* et Tn*GBS2* ciblent une position localisée 15 à 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 (Brochet *et al.*, 2009). Quant aux transposases à DDE des ICE de mycoplasmes, elles catalysent une intégration peu spécifique qui semble cibler les régions riches en AT (Roberts and Mullany, 2009). La transposition des ICE de ces deux familles provoque une duplication de la séquence ciblée de 8 pb. En conséquence, l'ICE est flanqué de répétitions directes de 8 pb.

2.1.2 Modules de transfert

2.1.2.1 Le module de conjugaison

Chez les ICE, le module de conjugaison code l'ensemble des enzymes et structures protéiques nécessaires au transfert de l'élément d'une cellule donneuse vers une receveuse. Deux

mécanismes de conjugaison sont bien connus. Dans l'un d'eux, rencontré uniquement chez les actinobactéries, l'ADN transféré est double brin (te Poele *et al.*, 2008; Bordeleau *et al.*, 2012). Ce mécanisme ne sera pas détaillé dans cette thèse puisqu'il n'est pas rencontré chez les Firmicutes. L'autre mécanisme, beaucoup plus répandu, est le seul connu chez les Firmicutes et implique le transfert d'ADN simple brin.

Comme pour les plasmides conjugatifs, l'initiation du transfert simple brin est probablement catalysée par un complexe multiprotéique appelé le relaxosome qui reconnaît l'origine de transfert (*oriT*) de l'ICE. La protéine principale du relaxosome est la relaxase, une enzyme du module de conjugaison, qui assure la **coupure** simple brin de (Alvarez-Martinez and Christie, 2009) l'ADN au niveau du site *nick* de l'*oriT*. Tandis que l'extrémité 3' de l'ADN clivé est utilisée pour initier la réplication par cercle roulant de l'élément circularisé, son extrémité 5' est covalemment attachée à la relaxase. Actuellement, sept superfamilles de relaxases non apparentées ou apparentées de façon lointaines sont connues chez les plasmides conjugatifs ou ICE de Firmicutes (MOB_F, MOB_H, MOB_Q, MOB_C, MOB_P, MOB_V, MOB_T) (Garcillán-Barcia *et al.*, 2009; Guglielmini *et al.*, 2011). Le **transfert** du complexe ADN-relaxase est assuré par une protéine de couplage dont le rôle serait d'initier son passage au travers du pore de conjugaison. Deux grandes superfamilles de protéines de couplage sont connues chez les Firmicutes : VirD4 et TcpA (Guglielmini *et al.*, 2013). Le pore de conjugaison est un complexe multiprotéique apparenté au système de sécrétion de type IV (T4SS). Il assure le transport du complexe ADN-relaxase au travers des enveloppes des cellules donneuse et receveuse. Ce transport nécessite de l'énergie fournie par des ATPases du T4SS, une des plus connues est la VirB4, une protéine très conservée dont une seule superfamille est connue (Alvarez-Martinez and Christie, 2009). Dans la cellule donneuse et la cellule receveuse, l'ADN de l'ICE est répliqué et recircularisé probablement grâce à la relaxase pour former un intermédiaire circulaire double brin. L'ICE est ensuite intégré dans le chromosome de la cellule hôte grâce à l'intégrase/recombinase codée par le module de recombinaison.

Le module de conjugaison assure donc deux fonctions différentes : la mobilisation de l'ADN pour le transfert vers la cellule receveuse et la mise en place du pore de conjugaison (aussi appelé MPF pour Mating Pore Formation). Ces fonctions sont codées par deux sous-modules

: le sous-module MOB et le sous-module MPF. Ces sous-modules sont rarement échangés, ainsi on trouve globalement un type de sous-module MOB qui sera associé à un type de sous-module MPF particulier (De La Cruz *et al.*, 2010).

2.1.2.2 Le module de mobilisation

À ce jour, aucune étude n'a décrypté la mécanistique de transfert d'un IME au niveau moléculaire. Dans ces grandes lignes, le mécanisme de transfert d'un IME semble similaire à celui des ICE, à la différence qu'un IME ne se transfère jamais de manière autonome. Ceci résulte du fait que le module de mobilisation des IME ne contient qu'une partie des gènes impliqués dans la conjugaison et en particulier ne code jamais de protéine VirB4. À l'instar des plasmides mobilisables (Ramsay and Firth, 2017), il est possible que les IME utilisent différentes stratégies de mobilisation en fonction des protéines de transfert qu'ils codent. Les modules de mobilisation des IME sont très variables. Chez les protéobactéries, la plupart des IME connus ne codent ni relaxase, ni protéines du T4SS. À minima, leur module de mobilisation contient une *oriT*. Cependant, la quasi-totalité des IME connus de Firmicutes codent leur propre relaxase (Guédon *et al.*, 2017). Une grande partie des IME connus de streptocoques code, en plus de la relaxase, leur propre protéine de couplage (Coluzzi *et al.*, 2017). Au total neuf superfamilles de relaxases sont codées par les IME de streptocoques incluant quatre familles non-retrouvées chez les ICE (Coluzzi *et al.*, 2017).

2.1.3 Modules d'adaptation

Les ICE et les IME possèdent des gènes qui n'interviennent pas dans leur cycle de vie et qui peuvent conférer différentes fonctions à l'organisme hôte. Ces fonctions peuvent apporter un avantage sélectif à l'hôte.

C'est par l'intermédiaire d'études sur la dissémination des résistances aux antibiotiques et de résistances aux métaux lourds que les premiers ICE ont été découverts. En effet, l'ICE Tn*916* de *Enterococcus faecalis* a été le premier ICE caractérisé, celui-ci porte des gènes codant la résistance à la tétracycline (Franke and Clewell, 1981). Quant aux résistances aux métaux lourds, les ICE ICE*St1* et ICE*St3* de *Streptococcus thermophilus* portent des gènes qui codent des systèmes de restriction-modification (Burrus *et al.*, 2001; Bellanger *et al.*, 2009).

Les fonctions d'adaptation apportées par les ICE et les IME sont très variées. En plus des résistances aux antibiotiques et aux métaux lourds, on trouve aussi des fonctions de virulence. Un exemple est un ICE de *Pseudomonas aeruginosa* connu sous le nom d'îlot de pathogénicité PAPI-1 (Carter *et al.*, 2010). L'acquisition de certains ICE peut permettre à l'hôte de rentrer en symbiose avec d'autres organismes. Cela a été observé grâce aux études de Sullivan *et al.* (Sullivan *et al.*, 1995; Sullivan and Ronson, 1998) où des bactéries du sol ont acquis la capacité d'entrer en symbiose avec des racines de lotier corniculé par l'acquisition de l'ICE ICEMlsym^{R7A}.

Il n'est pas rare que les ICE et les IME cumulent plusieurs fonctions d'adaptation. C'est le cas pour l'ICE Tn*5276* de *Lactococcus lactis* qui permet à son hôte de synthétiser des peptides à propriété antibiotique appelés bactériocine ainsi que d'exploiter une source alternative de carbone en lui permettant de fermenter le saccharose (Rauch and Vos 1992).

2.2 Classification des éléments

2.2.1 Classification des ICE

L'analyse phylogénétique des éléments conjugatifs montre de nombreux échanges de modules entre plasmides, ICE, voire entre ICE et plasmides ou entre ICE et IME (Burrus *et al.*, 2002b; Bellanger *et al.*, 2014; Ambroset *et al.*, 2016; Coluzzi *et al.*, 2017). Ceci rend illusoire toute classification basée sur l'ensemble des modules d'un plasmide ou d'un ICE.

Cependant une classification basée uniquement sur le module de conjugaison a été réalisée chez les plasmides conjugatifs et les plasmides mobilisables. Cette classification prend en compte la relaxase (Francia *et al.*, 2004; Garcillán-Barcia *et al.*, 2009, 2011) et des protéines du pore de conjugaison (Smillie *et al.*, 2010). Nous avons fait le choix de reprendre à notre compte cette méthode pour classifier les ICE.

La classification que nous utilisons repose sur la nature et la phylogénie de protéines de conjugaison et notamment des relaxases, protéines de couplage et VirB4. L'analyse phylogénétique de ces différentes protéines montre qu'elles évoluent presque toujours de façon coordonnée ce qui implique que les échanges de gènes entre les modules de conjugaison sont rares. C'est en particulier le cas pour les ICE de streptocoques, un genre des Firmicutes où ces éléments ont été le plus étudiés (Ambroset *et al.*, 2016).

Cette analyse a permis de définir trois superfamilles d'ICE chez les streptocoques :

- La superfamille Tn*916* se caractérise par la présence d'une relaxase de type MOB_T, d'une protéine de couplage de type TcpA et d'une VirB4. Ces VirB4 ont la caractéristique d'être toutes apparentées entre elles. Chez les streptocoques, la superfamille est subdivisée en deux familles : Tn*916* et ICE*St3.* Ces familles ont été définies en fonction du pourcentage d'identité de ces trois protéines (relaxase, protéine de couplage et VirB4). Il faut en effet que les trois protéines signatures du module de conjugaison d'une même famille présentes plus de 40 % d'identité.

- La superfamille Tn*5252* se caractérise par la présence d'une relaxase de type MOB_P, d'une protéine de couplage de type VirD4 et d'une VirB4. Chez les streptocoques, cette superfamille se décompose en quatre familles Tn*5252*, Tn*1549*, Tn*GBS2* et vanG.

- La superfamille Tn*GBS1* se caractérise par la présence d'une relaxase de type MOB_L, d'une protéine de couplage de type VirD4 et d'une VirB4 (Coluzzi, 2017; Coluzzi *et al.*, 2017). Chez les streptocoques, cette superfamille n'inclut qu'une famille.

Chez d'autres Firmicutes des éléments des superfamilles Tn*916* et Tn*5252*, appartenant ou non à d'autre familles connues, ont été caractérisés (Bellanger *et al.*, 2014).

2.2.2 Classification des IME

Le module de mobilisation des IME ne porte qu'un petit nombre de gènes dont celui de la relaxase et souvent celui de la protéine de couplage. L'analyse phylogénétique de ces deux protéines montre de nombreux échanges de ces gènes à l'intérieur des modules de mobilisation rendant difficile toute classification basée sur l'ensemble des gènes de ce module. Chez les streptocoques, neuf superfamilles de relaxases et deux superfamilles de couplage ont été identifiées. De plus, les IME présentent deux superfamilles d'intégrases avec au moins 17 spécificités d'intégration dont huit sont communes avec les ICE. Au global, l'analyse exhaustive de 124 génomes de streptocoques avait révélé sur 144 éléments, de très nombreuses combinaisons différentes tenant compte de ces trois protéines. Ceci nous a conduit à baser la classification des IME sur la seule relaxase (Coluzzi *et al.*, 2017).

2.3 Éléments composites

Les îlots génomiques incluant les ICE et les IME présentent souvent une structure composite (Wozniak and Waldor, 2010; Bellanger *et al.*, 2014), c'est-à-dire qu'ils sont, en fait, constitués d'assemblages d'éléments plus petits. Ces petits éléments peuvent, par exemple, être des éléments transposables telles que des transposons ou des séquences d'insertion (IS) ne codant aucune fonction de transfert mais codant leur propre transposase. Ils peuvent aussi être 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) ou des IME intégrés de façon spécifique dans des gènes de conjugaison dont le gène codant la protéine VirD4 (Coluzzi *et al.*, 2017). Ils forment ainsi des éléments composites, constitués de deux éléments, l'un inséré dans un autre. Ces emboîtements peuvent impliquer plus de deux éléments. Ainsi, Tn*6103* de *C. difficile* apparenté à Tn*1549*, héberge trois IME (Brouwer *et al.*, 2011). Il peut également y avoir plusieurs niveaux d'emboitements. Un bon exemple est l'ICE ICE*Sluvan* de *Streptococcus lutetiensis*, apparenté à Tn*5252*, qui porte un ICE de type Tn*1549* qui héberge lui-même un IME putatif (Bjørkeng *et al.*, 2013; Bellanger *et al.*, 2014).

De plus, les éléments qui s'intègrent de façon site-spécifique peuvent s'intégrer non seulement dans le site *attB* vide mais également dans le site *attL* ou *attR* flanquant un élément résident de même spécificité d'intégration. Ceci conduit à la formation d'un îlot génomique composite formé de deux éléments intégrés en accrétion (Bellanger *et al.*, 2014). L'élément composite produit par cette accrétion est alors formé des deux éléments intégratifs séparés par un site *attI* chimérique et peut avoir deux structures différentes (voir Figure 4 ci-dessous).

L'analyse des génomes de streptocoques a révélé des structures complexes composées de plus de deux éléments en accrétion et/ou d'éléments en accrétion qui hébergent d'autres éléments (Coluzzi *et al.*, 2017; Libante *et al.*, 2020).

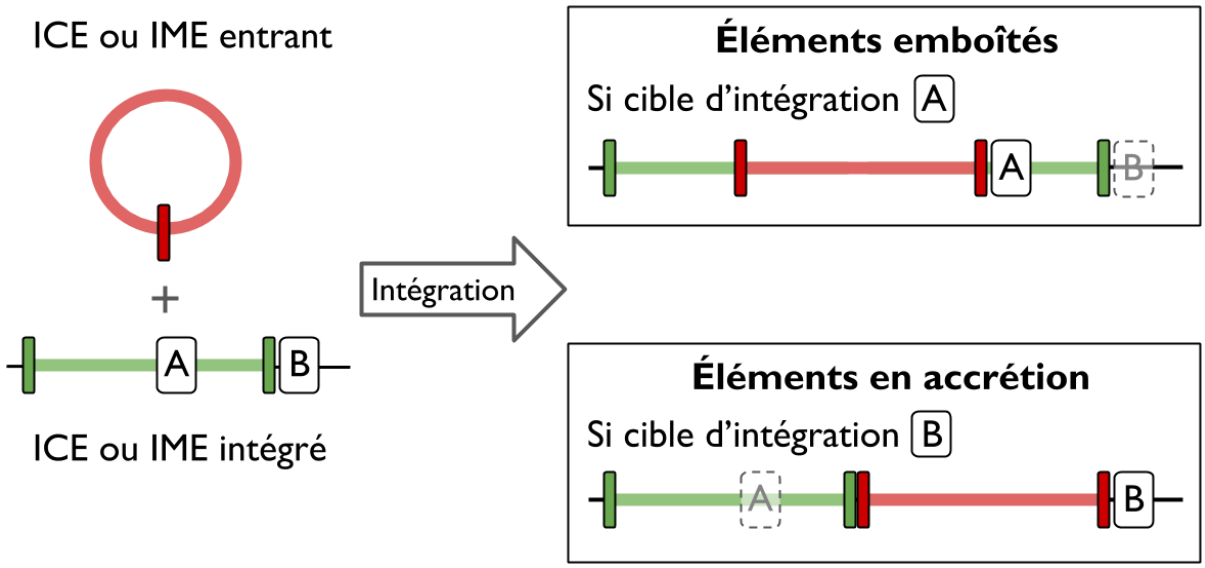

Figure 4 : Représentation schématique d'éléments emboîtés et d'éléments en accrétion. Si un ICE ou IME entrant (rouge) s'intègre à l'intérieur d'un ICE ou IME déjà présent dans la cellule (vert) en ciblant une séquence (A) interne de l'élément résident alors l'élément rouge sera emboîté dans l'élément vert. Si l'élément rouge cible la même séquence d'intégration que l'élément vert (B) alors les deux éléments seront en accrétion une fois l'élément rouge intégré.

3. Ressources disponibles pour la détection et l'annotation des ICE et IME

La recherche et l'annotation des ICE et IME dans les génomes bactériens par des méthodes bioinformatiques sont l'objet de travaux de recherche récents qui ont démarré il y a une dizaine d'années.

La première initiative dans ce domaine a été la publication d'une **base de données spécialisée ICE/IME :** la ressource ICEberg développée par l'équipe du professeur Hong-Yu Ou à l'Université Jiaotong de Shanghai en Chine. Il s'agit de la seule ressource publique dédiée aux éléments de type ICE et IME à ce jour (https://db-mml.sjtu.edu.cn/ICEberg2/index.php, Bi *et al.*, 2012).

Parallèlement, deux équipes de recherche ont développé des approches bioinformatiques permettant de détecter et d'annoter des éléments conjugatifs intégrés dans les génomes :

- l'équipe ICE-TeA, dirigée par Nathalie Leblond-Bourget de l'unité DynAMic de Nancy, a conçu et développé une **méthode de détection et d'annotation des ICE et des IME dans les génomes de streptocoques au nucléotide près** nommée ICE/IME Finder (Ambroset *et al.*, 2016; Coluzzi *et al.*, 2017).

- l'équipe d'Eduardo Rocha de l'Institut Pasteur Paris a conçu et développé une **méthode de détection et d'annotation des ICE au gène près dans tous les génomes bactériens** (Cury *et al.*, 2020). Cette méthode utilise le module CONJscan (Abby *et al.*, 2016) de l'outil MacSyFinder (Abby *et al.*, 2014) qui est un outil très performant pour la détection des modules de conjugaisons dans les génomes bactériens.

Enfin, un troisième outil nommé ICEfinder a été proposé récemment pour détecter les éléments dans tous les génomes bactériens lors de la publication de la version 2 de la base de données nommée ICEberg2 (Liu *et al.*, 2019) par l'équipe de Hong-Yu Ou.

Il est intéressant de mentionner ici que ces trois méthodes se basent toutes, dans une première étape, sur la recherche de protéines signatures, mais elles ne sont pas exactement les mêmes pour chaque outil.

Deux programmes sont utilisés pour l'annotation des protéines signatures par les trois outils disponibles pour l'annotation des ICE et des IME : BlastP (Altschul *et al.*, 1997) et HMMER (Eddy, 2011) qui permettent d'identifier des séquences protéiques par homologie.

Les programmes BlastP et HMMER permettent d'identifier des séquences similaires entre elles. Deux séquences significativement similaires peuvent correspondre à des séquences homologues, c'est-à-dire deux séquences dérivant d'une séquence ancestrale commune. Ainsi l'annotation fonctionnelle d'une séquence protéique connue peut être transférée à son homologue.

La plus grande différence entre une approche utilisant BlastP et une utilisant HMMER est l'entité comparée avec la séquence à identifier. BlastP compare une séquence protéique contre une autre séquence protéique en les alignant directement et calcule un score reflétant la similarité entre les deux séquences. Quant à HMMER, la séquence à identifier est comparée à un profil HMM (pour Hidden Markov Model).

Un profil HMM est construit à partir d'un alignement multiple de séquences et correspond généralement à un modèle d'une famille de protéines. Ainsi, contrairement à BlastP, une approche utilisant des profils HMM permet d'identifier une séquence protéique par rapport aux domaines conservés qu'elle possède.

Pour autant, les approches utilisant des profils HMM ne sont pas "meilleures" que celles d'alignement de séquences (Blast). En effet, il n'est pas toujours possible de construire des profils HMM de qualité à cause d'un manque de séquences représentantes. De plus, l'annotation fonctionnelle d'une séquence protéique peut être plus précise que celle d'un domaine protéique.

Dans la suite de ce chapitre nous allons décrire rapidement la base de données ICEberg et les trois outils existants pour l'annotation des éléments conjugatifs.

3.1 Base de données ICEberg

La base de données ICEberg rassemble des informations sur les ICE et IME de bactéries monoderme et diderme extraites essentiellement de la littérature scientifique (base bibliographique Pubmed) en utilisant des technologies de *fouille de données*.

Dans sa dernière version de 2019 la base de données ICEberg intègre 694 références bibliographiques permettant de répertorier 1032 ICE, dont 270 ont été validés expérimentalement et 762 viennent de prédiction bioinformatique. Les données sont organisées dans une base de données relationnelle PostgreSQL accompagnée d'une interface web.

Dans sa première version de 2012 (Bi *et al.*, 2012), la base de données ICEberg contenait 428 ICE de 124 espèces bactériennes. Cependant, un grand nombre d'erreurs relatives à la nature des éléments de Firmicutes ou à leur classification ont été identifiées par l'unité DynAMic (voir discussion de Ambroset *et al.*, 2016). La version 2 publiée en 2019 (Liu *et al.*, 2019) reconnaît des problèmes de fiabilité et la présence d'erreurs dans la base de ICEberg de 2012 et annonce plusieurs améliorations dont :

- la correction de la majorité des erreurs sur la nature et la classification d'éléments mentionnées dans (Ambroset *et al.*, 2016),
- l'ajout de nouveaux ICE et IME,
- l'ajout de graphes d'interactions entre ICE et IME,
- et la mise à disposition de l'outil ICEfinder permettant la détection d'ICE et d'IME au nucléotide près.

L'outil ICEfinder, mis à disposition de la communauté scientifique par l'équipe du Professeur Ou est décrit très succinctement dans une publication (Liu *et al.*, 2019), il sera présenté rapidement dans la section 3.4 de cette introduction.

3.2 Méthode ICE/IME Finder

Cette méthode, développée par l'unité DynAMic, a été utilisée pour caractériser 105 ICE (Ambroset *et al.*, 2016) et 144 IME (Coluzzi *et al.*, 2017) dans 27 espèces différentes de streptocoques (124 génomes). Il s'agit d'une méthode semi-automatique d'annotation au nucléotide près des ICE, des IME et d'autres éléments mobiles proches dans les génomes de streptocoques. Elle comporte deux étapes :

1) **Recherche automatisée des protéines signatures** :

Recherche par le logiciel BlastP de protéines signatures (SP) d'ICE et d'IME en utilisant une base de données de protéines curées. Les alignements obtenus sont ensuite filtrés afin d'obtenir une liste de SP validées. Pour cela une série de filtres (pourcentage d'identité, pourcentage de couverture et longueur des alignements) adaptés aux SP trouvées dans les génomes de streptocoques est utilisée. Ces résultats automatiques sont ensuite enrichis d'une annotation experte qui permet d'associer les SP à des ICE et des IME déjà identifiés ainsi que leur site d'intégration.

Pour rappel, les protéines signatures sont des protéines codées par les ICE et des IME qui assurent des fonctions essentielles. La méthode en recherche quatre types :

- l'**intégrase**, indispensable pour l'excision et l'intégration de l'élément ;
- la **relaxase** et la **protéine de couplage** sont nécessaires pour le transfert de l'élément ;
- la **VirB4** joue un rôle essentiel dans la mise en place et la fonctionnalité du pore de conjugaison assurant le passage de l'élément de la cellule donneuse à la cellule receveuse. Elle fait partie d'un cluster de gènes du MPF (Mating Pair Formation).

2) **Identification manuelle des éléments** :

Délimitation primaire des ICE et IME par co-localisation des protéines signatures puis recherche du site d'insertion des éléments et recherche de leurs limites précises. Le site d'insertion est trouvé en caractérisant l'intégrase de l'élément et le gène ciblé par l'intégrase lors de l'insertion dans le cas d'une intégration site-spécifique. Une intégration site-spécifique

entraîne la formation de répétitions directes (DR) ou de répétitions inversées (IR) qui bornent l'élément. La recherche de ces répétitions permet de borner l'élément au nucléotide près.

La méthode permet ainsi d'annoter quatre types d'éléments mobiles :

- **ICE** : Éléments codant une relaxase, une protéine de couplage, une protéine VirB4 et une intégrase (protéines signatures) et possédant des limites cohérentes avec l'intégrase codée sont considérés comme des ICE.
- **dICE** : Les éléments non fonctionnels dérivant d'un ICE par perte ou pseudogénisation d'un ou deux des gènes codant des protéines signatures ou par perte d'une des extrémités sont considérés comme des ICE dégénérés (dICE pour « decayed » ICE).
- **IME** : Les éléments codant une relaxase, une intégrase et éventuellement une protéine de couplage et possédant des limites cohérentes avec l'intégrase codée sont considérés comme des IME.
- **dIME** : Les éléments codant une relaxase apparentée à celle des IME mais aucune intégrase fonctionnelle ou dont l'une des extrémités est absente sont considérés comme des IME dégénérés (dIME pour « decayed » IME).

Cette méthode permet d'identifier précisément et de manière exhaustive les protéines signatures des éléments des streptocoques et constitue ainsi une aide précieuse pour l'annotation des éléments. Cependant elle ne les annote pas de façon automatique et un œil expert est nécessaire pour compléter leur annotation.

3.3 Méthode utilisant CONJscan

Cette méthodologie récemment décrite (Cury *et al.*, 2020) est une approche de génomique comparative en trois étapes permettant d'identifier et d'annoter les ICE au gène près dans les génomes de bactéries monoderme et diderme :

1) Détection des systèmes de conjugaison codés par les chromosomes et donc présumés comme appartenant à des éléments conjugatifs, notamment des ICE.

2) Recherche des gènes conservés présumés orthologues (« core » gènes) ne pouvant être issus de transferts horizontaux qui bornent les systèmes de conjugaison détectés. Un minimum de quatre génomes de la même espèce est nécessaire pour calculer le « core » génome de l'espèce.

3) Affinement des limites des éléments par synténie au gène prés (étape non automatisée).

Un ensemble de gènes codant un système de conjugaison complet et borné par des gènes « core » est considéré comme étant un ICE.

L'étape 1 est réalisée avec le module CONJscan (Abby *et al.*, 2016) du logiciel MacSyFinder (Abby *et al.*, 2014). Ce module permet la détection de systèmes de sécrétion de type IV (T4SS) et donc de systèmes de conjugaison ou de mobilisation grâce à des profils HMM des gènes du système et des règles sur la présence, l'absence et l'organisation spatiale de ces gènes.

Il existe des modèles prédéfinis pour la détection des modules de conjugaison ou de mobilisation :

- **Modèle CONJ** : Le système détecté doit obligatoirement posséder une relaxase, une protéine VirB4 et une protéine de couplage de T4SS (T4CP). Un élément possédant ces gènes est considéré comme un ICE.

- **Modèle MOB** : Le modèle MOB n'a cependant pas fait l'objet de publication. Le système détecté par le modèle MOB doit obligatoirement posséder une relaxase, peut éventuellement posséder une VirB4 ou une T4CP. Cet élément pourrait donc être un IME ou éventuellement un élément non fonctionnel dérivant d'ICE.

Un système de *Mating Pair Formation* (MPF) est indispensable pour les T4SS. Ainsi, les T4SS ont été caractérisés en fonction de leur type de MPF. Sous le modèle CONJ, la caractérisation des T4SS peut être raffinée en ajoutant l'information du type de MPF recherché (B, C, F, FA, FATA, G, I ou T). Chez les Firmicutes, deux types sont bien définis : FA et FATA (Guglielmini *et al.*, 2014), les éléments possédant ces types de modules de conjugaison

doivent coder en plus de la relaxase, de la protéine de couplage et de la protéine VirB4 au moins deux autres protéines du pore de conjugaison.

L'étape 2 est réalisée avec le logiciel Roary (Page *et al.*, 2015) et un Notebook Python. Roary est utilisé pour calculer l'ensemble des gènes conservés de l'espèce bactérienne analysée, aussi appelé « core » génome. Il est nécessaire d'avoir au moins quatre génomes complets proches dont le contenu en éléments mobiles est différent pour calculer ce core génome. Le calcul du core génome a pour but d'exclure tous les gènes pouvant appartenir à un îlot génomique (acquis par transfert horizontal) et donc d'exclure les gènes d'un ICE. Ainsi, les gènes bornant les hypothétiques ICE ne peuvent pas être issus d'un transfert horizontal.

La première délimitation des ICE se fait ensuite par croisement des résultats de la première et de la seconde étape et permet d'obtenir une première délimitation des ICE qui sont appelés à cette étape de la méthodologie, des « spots ».

L'étape 3 consiste à affiner davantage les limites des spots détectés. Pour cela, les gènes conservés de ces régions sont identifiés par clustering avec l'algorithme UCLUST du logiciel USEARCH (Edgar, 2010). Un gène est dit conservé s'il est retrouvé dans, au moins, N-1 des génomes (N, le nombre total de génomes). Si des gènes conservés sont contigus à un gène du core génome alors la limite de l'élément du spot correspondra au gène conservé et non aux gènes « core ». Dans l'article de (Cury *et al.*, 2017), une autre méthode de clustering a été utilisée avec l'outil SiLiX (Miele *et al.*, 2011). Les deux algorithmes sont différents et ne génèrent pas les mêmes résultats.

Ainsi, contrairement à la méthode ICE/IME Finder, la méthode utilisant CONJscan permet de repérer automatiquement les modules de conjugaison des ICE et délimite ces éléments aux gènes près. Cependant, elle nécessite la construction d'un core génome, ainsi elle ne peut être appliquée que si l'on dispose d'au moins quatre génomes proches. Enfin cette méthode n'a pas été conçue pour la recherche des IME.

3.4 Outil ICEfinder

Cet outil a été proposé lors de la mise à disposition de la base de données ICEberg2 (Liu *et al.*, 2019). La base de données rassemble des informations sur une grande diversité d'ICE et d'IME de bactéries monoderme et diderme.

L'outil ICEfinder permet la détection et l'annotation des ICE et des IME au nucléotide près dans des génomes bactériens. La méthode de détection est basée sur la co-localisation de résultats d'alignements de protéines signatures du module d'intégration et du module de transfert. Ces résultats sont obtenus avec trois outils différents : BlastP, HMMER et oriTfinder (Li *et al.*, 2018). Pour le bornage, seuls les éléments ayant pour cible d'intégration un tRNA sont traités, les cibles sont annotées avec le logiciel ARAGORN (Laslett and Canback, 2004) et la délimitation au nucléotide près de l'élément est effectuée par recherche des Direct Repeat (séquences répétées) qui le borne avec le logiciel Vmatch (http://vmatch.de/). Deux types d'ICE sont détectés par l'outil : les ICE de type T4SS et les AICE (Ghinet *et al.*, 2011) pour qui sont retrouvés dans les génomes des actinobactéries (Actinomycete integrative and conjugative element). Un seul type d'IME est recherché.

ICEfinder définit un ICE de type T4SS comme une région d'au maximum 600 kb du génome possédant une intégrase, une relaxase, une protéine de couplage, et des gènes du MPF incluant la protéine VirB4. Un IME est défini comme une région d'au maximum 50 kb du génome pouvant posséder une intégrase, une relaxase, une protéine de couplage et une *oriT* (origine de transfert) et ne possédant pas de gènes du MPF. Les critères de co-localisation de ces gènes ne sont malheureusement pas plus explicités.

L'intérêt de cet outil est qu'il annote automatiquement les éléments au nucléotide près. Toutefois, il ne recherche qu'une partie des IME.

4. Difficultés de l'identification des ICE et des IME

Actuellement, malgré le développement récent des outils CONJscan et ICEfinder, l'annotation des ICE mais surtout des IME de Firmicutes n'est toujours pas satisfaisante et reste encore un défi majeur. Ceci est dû d'une part au manque de connaissances concernant les éléments de Firmicutes et d'autre part à la nature intrinsèquement complexe des éléments recherchés :

- **De nombreux modules ne sont pas spécifiques des ICE et IME** : les ICE et IME sont définis par une combinaison de modules mais aucun d'eux n'est spécifique de ces éléments. Par exemple, leurs modules de recombinaison sont aussi présents chez les phages et leurs modules de transferts peuvent être présents chez les plasmides.

- **Les génomes contiennent souvent des éléments partiels** : de nombreux éléments portent des séquences d'insertions (IS), qui provoquent fréquemment des délétions des gènes adjacents, y compris dans les ICE et IME.

- **Beaucoup d'éléments ont une structure composite** : de nombreux éléments ne sont pas isolés mais sont, soit emboîtés les uns dans les autres, soit en tandem (accrétion). Ces structures complexes sont difficiles à résoudre et rendent l'identification précise des éléments complexes.

- **Les protéines signatures sont souvent difficiles à détecter** : tous les outils existants se basent sur une première étape de détection de protéines signatures qui conditionne la détection correcte des éléments. Or cette première étape est difficile pour plusieurs raisons :

    - **Les protéines signatures présentent une grande diversité** : il existe une très grande diversité des familles des SP qui composent ces modules. Il est aussi probable que les connaissances actuelles sur ces familles de protéines signatures soient encore très partielles, notamment chez les Firmicutes dont les éléments conjugatifs sont bien moins étudiés que ceux des protéobactéries.

    - **L'annotation fonctionnelle des protéines signatures est non homogène et souvent ambiguë ou erronée dans les banques publiques** : en conséquence, il est difficile d'extraire automatiquement les protéines signatures des

ressources publiques. Par exemple, la relaxase TcpM de Clostridioides est annotée comme appartenant à la superfamille des recombinases à tyrosine. Un autre exemple concerne la relaxase MOB_T qui est annotée comme appartenant à la famille des initiateurs de réplication Rep-trans impliqués dans la réplication de nombreux plasmides. Dans le cas de Tn*916* et d'ICE*Bs1*, ces relaxases MOB_T sont impliquées à la fois dans le transfert conjugatif de l'ICE et dans le maintien par réplication de l'ICE excisé.

- **Les gènes codant les protéines signatures peuvent exister sous forme de pseudogènes** : les protéines d'ICE et d'IME n'étant souvent pas essentielles à leur hôte, les séquences codantes (CDS) des gènes correspondants sont fréquemment dégradées par des mutations : mutations ponctuelles ou délétions/insertions engendrant des décalages de phase de lecture. Comme les pseudogènes sont souvent mal ou non annotés dans les banques publiques, cela complexifie la détection des protéines signatures.

5. Objectifs de la thèse

Mon travail de thèse s'inscrit dans la continuité des travaux réalisés chez les streptocoques par l'équipe de l'unité DynAMic (Ambroset *et al.*, 2016; Coluzzi *et al.*, 2017; Coluzzi, 2017) et se base sur la forte expertise de l'équipe sur les ICE et IME présents au sein de ce genre. Les travaux antérieurs à mon arrivée avaient ainsi permis :

- la mise en place d'une méthodologie semi-automatique de recherche des éléments dans les génomes de streptocoques et le développement du prototype ICE/IME Finder de l'unité DynAMic,
- l'établissement des règles de détection des protéines signatures de streptocoques par recherche d'homologues par BlastP ou par l'utilisation de profils HMM et la mise au point des paramètres et critères de filtrage permettant de limiter la détection de faux positifs,
- la constitution d'une banque de protéines signatures provenant majoritairement d'ICE et d'IME de streptocoques (Ambroset *et al.*, 2016; Coluzzi *et al.*, 2017).

Au début de mon travail seule la première étape de recherche des protéines signatures était automatisée.

Mon objectif de thèse était double :

- automatiser au maximum l'annotation des éléments de streptocoques.
- élargir la méthode à l'identification des ICE et des IME d'autres Firmicutes.

Ces deux objectifs m'ont conduit à travailler selon deux axes différents au cours de mon doctorat :

- acquérir une expertise sur la détection et l'annotation des ICE et IME chez les Firmicutes, en particulier en commençant par analyser en détail le contenu en ICE/IME des génomes de l'espèce *S. salivarius*, ce qui a conduit à une première publication (article du chapitre 1 des Résultats).

- concevoir, développer et évaluer un nouvel outil bioinformatique, ICEscreen, qui se base en partie sur le prototype ICE/IME Finder mais apporte plusieurs nouveautés : des nouveaux profils HMM dédiés à la détection d'IME, des règles supplémentaires pour caractériser automatiquement les types d'éléments et un nouvel algorithme récursif permettant la résolution des éléments composites et l'identification de nouveaux éléments chez les Firmicutes. Une partie importante de mon travail a consisté à concevoir cet outil et à le développer avec l'aide de Thomas Lacroix (unité MaIAGE), puis à évaluer ses performances sur un jeu de données de 40 génomes de Firmicutes : le jeu de données FirmiData dont les éléments ont été annotés manuellement par Gérard Guédon (unité DynAMic). Les performances de ICEscreen ont été comparées à celles de CONJscan et ICEfinder sur ce même jeu de données. Ce travail n'est pas encore publié mais donnera lieu à une publication en 2021.

Résultats

1. Caractérisation des éléments conjugatifs intégratifs de *Streptococcus salivarius*

1.1 Introduction

Au laboratoire DynAMic, les ICE de la famille ICE*St3* sont très étudiés et notamment ceux identifiés au sein de l'espèce S*treptococcus thermophilus*. Les ICE de cette famille s'intègrent et s'excisent de manière site-spécifique (Burrus *et al.*, 2002a; Pavlovic *et al.*, 2004). Ils se transfèrent de manière autonome entre souches de *S. thermophilus* mais des transferts interspécifiques (vers d'autres espèces de streptocoques ou vers *Enterococcus faecalis*) ont aussi été mis en évidence (Bellanger *et al.*, 2009). Certains ICE*St3* forment des structures composites et se transférent seuls ou sont capables de mobiliser d'autres éléments *en cis* (Bellanger *et al.*, 2011). En 2016, le laboratoire DynAMic a publié une étude démontrant que la famille ICE*St3* était répandue chez les streptocoques (Ambroset *et al.*, 2016) mais que les trois génomes analysés de *Streptococcus salivarius*, une espèce extrêmement proche de *S. thermophilus* (Delorme *et al.*, 2015), était dépourvue de ce type d'élément. Ce résultat surprenant nous a conduit à collecter des souches de *Streptococcus salivarius*, à séquencer leur génome afin d'étudier la prévalence des ICE au sein de cette espèce.

L'espèce *S. salivarius* regroupe des souches commensales qui colonisent la cavité buccale humaine juste après la naissance et y persiste comme espèce prédominante. Cette espèce est également présente dans le tube digestif (Qin *et al.*, 2010) mais colonise également d'autres tissus (peau, cavité urogénitale...). Quelques souches de *S. salivarius* sont aussi décrites comme des pathogènes opportunistes et sont associées à diverses maladies (cas sporadiques de méningite, d'abcès pancréatique, d'impétigo, de péritonite, de sinusite). Les génomes de *S. salivarius* sont connus pour leur évolution rapide, vraisemblablement grâce aux transferts horizontaux (Delorme *et al.*, 2007).

L'ensemble de ces données a conduit le laboratoire DynAMic à collecter à la fois des souches de *S. salivarius* issus d'individus sains et des souches cliniques pour lesquelles le contenu en ICE*St3* a été recherché. Au total 13 ICE de la famille ICE*St3* ont été identifiés et le transfert de

deux d'entre eux démontré (Dahmane *et al.*, 2017). C'est dans ce contexte que se situe mon travail. L'objectif de ce travail était multiple :

- Tester une première stratégie, proche de celle ensuite implémentée dans ICEscreen, sur un large panel de souches de *S. salivarius* afin d'identifier l'ensemble des ICE et IME portés par ces génomes.
- Décrire la structure et délimiter les éléments en recherchant le site d'insertion.
- Caractériser fonctionnellement les gènes du module d'adaptation pour identifier des fonctions d'intérêt associées aux éléments.
- Évaluer globalement la nature, la diversité et l'impact des éléments sur l'évolution des génomes de *S. salivarius.*
- Acquérir une expertise sur l'annotation et la caractérisation fine des éléments d'une espèce d'intérêt

1.2 L'article

genes

MDPI

Article

Abundance, Diversity and Role of ICEs and IMEs in the Adaptation of *Streptococcus salivarius* to the Environment

Julie Lao [1,2], Gérard Guédon [1], Thomas Lacroix [2], Florence Charron-Bourgoin [1], Virginie Libante [1], Valentin Loux [2], Hélène Chiapello [2], Sophie Payot [1] and Nathalie Leblond-Bourget [1,*]

[1] Université de Lorraine, INRAE, DynAMic, F-54000 Nancy, France; julie.lao@univ-lorraine.fr (J.L.); gerard.guedon@univ-lorraine.fr (G.G.); florence.charron@univ-lorraine.fr (F.C.-B.); virginie.libante@univ-lorraine.fr (V.L.); sophie.payot-lacroix@inrae.fr (S.P.)

[2] Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France; thomas.lacroix@inrae.fr (T.L.); valentin.loux@inrae.fr (V.L.); helene.chiapello@inrae.fr (H.C.)

* Correspondence: nathalie.leblond@univ-lorraine.fr; Tel.: +33-3-72-74-51-46

Received: 30 June 2020; Accepted: 21 August 2020; Published: 26 August 2020

check for updates

Abstract: *Streptococcus salivarius* is a significant contributor to the human oral, pharyngeal and gut microbiomes that contribute to the maintenance of health. The high genomic diversity observed in this species is mainly caused by horizontal gene transfer. This work aimed to evaluate the contribution of integrative and conjugative elements (ICEs) and integrative and mobilizable elements (IMEs) in *S. salivarius* genome diversity. For this purpose, we performed an in-depth analysis of 75 genomes of *S. salivarius* and searched for signature genes of conjugative and mobilizable elements. This analysis led to the retrieval of 69 ICEs, 165 IMEs and many decayed elements showing their high prevalence in *S. salivarius* genomes. The identification of almost all ICE and IME boundaries allowed the identification of the genes in which these elements are inserted. Furthermore, the exhaustive analysis of the adaptation genes carried by these elements showed that they encode numerous functions such as resistance to stress, to antibiotics or to toxic compounds, and numerous enzymes involved in diverse cellular metabolic pathways. These data support the idea that not only ICEs but also IMEs and decayed elements play an important role in *S. salivarius* adaptation to the environment.

Keywords: integrative and conjugative elements; integrative and mobilizable elements; conjugation; antibiotic resistance; metabolic functions

1. Introduction

Conjugation is a horizontal gene transfer (HGT) mechanism that massively contributes to the evolution of prokaryotic genomes [1–3]. It is mediated not only by extrachromosomal elements (i.e., plasmids), but also by other mobile genetic elements (MGEs) that are integrated into the chromosome or plasmids of their host (for a review [2]). Integrated elements that transfer by conjugation include: (i) the integrative and conjugative elements (ICEs), (ii) the integrative and mobilizable elements (IMEs) and (iii) decayed elements deriving from ICEs or IMEs, such as cis-mobilizable elements (CIMEs) (for reviews see [2–4]). ICEs are autonomous conjugative elements: they carry a recombination module and a conjugation module that together ensure the excision of the element, its transfer by conjugation, its replication during transfer and its integration in donor and recipient genomes (for reviews see [2–6]).

The ICE recombination module includes the genes and sequences dedicated to the excision from and integration into the bacterial chromosome or plasmid. It encodes one (or several) protein(s) belonging to one of the three phylogenetically and structurally unrelated families of enzymes: tyrosine

integrases, serine integrases and DDE transposases [2–7]. Tyrosine and serine integrases generally catalyze excision by site-specific recombination between short direct repeats of the *att*L and *att*R flanking sites. This leads to an *attI* site that includes a single copy of this sequence. After transfer, most of them promote integration by catalyzing a site-specific recombination between this sequence from *attI* and another copy carried by the *attB* chromosomal site. As a consequence, the integrated ICE is flanked by DRs. The DDE transposases recognize terminal inverted repeats. The integration of elements encoding a DDE transposase generates target duplication, also leading to short DRs flanking the element.

The conjugation module of all ICEs from Firmicutes is dedicated to their transfer as single-strand DNA. It encodes a relaxase, a coupling protein (CP) and a "type IV secretion system" (T4SS) [8,9], including VirB4, a conserved ATPase providing energy [10]. Three distinct superfamilies of relaxases have been identified thus far in ICEs from Firmicutes: the MobP, MobC and MobT [2,11,12]. The relaxase recognizes and cleaves one of the DNA strands of the circular element, specifically at the *oriT* site [13–15]. It is then recognized by a membrane-associated CP belonging to either the VirD4 or TcpA superfamilies. Together, CP and T4SS ensure the translocation of the relaxase-tethered DNA from the donor to the recipient bacteria. 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 [2]).

IMEs are mobile elements that dispose of a recombination module similar to that of ICEs that allows their autonomous integration and excision. However, they cannot self-transfer. Instead of a conjugation module, they carry a mobilization module that does not result from a recent decay of a conjugation module. It ensures their mobilization in trans: IMEs subvert the conjugative machinery of a co-resident conjugative element (plasmid or ICE) to promote their own transfer [3]. Known IMEs use many different mobilization strategies and therefore exhibit diverse mobilization modules [3]. In this work, we consider as IMEs only elements whose mobilization module encodes a relaxase, eventually a CP but no T4SS protein (no VirB4). In Firmicutes, IMEs encode relaxases that belong to superfamilies found in ICEs (MobP, MobC or MobT) and conjugative plasmids (MobV or MobQ). We recently proposed that IMEs could also encode relaxases belonging to superfamilies of initiators of rolling circle replication harboring PF01719, PF01719-PF00910, PF02407 or PHA00330 domains [16].

CIMEs are decayed elements deriving from ICEs and IMEs by deletion that retained their *att* recombination sites but not their conjugation/mobilization modules and all their genes involved in recombination (for a review see [2]). Hence, CIMEs can only transfer by cis-mobilization resulting from an accretion-mobilization process [17]. This latter takes place when an ICE or an IME integrates in one of the attachment sites flanking the element resulting in a composite element that can excise and transfer by conjugation.

The vast majority of these integrated mobile elements carry adaptive genes that may confer on their host a significant selective advantage or may change their lifestyle (e.g., antibiotic, heavy metal or phage resistance, sucrose catabolism, bacteriocin synthesis, pathogenicity or symbiosis) [2,3,10,18–21]. It is therefore important to study the prevalence of these elements and to identify the adaptive function they tend to disseminate among bacterial populations.

In this work, we focused on *S. salivarius*, a species of Firmicutes that belongs to microbiomes of all humans and contributes to the maintenance of oral, pharyngeal and gut health [22–24]. Some *S. salivarius* strains are also described as opportunistic pathogens since they have been associated with cases of meningitis [25,26], endocarditis [27] and bacteremia in immunocompromised patients [28,29]. *S. salivarius* genomes are known to evolve rapidly, presumably through HGT [30]. Indeed, there is accumulating evidence of the pivotal role played by conjugative and/or mobilizable elements in *S. salivarius* HGT [31]. In a previous study, we highlighted the occurrence of ICEs belonging to the ICE*St3* family in 13 *S. salivarius* genomes [32]. In this work, we enlarged the number of *S. salivarius* genomes analyzed (n = 75) and searched not only for ICE*St3* elements but also made an exhaustive search of the diversity and abundance of all ICE and IME families. We delimited almost all the identified

elements, precisely defined their insertion site and searched for CIMEs that are integrated in tandem with ICEs and IMEs. We also characterized the adaptive functions carried by these mobile elements. Altogether, these data shed light on the diversity and prevalence of ICEs and IMEs in *S. salivarius*. They also make a comprehensive picture of the role of these mobile elements in *S. salivarius* adaptation to the environment.

2. Materials and Methods

2.1. S. salivarius Genomes and Phylogenetic Analysis

2.1.1. *S. salivarius* Strains and Genome Analysis

In this work, the genomes of 75 *S. salivarius* strains were analyzed. Dates and sites of sampling of the strains are given in Table S1. Twenty of the 21 strains isolated from our strain collection [33] were subjected to whole genome sequencing using an Illumina HiSeq2000 sequencer by Beckman Coulter Genomics (2 × 100 bp after paired-end library construction, at least 60 × coverage). De novo assemblies were performed using CLC Genomics Workbench (CLC Bio) using default parameters. Scaffold of the genomes was built by using the Genome Finishing module of CLC Genomics Workbench with the *S. salivarius* JIM8777 genome as Guédon et al. [34]. Some assembly gaps were filled by PCR and sequencing. Automatic annotation for each genome utilized the pipeline AGMIAL (Bryson 2006 agmial). The sequences (raw reads and assembled scaffolds) of the 20 strains have been deposited in the EBI-ENA database under the study number PRJEB37543. Accession numbers of the assemblies are also indicated in Table S1. The coordinates of the elements are included in Table S2. The remaining 54 genomes were retrieved from the NCBI genome databank either as complete genomes ($n = 7$) or as scaffolds of WGS ($n = 47$) (last accessed may 2019). Pseudocontigs were generated using CONTIGuator (http://combo.dbe.unifi.it/contiguator, with default parameters) with the *S. salivarius* JIM8777 genome as reference. Unmapped contigs were added at the end of the pseudocontigs. When available, the annotations were transferred to contigs using Geneious prime 2020.1.1. The nucleotidic sequences of elements are available in the Supplementary File S1—55.

2.1.2. Phylogenetic Tree Based on Single Nucleotide Polymorphisms (SNPs)

Phylogenetic relationships among *S. salivarius* strains were evaluated by analysis of single nucleotide polymorphisms (SNPs) using the CSI Phylogeny software (version 1.4) on the CGE website (https://cge.cbs.dtu.dk/services/CSIPhylogeny/). A multifasta file was generated using the NCTC8618 type strain as reference and including SNPs of all the 75 *S. salivarius* aligned genome assemblies. The phylogenetic tree of the strains was inferred using the Maximum Likelihood method based on the Tamura-Nei model implemented in the Mega7 software [35].

2.2. Detection and Delineation of the Integrative Elements

The workflow used to detect, delineate and classify ICEs was described previously [12]. Briefly, for ICE identification, signature proteins of the recombination module (integrase) and the conjugation module (relaxase, CP and VirB4) were searched by BLASTp comparison (BLAST 2.9.0+) against a curated database of 1029 signature proteins extracted from Firmicutes ICEs and IMEs. False positives were then filtered out by retaining only candidates that meet the four following criteria: the identity percentage (≥25%), the alignment coverage (≥40%, must cover the functional domain), the *E*-value ($\leq 1 \times 10^{-5}$ for CP and VirB4, $\leq 1 \times 10^{-4}$ for relaxase and integrase) and the hit length (≥320 aa for integrase, ≥180 aa for relaxase, ≥500 aa for VirB4, between 180 aa and 700 aa for short CP and between 1000 aa and 1200 aa for long CP). Genes encoding signature proteins were then co-localized. If all four proteins were present, the element was considered as ICE. Its delineation was done by searching for DRs at their two ends by BLASTn analysis (word = 7 bp) using either the 3′ or the 5′ end of the potential target gene as query. When DRs were absent, too short or too degenerated to be detected

by BLASTn, the sequence of the region containing signature proteins was compared to chromosomal sequences of *S. salivarius* strains devoid of the analyzed element and/or to sequences of elements sharing a very closely integrase using Megablast (word = 16 bp). The DRs were then identified by manually comparison of the ends of syntenic regions. To resume, all elements flanked by recombination sites and/or DRs and which encode an integrase, a relaxase, a CP and a VirB4 were considered as ICEs. Elements that lack one or two of these characteristics and clearly derive from a closely related ICE were counted as dICEs (decayed ICEs). The classification of ICEs/dICEs into families was based on the nature of the proteins of the conjugation module and was carried out as previously described [12].

IME identification and delineation were done as previously defined [16]. Briefly, IME detection is based on the combined presence of a recombination module (detected by its integrase) and a mobilization module (dedicated relaxase, eventually a CP and absence of other proteins of the conjugation module). The detection of genes encoding these signature proteins and the filtration of false positives were performed as described above. Even if various genuine IMEs from proteobacteria and one of firmicutes do not encode their own relaxase [3], elements were considered as IMEs in this work if: (i) they encode their own putative relaxase, regardless of their ability to encode a putative CP, (ii) they do not carry any putative VirB4 gene or pseudogene and (iii) their relaxase (and eventually their CP) is very distantly related (<40% identity) or unrelated to any relaxase (or CP) of any ICE or conjugative plasmid. Elements were identified as dIMEs (decayed IMEs) if a signature protein or *att* site was predicted non-functional. Integrated elements that encode their own integrase but do not encode other signature proteins (nor prophage signature genes) were counted as mobile genomic islands (MGIs). Decayed elements were counted as CIMEs if they were found devoid of functional genes encoding all the signature proteins. The method used in this work (search of signature genes and flanking recombination sites/DRs) allows the detection of CIMEs only if they are integrated in tandem with ICEs, dICEs or IMEs. As a reminder, the Figure S1 schematizes of the attributes used to discriminate the different types of elements.

To evaluate the number of these elements, we chose to count separately the elements that appear integrated in tandem, whatever the nature of the element. The denomination of the elements includes the putative nature of the element (ICE, IME, CIME or MGI), the host strain and its insertion site or specificity, for example IME_*SsalL25_oriT*.

Circos6 [36] was used to show associations of signature proteins in elements and the content of MGEs of each strain. Manual editing of the figures was done using Inkscape.

2.3. Characterization of Cargo Genes Encoded by ICEs and IMEs

Taking into account data on cargo genes in the literature, 31 functional categories related to fitness and 10 category tags were defined. The attribution of cargo genes encoded by ICEs and IMEs to these categories was based on keywords found in the functional annotation of the genes or on alignment with six external resources with significant hits: AMRFinderPlus [37], BACTIBASE [38], VFDB [39], REBASE [40], NORINE [41] and MEROPS [42]. AMRFinderPlus is packaged with its own search engine (AMRfinder) and was used with default parameters. BLASTp was used to query the other resources. Alignments were considered significant if they met the following stringent criteria: (i) for protein of size ≤100 aa: coverage ≥ 80%, identity ≥ 80%, e-value ≤ 1×10^{-10}; for protein of size >100 aa and ≤250 aa: coverage ≥ 70%, identity ≥ 70%, e-value ≤ 1×10^{-20}; (ii) for protein of size >250 aa and ≤500 aa: coverage ≥ 65%, identity ≥ 65%, e-value ≤ 1×10^{-40}; (iii) for protein of size >500 aa: coverage ≥ 60%, identity ≥ 60%, e-value ≤ 1×10^{-60}. The assignments in the categories were carried out manually.

3. Results

3.1. ICEs in S. salivarius Genomes

3.1.1. ICE Prevalence and Diversity

Among the 75 *S. salivarius* genomes analyzed, more than 2/3 (n = 53) carry at least one ICE or dICE and only 22 are devoid of ICEs Table S3, Figure 1. A total of 69 ICEs (and eight dICEs) were identified and classified into superfamilies and families. This classification was carried out as previously described [13] and takes into account the nature of the domains carried by signature proteins for superfamilies (relaxase, VirB4 and CP) and the 40% identity clustering of these signature proteins for families. Among the seven families of ICEs described thus far in streptococci [12], four are present in the genomes of *S. salivarius*: Tn*916*, ICE*St3*, Tn*GBS2* and Tn*1549* families.

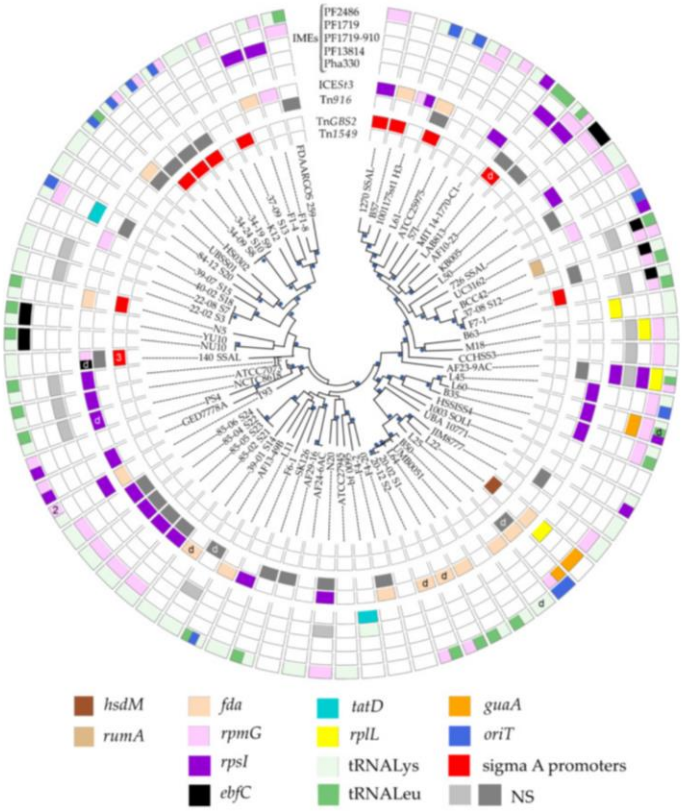

Figure 1. Integrative and conjugative elements (ICEs) and integrative and mobilizable elements (IMEs) carried by *Streptococcus salivarius* genomes. Strains are indicated in the middle of the circle and were grouped according to their phylogenetic relatedness. Groupings with bootstrap values >90 are marked with blue dots. Boxes located in front of a strain name are either empty (absence of element) or colored (presence of an ICE or an IME). The different families of ICEs and IMEs are indicated at the top in the opening of the circle. The first four inner circles indicate the presence of elements belonging to the four distinct families of ICEs retrieved in this study. The five outer circles show the presence of IMEs, where each line corresponds to one IME superfamily. Defective ICEs or IMEs are indicated by a "d." The box is split in two (or more) when two elements (or more) of the same superfamily are present at distinct integration sites. If several elements of the same superfamily are present at the same integration site, the number of elements is indicated in the box (for example, two IMEs with a PF02486 domain in strain T93). The colored boxes indicate the genes where the elements are inserted. The medium-grey boxes materialize low-specific integration (NS).

In a previous work, we already explored the presence of ICE*St3*-related elements for 13 of the 21 strains of our collection [32]. For the sake of completeness, we show here all the elements present in these strains Table S3, Figure 1. ICE*St3*-related elements are not only present in 18 of the 21 strain of our collection but also in other *S. salivarius* genomes. Thus, this family of ICEs is the most prevalent one (n = 38 including five dICEs). Two other families of ICEs were also found in abundance: Tn*916* family (n = 24 including two dICEs) and Tn*GBS2* family (n = 13 including one dICE). Two Tn*916* (including one dICE) were found on megaplasmids. Only two strains carry a Tn*1549*-related element.

Most genomes (n = 34) exhibit only one ICE or dICE, but 18 genomes carry several ICEs and/or dICEs, usually two that belong to two distinct ICE families Figure 1. The most frequent co-occurrence is Tn*916*-related element with ICE*St3*-related element (n = 10). Only two strains encode two ICEs belonging to the same family (strains 140 SSAL and 1001175st1_H3 with two ICE*St3*-related elements). Strain 140 SSAL differentiates from other strains by its richness in ICEs since it harbors six ICEs or dICEs belonging to Tn*916* (n = 1), ICE*St3* (n = 2 including one dICE) and Tn*GBS2* (n = 3) families.

Most of the Tn*916*-related dICEs or ICEs harbor regulation, conjugation and integration modules are almost identical to those of Tn*916* (>99% identity). Many of these elements carry some transposon insertion(s), in particular in *orf*9, that were previously identified in closely related ICEs [43]. However, two very closely related ICEs (ICE_*SsalL61_Tn916* and ICE_*SsalLAB813_Tn916* showing >99% identity), which do not carry any transposon, share only 90% identity with Tn*916* over their entire length. Two other elements seem to be chimerical. Thus, the left part of dICE_*SsalAF13-49B_Tn916* (from *attL* to an internal position of *orf16*) is almost identical to ICE_*SsalLAB813_Tn916*, whereas its right part is almost identical to that of Tn*916* (>99% identity) with a transposon insertion in *orf*9. This suggests that this element is a chimera resulting from a homeologous recombination between two Tn*916*-related elements sharing only 90% identity. Furthermore, the left part of ICE_*SsalSK126_Tn916* (from *attL* to the *nic* site cut by the relaxase) shares only 91% identity with Tn*916*, whereas its right part is almost identical to Tn*916* (>99% identity) with an insertion in *orf*9. This suggests that this element is a chimera resulting from a recombination catalyzed by the relaxase of the element between two Tn*916*-related elements sharing only 91% identity. Such recombinations were previously reported for other ICEs but not for Tn*916*-related elements (for a review see [2]).

3.1.2. ICE Integrases and Integration Sites

In ICEs/dICEs found in *S. salivarius*, the most prevalent integrases are tyrosine integrases (n = 62) that are found in 80% of them. As expected, the tyrosine integrase encoded by the elements of the Tn*916* family has a low specificity of integration. All ICE*St3*-related elements (including dICE) encode a tyrosine integrase and are integrated in four well-conserved housekeeping genes of *S. salivarius* genomes, i.e., in the 3′ end of *rpsI* (encoding the S9 ribosomal protein) (n = 17 including one dICE), *fda* (fructose-1,6-bisphosphate aldolase gene) (n = 15 including three dICEs) and *rpmG* (encoding the L33 ribosomal protein) (n = 5) and in the 5′ end of *ebfC* (encoding a nucleoid associated protein) (n = 1 dICE) Table S3, Figure 1.

The second most prevalent integrases in ICEs/dICEs of *S. salivarius* are the DDE transposases belonging to the IS*Lre2* transposase family. These integrases were exclusively found in ICEs belonging to the Tn*GBS2* family (n = 13 including 1 dICE) Table S3, Figure 1. DDE transposases of this family target diverse sigma A promoters and therefore avoid integration of elements into various genes [44,45]. Our results are consistent with these findings.

The two ICEs belonging to the Tn*1549* family rely on serine integrase(s) for their integration/excision. They are specific of sites located within some widespread but dispensable genes. One ICE encodes a serine integrase that targets *rumA* (encoding a 23S rRNA (uracil-5-) methyltransferase). The other encodes a triplet of serine integrases in the same orientation and is inserted in *hsdM* (encoding a methyltransferase subunit of a type I restriction-modification system) Table S3, Figure 1.

3.1.3. Slightly Decayed Elements Deriving from ICEs

Eight dICEs which are closely related to ICEs belonging to the ICE*St3*, Tn*916* or Tn*GBS2* families were found. Four of them result from the pseudogenization of one of the genes encoding a signature protein and another cannot be precisely delimited but seems to lack relaxase and CP encoding genes.

Two strains, Ssal20-12 S2 and Ssal20-02 S1, carry ICE*St3*-related identical dICEs integrated into the 3′ end of their *fda* gene. In ICE*St3*, the conjugation and recombination modules are transcribed as a unique operon (*orfONMLK-oriT-orfJIHGFEDBA-xis-int*) [46]. The related region of both dICEs differs by a large but precise internal deletion of *orfIHGFEDBA* that probably encodes the T4SS. These dICEs contain all other genes resulting in a putative operon *orfONMLK-oriT-orfJ-xis-int*, which shares 94.1% identity with the related sequences of ICE*St3*. Another element, dICE *Ssal39-01_fda*, has also this large deletion of *orfIHGFEDBA* (located exactly at the same position as the one found in dICE_*Ssal20-12_fda*). However, it is different from the other two since it has two additional deletions in *orfK* (one including the 5′ end of the gene and another within the gene). Its *orfONMLψorfK-oriT-orfJ-xis-int* putative operon shares only 94.2% identity with the related region of dICE_*Ssal20-12_fda*. These three elements could be derived from ICE*St3*related elements by deletion and can no longer transfer autonomously. They could be mobilizable by using the functional T4SS of ICE*St3*-related ICE.

3.2. IMEs in S. salivarius Genomes

3.2.1. IME Prevalence and Diversity

Among the 75 *S. salivarius* genomes, more than 90% (n = 69) carry at least one IME/dIME and only six genomes are devoid of it. Among the latter, two are also devoid of ICEs (*S. salivarius* HSISS4 and *S. salivarius* F4-20). A total of 165 IMEs and two dIMEs were identified, one of which is carried by a plasmid (IME_*SsalM18_rpmG*) (Table S3). As seen in Table S3, the occurrence of IMEs/dIMEs varied within strains. Numerous chromosomes carry two IMEs/dIMEs (n = 30) but this number can go up to seven in strain AF23-9AC.

These 167 IMEs/dIMEs were classified into five superfamilies in Table 1 by taking into account the domain composition of their relaxase that is the main protein of the mobilization module. The most prevalent superfamily of IMEs/dIMEs is the IME_PF02486 superfamily (n = 113, including two dIMEs) that comprises elements with a relaxase exhibiting a PF02486 domain Table 1, also known as Rep_trans. The four other superfamilies of *S. salivarius* IMEs/dIMEs are IME_PF01719 (n = 28), IME_PF13814 (n = 13), IME_PHA00330 (n = 6) and IME_PF01719-PF00910 (n = 7).

The 40% sequence identity clustering of the relaxases sharing the same catalytic domain allows to classify relaxase families. In total, 11 families of relaxases were retrieved Table 1, column 2. The comparison of these families with those previously described in streptococci [16] reveals the existence of a new family of relaxases, Rel_PF01719-PF0910_5.

In this previous study [15], we also found that half of the streptococcal IME mobilization modules include a CP and that all these CPs except two do not belong to the canonical VirD4 superfamily but to the TcpA superfamily. Here, the percentage of *S. salivarius* IMEs/dIMEs encoding a CP is somewhat lower (33%, n = 54) and the majority of these CPs also belong to the TcpA superfamily, but the fraction is somewhat lower (77%, n = 42). A 40% sequence identity clustering of CP allows to subdivide the TcpA superfamily into seven families, see Table 1, column 3. The TcpA_2, TcpA_6 and TcpA_12 families were already described [15], whereas TcpA_13, TcpA_14 and TcpA_15 are novel families. The most prevalent family is the TcpA_12 (n = 32), as previously described for streptococcal IMEs [16].

The VirD4 proteins identified in this work (n = 12) all belong to the IME_PF13814 family. The association of a VirD4 protein with a Rel_PF13814 was already described in IME mobilization modules [16]. However, one Rel_PF13814 is not associated with a CP, which is unusual.

Table 1. Diversity of the relaxases and CPs associated with serine and tyrosine integrases in IMEs/dIMEs.

Integrase Type (Number of IMEs/dIMEs)	Relaxase Families (Number of Relaxases)	CP Families (Number of CPs)	IME Superfamilies [1]
Tyrosine (n = 155)	Rel_PF02486_2 (n = 65)	none (n = 65)	IME_PF02486 (IME_class_1)
	Rel _PF02486_4 (n = 4)	TcpA_2 (n = 4)	
	Rel _PF02486_5 (n = 21)	none (n = 22)	
	Rel _PF02486_6 (n = 22)	none (n = 22)	
	Rel _PF01719 (n = 28)	TcpA_12 (n = 28)	IME_PF01719 (IME_class_2)
	Rel_PF01719-PF0910_4 (n = 3) **Rel_PF01719-PF0910_5 (n = 4)**	TcpA_6 (n = 1) **TcpA_14** (n = 1) none (n = 1) TcpA_12 (n = 4)	IME_PF01719-PF0910 (IME_class_4)
	Rel_PHA00330_1 (n = 3)	**TcpA_15** (n = 1) none (n = 2)	IME_PHA00330 (IME_class_3)
	Rel_PHA00330_2 (n = 2)	TcpA_8 (n = 2)	
	Rel_PHA00330_3 (n = 1)	**TcpA_13 (n = 1)**	
	Rel_PF13814 (n = 1)	none (n = 1)	IME_PF13814 (IME_class_8)
Serine (n = 12)	Rel_PF13814 (n = 12)	VirD4 (n = 12)	

In bold, new families of relaxases and CPs. [1] The IME superfamily name comprises the pfam accession number of the relaxase catalytic domain(s). The correspondence with the names given in [16] is indicated in brackets.

3.2.2. IME Integrases and Integration Sites

S. salivarius IMEs/dIMEs encode integrases belonging to two unrelated superfamilies of recombinases, tyrosine integrases and serine integrases. Serine integrases were found in 12 IMEs that all belong to the IME_PF13814 superfamily Table 1. All these elements are integrated in intergenic regions, suggesting that these integrases are specific of elements or structures found in promoters or terminators, as the DDE transposases encoded by ICEs belonging to Tn*GBS1* and Tn*GBS2* families.

Tyrosine integrases were detected in more than 92% of the IMEs (n = 155/167). Most of the tyrosine integrases (n = 133) specifically target the 3′ end of housekeeping genes such as genes encoding tRNAs (tRNALys n = 55; tRNALeu n = 22) or ribosomal proteins being either *rpsI* (n = 13), *rpmG* (n = 39) or *rplL* (n = 4). More rarely, tyrosine integrases from IMEs catalyze integration at the 3′ or 5′ end of other protein-encoding genes (*guaA* (n = 3), *tatD* (n = 2) or *ebfC* (n = 6)). The last 11 IMEs encoding closely related tyrosine integrases are specifically integrated in the *oriT* sequence of ICEs belonging to the ICE*St3* (n = 6) or Tn*916* (n = 1) families or are integrated in secondary sites (n = 4) (see [47] for more details).

3.2.3. Diversity of Integrase-Relaxase-CP Combinations within IMEs

The analysis of the co-occurrence of the superfamilies and specificity of integrases, families of CPs and families of relaxases did not reveal exclusive associations, see Table 1 and Figure 2. Altogether, 21 different ternary associations were observed, suggesting a high frequency of shuffling between signature proteins. Among these associations, five have never been observed before: those involving the three newly identified superfamilies of CPs (TcpA_13; TcpA_14 and TcpA_15), those involving the newly discovered Rel_PF01719-PF0910_5 family of relaxases associated with CPs of the TcpA_12 family and those from the IME_PF13814 superfamily that are composed of a Rel_PF13814 associated with a tyrosine integrase in the absence of CP.

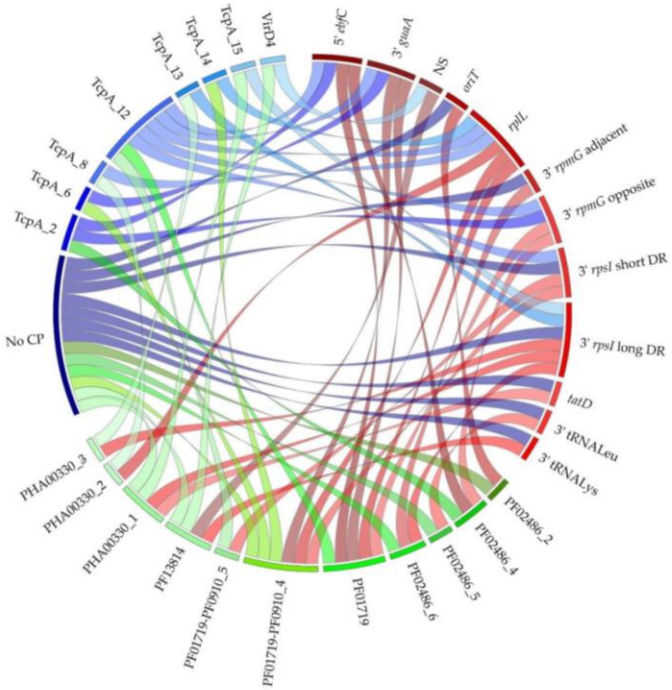

Figure 2. Diversity of integrase-relaxase-CP combinations for IMEs/dIMEs. Arcs show the group proteins belonging to the same family according to phylogenetic analysis and 40% sequence identity clustering: red, green and blue arcs for clustered integrases, relaxases and CPs, respectively. Ribbons indicate the association between integrases and relaxases in red, relaxases and CPs in green and CPs and integrases in blue.

3.3. Composite Elements in S. salivarius Genomes

In the 75 analyzed genomes, we found 33 complex genomic islands consisting of two elements integrated in tandem and four genomic islands composed of three elements. These composite genomic islands are integrated in the 3′ end of *rpmG* (n = 12), *rpsI* (n = 16), *fda* (n = 8) and *rplL* (n = 1).

CIMEs found in these structures can be divided in two classes according to their size and gene content: (i) classical CIMEs which carry cargo genes and have a size from 1.2 to 13 kb and (ii) microCIMEs, which do not carry any gene and have a size from 130 to 220 bp.

The 37 composite structures include 16 ICEs, three dICEs, 27 IMEs, 19 classical CIMEs and 11 microCIMEs. One of these composite structures include another mobile genetic element, MGI_*SsaL22_rplL*, which encodes its own tyrosine integrase, but does not carry any gene (or pseudogene) encoding relaxase, CP or VirB4. The analysis of this large element (32.3 kb) suggests that it does not derive from any ICE or IME. We found single related large elements (28–38 kb) with similar features in six other strains, suggesting that they are not decayed elements.

The order and the nature of the elements in these 37 composite structures resulting from accretion are highly variable. Table 2 indicates the different combinations of the tandem elements relative to the position of the target gene. Several rules seem to emerge from this analysis: (i) two ICEs are never integrated in tandem, (ii) two IMEs belonging to the same superfamily are never integrated in tandem, (iii) the most decayed elements, i.e., CIMEs and microCIMEs, are always the most distant from the target genes.

Complex elements can also result from the integration of an element within another, resulting in a matryoshka element. Ten matryoshka elements were identified: an ICE carrying an IME integrated in its *oriT* (see [47] for further details) (n = 7) and conjugative plasmids carrying a Tn*916*-related ICE (n = 1) or a Tn*916*-related dICE (n = 1), or an IME (n = 1).

Table 2. Composite structures resulting from tandem integration.

Structure of Composite Regions	Prevalence
CIME-IME-target gene	7
microCIME-ICE-target gene	6
CIME-ICE-target gene	4
IME-IME-target gene	3
CIME-IME-target gene	3
microCIME-IME-target gene	2
2 microCIMEs-dICE-target gene	2
IME-ICE-target gene	2
ICE-IME-target gene	2
CIME-IME-IME-target gene	2
microCIME-CIME-ICE-target gene	1
CIME-dICE-target gene	1
CIME-ICE-IME-target gene	1
MGI-IME-target gene	1

3.4. Function Encoded by the Cargo Genes

In addition to functions that are essential for their mobility, almost all conjugative and mobilizable elements carry cargo genes, i.e., genes not involved in the transfer of the element. In this work, blast analysis against diverse databases (AMRFinderPlus, BACTIBASE, VFDB, REBASE, NORINE and MEROPS) were undertaken to identify cargo genes encoded by *S. salivarius* elements. This allowed assigning the biological function of 667 potential cargo genes Figure 3. Figure S2 lists the elements carrying cargo genes and indicates their presumed function. The most frequent function was the one corresponding to "Signal transduction and regulatory system" that includes 230 genes of which 168 encode transcriptional regulators. This number is overestimated since it is not possible to precisely distinguish cargo regulatory genes from those dedicated to the control of ICEs and IMEs transfer. Nevertheless, many of these transcriptional regulators can be considered as cargo genes such as the genes found in seven *S. salivarius* elements that encode proteins homologous to the DeoR/GlpR-type known as a regulator of sugar metabolism [48]. These genes are all located next to a cluster of genes involved in carbohydrate metabolism in accordance with their presumed function. Another example is the *cadX*-like genes found in four elements that likely act as cadmium transcriptional repressors [49] of a gene cluster located nearby and that likely confers resistance on Cd^{2+} and Zn^{2+}.

Cargo regulatory genes also include 34 genes encoding transcriptional regulators and kinases probably composing two component systems (TCSs). These allow bacteria to sense and respond to changes in their environment [50]. They are carried by 12 *S. salivarius* elements. Five of the kinases carried by MGIs are related to ComD, a protein that leads to the activation of the competence regulon [51]. Eight CIMEs encode proteins with PAS/PAC domains that are commonly involved in environmental sensing (presumably of oxygen, redox, light or metals) [52,53].

Transporters is the second class of cargo genes (n = 148). This class comprises ABC transporters (n = 59) that couple the energy stored in adenosine triphosphate (ATP) to the movement of molecules across the membrane [41]. Six of them likely belong to the major facilitator superfamily (MFS) of transporters that are involved in the transport of a variety of substrates including antibiotics. Other transporters homologous to FtsX-like permeases (n = 6) or characterized by a YeiH domain (COG2855) of unknown function (n = 42) were retrieved. Diverse other membrane proteins (n = 32) were also identified but we were not able to assign them a precise function.

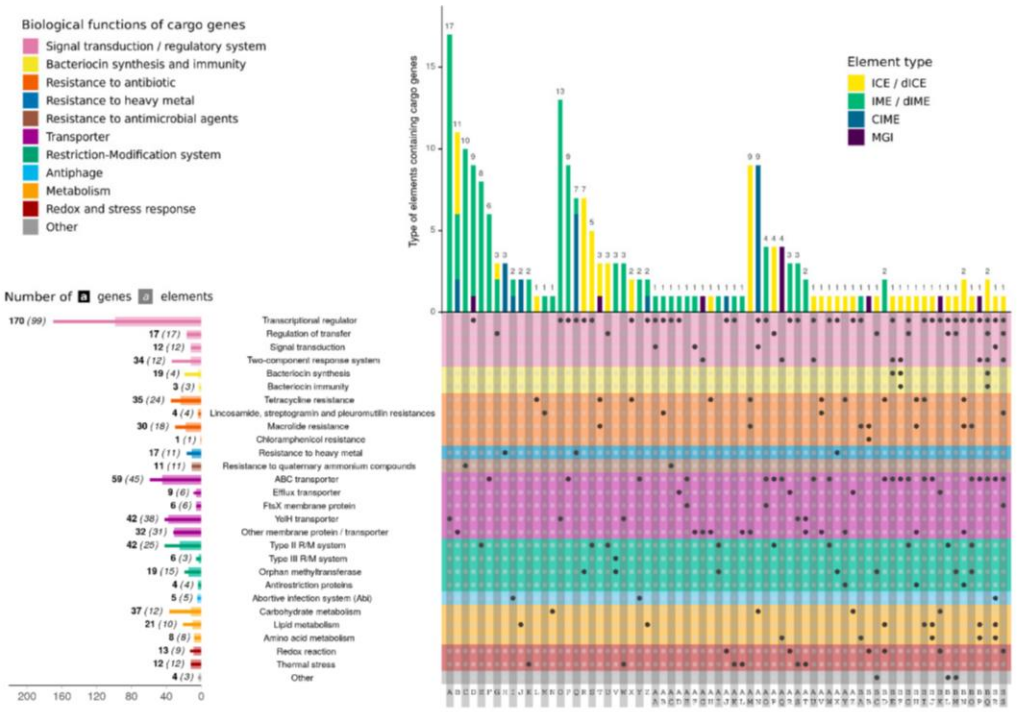

Figure 3. Function of cargo genes encoded by ICEs, IMEs and other MGEs from *S. salivarius*. The upper left indicates the different biological functions of the cargo genes. The upper right indicates the type of elements carrying the cargo genes. The lower left indicates the nature of cargo genes, their number (in bold) and the number of elements that encodes them (in brackets). The lower right connects the cargo functions with the different types of elements among them, IME_*Ssal1003-SOLI_rpsI* carries a cluster of four genes probably involved in the synthesis of polyketide fatty acid and ICE_*SsalL11_fda* displays a cluster of six contiguous genes that are potentially involved in the synthesis of fatty acids. Furthermore, two ICEs (ICE_*SsalF6-1_rpsI* and ICE_*SsalL25_fda*), one IME (IME_*SsalAF10-23_tRNAleu*) and five MGIs each encode one protein that potentially functions as an asparagine synthetase ($n = 5$), a 2-amino-3-ketobutyrate coenzyme A ligase ($n = 2$) or an aminotransferase ($n = 1$) that could potentially be involved in amino acid metabolism.

The third most frequent function encoded by cargo genes corresponds to restriction/modification (RM) systems. A total of 67 genes encoding restriction or modification proteins were retrieved. These include 42 type II RM proteins that are found on ICEs and IMEs. Type III RM ($n = 6$) are exclusively present on IMEs. Additional orphan methyltransferases ($n = 19$) were also retrieved frequently on ICEs or IMEs.

S. salivarius conjugative and mobilizable elements also encode proteins conferring antimicrobial resistance to their host cell. These include QacE efflux transporters ($n = 11$) that are carried by eight IMEs and likely confer resistance to quaternary ammonium compounds (QACs). They also comprise 70 proteins conferring resistance to diverse antibiotics. Indeed, 35 cargo genes were homologous to genes involved in resistance to tetracycline (*tet*(M)), 30 to macrolides (*erm*(B), *erm*(C), *erm*(F), *mef*(A), *msr*(D)), four to lincosamines (*lsa*(C)) and one to chloramphenicol (*catA*). At least 22 mobile elements of *S. salivarius* encode genes involved in resistance to antibiotics. The tetracycline and erythromycin resistance genes were mostly found on Tn*916*-related elements ($n = 14$). Four elements, two ICEs and two IMEs integrated in *oriT* of conjugative elements (IME_*SsaHS0302_oriT* and IME_*Ssa1001175st1_H3_oriT*, see [47]), encode a *lsa*(C) gene that is known to confer cross-resistance to lincosamides, streptogramin A and pleuromutilins in *S. agalactiae* [54]. Lastly, one MGI (MGI_*SsaF4-20_rumA*) encodes a chloramphenicol acetyltransferase that confers resistance to chloramphenicol. In addition to antibiotic resistance genes,

21 genes were found to be likely involved in bacteriocin synthesis and immunity. These were found on three ICEs and one dICE Figure S1. In addition, genes involved in resistance to the cadmium heavy metal (n = 17) were quite well represented in *S. salivarius* elements, mostly within CIMEs.

Elements also encode many proteins (n = 64) that possibly play a role in the cellular metabolism Figure S2. Indeed, 14 elements, mostly CIMEs (n = 10), each encode one to three proteins homologous to enzymes involved in carbohydrate metabolism. These enzymes potentially catalyze diverse functions such as 1-phosphofructokinase (n = 8), galactose-6-P isomerase (n = 4), tagatose-2P aldolase (n = 2), UDP-N-acetylglucosamine 2-epimerase (n = 7) and glycosyl transferase (n = 13). Several of these enzymes encoded by an MGI in the strain *S. salivarius* SK12 are likely responsible for the lactose assimilation via the tagatose-6 phosphate pathway [37]. In addition, 21 other cargo proteins are homologous to proteins involved in lipid metabolism. They are likely involved in diverse functions: synthase (e.g., 3-oxoacyl-[acyl-carrier-protein] synthase (n = 5), β-ketoacyl synthase (n = 1)), reductase (e.g., 3-oxoacyl-[acyl-carrier-protein] reductase (n = 1)) or kinase (diacylglycerol kinase (n = 3)). These enzymes are encoded by 10 elements (four ICEs, two IMEs, one MGI and three CIMEs). Lastly, *S. salivarius* integrative elements also encoded proteins that are required for stress tolerance. Indeed, nine elements (two CIMES, five IMEs and two ICEs) encode proteins (n = 13) that could catalyze redox reactions. Twelve other elements harbor one gene encoding a protein homologue to the LtrA protein which has been found to be essential for growth at low temperature (4 °C) in *Listeria monocytogenes* [55].

4. Discussion

Although ICEs belonging to the ICE*St3* and Tn*916* families were previously reported in some strains of *S. salivarius* [12,32,33], the real prevalence and diversity of ICEs in this species was still unknown. The present analysis identified 69 ICEs and eight slightly decayed ICEs (dICEs) in the 75 analyzed genomes, revealing a high prevalence of ICEs in this species. All these elements belong to the superfamilies (based on the domain content of their relaxases and CPs) and families (based on identities >40% of relaxases, CPs and VirB4) previously identified in the *Streptococcus* genus [12]. Most of them belong to the two previously reported families in *S. salivarius*, ICE*St3* (n = 38) and Tn*916* (n = 24), that encode MobT relaxases, TcpA CPs, MPF_{FA} T4SSs and tyrosine integrases. One of these ICE*St3* elements is specifically integrated in the 5′ end of *ebfC*, an integration site that has not yet been identified in *S. salivarius*. Among the *S. salivarius* set of genomes, 13 have been initially screened and selected because they carry members of the ICE*St3* family [32]. However, ICE*St3*-related elements are also present in many other genomes (25 ICEs or dICEs in 62 strains), showing the high prevalence of this ICE family in this species. The complete genomes retrieved from Genbank that were analyzed in this study include many strains that have been initially sequenced and studied because they are resistant to antibiotics. Therefore, since Tn*916*-related ICEs from streptococci confer resistance to antibiotics, the high frequency of the Tn*916* family found in this work is probably overestimated. Our analysis also revealed ICEs or dICEs belonging to two other families, Tn*GBS2* family (n = 13) and Tn*1549* family (n = 2), that have not been previously reported in *S. salivarius*. These two distantly related families encode MobP relaxases, VirD4 CPs and MPF_{FATA} T4SSs. Their DDE transposases or serine integrases have specificities that have never been reported in *S. salivarius*: (i) sigma A promoters for DDE transposases of Tn*GBS2*-related ICEs and (ii) internal sites of *rumA* or *hsdM* for serine integrases of Tn*1549*-related ICEs. Overall, this study extends the repertoire of ICEs and of their integration sites in *S. salivarius*.

An initial study of 1124 prokaryotic genomes based on the chromosomal location of complete conjugation modules (probably carried by ICEs) or of relaxase genes devoid of accompanying T4SS genes (probably carried by IMEs) suggested that IMEs slightly outnumber ICEs [1]. Thereafter, this prediction was corroborated by the only exhaustive searches of ICEs and IMEs encoding a relaxase in a large amount of strains: on 124 genomes of various streptococci [12–16] and on 214 genomes of *Streptococcus suis* [18]. Unexpectedly, our exhaustive search of IMEs in a large set of *S. salivarius* revealed many more IMEs (n = 165) than ICEs/dICEs (n = 77). It should also be emphasized that many

IMEs from proteobacteria and a few from Firmicutes carry their own integrase gene and *oriT* but do not encode any relaxase. In this study, elements encoding their own integrase and devoid of relaxase were not considered as IMEs but as MGIs. Therefore, the high prevalence of IMEs that we have found in *S. salivarius* could be underestimated.

This work revealed not only a large number of IMEs but also a huge diversity of their: (i) integration specificities (10 different integration specificities compared to seven for ICEs), (ii) relaxases (five superfamilies and 11 families compared to two superfamilies and four families for ICEs), and (iii) CPs (two superfamilies and eight families compared to only two superfamilies and four families for ICEs). Some families (three families of CPs and one family of relaxases) are reported here for the first time. In total, IMEs exhibit 21 different combinations of integrase-relaxase-CP families and insertion specificities Table S3 compared to seven for ICEs. Overall, this study greatly extends the repertoire of IMEs in *S. salivarius*.

This work is the first exhaustive search of *cis*-mobilizable elements (CIMEs) integrated in tandem with ICEs and IMEs in a large set of strains. It revealed 32 CIMEs or microCIMEs. However, our method did not allow to identify single CIMEs, and therefore probably missed many elements. The amount of detected CIMEs (n = 30) is similar to that of IMEs integrated in tandem and largely outnumbers the amount of ICEs integrated in tandem. Therefore, it suggests that, similarly to IMEs, the prevalence of CIMEs is much larger than the one of ICEs in *S. salivarius*. The only other published search of CIMEs in a large array of genomes (303 of *Streptococcus agalactiae*) concerns CIMEs integrated in the 3′ end of the tRNALys CTT gene alone or in accretion with ICEs or IMEs [51]. It demonstrated the presence of 215 CIMEs deriving from ICEs and IMEs besides 88 ICEs and 66 IMEs integrated in this locus. Taken at whole, these previous results and the present work suggest that the prevalence of CIMEs is very high.

The detected ICEs, IMEs and CIMEs are expected to transfer by conjugation. This has been confirmed for ICE*St3*-related ICEs ([32,47] and for one IME [47]. Furthermore, apart from the case of identical or almost identical strains (such as Nu10 and Yu10), the comparison of the distribution patterns of related ICEs or IMEs and of the phylogenetic tree Figure 1 shows that these elements were horizontally transferred between strains. In this work, we searched for signature genes encoding integrase, CP, relaxase and VirB4 proteins but we did not search for other genes needed for conjugative transfer of ICEs nor for *trans* mobilization of IMEs. Therefore, some of the elements that are reported as ICEs or IMEs may actually be decayed ICEs that do not encode their own transfer or decayed IMEs unable to subvert conjugation apparatus of conjugative elements. However, most of the decayed ICEs probably keep their transfer ability by *trans*-mobilization by related conjugative elements. The best example is dICE_*Ssal39-01_fda* whose "conjugation" module differs from that of ICE*St3* by the precise deletion of the genes encoding the T4SS proteins but not of all other genes and sequences involved in transfer. Furthermore, decayed elements could be cis-mobilizable if they are integrated in tandem with a related or distantly related functional IME or ICE encoding an integrase able to recognize the *attL* and *attR* sites of the composite structure. *Cis*-mobilization of distantly related elements is likely rare since their *att* sites are generally very different. However, tandem structures can have excision patterns that can be somewhat surprising. For example, IME_*Sag2603_tRNAlys* from *S. agalactiae* is integrated in the 3′ end of a tRNALys gene in tandem with a dICE, generating the composite structure $attL_{ICE}$-dICE-*attI*-IME-$attR_{IME}$-3′ end of a tRNALys gene. Although these two integrated elements encode very different tyrosine integrases (<30% identity) and have very different *att* sites, the IME excises by site-specific recombination between the chimerical *attI* site and $attR_{IME}$ and the whole composite element excises by site-specific recombination between $attL_{ICE}$ and $attR_{IME}$ [37].

ICEs and/or IMEs were found to be integrated specifically in 12 different target genes among which tRNALeu, tRNALys, *rpmG*, *rpsI* and *fda* are the most frequent. Integrations in tandem were observed only in three of these genes (*rpmG*, *rpsI* and *fda*). This suggests that IMEs that integrate specifically in the two frequently target genes encoding tRNALeu and tRNALys cannot integrate in the *att* sites flanking a resident element. However, it should be noticed that in *S. agalactiae*, IMEs related to the IMEs of *S. salivarius* that are specific of tRNALys genes are frequently integrated in tandem in

tRNALys genes with ICEs distantly related to ICE*St3* and with CIMEs [56]. Although no tandem of ICEs has been identified in *S. salivarius*, tandems of ICEs belonging to different families were found in *Clostridium difficile* [57] and in *S. suis* [18]. In this work, we did not detect tandems of IMEs or ICEs belonging to the same family, as previously described in streptococcal genomes [16–18]. This could be due to an inhibition of conjugation or a surface exclusion by a resident element in the recipient strain as testified for various ICEs (for a review see [2]). It could also be due to an instability of tandems resulting from recombination or interactions between related elements, as shown for various ICEs including ICE*St3* [17,58,59].

We found that, in all composite regions, the most decayed elements, i.e., CIMEs, are always the most distant elements from the target genes. In the same way, the less decayed elements are located at the 3′ end of the target gene (*attR* end) in tandems integrated in the 3′ end of tRNALys genes from *S. agalactiae* [56] and in the 3′ end of the tmRNA genes of *Escherichia coli* and *Salmonella enterica* [60]. This structure probably results from the integration of an incoming element in the *attR* site of a decayed resident element that retains its *attL* and *attR* sites. We also found tandems of two functional unrelated elements (ICE and IME, or two IMEs). This structure probably results from the integration of an incoming element in the *attR* site flanking an unrelated functional resident element. Another scenario involving the acquisition of the whole composite element cannot be excluded. It was actually observed for composite structures including CIMEs and ICEs [2,61,62] or two unrelated ICEs [63]. Comparison of module compositions also suggests that recombination between related or unrelated ICEs, IMEs and CIMEs, likely integrated in tandem, plays a major role in the evolution and plasticity of ICEs and IMEs [2].

This work is one of the very few studies that identifies precisely the boundaries of a large number of various classes and families of integrated elements able to transfer by conjugation. All the integration specificities found in ICEs and IMEs from *S. salivarius* except one were previously described in other *Streptococci* [12–16] and their possible impact on the host fitness were previously discussed [3]. Globally, it seems that these integration specificities have evolved to reduce their impact on the host fitness to allow integration into a large array of strains and species and often to allow their mobilization *in cis* or *in trans* by other mobile elements ([47] for this last point). It appears that most elements encoding tyrosine recombinases integrate in the 3′ end of essential housekeeping genes and some in their 5′ end but without changing the sequence of the functional product (either tRNA or protein). By contrast, ICEs encoding serine integrases target internal sites of conserved genes encoding dispensable proteins. It was hypothesized that stimuli that induce the expression of the target gene also induce the excision of the integrated element and that the excised element controls its provisional maintenance by replication as an extrachromosomal element. Some ICEs target intergenic regions: (i) Tn*916*-related ICEs integrate preferentially in AT-rich short sequences found mainly in intergenic regions and/or in other mobile genetic elements and (ii) Tn*GBS2*-related elements encoding DDE transposases integrate 15 or 16 bp upstream from the −35 box of promoters recognized by sigma A [45], probably without modifying the expression of the downstream gene. IMEs (n = 11) encoding a serine recombinase related to the one of IME_*Sol3089_ND* described previously [16] are also integrated in various intergenic regions. Therefore, as for Tn*GBS2*-related elements, these elements seem to target transcription signals that belong either to promoters or terminators and are present in intergenic regions.

Therefore, the main impact on host fitness probably results from the expression of the cargo genes carried by the element. Some of these cargo genes may have dual properties being: (i) in certain circumstances, advantageous for the strain and therefore for the element or (ii) in other circumstances, advantageous for the element and consequently disadvantageous for the strain. For instance, five elements encode putative abortive infection systems (Abi) that would cause the death of cells infected by bacteriophages to prevent phage propagation. Therefore, Abi systems are advantageous for bacterial strains. However, recent studies have shown that various Abi systems are also toxin-antitoxins (TA) systems [64–66]. These systems kill the cells that lost the MGE that encodes

them. They are also involved in the competition between incompatible conjugative elements and in their maintenance in bacterial population.

Furthermore, 28 elements encode RM systems that are advantageous for the cell since they confer resistance to bacteriophage. Most of them are type II RM systems, carried by ICE*St3* elements and located in a conserved position at the left of the regulation module. Since various type II RM systems are TA systems involved in plasmid maintenance [67,68], this suggests that type II RM are not only involved in bacteriophage resistance [69], but also in the selfish maintenance of ICE*St3*-related elements. Besides their advantageous function in strain competition, as for type II RM and TA systems, the production of bacteriocin that is encoded by four elements can also be viewed as an addiction system: harmless for cells harboring MGEs and harmful for cells that lost the elements.

Many elements from *S. salivarius* carry other cargo genes that have no role in the maintenance of the element in the cell but may increase bacterial fitness. Hence, 45 of them encode resistance to antimicrobial compounds. These include all ICEs, IMEs and MGIs that encode resistance to antibiotics. Resistance to tetracycline and erythromycin are carried mainly by the well-known Tn*916*-related elements. An MGI deriving from an ICE related to Tn*1549* (MGI_*SsaF4-20_rumA*) encodes a chloramphenicol acetyltransferase that we demonstrated to confer resistance to chloramphenicol [33]. Furthermore, two IMEs integrated in *oriT* of conjugative elements [47] encode a *lsa*(C) gene that is known to confer cross-resistance to lincosamides, streptogramin A and pleuromutilins in *S. agalactiae* [54]. In addition to these resistance genes, several ICEs and IMEs encode multidrug efflux transporter (MFS) that might also be involved in resistance to antibiotics. These data indicate that conjugative and mobilizable elements from *S. salivarius* are likely to be major determinants for the spreading of antibiotic resistance genes.

Moreover, 11 IMEs from *S. salivarius* encode *qacE* genes that are known to confer efflux-mediated resistance to QACs. QACs are disinfectants used in hospitals and food-processing environments to ensure microbiological safety [70]. QAC resistance genes are generally carried on plasmids and/or integrons. Here, we demonstrate that QAC genes are also carried by IMEs and are therefore likely to spread by mobilization. Several *S. salivarius* elements, mainly CIMEs, carry resistance to cadmium, a highly poisonous metal air pollutant [71]. Several elements from *S. salivarius* also encode proteins required for growth at 4 °C or involved in redox reactions or in the metabolism of amino acids, lipids or carbohydrates. For instance, Delorme et al. [31] described that *S. salivarius* K12 encodes all enzymes devoted to lactose utilization by the tagatose-6P pathway. Our analysis indicates that these enzymes are encoded by an MGI in this strain and that other elements (one dICE and two CIMEs) deriving from ICE*St3* also encode this metabolism. Hence, acquisition of such fitness mobile elements can be viewed as an important adaptive mechanism enabling survival of *S. salivarius* in a changing environment.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/11/9/999/s1, Table S1: Date and origin of *Streptococcus salivarius* sampling and genbank accession number. Table S2: Coordinates of *Streptococcus salivarius* integrative elements transferable by conjugation. Table S3: Diversity and abundance of *Streptococcus salivarius* integrative elements transferable by conjugation. Figure S1: Schematic representation of ICEs, IMEs, MGIs and CIMEs. Figure S2: Function of the cargo genes carried by the conjugative elements integrated in the genomes of *S. salivarius*. File S1: nucleotidic sequences of elements.

Author Contributions: Conceptualization, G.G., S.P. and N.L.-B.; methodology, J.L., T.L., H.C., G.G., S.P. and N.L.-B.; validation, G.G., S.P., T.L., J.L. and N.L.-B.; sequence analysis and genome submission, S.P., V.L. (Valentin Loux), T.L., J.L., H.C. and N.L.-B.; investigation, J.L., T.L., G.G., F.C.-B. and V.L. (Virginie Libante); writing—original draft preparation, G.G. and N.L.-B.; review and editing, all authors; visualization, S.P. and N.L.-B.; supervision, N.L.-B.; project administration, N.L.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Guglielmini, J.; Quintais, L.; Garcillán-Barcia, M.P.; de la Cruz, F.; Rocha, E.P.C. The Repertoire of ICE in Prokaryotes Underscores the Unity, Diversity, and Ubiquity of Conjugation. *PLoS Genet.* **2011**, *7*, e1002222. [CrossRef] [PubMed]

2. Bellanger, X.; Payot, S.; Leblond-Bourget, N.; Guédon, G. Conjugative and mobilizable genomic islands in bacteria: Evolution and diversity. *FEMS Microbiol. Rev.* **2014**, *38*, 720–760. [CrossRef]
3. Guédon, G.; Libante, V.; Coluzzi, C.; Payot, S.; Leblond-Bourget, N. The Obscure World of Integrative and Mobilizable Elements, Highly Widespread Elements that Pirate Bacterial Conjugative Systems. *Genes* **2017**, *8*, 337. [CrossRef] [PubMed]
4. Delavat, F.; Miyazaki, R.; Carraro, N.; Pradervand, N.; van der Meer, J.R. The hidden life of integrative and conjugative elements. *FEMS Microbiol. Rev.* **2017**, *41*, 512–537. [CrossRef]
5. Toussaint, A.; Merlin, C. Mobile Elements as a Combination of Functional Modules. *Plasmid* **2002**, *47*, 26–35. [CrossRef]
6. Burrus, V.; Pavlovic, G.; Decaris, B.; Guédon, G. Conjugative transposons: The tip of the iceberg. *Mol. Microbiol.* **2002**, *46*, 601–610. [CrossRef]
7. Wozniak, R.A.F.; Waldor, M.K. Integrative and conjugative elements: Mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat. Rev. Microbiol.* **2010**, *8*, 552–563. [CrossRef]
8. Goessweiner-Mohr, N.; Arends, K.; Keller, W.; Grohmann, E. Conjugative type IV secretion systems in Gram-positive bacteria. *Plasmid* **2013**, *70*, 289–302. [CrossRef]
9. Guglielmini, J.; Néron, B.; Abby, S.S.; Garcillán-Barcia, M.P.; de la Cruz, F.; Rocha, E.P.C. Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.* **2014**, *42*, 5715–5727. [CrossRef]
10. Johnson, C.M.; Grossman, A.D. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annu. Rev. Genet.* **2015**, *49*, 577–601. [CrossRef] [PubMed]
11. Garcillán-Barcia, M.P.; Redondo-Salvo, S.; Vielva, L.; de la Cruz, F. MOBscan: Automated Annotation of MOB Relaxases. In *Horizontal Gene Transfer: Methods and Protocols*; De la Cruz, F., Ed.; Methods in Molecular Biology; Springer US: New York, NY, USA, 2020; pp. 295–308.
12. Ambroset, C.; Coluzzi, C.; Guédon, G.; Devignes, M.-D.; Loux, V.; Lacroix, T.; Payot, S.; Leblond-Bourget, N. New Insights into the Classification and Integration Specificity of Streptococcus Integrative Conjugative Elements through Extensive Genome Exploration. *Front. Microbiol.* **2016**, *6*. [CrossRef] [PubMed]
13. Garcillán-Barcia, M.P.; Cuartas-Lanza, R.; Cuevas, A.; Cruz, F. de la Cis-Acting Relaxases Guarantee Independent Mobilization of MOBQ4 Plasmids. *Front. Microbiol.* **2019**, *10*. [CrossRef] [PubMed]
14. 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.; et al. Discovery of a new family of relaxases in *Firmicutes* bacteria. *PLoS Genet.* **2017**, *13*. [CrossRef] [PubMed]
15. Soler, N.; Robert, E.; Chauvot de Beauchêne, I.; Monteiro, P.; Libante, V.; Maigret, B.; Staub, J.; Ritchie, D.W.; Guédon, G.; Payot, S.; et al. Characterization of a relaxase belonging to the MOBT family, a widespread family in *Firmicutes* mediating the transfer of ICEs. *Mob. DNA* **2019**, *10*, 18. [CrossRef] [PubMed]
16. Coluzzi, C.; Guédon, G.; Devignes, M.-D.; Ambroset, C.; Loux, V.; Lacroix, T.; Payot, S.; Leblond-Bourget, N. A Glimpse into the World of Integrative and Mobilizable Elements in Streptococci Reveals an Unexpected Diversity and Novel Families of Mobilization Proteins. *Front. Microbiol.* **2017**, *8*, 443. [CrossRef]
17. Bellanger, X.; Morel, C.; Gonot, F.; Puymege, A.; Decaris, B.; Guédon, G. Site-specific accretion of an integrative conjugative element together with a related genomic island leads to cis mobilization and gene capture. *Mol. Microbiol.* **2011**, *81*, 912–925. [CrossRef]
18. Libante, V.; Nombre, Y.; Coluzzi, C.; Staub, J.; Guédon, G.; Gottschalk, M.; Teatero, S.; Fittipaldi, N.; Leblond-Bourget, N.; Payot, S. Chromosomal Conjugative and Mobilizable Elements in *Streptococcus suis*: Major Actors in the Spreading of Antimicrobial Resistance and Bacteriocin Synthesis Genes. *Pathogens* **2019**, *9*, 22. [CrossRef] [PubMed]
19. Juhas, M.; van der Meer, J.R.; Gaillard, M.; Harding, R.M.; Hood, D.W.; Crook, D.W. Genomic islands: Tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol. Rev.* **2009**, *33*, 376–393. [CrossRef]
20. Hegstad, K.; Mylvaganam, H.; Janice, J.; Josefsen, E.; Sivertsen, A.; Skaare, D. Role of Horizontal Gene Transfer in the Development of Multidrug Resistance in *Haemophilus influenzae*. *mSphere* **2020**, *5*, e00969-19. [CrossRef]
21. Mullany, P.; Allan, E.; Roberts, A.P. Mobile genetic elements in Clostridium difficile and their role in genome function. *Res. Microbiol.* **2015**, *166*, 361–367. [CrossRef]

22. Fernandez-Gutierrez, M.M.; Roosjen, P.P.J.; Ultee, E.; Agelink, M.; Vervoort, J.J.M.; Keijser, B.; Wells, J.M.; Kleerebezem, M. *Streptococcus salivarius* MS-oral-D6 promotes gingival re-epithelialization in vitro through a secreted serine protease. *Sci. Rep.* **2017**, *7*, 11100. [CrossRef] [PubMed]
23. Couvigny, B.; de Wouters, T.; Kaci, G.; Jacouton, E.; Delorme, C.; Doré, J.; Renault, P.; Blottière, H.M.; Guédon, E.; Lapaque, N. Commensal *Streptococcus salivarius* Modulates PPARγ Transcriptional Activity in Human Intestinal Epithelial Cells. *PLoS ONE* **2015**, *10*, e0125371. [CrossRef] [PubMed]
24. Kaci, G.; Goudercourt, D.; Dennin, V.; Pot, B.; Doré, J.; Ehrlich, S.D.; Renault, P.; Blottière, H.M.; Daniel, C.; Delorme, C. Anti-Inflammatory Properties of *Streptococcus salivarius*, a Commensal Bacterium of the Oral Cavity and Digestive Tract. *Appl. Environ. Microbiol.* **2014**, *80*, 928–934. [CrossRef] [PubMed]
25. Wilson, M.; Martin, R.; Walk, S.T.; Young, C.; Grossman, S.; McKean, E.L.; Aronoff, D.M. Clinical and laboratory features of *Streptococcus salivarius* meningitis: A case report and literature review. *Clin. Med. Res.* **2012**, *10*, 15–25. [CrossRef]
26. Shirokawa, T.; Nakajima, J.; Hirose, K.; Suzuki, H.; Nagaoka, S.; Suzuki, M. Spontaneous meningitis due to *Streptococcus salivarius* subsp. salivarius: Cross-reaction in an assay with a rapid diagnostic kit that detected Streptococcus pneumoniae antigens. *Intern. Med. Tokyo Jpn.* **2014**, *53*, 279–282. [CrossRef]
27. Kitten, T.; Munro, C.L.; Zollar, N.Q.; Lee, S.P.; Patel, R.D. Oral streptococcal bacteremia in hospitalized patients: Taxonomic identification and clinical characterization. *J. Clin. Microbiol.* **2012**, *50*, 1039–1042. [CrossRef]
28. Corredoira, J.C.; Alonso, M.P.; García, J.F.; Casariego, E.; Coira, A.; Rodriguez, A.; Pita, J.; Louzao, C.; Pombo, B.; López, M.J.; et al. Clinical characteristics and significance of *Streptococcus salivarius bacteremia* and *Streptococcus bovis bacteremia*: A prospective 16-year study. *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.* **2005**, *24*, 250–255. [CrossRef]
29. Han, X.Y.; Kamana, M.; Rolston, K.V.I. *Viridans streptococci* isolated by culture from blood of cancer patients: Clinical and microbiologic analysis of 50 cases. *J. Clin. Microbiol.* **2006**, *44*, 160–165. [CrossRef]
30. Delorme, C.; Poyart, C.; Ehrlich, S.D.; Renault, P. Extent of Horizontal Gene Transfer in Evolution of Streptococci of the *Salivarius* Group. *J. Bacteriol.* **2007**, *189*, 1330–1341. [CrossRef]
31. Delorme, C.; Abraham, A.-L.; Renault, P.; Guédon, E. Genomics of *Streptococcus salivarius*, a major human commensal. *Infect. Genet. Evol.* **2015**, *33*, 381–392. [CrossRef]
32. Dahmane, N.; Libante, V.; Charron-Bourgoin, F.; Guédon, E.; Guédon, G.; Leblond-Bourget, N.; Payot, S. Diversity of Integrative and Conjugative Elements of *Streptococcus salivarius* and Their Intra- and Interspecies Transfer. *Appl. Environ. Microbiol.* **2017**, *83*. [CrossRef]
33. Chaffanel, F.; Charron-Bourgoin, F.; Libante, V.; Leblond-Bourget, N.; Payot, S. Resistance Genes and Genetic Elements Associated with Antibiotic Resistance in Clinical and Commensal Isolates of *Streptococcus salivarius*. *Appl. Environ. Microbiol.* **2015**, *81*, 4155–4163. [CrossRef] [PubMed]
34. Guedon, E.; Delorme, C.; Pons, N.; Cruaud, C.; Loux, V.; Couloux, A.; Gautier, C.; Sanchez, N.; Layec, S.; Galleron, N.; et al. Complete Genome Sequence of the Commensal *Streptococcus salivarius* Strain JIM8777. *J. Bacteriol.* **2011**, *193*, 5024–5025. [CrossRef] [PubMed]
35. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [CrossRef] [PubMed]
36. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: An information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645. [CrossRef] [PubMed]
37. Feldgarden, M.; Brover, V.; Haft, D.H.; Prasad, A.B.; Slotta, D.J.; Tolstoy, I.; Tyson, G.H.; Zhao, S.; Hsu, C.-H.; McDermott, P.F.; et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob. Agents Chemother.* **2019**, *63*. [CrossRef]
38. Hammami, R.; Zouhir, A.; Le Lay, C.; Ben Hamida, J.; Fliss, I. BACTIBASE second release: A database and tool platform for bacteriocin characterization. *BMC Microbiol.* **2010**, *10*, 22. [CrossRef]
39. Liu, B.; Zheng, D.; Jin, Q.; Chen, L.; Yang, J. VFDB 2019: A comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* **2019**, *47*, D687–D692. [CrossRef]
40. Roberts, R.J.; Vincze, T.; Posfai, J.; Macelis, D. REBASE—A database for DNA restriction and modification: Enzymes, genes and genomes. *Nucleic Acids Res.* **2015**, *43*, D298–D299. [CrossRef]
41. Flissi, A.; Ricart, E.; Campart, C.; Chevalier, M.; Dufresne, Y.; Michalik, J.; Jacques, P.; Flahaut, C.; Lisacek, F.; Leclère, V.; et al. Norine: Update of the nonribosomal peptide resource. *Nucleic Acids Res.* **2020**, *48*, D465–D469. [CrossRef]

42. Rawlings, N.D.; Barrett, A.J.; Thomas, P.D.; Huang, X.; Bateman, A.; Finn, R.D. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* **2018**, *46*, D624–D632. [CrossRef]
43. Roberts, A.P.; Mullany, P. Tn916-like genetic elements: A diverse group of modular mobile elements conferring antibiotic resistance. *FEMS Microbiol. Rev.* **2011**, *35*, 856–871. [CrossRef] [PubMed]
44. Brochet, M.; Couvé, E.; Glaser, P.; Guédon, G.; Payot, S. Integrative Conjugative Elements and Related Elements Are Major Contributors to the Genome Diversity of *Streptococcus agalactiae*. *J. Bacteriol.* **2008**, *190*, 6913–6917. [CrossRef] [PubMed]
45. Guérillot, R.; Siguier, P.; Gourbeyre, E.; Chandler, M.; Glaser, P. The Diversity of Prokaryotic DDE Transposases of the Mutator Superfamily, Insertion Specificity, and Association with Conjugation Machineries. *Genome Biol. Evol.* **2014**, *6*, 260–272. [CrossRef] [PubMed]
46. Carraro, N.; Libante, V.; Morel, C.; Decaris, B.; Charron-Bourgoin, F.; Leblond, P.; Guédon, G. Differential regulation of two closely related integrative and conjugative elements from *Streptococcus thermophilus*. *BMC Microbiol.* **2011**, *11*, 238. [CrossRef] [PubMed]
47. Libante, V.; Sarica, N.; Mohamad Ali, A.; Gapp, C.; Oussalah, A.; Guédon, G.; Leblond-Bourget, N.; Payot, S. Mobilization of IMEs integrated in the oriT of ICEs involves their own relaxase belonging to the Rep-trans family of proteins. *Genes* **2020**, in press.
48. Rawls, K.S.; Yacovone, S.K.; Maupin-Furlow, J.A. GlpR Represses Fructose and Glucose Metabolic Enzymes at the Level of Transcription in the Haloarchaeon *Haloferax volcanii*. *J. Bacteriol.* **2010**, *192*, 6251–6260. [CrossRef]
49. Chen, Y.-Y.M.; Feng, C.W.; Chiu, C.F.; Burne, R.A. cadDX Operon of *Streptococcus salivarius* 57.I. *Appl. Environ. Microbiol.* **2008**, *74*, 1642–1645. [CrossRef]
50. Stock, A.M.; Robinson, V.L.; Goudreau, P.N. Two-Component Signal Transduction. *Annu. Rev. Biochem.* **2000**, *35*, 183–215. [CrossRef]
51. Martin, B.; Granadel, C.; Campo, N.; Hénard, V.; Prudhomme, M.; Claverys, J.-P. Expression and maintenance of ComD–ComE, the two-component signal-transduction system that controls competence of *Streptococcus pneumoniae*. *Mol. Microbiol.* **2010**, *75*, 1513–1528. [CrossRef]
52. Zhulin, I.B.; Taylor, B.L.; Dixon, R. PAS domain S-boxes in Archaea, Bacteria and sensors for oxygen and redox. *Trends Biochem. Sci.* **1997**, *22*, 331–333. [CrossRef]
53. Ponting, C.P.; Aravind, L. PAS: A multifunctional domain family comes to light. *Curr. Biol. CB* **1997**, *7*, R674–R677. [CrossRef]
54. Malbruny, B.; Werno, A.M.; Murdoch, D.R.; Leclercq, R.; Cattoir, V. Cross-Resistance to Lincosamides, Streptogramins A, and Pleuromutilins Due to the *lsa* (C) Gene in *Streptococcus agalactiae* UCN70. *Antimicrob. Agents Chemother.* **2011**, *55*, 1470–1474. [CrossRef] [PubMed]
55. Zheng, W.; Kathariou, S. Transposon-induced mutants of *Listeria monocytogenes* incapable of growth at low temperature (4Â°C). *FEMS Microbiol. Lett.* **1994**, *121*, 287–291. [CrossRef]
56. Puymège, A.; Bertin, S.; Guédon, G.; Payot, S. Analysis of *Streptococcus agalactiae* pan-genome for prevalence, diversity and functionality of integrative and conjugative or mobilizable elements integrated in the tRNALys CTT gene. *Mol. Genet. Genomics* **2015**, *290*, 1727–1740. [CrossRef]
57. Van Eijk, E.; Anvar, S.; Browne, H.P.; Leung, W.; Frank, J.; Schmitz, A.M.; Roberts, A.P.; Smits, W. Complete genome sequence of the Clostridium difficile laboratory strain 630Δerm reveals differences from strain 630, including translocation of the mobile element CTn5. *BMC Genom.* **2015**, *16*, 31. [CrossRef]
58. Hochhut, B.; Beaber, J.W.; Woodgate, R.; Waldor, M.K. Formation of Chromosomal Tandem Arrays of the SXT Element and R391, Two Conjugative Chromosomally Integrating Elements That Share an Attachment Site. *J. Bacteriol.* **2001**, *183*, 1124–1132. [CrossRef]
59. Possoz, C.; Ribard, C.; Gagnat, J.; Pernodet, J.-L.; Guérineau, M. The integrative element pSAM2 from Streptomyces: Kinetics and mode of conjugal transfer. *Mol. Microbiol.* **2001**, *42*, 159–166. [CrossRef]
60. Song, L.; Jiang, Y.; Zhang, X. Chronology and pattern of integration of tandem genomic islands associated with the tmRNA gene in *Escherichia coli* and *Salmonella enterica*. *Chin. Sci. Bull.* **2011**, *56*, 3836–3843. [CrossRef]
61. Chapleau, M.; Guertin, J.F.; Farrokhi, A.; Lerat, S.; Burrus, V.; Beaulieu, C. Identification of genetic and environmental factors stimulating excision from *Streptomyces scabiei* chromosome of the toxicogenic region responsible for pathogenicity. *Mol. Plant Pathol.* **2016**, *17*, 501–509. [CrossRef]

62. Wang, P.; Zeng, Z.; Wang, W.; Wen, Z.; Li, J.; Wang, X. Dissemination and loss of a biofilm-related genomic island in marine Pseudoalteromonas mediated by integrative and conjugative elements. *Environ. Microbiol.* **2017**, *19*, 4620–4637. [CrossRef] [PubMed]
63. León-Sampedro, R.; Fernández-de-Bobadilla, M.D.; San Millán, Á.; Baquero, F.; Coque, T.M. Transfer dynamics of Tn6648, a composite integrative conjugative element generated by tandem accretion of Tn5801 and Tn6647 in *Enterococcus faecalis*. *J. Antimicrob. Chemother.* **2019**, *74*, 2517–2523. [CrossRef] [PubMed]
64. Goeders, N.; Chai, R.; Chen, B.; Day, A.; Salmond, G.P.C. Structure, Evolution, and Functions of Bacterial Type III Toxin-Antitoxin Systems. *Toxins* **2016**, *8*, 282. [CrossRef] [PubMed]
65. Fraikin, N.; Goormaghtigh, F.; Van Melderen, L. Type II Toxin-Antitoxin Systems: Evolution and Revolutions. *J. Bacteriol.* **2020**, *202*, e00763-19. [CrossRef] [PubMed]
66. Harms, A.; Brodersen, D.E.; Mitarai, N.; Gerdes, K. Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology. *Mol. Cell* **2018**, *70*, 768–784. [CrossRef] [PubMed]
67. Mruk, I.; Kobayashi, I. To be or not to be: Regulation of restriction–modification systems and other toxin–antitoxin systems. *Nucleic Acids Res.* **2014**, *42*, 70–86. [CrossRef]
68. Kobayashi, I. Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.* **2001**, *29*, 3742–3756. [CrossRef]
69. Burrus, V.; Bontemps, C.; Decaris, B.; Guédon, G. Characterization of a Novel Type II Restriction-Modification System, Sth368I, Encoded by the Integrative Element ICESt1 of Streptococcus thermophilusCNRZ368. *Appl. Environ. Microbiol.* **2001**, *67*, 1522–1528. [CrossRef]
70. Chaidez, C.; Lopez, J.; Castro-del Campo, N. Quaternary ammonium compounds: An alternative disinfection method for fresh produce wash water. *J. Water Health* **2007**, *5*, 329–333. [CrossRef]
71. Järup, L.; Åkesson, A. Current status of cadmium as an environmental health problem. *Toxicol. Appl. Pharmacol.* **2009**, *238*, 201–208. [CrossRef]

Des données supplémentaires de l'article sont disponibles sous forme d'archive ZIP à ce lien : http://www.mdpi.com/2073-4425/11/9/999/s1. Cette archive contient :

- Des informations sur les échantillons des 75 génomes de *S. salivarius* étudiés (source, date, niveau d'assemblage, numéro d'accession Genbank, …) (*Table S1*).
- Les coordonnées génomiques des éléments intégrés se transférant par conjugaison des 75 génomes de *S. salivarius* de l'étude (*Table S2*).
- La diversité et l'abondance des éléments intégrés se transférant par conjugaison de l'étude (*Table S3*).
- Une représentation schématique des différents types d'éléments intégrés se transférant par conjugaison identifiés dans l'étude (*Figure S1*).
- Les fonctions des gènes cargo portés par les d'éléments intégrés se transférant par conjugaison de l'étude (*Figure S2*).
- La séquence nucléotidique des éléments intégrés se transférant par conjugaison identifiés dans l'étude sous format FASTA (*File S1*).

1.3 Discussion

Au total, 75 génomes de *S. salivarius* ont été analysés et un ensemble de 69 ICE, 165 IME ont été répertoriés. Une partie non négligeable des ICE (27,5 %) et des IME (16,4 %) sont retrouvés dans différentes structures composites d'éléments. Ainsi, 37 structures différentes constituées uniquement d'éléments en accrétion ont été caractérisées dont près de la totalité (34 sur 37) sont des structures composites dites « mixtes », c'est-à-dire constituées d'éléments de différents types, avec cinq d'entre-elles étant constitués à la fois d'ICE et d'IME. En plus des accrétions, dix structures complexes d'éléments en emboîtement ont aussi été détectées.

Ces premiers résultats nous ont fait prendre conscience de la nécessité de développer un module d'ICEscreen spécifiquement dédié à la résolution des éléments composites pour pouvoir annoter correctement les ICE et les IME d'un génome.

Une procédure d'annotation des gènes d'adaptation portés par ces éléments a été mise au point et automatisée par Thomas Lacroix (ingénieur du laboratoire MaIAGE). Il ne s'agit pas ici de déterminer la fonction de tous les gènes d'adaptation mais d'annoter les gènes codant des fonctions (31 catégories de fonctions) fréquemment portés par des ICE. L'analyse fonctionnelle de ces gènes d'adaptation des ICE et IME de *S. salivarius* montre que leur nature varie d'un élément à l'autre et qu'ils codent une très grande variété de fonctions distinctes. Une partie non négligeable de ces gènes codent des résistances à différents stress (stress thermique et redox), des résistances à divers antibiotiques (tétracycline, érythromycine, chloramphenicol, lincosamides) ou à des composés antibactériens tels des désinfectants. Ces gènes codent également de nombreuses enzymes impliquées dans diverses voies métaboliques cellulaires telles que le métabolisme des acides aminés, des lipides ou des souches. Ainsi, ces données confirment notre hypothèse que non seulement les ICE mais aussi les IME jouent un rôle important dans l'adaptation de *S. salivarius* à son environnement. L'absence de métadonnées consolidées concernant le caractère commensale ou pathogène des souches ne nous a pas permis d'établir de corrélations fiables entre ce phénotype et la présence de gènes d'adaptation particuliers portés par des ICE et des IME.

2. Mise au point de ICEscreen

Nous avons conçu et développé une méthodologie permettant la détection et l'identification d'ICE et d'IME dans les génomes de Firmicutes. Ces ICE et IME sont identifiés par la recherche et la co-localisation de quatre types de protéines signatures appartenant à des modules indispensables pour le fonctionnement de ces éléments. Notre procédure de co-localisation permet de repérer les éléments intégrés dans les chromosomes bactériens, y compris ceux intégrés de manière complexe, tels que des éléments emboîtés ou en accrétion (Figure 4).

2.1 Principe général

L'approche ICEscreen identifie en trois grandes étapes, les ICE, les IME et les éléments composites (plusieurs ICE ou IME emboités et/ou en accrétion) dans un génome de Firmicutes donné en entrée (voir Figure 5) :

1) Détection des protéines signatures

   Les quatre protéines signatures (SP pour « signature protein ») des modules de transfert ou d'intégration d'ICE ou d'IME sont recherchées le long du génome : relaxase, protéine de Couplage, VirB4 et intégrase (étape A de la figure 5). À la fin de cette étape, la liste des SP est segmentée en régions contenant des SP distantes de moins de 100 CDS.

2) Détection des éléments potentiels

   Les éléments sont détectés par co-localisation des SP dans les segments définis lors de l'étape 1 en trois sous-étapes :

   a) Recherche de tous les modules de transfert complets ou partiels (appelés ici *ancres*) par co-localisation des SP par une stratégie dite de « création et extension d'ancres » (étape B de la figure 5) ;

   b) Résolution de régions d'éléments composites par fusion récursive de modules de transfert potentiels partiels détectés à l'étape (2a) (étape C de la figure 5) ;

c) Recherche et affectation du module d'intégration aux modules de transfert potentiels détectés aux étapes (2a) et (2b) (étape D de la figure 5).

3) Caractérisation des éléments mobiles

Caractérisation des éléments détectés de l'étape 2 grâce à son contenu en SP selon le tableau indiqué à l'étape E de la figure 5.

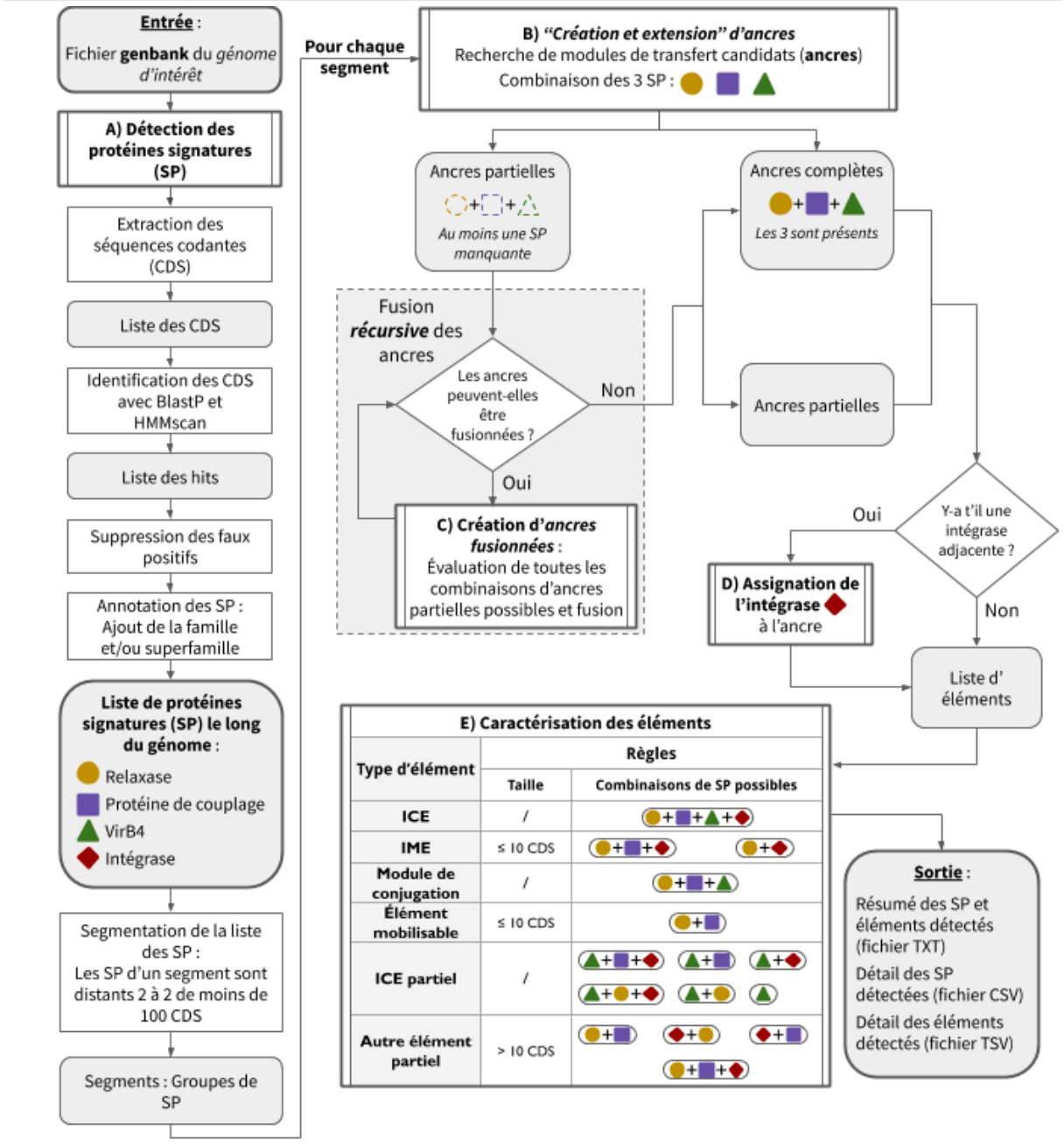

Figure 5 : Description de l'approche ICEscreen de recherche des éléments mobiles dans un génome de Firmicutes. Le génome d'intérêt doit être au format Genbank. A : Les quatre types de protéines signatures (relaxases, protéines de couplages, VirB4 et intégrases) sont recherchées (étape 1). Pour cela, les séquences codantes du génome sont extraites à partir de l'annotation Genbank. Puis les SP potentielles sont identifiées par BlastP et HMMscan. La liste de SP obtenue est ensuite découpée en segments afin de limiter les associations de SP superflues. B : L'algorithme recherche ensuite des modules de transfert potentiels par "création et extension" d'ancres (étape 2). Une ancre est créée lorsqu'une relaxase, une protéine de couplage ou une VirB4 est détectée. Puis l'ancre est étendue si les types de SP suivantes ne sont pas des intégrases et n'ont pas déjà été associés à l'ancre. C : Lorsque toutes les ancres potentielles ont été créées, l'algorithme évalue les combinaisons d'ancres partielles possibles afin de résoudre les cas d'éléments scindés en plusieurs morceaux (cas d'emboîtements d'éléments). D : L'intégrase est ensuite recherchée de part et d'autre des ancres. E : Finalement, les éléments sont caractérisés en fonction de leur taille en CDS (gènes codants) et contenu en SP (étape 3).

2.2 Détection des protéines signatures

La première étape de la méthode ICEscreen est la recherche de protéines signatures (SP) d'ICE et d'IME à partir des séquences codantes (CDS) du génome d'intérêt. Quatre types de SP faisant partie des modules de transfert et des modules de recombinaison de ces éléments sont recherchées. Pour le module de transfert, les relaxases, protéines qui initient le transfert de l'élément et des protéines du système de sécrétion de type IV sont recherchées (les protéines de couplages et les ATPases VirB4). Pour le module de recombinaison, seule, l'intégrase est recherchée.

Afin d'identifier ces SP, nous avons construit une banque de séquences protéiques de référence ainsi qu'une banque de profils HMM de référence. La recherche de hits avec BlastP (Altschul *et al.*, 1997) contre notre banque de référence permet d'identifier des SP proches de ceux des ICE et IME de streptocoques. Quant à la banque de profils HMM, elle est utilisée avec HMMscan de la suite HMMER3 (Eddy, 2011) pour détecter des homologues moins conservés de SP d'ICE et d'IME de Firmicutes.

Les alignements obtenus sont ensuite filtrés en deux temps afin de supprimer les faux positifs (voir l'article 2.2.3 sur la filtration des alignements) : (1) validation des SP grâce à des filtres ajustés et (2) suppression de faux positifs connus (expertisé par Gérard Guédon) grâce à des séquences protéiques et profils HMM dédiés.

La recherche des protéines signatures est complétée par l'annotation, si possible, de leur famille et superfamille. Pour cela, l'annotation de la protéine ou du profil HMM de référence utilisée pour l'identification de la SP est transférée.

2.2.1 Banque de séquences protéiques

Le laboratoire DynAMic a constitué une banque de séquences de protéines signatures à partir d'ICE et d'IME de génomes de Firmicutes, majoritairement des streptocoques. Cette banque avait déjà été utilisée dans plusieurs recherches extensives des ICE et IME de génomes de streptocoques (Ambroset *et al.*, 2016; Coluzzi *et al.*, 2017) via la méthodologie ICE/IME Finder et enrichie à partir des données obtenues au cours de ma thèse.

Pour l'intégration de cette banque à l'outil ICEscreen, nous l'avons standardisée par l'ajout systématique d'informations sur la famille et la superfamille du module de conjugaison dans le cas des SP d'ICE. Les familles et superfamilles de modules de conjugaison d'ICE sont décrites dans un article antérieur (Ambroset *et al.*, 2016) et représentées dans la figure 6. Pour les SP d'IME, leur annotation a été enrichie avec l'indication de la superfamille de la relaxase de l'élément selon (Coluzzi *et al.*, 2017). Ces superfamilles sont présentées dans la figure 7. Actuellement, cette banque contient **1029** protéines signatures, dont **318** relaxases, **232** protéines de couplages, **140** VirB4 et **339** intégrases (**240** intégrases à tyrosine, **77** intégrases à sérine et **22** transposases à DDE).

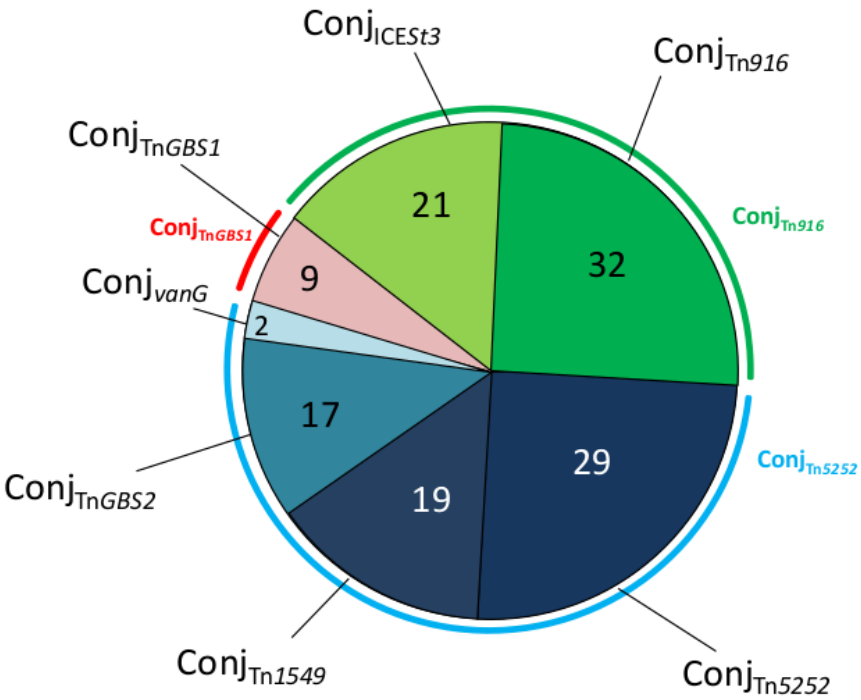

Figure 6 : Superfamilles et familles de modules de conjugaison d'ICE de streptocoques décrits dans (Ambroset *et al.*, 2016). Celles-ci sont basées sur leur contenu en relaxase, protéine de couplage et VirB4. Les valeurs correspondent aux nombres d'ICE caractérisés de ces familles dans un jeu de 124 génomes de streptocoques. Trois superfamilles ont été caractérisées, la superfamille Tn*916* composée de deux familles (ICE*St3* et Tn*916*), la superfamille Tn*5252* composée de quatre familles (Tn*5252*, Tn*1549*, Tn*GBS2* et *VanG*) et la superfamille Tn*GBS1* qui comporte une seule famille du même nom.

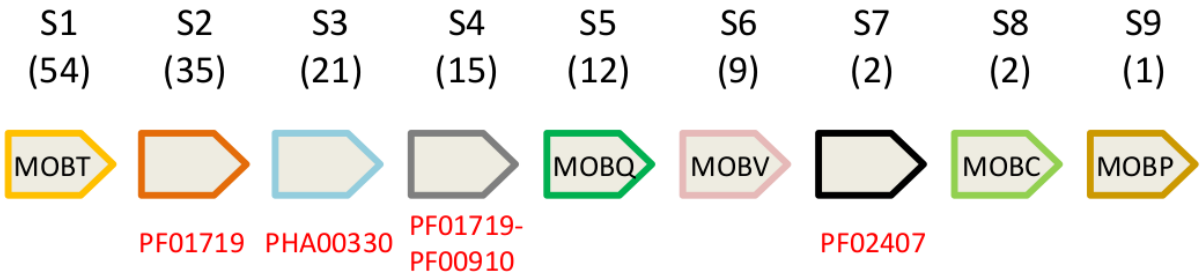

Figure 7 : Superfamilles de relaxases d'IME de streptocoques décrites dans (Coluzzi, 2017). Celles-ci sont basées sur leur contenu en domaines fonctionnels. Les valeurs entre parenthèses correspondent au nombre de relaxases caractérisées de ces superfamilles dans un jeu de 124 génomes de streptocoques.

2.2.2 Banque de profils HMM

Pour la recherche par HMMscan, nous avons construit une banque de profils HMM d'intérêt décrits dans le tableau 1. Pour cela nous avons récupéré directement plusieurs profils déjà disponibles :

- de la base de données généraliste de profils HMM de domaines protéiques PFAM (El-Gebali *et al.*, 2019),
- du module TXSScan (Abby *et al.*, 2016) qui regroupe des profils HMM de systèmes de sécrétions,
- de la base de données MOBfamDB de la ressource spécialisée MOBscan (Garcillán-Barcia *et al.*, 2020). Il s'agit d'une base de données spécialisée regroupant des profils HMM de relaxases.

Nous avons aussi créé des nouveaux profils pour détecter les relaxases MOB_L ainsi que les nouvelles relaxases d'IME de streptocoques décrites dans l'étude de (Coluzzi *et al.*, 2017). Il s'agit des relaxases PF01719-*like*, relaxases PHA00330-*like* et relaxases PF02407-*like*. Actuellement, cette banque contient **22** profils HMM, dont **15** de relaxases, **3** de protéines de couplages, **1** de VirB4 et **3** d'intégrases.

L'analyse des relaxases d'IME de streptocoques a permis l'identification de neuf superfamilles de relaxases sur la base de leur contenu en domaines (Table 2, Coluzzi *et al.*, 2017). Des profils HMM existaient déjà pour des domaines spécifiques de cinq des neuf superfamilles. Les quatre superfamilles restantes sont caractérisées respectivement par les domaines PHA00330, PF02407, PF01719 et la présence conjointe de PF01719 et PF00910 (nommé PF01719+PF00910). Nous avons donc créé des nouveaux profils HMM pour les domaines PHA00330, PF02407 et PF01719 de ces quatre superfamilles d'IME (voir Tableau 1, les domaines avec Source=ICEscreen).

Tableau 1 : Banque de profils HMM intégrés dans l'outil ICEscreen.

Type ou famille de protéine signature	Source	Domaines	Nom du profil
Intégrases			
Intégrase à tyrosine	PFAM	Phage_integrase (PF00589)	Tyr_Phage_integrase
Intégrase à sérine	PFAM	Recombinase (PF07508)	Ser_Recombinase
Transposase à DDE	PFAM	UPF0236 (PF06782)	DDE_UPF0236
Protéines de couplage			
TcpA	TXSScan	FtsK_SpoIIIE (PF01580)	tcpA
VirD4	TXSScan	T4SS-DNA_transf (PF02534)	t4cp1
		TraG-D_C (PF12696)	t4cp2
VirB4			
VirB4	TXSScan		T4SS_virb4
Relaxases			
MOB_L	**ICEscreen**		**firmi_MOBL**
PF01719-*like*	**ICEscreen**	**Rep_2 (PF01719)**	**firmi_Rep_2**
PHA00330-*like*	**ICEscreen**	**PHA00330**	**PHA_IME_A1** **PHA_IME_B**
PF02407-*like*	**ICEscreen**	**Viral_Rep (PF02407)**	**firmi_Viral_Rep_A** **firmi_Viral_Rep_B1** **firmi_Viral_Rep_B2**
MOB_T	TXSScan	Rep_trans (PF02486)	T4SS_MOBT
	MOBscan	Rep_trans (PF02486) HTH_3 (PF01381)	profile_MOBT
MOB_P	TXSScan	Relaxase (PF03432)	T4SS_MOBP1 T4SS_MOBP2 T4SS_MOBP3
MOB_Q	TXSScan	MobA_MobL (PF03389)	T4SS_MOBQ
MOB_V	TXSScan	Mob_Pre (PF01076)	T4SS_MOBV
MOB_C	TXSScan	Replic_Relax (PF13814)	T4SS_MOBC

Pour la création de ces profils HMM nous avons utilisé la méthodologie décrite ci-dessous :

1) Création du jeu de séquences de relaxases : une ou plusieurs séquences de références ont été sélectionnées dans la base de données de ICE/IME Finder ainsi que dans les banques publiques lorsque cela a été possible. L'architecture en domaines fonctionnels de ces relaxases a ensuite été identifiée en utilisant CD-Search (Marchler-Bauer and Bryant, 2004) et SPARCLE (Marchler-Bauer *et al.*, 2017). Cette architecture a ensuite été utilisée pour sélectionner toutes les relaxases de Firmicutes dans NCBI possédant une architecture proche en domaine ;

2) Réduction de la redondance du jeu de séquence : la redondance du jeu de séquence a été réduite par un clustering à 95 % d'identité sur 100 % de la longueur avec CD-HIT (Li and Godzik, 2006; Huang *et al.*, 2010; Fu *et al.*, 2012). Lorsque le nombre de séquences restant était encore trop élevé, un second clustering à 40 % d'identité sur 100 % de la longueur a été effectué ;

3) Construction de l'alignement multiple : l'alignement a été construit itérativement à partir des relaxases représentatives des clusters. Pour cela, un alignement est construit avec MAFFT v7.407 avec l'algorithme FFTNS1 (Katoh *et al.*, 2002) et est ensuite curé de façon à retirer les relaxases trop divergentes en prenant comme guide un arbre phylogénétique construit par Neighbor-Joining avec la méthode BioNJ (gaps non compris ; distance de Poisson ; bootstrap de 100 itérations) avec le logiciel SeaView 4.2 (Gouy *et al.*, 2010). L'étape de curage a été répétée jusqu'à obtenir un alignement correct. Lorsqu'un ensemble de relaxases étaient trop distants, l'alignement a été séparé en sous-alignements pour créer par la suite de multiples profils HMM. Le domaine fonctionnel des relaxases se trouve en partie N-terminale de ces protéines (Garcillán-Barcia *et al.*, 2009), ainsi l'alignement obtenu avec MAFFT a été réaligné avec Clustal Omega v1.2.4 avec les paramètres par défaut (Sievers *et al.*, 2011) afin de minimiser les gaps dans la partie N-terminal de l'alignement. Les extrémités des alignements ont ensuite été élaguées avec BMGE (Criscuolo and Gribaldo, 2010) avec une matrice de substitution BLOSUM 30. L'utilisation de la matrice BLOSUM 30 permet de supprimer les extrémités de l'alignement qui possèdent peu d'information sur le

domaine d'intérêt, sans pour autant élaguer une partie du domaine d'intérêt ce qui aurait été provoqué par une matrice de substitution moins profonde (ex : BLOSUM 64).

4) Construction du profil HMM : les profils HMM ont ensuite été construits avec HMMbuild de la suite HMMER 3.2.1 (Eddy, 2011) avec les paramètres par défaut.

2.2.3 Filtration des alignements

2.2.3.1 Validation de protéines signatures

Cette étape permet de supprimer un maximum de protéines non apparentées aux SP recherchées ainsi que des protéines apparentées mais ne possédant pas les fonctions recherchées (faux positifs).

Les paramètres utilisés pour filtrer les hits obtenus par BlastP sont adaptés de ceux de la méthodologie ICE/IME Finder décrits dans (Coluzzi, 2017) et décrits dans le tableau 2. Quatre filtres sont utilisés :

- Un *pourcentage d'identité minimal* et un *seuil maximal de E-value* afin de supprimer les résultats non significatifs et/ou les protéines non homologues.

- Un *taux de couverture minimal* permettant de valider que l'alignement comprend le domaine fonctionnel recherché.

- Et enfin, une *longueur minimale et maximale* est fixée pour les différents types de protéines signatures possibles, ce qui permet d'éliminer des « protéines » erronées provenant de la traduction de pseudogènes ainsi que des protéines qui présentent des domaines additionnels.

Tableau 2 : Filtres et paramètres utilisés pour valider les alignements de protéines signatures obtenus avec BlastP de l'outil ICEscreen. (*) : paramètres utilisés pour les SP de modules de conjugaison de famille Tn*GBS1*.

	E-value	Taux de couverture	Pourcentage d'identité	Longueur du CDS
Protéine de couplage	≤ 10^{-5}	≥ 40 %	≥ 25 %	≥ 180 aa ; ≤ 700 aa ≥ 1000 aa ; ≤ 1200 aa*
Relaxase	≤ 10^{-4}	≥ 40 %	≥ 25 %	≥ 180 aa
Intégrase	≤ 10^{-4}	≥ 40 %	≥ 25 %	≥ 320 aa
VirB4	≤ 10^{-5}	≥ 40 %	≥ 25 %	≥ 500 aa

Les alignements de protéines signatures obtenus avec HMMscan correspondent à des domaines homologues à ceux d'un profil HMM. Pour filtrer les faux positifs parmi ces résultats de domaines, quatre filtres sont aussi utilisés. Les paramètres de ces filtres ont été ajustés à partir des résultats obtenus sur les premières versions de ICEscreen sur le jeu de données FirmiData (décrit dans la section 3.3). Les choix de paramètres sont les suivants :

- Longueur du CDS : le seuil minimal de longueur des CDS est le même que celui utilisé pour BlastP (voir tableau 2). La longueur maximale des CDS n'est pas seuillée afin de permettre la détection de protéines signatures plus distantes.

- i-Evalue (independant E-value) : une i-Evalue maximale de 10^{-5} est acceptée pour tous les hits obtenus avec HMMscan. Un CDS donné peut avoir de multiples résultats pour un même domaine, la i-Evalue permet d'évaluer la significativité de chaque alignement indépendamment des autres.

- Pourcentage de couverture du domaine du profil HMM : une couverture minimale du domaine du profil HMM est requise afin d'obtenir des alignements de domaines homologues. Ce pourcentage a été ajusté à 40 % pour tous les profils HMM de protéines signatures excepté pour le profil des relaxases PF01719-*like* qui est ajusté à 80 %.

- Pourcentage de couverture du domaine du CDS : une couverture minimale du domaine du CDS a été fixée à 40 % pour les profil des protéines de couplage TcpA uniquement,

aucun seuil n'a été utilisé pour les autres profils. Ce choix de paramètre a été effectué afin de ne pas limiter la longueur maximale des protéines signatures.

2.2.3.2 Suppression de faux positifs grâce à des séquences protéiques et profils HMM dédiés

Certains faux positifs apparentés aux protéines signatures recherchées ne peuvent être supprimés lors de la première étape de filtration des résultats. Il s'agit de protéines apparentées dont au moins certains des domaines fonctionnels sont très similaires mais pas identiques à ceux des SP recherchées.

Ainsi, par exemple des protéines transporteurs à cassettes liant l'ATP (ATP Binding Cassette) possédant les domaines fonctionnels COG1126 ou COG4586 (domaines de la base de données NCBI COG (Tatusov *et al.*, 1997; Galperin *et al.*, 2015)) peuvent être détectées en tant que protéines de couplage. Certaines protéines de fonction inconnue DUF853 sont aussi détectées en tant que protéines de couplage VirD4 ou en tant que protéines VirB4.

Afin de supprimer ces faux positifs connus, une deuxième étape de filtration a été mise en place. Pour la détection des protéines transporteurs à cassettes liant l'ATP deux profils HMM ont été construits pour les domaines COG1126 et COG4586 respectivement. Les alignements de ces profils sont validés si la i-Evalue $\leq 10^{-5}$, si la couverture avec le CDS ≥ 80 % et si la couverture avec le profil est ≥ 40 %. Pour les protéines DUF853, le profil HMM du PFAM PF05872 est utilisé. Les hits de ce profil sont validés si la i-Evalue $\leq 10^{-5}$ et si la couverture avec le profil est ≥ 40 %.

De la même façon, les recombinases à tyrosine site-spécifiques XerS ont nécessité la mise en place d'une stratégie de filtration spécifique. Chez les streptocoques, elles sont impliquées dans la résolution en monomère des dimères du chromosome, leur implication en tant qu'intégrase d'ICE ou d'IME n'a jamais été décrite dans la littérature. Cependant, leur fonction potentielle d'intégrase ne peut pas être exclue chez les autres Firmicutes. Les séquences protéiques des protéines XerS ACO17137 et WP_011835230 sont utilisées pour discriminer les XerS de streptocoques de celles des autres Firmicutes. Les alignements ayant plus de 70 % d'identité avec ACO17137 sont classés comme des XerS de streptocoques et donc supprimés de la liste des protéines signatures détectées. Si un CDS possède un résultat avec

WP_011825230, il est classé comme XerS de Firmicutes et est donc annoté comme intégrase à tyrosine.

2.3 Algorithme de détection des ICE et IME par co-localisation des protéines signatures

Pour rappel, afin de réduire le nombre d'opérations dans les étapes suivantes, la liste des SP détectées est segmentée de façon à ce que deux SP ne soient pas distantes de plus de 100 CDS dans un segment donné.

Les différentes étapes de notre algorithme d'annotation des éléments de chaque segment sont décrites dans la figure 8.

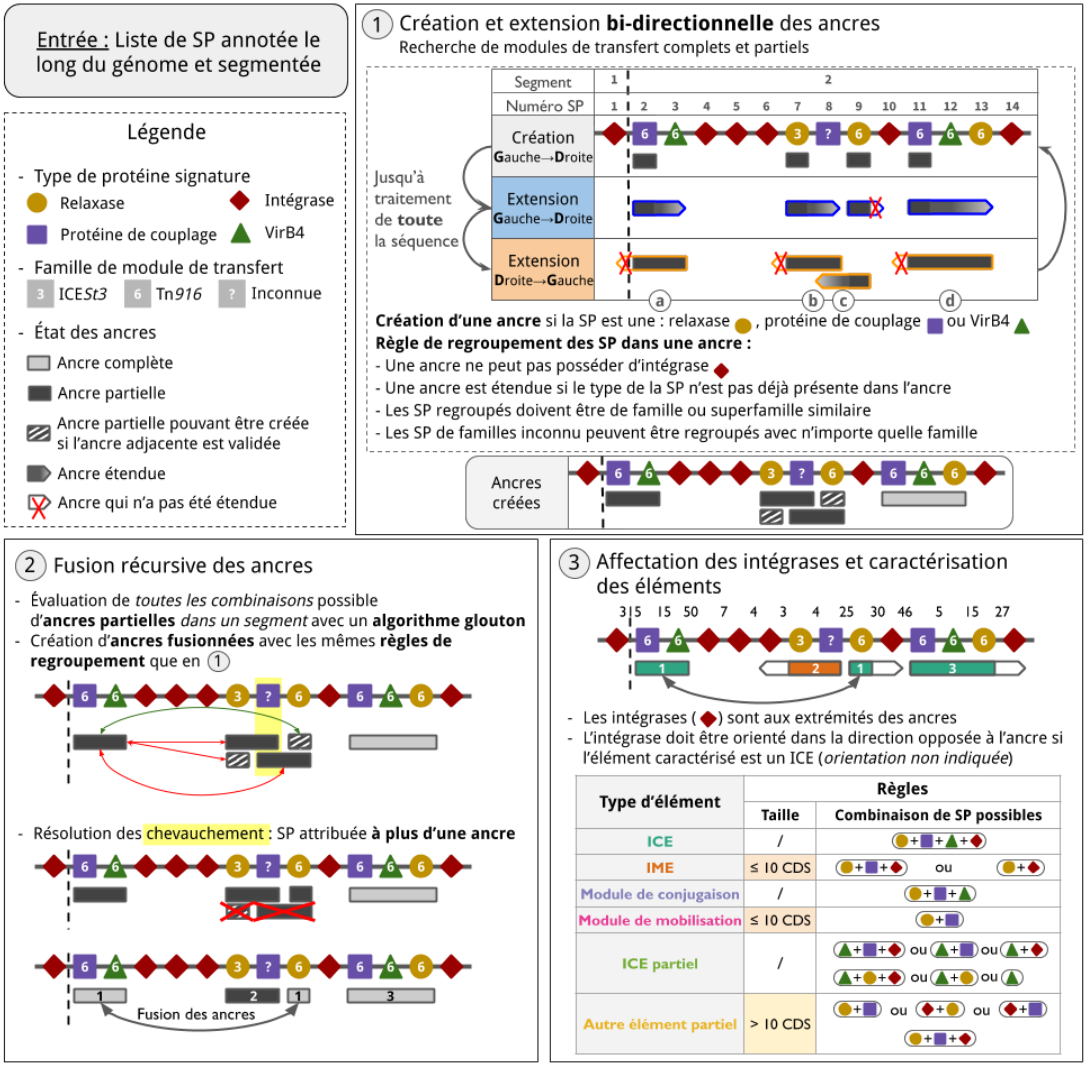

Figure 8 : Description de l'algorithme de détection des ICE et IME par co-localisation des protéines signatures (SP) de l'approche ICEscreen. Une liste de SP ordonnées suivant la position des protéines le long du génome et segmentée de sorte que deux SP d'un même segment ne peuvent être séparées de plus de 100 séquences codantes est prise en entrée. Puis dans chaque segment, trois étapes sont effectuées : 1) Recherche des modules de transfert potentiels complets et partiels par création d'ancres, correspondant aux modules potentiels, le long de la liste de SP. a) Une ancre est créée avec la SP n°2 qui est étendue vers la droite avec une VirB4 de même famille Tn*916* mais ne peut pas être étendue une 2ème fois car la SP n°4 est une intégrase. L'extension vers la gauche est impossible car la SP n'est pas dans le même segment que l'ancre. b) Une ancre est créée avec la relaxase n°7, celle-ci est étendue vers la droite avec une protéine de couplage (CP) de famille inconnue, donc compatible avec la famille de la relaxase. Une 2nde extension n'est pas possible car une relaxase est déjà dans l'ancre. La SP à gauche de l'ancre est une intégrase, l'extension vers la gauche n'est donc pas possible. c) La 3ème ancre est créée à partir de la relaxase n°9 de manière semblable à l'ancre (b), elle est étendue du côté gauche une fois avec la CP n°8 mais non une 2ème fois à cause d'une relaxase. Elle n'est pas étendue du côté droit à cause d'une intégrase. d) L'ancre créée avec la SP n°11, est étendue deux fois vers la droite car les deux SP successives sont de types différents et de familles identiques. 2) Les modules de transfert interrompus par d'autres éléments sont recherchés par fusion récursive des ancres partielles. Les ancres partielles (b) et (c) sont en chevauchement, ainsi la validation d'une de ces ancres entraînera la création d'une autre ancre partielle par "soustraction" de la SP en commun. La validation de l'ancre (b) permet de reconstituer une ancre complète par fusion de l'ancre (a) avec l'ancre partielle créée par soustraction de l'ancre partielle (b) et (c). 3) L'intégrase des modules de transfert détectés lors des étapes précédentes est recherchée puis l'élément est caractérisé grâce à son contenu en SP.

2.3.1 Création et extension des ancres

Dans chaque segment, des ***ancres*** sont recherchées par co-localisation de SP spécifiques. Nous appelons ***ancre***, un module de transfert complet ou partiel. Ces ancres sont recherchées par "création et extension" le long de la séquence de protéines signatures. Une première ancre est créée si la SP considérée est une relaxase, une protéine de couplage ou une protéine VirB4. L'algorithme procède ensuite à une étape d'***extension de l'ancre*** : la SP suivante est ajoutée à l'ancre en cours de création selon les règles décrites ci-dessous. L'extension de l'ancre est **bi-directionnelle**, ainsi une fois qu'une ancre a été complètement étendue *vers la droite*, elle est ensuite étendue *vers la gauche* en considérant les mêmes règles.

Pour résumer, voici les différentes règles concernant la création, l'extension et l'arrêt des ancres :

- Création de l'ancre : une ancre est créée uniquement si la SP considérée est une relaxase, une protéine de couplage ou une protéine VirB4. Une ancre ne peut pas contenir d'intégrase et ne peut pas appartenir à deux segments différents.

- Extension de l'ancre : Une fois créée, l'ancre est étendue vers la droite, puis vers la gauche à partir de la première SP de l'ancre. Les extensions suivent les règles ci-dessous :

    (i) Si le type de la SP suivante n'est pas déjà présente dans l'ancre alors l'ancre est étendue avec cette nouvelle SP ;

    (ii) Si le type de la SP suivante est déjà présente dans l'ancre et que c'est une relaxase : l'ancre peut être étendue si la dernière SP de l'ancre est aussi une relaxase et si les deux relaxases sont de superfamille ou famille semblables ou compatibles (dans le cas où une des superfamilles/familles est inconnue et l'autre est connue par exemple).

- Conditions pour arrêter l'extension d'une ancre :

  (i) la SP suivante est distante de plus de 100 CDS de la dernière SP de l'ancre (donc dans un segment différent) ;

  (ii) la SP suivante est une VirB4 alors qu'une est déjà présente dans l'ancre ;

  (iii) la SP suivante est une protéine de couplage alors qu'une est déjà présente dans l'ancre ;

  (iv) la SP suivante est une relaxase alors qu'une est déjà présente dans l'ancre, **sauf si les deux relaxases sont adjacentes sur le génome ou séparées par au plus un CDS** ;

  (v) la SP suivante est une intégrase (elle est traitée à un stade ultérieur) ;

  (vi) la SP suivante est annotée avec une superfamille ou famille d'éléments différente des SP précédentes (i.e. ICE*St3*, Tn*916*, etc.).

Les familles des SP sont documentées pour les SP détectées par BlastP et correspondent à des éléments connus chez les streptocoques (voir l'article 2.2.1). Les SP de familles inconnues car détectées par les profils HMM peuvent être ajoutées à une ancre, quelle que soit la famille des SP déjà incluses dans l'ancre.

Une fois qu'une ancre a été créée, l'algorithme essaie alors de l'étendre de droite à gauche (selon les mêmes conditions pour arrêter l'extension de gauche à droite). Les ICE et les IME ne sont pas orientés sur le génome, c'est pour cela que l'algorithme est indépendant du choix de la direction de balayage initial. Certaines SP peuvent donc être attribuées à deux ancres différentes à ce stade. Les étapes pour trouver les ancres des modules de transfert et les étendre séquentiellement et bidirectionnellement sont répétées jusqu'à ce que toutes les SP du génome en entrée soient traitées.

À la fin de cette étape, toutes les SP du génome (sauf les intégrases) sont affectées à une ou plusieurs ancres partielles ou complètes.

2.3.2 Fusion récursive des ancres

Cette étape permet de **fusionner des ancres partielles** (incluant une à deux des trois protéines signatures du module de transfert), pour identifier des éléments composites imbriqués (emboîtés). Ces ancres distantes partielles **sont interrompues par une ou plusieurs ancres complètes ou partielles, mais appartiennent forcément à un même segment**.

Les éléments composites imbriqués sont recherchés grâce à un algorithme glouton. Ainsi, toutes les combinaisons entre les ancres partielles séparées par au moins une ancre au sein d'un même segment sont recherchées par récursivité et la meilleure solution est choisie pour chaque niveau d'imbrication jusqu'à résolution complète de la structure d'emboîtements. Les règles utilisées pour la fusion des ancres partielles sont les suivantes :

- les conditions de fusion des ancres partielles sont identiques aux conditions d'extension d'une ancre décrites dans l'article 2.3.1,
- après évaluation exhaustive de toutes les combinaisons d'ancres partielles et dans le cas où plusieurs fusions sont possibles, la fusion incluant les ancres partielles les plus proches entre elles sur le génome est choisie,
- la récursivité de l'algorithme de fusion des ancres s'arrête lorsqu'il n'est plus possible de fusionner aucune ancre d'un segment,
- les ancres complètes isolées ne sont pas traitées ici.

La fusion d'ancres partielles éloignées sert à identifier les éléments imbriqués et peut parfois aider à résoudre des cas de SP précédemment attribuées à deux ancres différentes (chevauchement d'ancres).

2.3.3 Affectation des intégrases et caractérisation des éléments

La dernière étape pour déduire les structures des ICE/IME est d'affecter les intégrases aux ancres (modules de transfert) détectées aux deux étapes précédentes. Les gènes d'intégrases sont toujours localisés à une extrémité de l'élément mobile et peuvent être de part et d'autre

du module de transfert. Les **règles pour associer une intégrase** à un module de transfert sont les suivantes :

1) seules sont considérées les intégrases à l'intérieur du segment où les ancres se trouvent ;

2) les intégrases ne sont pas associées à une famille particulière et peuvent être rattachée à n'importe quelle famille de module de transfert ;

3) un duo ou un trio d'intégrases adjacentes de même sens sur le génome ou séparées par un CDS de même sens peuvent être attribuées à une même ancre lorsqu'elles sont du même type ;

4) une intégrase ne peut être affectée à une ancre incluant une SP de type VirB4 (donc un élément ICE) que si le gène qui la code est orienté vers l'extérieur de la structure. En d'autre termes, si le gène de l'intégrase est en aval du module de conjugaison de l'ICE, elle doit être sur le brin + ; si l'intégrase est en amont du module de conjugaison de l'ICE, elle doit être sur le brin -.

Le **principe de l'algorithme** est ensuite le suivant ;

1) Dans une première étape, l'algorithme va attribuer aux ancres (modules de transfert) les intégrases qui sont conformes aux règles ci-dessus et qui sont considérées comme non ambiguës du point de vue du module de transfert.

2) De façon itérative, l'algorithme va ensuite attribuer les intégrases encore non attribuées en « masquant » les intégrases déjà attribuées lors de la première étape.

3) Cette méthodologie est appliquée de manière itérative (**attribution des intégrases en cascade**). Si après cette phase plusieurs choix d'intégrases sont encore possibles pour un même module de transfert et que l'algorithme ne peut pas lever l'ambiguïté sur l'attribution, alors les intégrases ambigües sont gardées et notées comme étant à vérifier manuellement pour ce module de transfert.

4) Enfin, s'il reste des intégrases non attribuées et des modules de transfert sans intégrase, l'algorithme va considérer l'attribution d'une intégrase non adjacente à un module de transfert dans la séquence de SP. C'est en particulier le cas si un autre élément est inséré entre l'intégrase et le module de transfert d'un même élément. Afin de limiter les attributions de faux positifs lors de cette phase, un maximum de deux ancres entre l'intégrase et le module de transfert est imposé par l'algorithme.

Les structures détectées sont ensuite caractérisées et classées en différentes catégories en fonction du contenu de leur module de transfert et de l'attribution ou non d'une intégrase. Les différentes catégories sont :

- « **ICE** » : $R^2 + C^3 + V^4 + I^5$;
- « **IME** » : R + I ou R + C + I avec distance ≤ 10 CDS ;
- « **module de conjugaison** » : R + C + V ;
- « **élément mobilisable** » : R + C avec distance ≤ 10 CDS ;
- « **ICE partiel** » : toute structure qui contient au moins une VirB4 et qui n'est pas un ICE complet,
- « **autre élément partiel** » : toute structure qui contient au moins deux SP et qui n'entre pas dans les catégories précédentes.

Des SP isolées sont aussi indiquées dans la sortie de ICEscreen comme trace potentielle d'un élément encore inconnu et sont classées dans la catégorie « à vérifier manuellement ». Ces SP isolés peuvent être des relaxases, des protéines de couplage ou des VirB4.

[2] Relaxase
[3] Protéine de couplage
[4] VirB4
[5] Intégrase

2.4 Implémentation de la méthode

L'outil ICEscreen a été implémenté sous la forme d'un pipeline de différents programmes et scripts Python 3 et Bash. Le pipeline est géré avec le gestionnaire de pipeline Snakemake (Köster and Rahmann, 2012). Il prend en entrée les fichiers Genbank des génomes à analyser et génère les résultats sous forme résumée, sous forme détaillée par SP détectées et sous forme détaillée par éléments détectés. Deux programmes sont utilisés pour la détection des SP par homologie, BlastP (Altschul *et al.*, 1997) pour l'identification d'homologues proches connus et HMMscan de la suite HMMER3 (Eddy, 2011) pour la reconnaissance par domaines fonctionnels d'homologues moins conservés. La filtration et la validation des résultats obtenus est réalisée grâce à divers scripts codés en Python 3 et Bash (les filtres et critères de validation sont détaillés dans l'article 2.2.3). La caractérisation des groupes de SP en éléments mobiles potentiels est également implémentée en Python 3. Le code de ICEscreen est disponible sur le gitlab du département MathNum de l'INRAE et sera rendu public dès que l'outil sera finalisé.

3. Résultats de ICEscreen

Afin d'évaluer précisément les performances de notre méthode ICEscreen, nous avons établi une stratégie en deux étapes :

1) Création d'un jeu de données de génomes de Firmicutes de référence avec annotation expertisée manuelle des éléments conjugatifs : le jeu de données FirmiData.

2) Mise au point d'une stratégie pour comparer les résultats de ICEscreen aux deux outils existants CONJscan et ICEfinder sur le jeu de référence FirmiData. Nous avons choisi de ne pas évaluer la stratégie de génomique comparative mise au point par (Cury *et al.*, 2020) pour cette comparaison car nous voulons évaluer la détection et l'annotation des éléments sans la recherche de leurs bornes. De plus, comme cette méthode se base sur la comparaison d'au moins quatre génomes proches, nous n'aurions pas pu l'appliquer pour les espèces de FirmiData pour lesquelles nous avions moins de quatre représentants (*Roseburia hominis*, *Lachnoclostridium* sp. *YL32*, *Dehalobacterium formicoaceticum* et *Lachnoclostridium phocaeense*).

3.1 Rappel sur les outils CONJscan et ICEfinder

Les outils utilisés dans cette comparaison sont décrits dans l'introduction (sections 3.3 et 3.4). Pour rappel, l'outil CONJscan est un module du logiciel MacSyFinder. Son objectif est de détecter les systèmes de sécrétion de type IV (T4SS) grâce à des profils HMM des gènes des T4SS et de règles concernant la présence/absence et l'organisation spatiale de ces gènes. Ainsi cet outil permet de rechercher des modules de conjugaison d'ICE et peut aussi être adapté à la recherche de modules de mobilisation d'IME.

L'outil ICEfinder a pour but de détecter et de délimiter les ICE et les IME au nucléotide près dans des génomes bactériens. La méthode de détection est basée sur la co-localisation de protéines et séquences signatures du module d'intégration et du module de transfert. La détection des signatures est obtenue avec trois outils différents : BlastP, HMMER et oriTfinder. Pour le bornage, seuls les éléments ayant pour cible d'intégration un tRNA sont traités. Les cibles sont annotées avec le logiciel ARAGORN et la délimitation au nucléotide près de l'élément est effectuée avec Vmatch par la recherche des Direct Repeats (séquences répétées)

qui le bornent. La comparaison des méthodes de détection des protéines signatures des outils disponibles est indiquée dans le tableau 3 ci-dessous.

Tableau 3 : Comparaison entre les logiciels ICEscreen, CONJscan et ICEfinder des méthodes de détection utilisées pour la recherche des protéines signatures et séquences nucléotidiques permettant de caractériser un ICE ou IME.

			ICEscreen	CONJscan	ICEfinder
Module de recombinaison	Intégrases		BlastP & HMMER	-	HMMER
Module de transfert	Relaxases		BlastP & HMMER	HMMER	HMMER
	Protéines de couplages		BlastP & HMMER	HMMER	HMMER
	Gènes du MPF	VirB4	BlastP & HMMER	HMMER	HMMER
		Autres gènes	-	HMMER	HMMER
	oriT		-	-	oriTfinder
Délimitation	Cible d'insertion		-	-	ARAGORN (tRNA seulement)
	Direct Repeats		-	-	Vmatch

3.2 La stratégie de comparaison

Nous nous intéressons aux performances de l'outil ICEscreen pour la réalisation de deux tâches :

1) Détection et annotation des protéines signatures
2) Détection et caractérisation des ICE et des IME, complets ou dégénérés, y compris dans des structures composites.

Pour cela, nous comparons les résultats d'ICEscreen à ceux d'une annotation de référence expertisée manuellement (jeu de données FirmiData) afin d'évaluer les performances de l'outil. De la même façon, nous évaluons les résultats de CONJscan et ICEfinder sur ce jeu de données FirmiData, pour comparer les performances d'ICEscreen à celles de ces deux outils.

Ainsi, nous analysons les résultats produits par les trois outils par rapport à l'annotation manuelle de référence et déterminons successivement si :

1) les éléments de la référence sont tous détectés et si des éléments supplémentaires sont identifiés ;
2) si le type de l'élément (ICE ou IME) a été correctement identifié
3) si la structure de ces éléments a été correctement résolue (cas de structures composites de type emboîtement et/ou accrétion).

Comme les objectifs, et par conséquent les résultats, de ces trois logiciels ne sont pas identiques, nous avons été obligés d'ajuster les paramètres pour effectuer les comparaisons.

Chez les Firmicutes, CONJscan détecte des modules de conjugaison constitués de relaxases, CP, VirB4 et d'au moins deux autres gènes du MPF (Mating Pore Formation). Comme CONJscan ne recherche pas l'intégrase, nous évaluerons les résultats de cet outil uniquement sur les trois types de SP des modules de transfert (relaxase, CP et VirB4). Le paramétrage utilisé pour CONJscan est donc :

- recherche des systèmes FA et FATA (Guglielmini *et al.*, 2013), qui correspondent à des modules de conjugaison que l'on sait être présents dans la grande majorité des ICE déjà identifiés de Firmicutes. Les autres systèmes utilisés par CONJscan n'ont pas été recherchés car les analyses précédentes réalisées sur un grand nombre de génomes n'avaient détecté aucun module de conjugaison d'un autre type chez les Firmicutes (Guglielmini *et al.*, 2014) ;

- recherche des modules de mobilisation des IME avec le modèle MOB modifié pour que la protéine signature VirB4 soit interdite comme dans notre définition des IME.

Le logiciel ICEfinder a été utilisé avec les paramètres par défaut.

3.3 Création d'une annotation de référence : FirmiData

3.3.1 Choix des génomes de FirmiData

Afin d'évaluer les performances de la méthode ICEscreen, nous avons sélectionné 40 génomes de Firmicutes incluant :

- 26 génomes de streptocoques choisis pour leur grande diversité d'éléments composites (Coluzzi *et al.*, 2017; Lao *et al.*, 2020; Libante *et al.*, 2020) ;

- une sélection de 14 souches de Firmicutes. Elle inclut des souches pour lesquelles des ICE et IME potentiels présents dans les génomes complets et assemblés avaient été documentés dans la littérature, vérifiés et/ou complétés par une analyse par le laboratoire DynAMic. Elle inclut également des souches dont le contenu en éléments n'avait pas été décrit mais a été complètement analysé par le laboratoire DynAMic.

L'arbre phylogénétique des 40 souches de FirmiData basé sur la protéine gyrA est présenté dans la figure 9. Parmi les 40 génomes de FirmiData, plus de la moitié font partie du genre *Streptococcus* (26/40). Nous avons ensuite sélectionné trois autres génomes, plus ou moins proches des streptocoques et appartenant au même ordre, les *Lactobacillales* : *Lactococcus lactis* subsp. *lactis* IO-1 dans lequel un ICE avait été identifié (Shimizu-Kadota *et al.*, 2013), *Enterococcus faecalis* V583 dans lequel sept ICE ou modules de

conjugaison avaient été identifiés (Burrus *et al.*, 2002a) et *Lactobacillus paracasei* LOCK919 dans lequel un ICE et un IME avaient été identifiés (Coluzzi, 2017).

Nous avons également sélectionné trois génomes de l'ordre des Bacillales : *Listeria monocytogenes* sv. 4^{e} SLCC 2378 dont un seul ICE est documenté (Kuenne *et al.*, 2013), ainsi que *Staphylococcus epidermidis* ATCC 12228 et *Staphylococcus pseudintermedius* HKU10-03 dont les contenus en ICE et IME ne sont pas décrits dans des publications mais pour lesquels une étude préliminaire que nous avons réalisée indique la présence de modules de conjugaison.

Finalement, les huit génomes restants font partie de la classe Clostridia et sont plus distants des streptocoques. Nous disposons de l'annotation d'ICE et IME de trois génomes de *Clostridioides difficile* (Brouwer *et al.*, 2011) qui a été vérifiée et/ou complétée par le laboratoire DynAMic. Les cinq autres génomes appartenant à des groupes éloignés de *C. difficile* ont été sélectionnés pour leur richesse potentielle en ICE et IME (richesse établie par notre étude préliminaire) : *Lachnoclostridium phocaeense* Marseille-P3177, *Lachnoclostridium* sp. YL32, *Dehalobacterium formicoaceticum* DMC, *Roseburia hominis* A2-183 et *Faecalibacterium prausnitzii* A2-165.

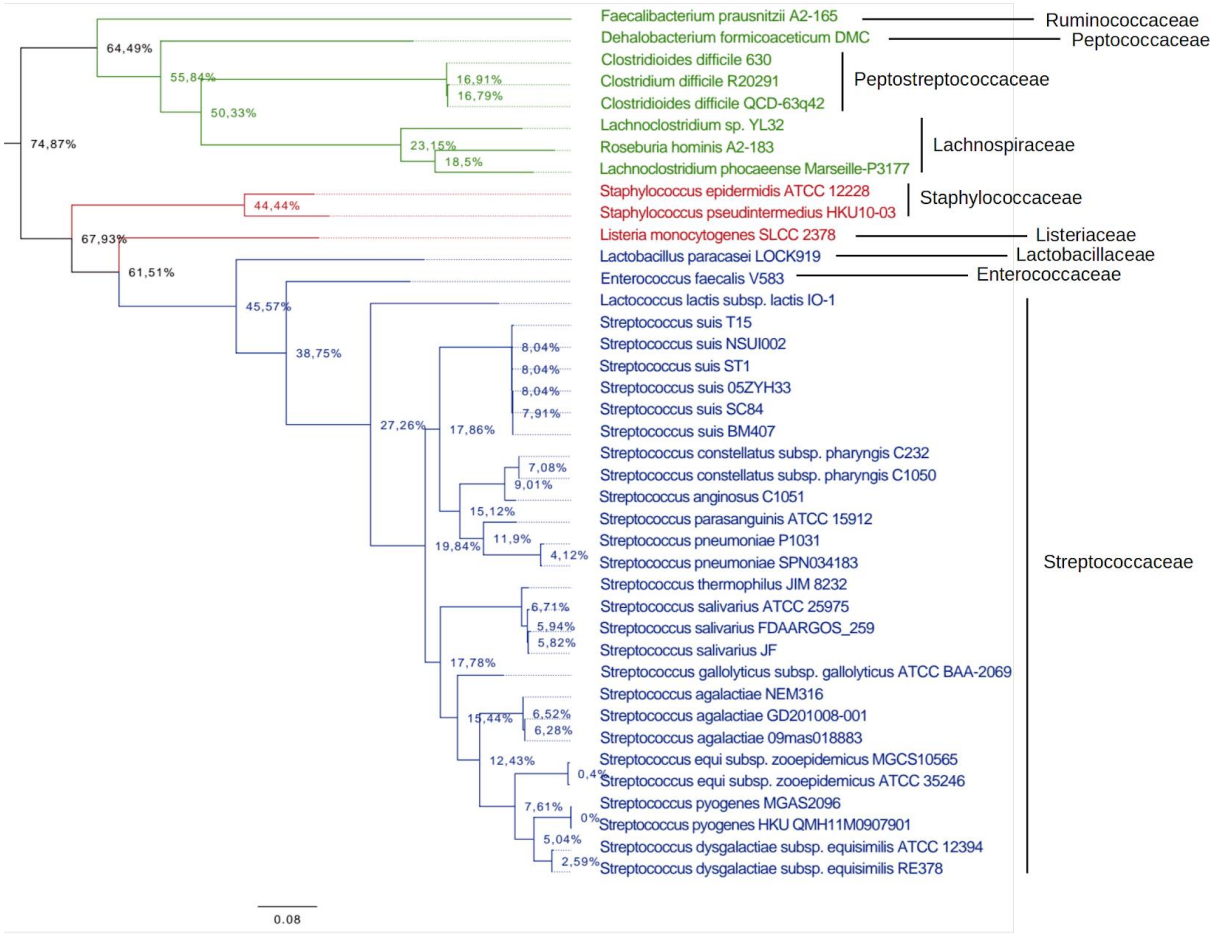

Figure 9 : Arbre phylogénétique des 40 souches de FirmiData. L'arbre est construit par la méthode du maximum de vraisemblance à partir de l'alignement de la protéine gyrA réalisé avec le logiciel Muscle (v3.8.1551). L'arbre a été construit en utilisant le logiciel FastTree (v2.1.10) avec les paramètres par défaut (modèle JTT+CAT, la robustesse de la topologie a été évaluée localement avec le test de Shimodaira-Hasegawa). L'arbre a été enraciné au barycentre. Les souches des trois ordres sont représentées avec différentes couleurs : bleu pour les Lactobacillales, rouge pour les Bacillales et vert pour les Clostridia. Le nom des familles a été ajouté sur la partie droite de l'arbre.

3.3.2 Annotation des éléments mobiles conjugatifs de FirmiData

L'annotation et la caractérisation des ICE et IME ont été réalisées en collaboration avec Gérard Guédon (unité DynAMic). La validité et/ou le contexte de chaque gène signature ont été expertisés manuellement, y compris ceux pour lesquels aucun élément n'était proposé. Parmi les 40 génomes, seuls deux génomes n'ont été que partiellement expertisés et ne seront pas détaillés ici (*Faecalibacterium prausnitzii* A2-165 et de *Lachnoclostridium* sp. *YL32*). Les indices quant à la présence d'ICE et d'IME ont été relevés par l'analyse de la littérature, par la recherche de modules de conjugaison potentiels avec CONJscan et en s'appuyant sur les connaissances expertes de Gérard Guédon sur les éléments connus chez les firmicutes. Les ICE et IME ont été annotés par une méthode semi-manuelle en se basant sur le contenu en gènes

potentiels d'ICE et IME. Selon les cas, la délimitation des éléments au nucléotide près se base sur la parenté des intégrases, la présence de gènes cibles potentiels, l'analyse de contexte génétique, la recherche de DR et/ou l'analyse du contexte biologique par synténie. Dans la suite du document nous distinguerons deux choses : l'**annotation selon la bibliographie** (l'annotation décrite dans la littérature sur le génome considéré) et **l'annotation de référence** (expertisée et validée manuellement par Gérard Guédon). Les sources utilisées pour l'annotation des génomes de FirmiData sont détaillées dans le tableau 1 en annexe. Les figures décrites dans la suite du document intègrent les deux types d'annotation.

3.3.2.1 Composition en protéines signatures

Au total, 551 protéines signatures ont été identifiées et annotées manuellement dans les 38 génomes de FirmiData complètement analysés (les génomes de *Faecalibacterium prausnitzii* A2-165 et de *Lachnoclostridium* sp. *YL32* sont encore en cours d'analyse). Certaines de ces protéines sont le produit de traduction de gènes interrompus par des introns de type II, des IME et/ou des ICE. Parmi les 551 protéines nous identifions :

- 183 intégrases dont 116 sont des intégrases à tyrosine (le gène correspondant inclut uniquement le feature "gene" et pas "CDS" annoté dans RefSeq (Haft *et al.*, 2018) et une autre qui est segmentée en deux) et 54 sont des intégrases à sérine (dont une est segmentée en deux). Les treize restantes sont des transposases à DDE.

- 172 relaxases dont deux segmentées en deux.

- 121 protéines de couplages dont six segmentées en deux, trois segmentées en trois et une segmentée en quatre.

- 75 VirB4 dont cinq segmentées en deux.

3.3.2.2 Composition en éléments mobiles conjugatifs

Les 38 génomes de FirmiData annotés manuellement par Gérard Guédon contiennent 89 ICE, 109 IME et 11 éléments partiels et/ou très dégénérés soit un total de 209 éléments.

Nous avons choisi d'inclure dans les ICE, les dICE, c'est-à-dire des dérivés d'ICE présentant au maximum deux défauts (tels qu'une protéine signature codée par un pseudogène ou l'absence d'un site *att*) selon la définition déjà adoptée (Ambroset *et al.*, 2016). De manière similaire, nous avons choisi d'inclure dans les IME, les dIME, c'est-à-dire des dérivés d'IME présentant au maximum un défaut.

Ainsi, 43/89 des ICE, 50/109 des IME et 9/11 des éléments partiels ou très dégénérés (appelés aussi éléments *remnant*) sont retrouvés dans des structures complexes d'éléments composites. Les éléments composites sont des ICE, des IME ou des éléments remnants en emboîtements et/ou en accrétions. Ainsi, 48,8 % des éléments de FirmiData sont retrouvés dans des structures complexes.

Le détail des structures des éléments mobiles et leurs nombres sont décrits dans le tableau 4 ci-dessous.

Tableau 4 : Liste des différents types d'éléments et de leurs structures annotées manuellement dans le jeu de génomes FirmiData. (*) L'ordre des éléments emboîtés est donné en partant du gène cible.

Type de structure	Description	Nombre
	Eléments isolés	**107**
IME	IME ne faisant pas partie d'éléments composites	59
ICE	ICE ne faisant pas partie d'éléments composites	46
remnant	Élément remnant ne faisant pas parti d'éléments composites	2
	Eléments composites constitués uniquement d'ICE	**11**
ICE(ICE)	Deux ICE emboîtés	7
ICE-ICE	Deux ICE en accrétion	1
ICE-remnant	Un ICE en accrétion avec un élément remnant*	1
ICE-ICE-ICE	Trois ICE en accrétion	1
ICE(ICE(remnant,remnant))	ICE hôte de deux éléments remnants et est inséré dans un ICE	1
	Eléments composites constitués uniquement d'IME	**6**
IME-IME	Deux IME en accrétion	3
IME(remnant)	Un élément remnant inséré dans un IME	1
remnant(IME,IME)	Élément remnant hôte de deux IME	1
remnant-remnant-IME	Deux éléments remnants et un IME en accrétion	1
	Eléments composites constitués d'ICE et d'IME	**20**
ICE(IME)	IME inséré dans un ICE	5
ICE-IME	Un ICE et un IME en accrétion*	2
ICE-IME-IME	ICE en accrétion avec deux IME*	2
IME-ICE-IME	IME en accrétion avec un ICE et un IME*	1
ICE-ICE(IME)	ICE en accrétion avec un ICE hôte d'un IME	1
ICE(IME,IME)	ICE hôte de deux IME	3
IME(IME,ICE)	IME hôte d'un IME et d'un ICE	1
ICE(IME,IME,IME)	ICE hôte de trois IME	2

Tableau 4 (Suite)

Type de structure	Description	Nombre
IME-ICE(IME)	IME en accrétion avec un ICE*, l'ICE étant hôte d'un IME	1
IME remnant (ICE(IME,IME,IME(IME)))	Un IME remnant, dans lequel est intégré un ICE qui est lui-même hôte de 3 IME. Un de ces IME héberge lui-même un IME.	1
ICE(IME(IME),IME,IME,IME)-remnant	Un ICE, dans lequel est intégré quatre IME. Un des IME héberge lui-même un IME. L'ICE est aussi en accrétion avec un élément remnant	1

3.4 Résultats d'ICEscreen, CONJscan et ICEfinder sur FirmiData

3.4.1 Détection et annotation des protéines signatures

Nous avons comparé la détection et l'annotation des quatre types de protéines signatures par les trois outils. La figure 9 montre de grandes disparités de détection selon les outils et les catégories de protéines signatures considérées :

- L'outil CONJscan détecte globalement très bien les protéines VirB4 (94 % de la référence), mais moins bien les protéines de couplage (72%) et les relaxases (67 %). Au total, il identifie 74 % des SP (relaxase, protéine de couplage, VirB4) des modules de transfert. Le taux d'intégrases retrouvées est de 0 % car elles ne sont pas recherchées par l'outil.

- L'outil ICEfinder détecte 47% des SP de notre jeu de données : 50,3 % des intégrases, 50 % des relaxases et des VirB4 et seulement 38 % des protéines de couplage.

- Enfin, l'outil ICEscreen retrouve la quasi-totalité des SP de FirmiData (98 %). Il est ainsi le plus performant des trois outils sur ce jeu de données. De plus, nous avons noté que toutes les protéines signatures de la référence détectées par CONJscan et ICEfinder le sont aussi par ICEscreen.

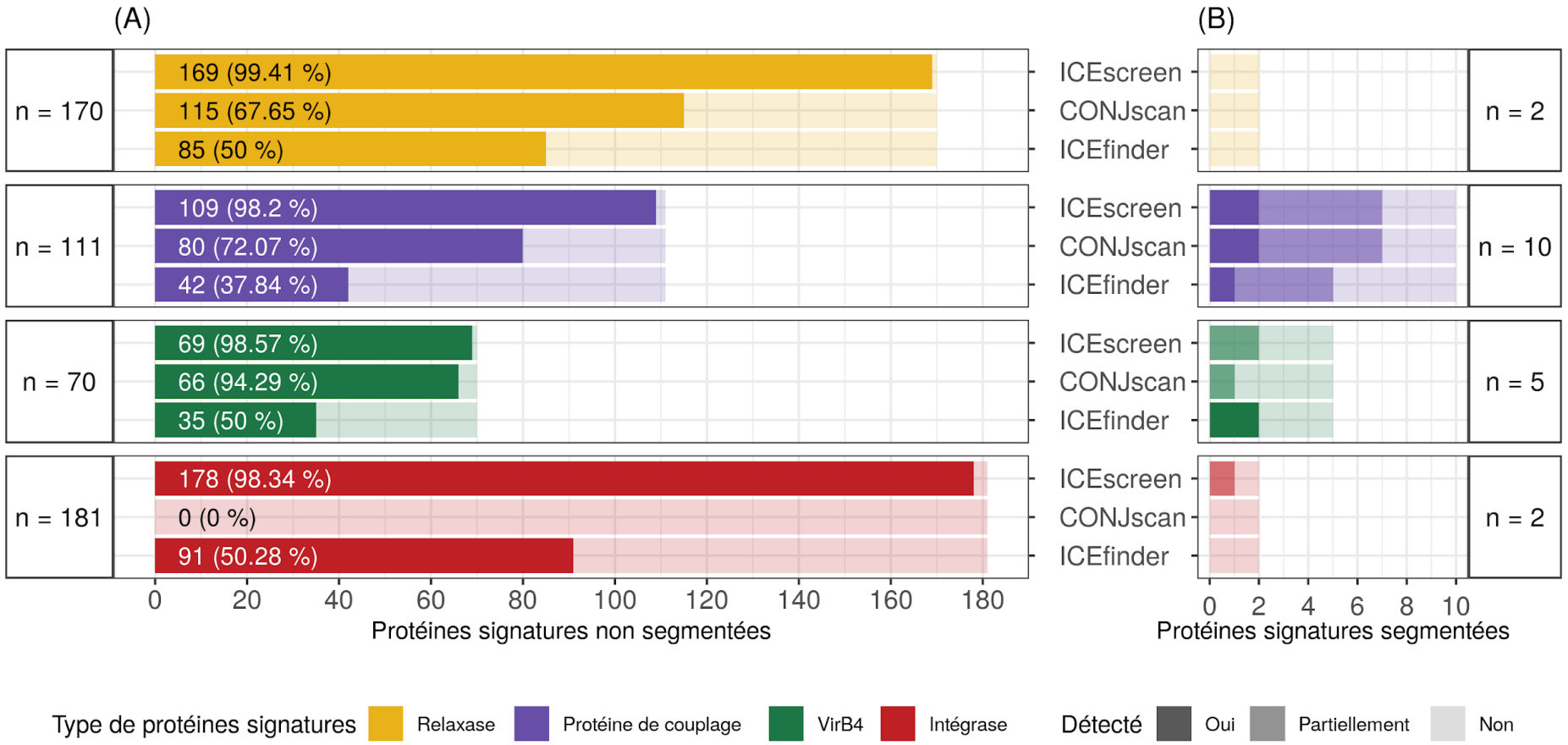

Figure 10 : Comparaison des protéines signatures de FirmiData détectées par les outils ICEscreen, CONJscan et ICEfinder. Chaque type de protéine signature est indiqué par une couleur distincte. Les chiffres indiquent de gauche à droite leur nombre ou leur pourcentage (entre parenthèses) (A) : Proportion des protéines signatures non segmentées détectées. (B) : Nombre de protéines signatures segmentées détectées totalement, détectées partiellement ou non détectées. Aucune intégrase n'est détectée avec CONJscan car cet outil ne les recherche pas.

3.4.2 Protéines signatures non détectées par l'outil ICEscreen

L'outil ICEscreen identifie 169/170 des relaxases de FirmiData. La relaxase non détectée est un pseudogène auquel la partie N-terminal est manquante. La protéine fait 143 aa de longueur, ce qui ne lui permet pas de passer les filtres sur la longueur (les relaxases fonctionnelles doivent faire ≥ 180 aa) de l'outil.

Concernant les autres SP, 109/111 des protéines de couplages et 69/70 des VirB4 ont été détectées. Les trois protéines manquantes s'avèrent être correctement détectées par ICEscreen mais ignorées parce que le taux de couverture de l'alignement avec la séquence requête est trop faible. En effet, ces trois protéines signatures sont « codées » par des pseudogènes.

Au total, 178/181 intégrases sont détectées par ICEscreen. La non détection de trois intégrases s'explique pour l'une d'elles par l'absence du gène la codant dans le fichier d'annotation donné en entrée (fichier au format Genbank), ainsi notre outil n'a pas pu l'identifier. Une autre est une transposase à DDE qui possède un domaine très différent de celles de notre banque de séquence et par conséquent qui ne peut pas être reconnu par notre profil HMM. La dernière intégrase est détectée mais écartée à cause d'une couverture insuffisante.

3.4.3 Comparaison des types d'éléments détectés

Nous présentons les résultats d'annotation obtenus par les trois logiciels comparés aux éléments annotés de FirmiData dans les figures 11a et 11b. La figure 11a présente les résultats d'annotation des 26 génomes de streptocoques de FirmiData dont le contenu en ICE et en IME est très bien connu. La figure 11b présente les résultats d'annotation des 14 autres génomes de Firmicutes de FirmiData qui ont été ré-annotés suivant la méthodologie décrite dans la section 3.3 et dont les éléments sont beaucoup moins bien connus.

Globalement, les résultats obtenus montrent les tendances suivantes :

L'outil ICEscreen :

- détecte près de la totalité des ICE et des IME de FirmiData. Il détecte la totalité des 89 ICE et la structure de ces éléments a été correctement résolue. Seuls sept ICE ne sont pas annotés en tant que tel dans les génomes de *Streptococcus anginosus* C1051, *S. dysgalactiae* subsp. e*quisimilis* ATCC 12394, *S. equi* subsp. z*ooepidemicus* ATCC 35246, *S. suis* NSUI002 et *C. difficile* R20291.

- détecte également presque tous les IME. Le seul non détecté est dIME_*EfaV583_rpsL* de *Enterococcus faecalis* V583 (figure 11b). La structure des IME retrouvés a été correctement résolue sauf pour deux IME qui font tous partie d'une même structure composite du génome *Streptococcus equi* subsp. *zooepidemicus* ATCC 35246 (figure 11a).

Figure 11a : Comparaison de l'annotation des éléments des 26 génomes de streptocoques de FirmiData détectés par les outils ICEscreen, CONJscan et ICEfinder. Les éléments remnants ne possédant pas de protéines signatures du type relaxase, protéine de couplage ou VirB4 sont omis.

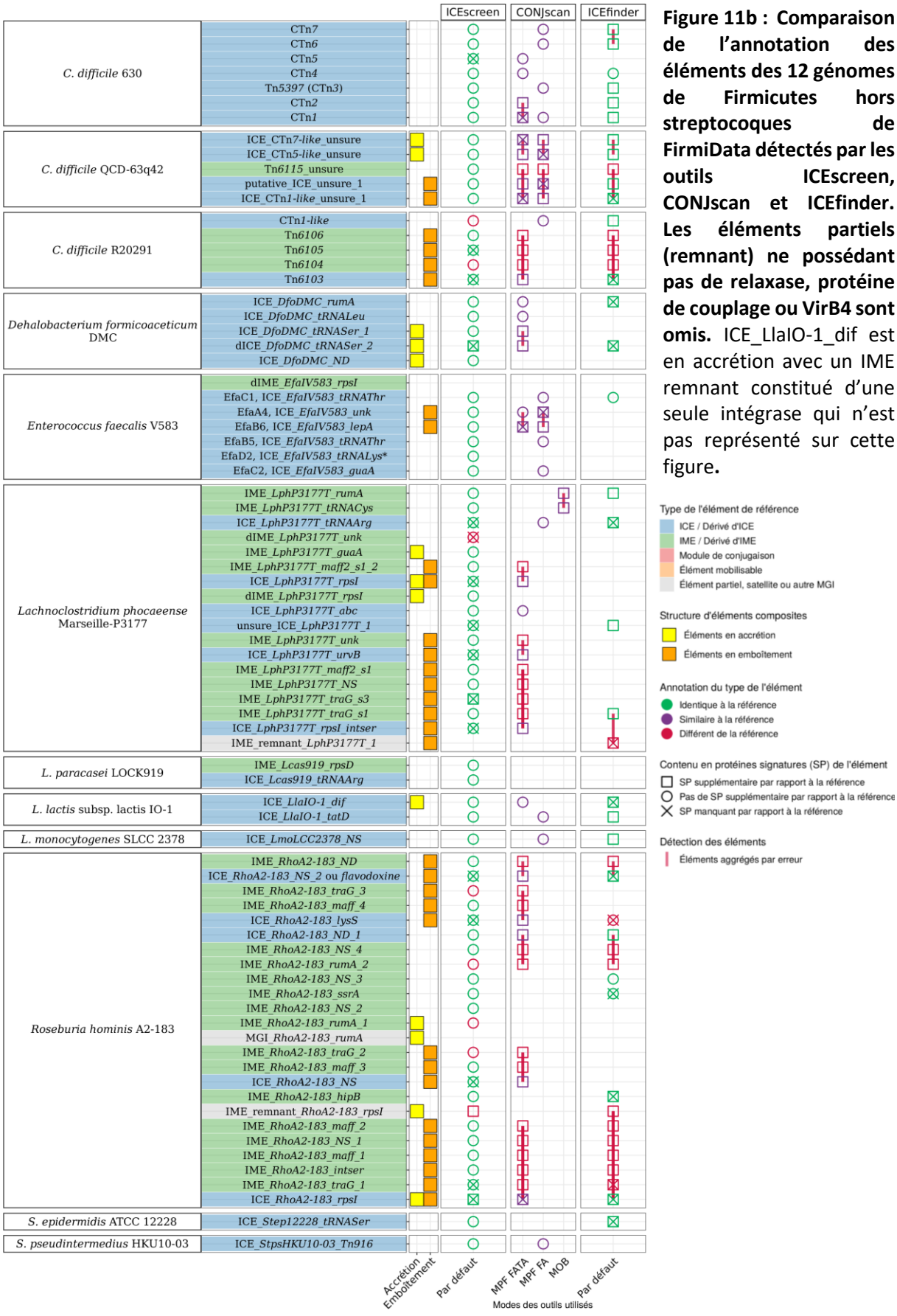

Figure 11b : Comparaison de l'annotation des éléments des 12 génomes de Firmicutes hors streptocoques de FirmiData détectés par les outils ICEscreen, CONJscan et ICEfinder. Les éléments partiels (remnant) ne possédant pas de relaxase, protéine de couplage ou VirB4 sont omis. ICE_LlaIO-1_dif est en accrétion avec un IME remnant constitué d'une seule intégrase qui n'est pas représenté sur cette figure.

L'outil CONJscan :

- détecte près de la totalité des ICE de la référence (80 ICE soit 89,9 %). Cependant, il ne résout pas correctement les structures des éléments composites. En effet, seulement sept ICE (sur 43) en structure composites ont été identifiés correctement, majoritairement des ICE en accrétion.

- retrouve très peu d'IME (38 IME soit 35 %), dont la majorité n'est pas identifiée spécifiquement en tant qu'IME. En effet, 35 sont agrégés par erreur avec des ICE dans le cas de structures composites et dont 28 sont caractérisés comme module de conjugaison (voir les génomes de *S. constellatus* subsp. *pharingis* C1050, *S. constellatus* subsp. *pharingis* C232, *S. dysgalactiae* subsp. *equisimilis* RE378, *S. pyogenes* MGAS2096, *S. salivarius* JF, *S. suis* NSUI002, *S. suis* T15 de la figure 11a). Seuls trois IME appartenant à des structures composites sont retrouvés sans erreur d'agrégation (dIME_*Seq35246_traG_s1* du génome *S. equi* subsp. *zooepidemicus* ATCC 35246 et IME_*SolAS1.3089_NS* et IME_*SsuNSUI002_tRNALeu_1* du génome de *S. suis* NSUI002) ;

L'outil ICEfinder :

- détecte un peu plus de la moitié des ICE de FirmiData (53 soit 59,6 %). Toutefois, ICEfinder n'identifie pas correctement les structures composites. Ainsi, seulement quatre ICE sur les 43 appartenant à ce type de structures ont été résolus correctement. Parmi les quatre, trois sont en accrétion.

- retrouve aussi près de la moitié des IME de FirmiData (54 soit 49,5 %), mais 30 IME sont agrégés avec d'autres éléments par erreur.

- agrège systématiquement les éléments qui sont en emboîtement, que ce soit pour les ICE ou les IME, mais résout correctement une partie des cas d'accrétion. Des exemples de cas d'accrétion bien résolus sont : IME_*SagNEM316_rpsI* du génome *S. agalactiae* NEM316 et IME_*SsaFDA259_rpmG_1* du génome *S. salivarius* ATCC 25975 de la figure 11a.

- sur les 47 ICE et IME détectés sans erreur d'agrégation, seulement cinq ont été mal annotés. Ainsi, dICE_*Sdy12394_rplL* de *S. dysgalactiae* subsp. *equisimilis* ATCC 12394 (voir la figure 1 de l'annexe), est annoté comme IME car sa VirB4 est codée par un pseudogène et le gène codant sa CP n'est pas détecté.

En conclusion, les premiers résultats obtenus sur le jeu de données "streptocoques" inclus dans FirmiData indiquent que :

- **ICEscreen** détecte et résout correctement la quasi-totalité des éléments de la référence, que ce soit des ICE isolés, des IME isolés ou des éléments composites.

- **CONJscan** détecte plutôt bien les ICE, mais ne détecte qu'une faible proportion des IME, même lorsqu'on utilise le modèle MOB. De plus, il agrège à tort les éléments en structure composite. Les IME sont détectés par CONJscan lorsqu'ils sont associés à un ICE dans une structure composite cependant ils sont systématiquement confondus avec l'ICE.

- **ICEfinder** est l'outil qui détecte le moins d'ICE et d'IME. Cependant, contrairement à CONJscan, il détecte spécifiquement des IME. Mais comme CONJscan, ICEfinder agrège ensemble les éléments d'une structure composite.

3.4.4 Exemples de détection de nouveaux éléments

Dans ce paragraphe, nous présentons, sous forme de figures, des exemples de résultats obtenus avec les trois outils et comparés à la référence (jeu FirmiData). Par ces exemples nous illustrons les performances des trois outils.

3.4.4.1 Affectation correcte des SP aux éléments

La performance des outils pour l'affectation correcte des SP aux éléments est dépendante de trois effets.

Effet des structures composites

L'outil ICEscreen affecte en général mieux les SP aux éléments que les outils CONJscan et ICEfinder. En effet, une des plus-values de l'outil est la recherche des éléments composites via un algorithme dédié récursif. Cela est particulièrement visible dans les résultats du génome de *Clostridioides difficile* QCD-63q42 (Figure 12). Selon la référence, ce génome possède deux structures d'éléments composites :

- un ICE de la superfamille Tn*5252* intégré dans un un ICE de la superfamille Tn*916* ce qui est un emboîtement inhabituel ;
- et un ICE de la superfamille Tn*5252* en accrétion avec un ICE de la superfamille Tn*916*.

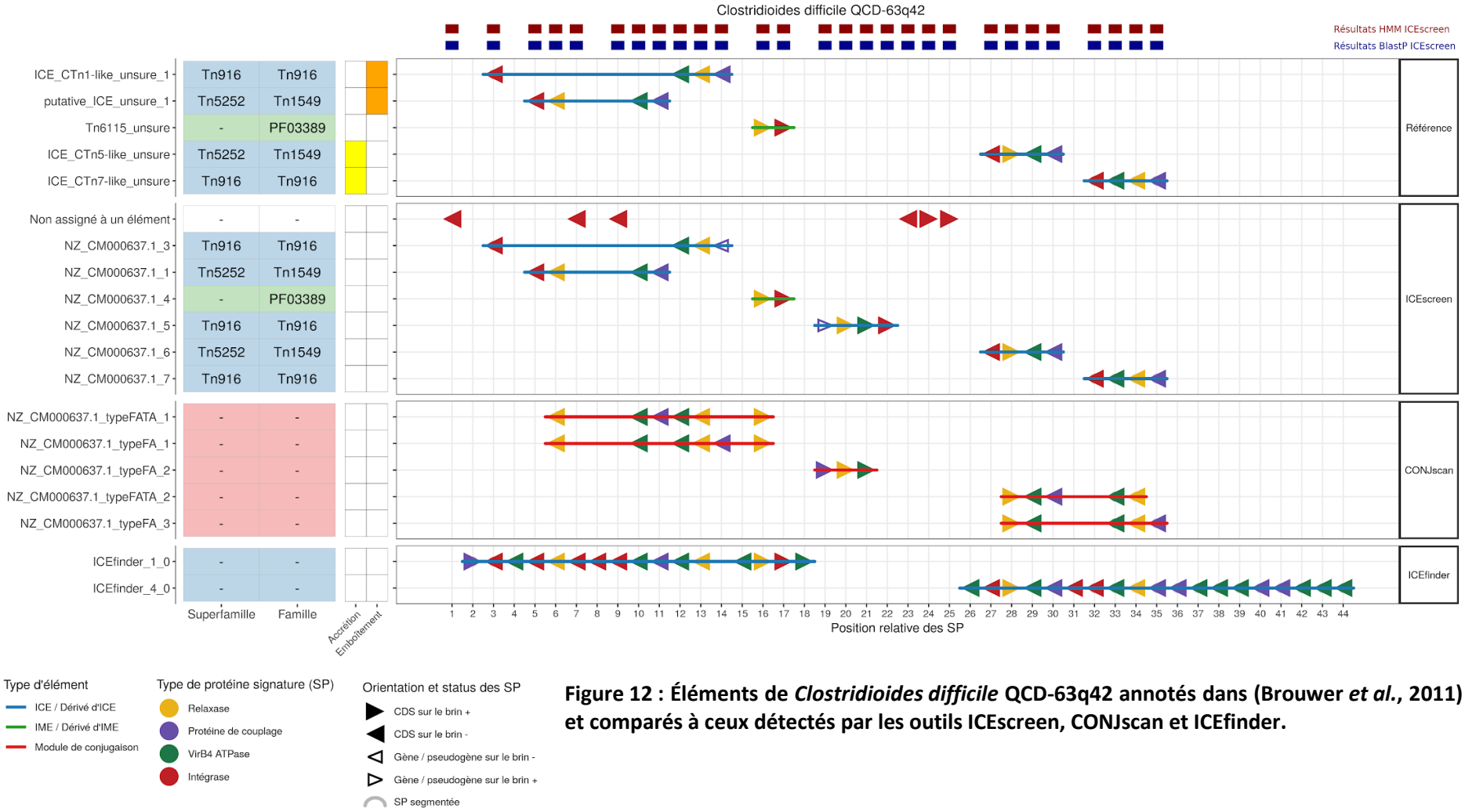

Figure 12 : Éléments de *Clostridioides difficile* QCD-63q42 annotés dans (Brouwer *et al.*, 2011) et comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder.

Tous les outils ont détecté l'ensemble des SP des ICE et IME annotés dans la référence. Cependant seul l'outil ICEscreen a résolu correctement la composition des différents éléments et annoté sans erreur leur famille. L'outil CONJscan détecte ces structures composites mais ne les résout pas, cela se traduit par la superposition de la prédiction d'un système FA (superfamille Tn*916*) et d'un hit d'un système FATA (superfamille Tn*5252*). Enfin, l'outil ICEfinder détecte extrêmement mal ces structures car il les agrège avec une multitude de SP erronées.

Effet de la proximité des éléments sur le génome

Les erreurs de d'affectation des SP des éléments peuvent être dues à la proximité des éléments sur le génome. Un exemple est celui du génome de *Clostridioides difficile* 630 (Figure 13) qui possède sept ICE isolés selon la référence. La totalité de ces ICE est détectée correctement par ICEscreen. CONJscan en détecte cinq et fusionne les deux premiers (C*Tn1* et C*Tn2*) en raison de leur proximité sur le génome (distants de moins de 21 kb soit ici 17 CDS). ICEfinder agrège également deux ICE (les deux derniers), proches l'un de l'autre sur le génome. Il affecte également presque systématiquement des SP erronées aux éléments, avec seulement un ICE détecté sans erreur d'agrégation. En plus de ces problèmes, ICEfinder détecte un IME supplémentaire composé de deux SP annotés comme étant une intégrase et une relaxase. L'intégrase détectée est une intégrase à tyrosine qui est aussi détectée par ICEscreen, cependant la protéine annotée comme relaxase est un faux positif. Le CDS en question code un répresseur transcriptionnel LexA qui réprime les gènes de réponse SOS codant entre autres pour les ADN polymérases.

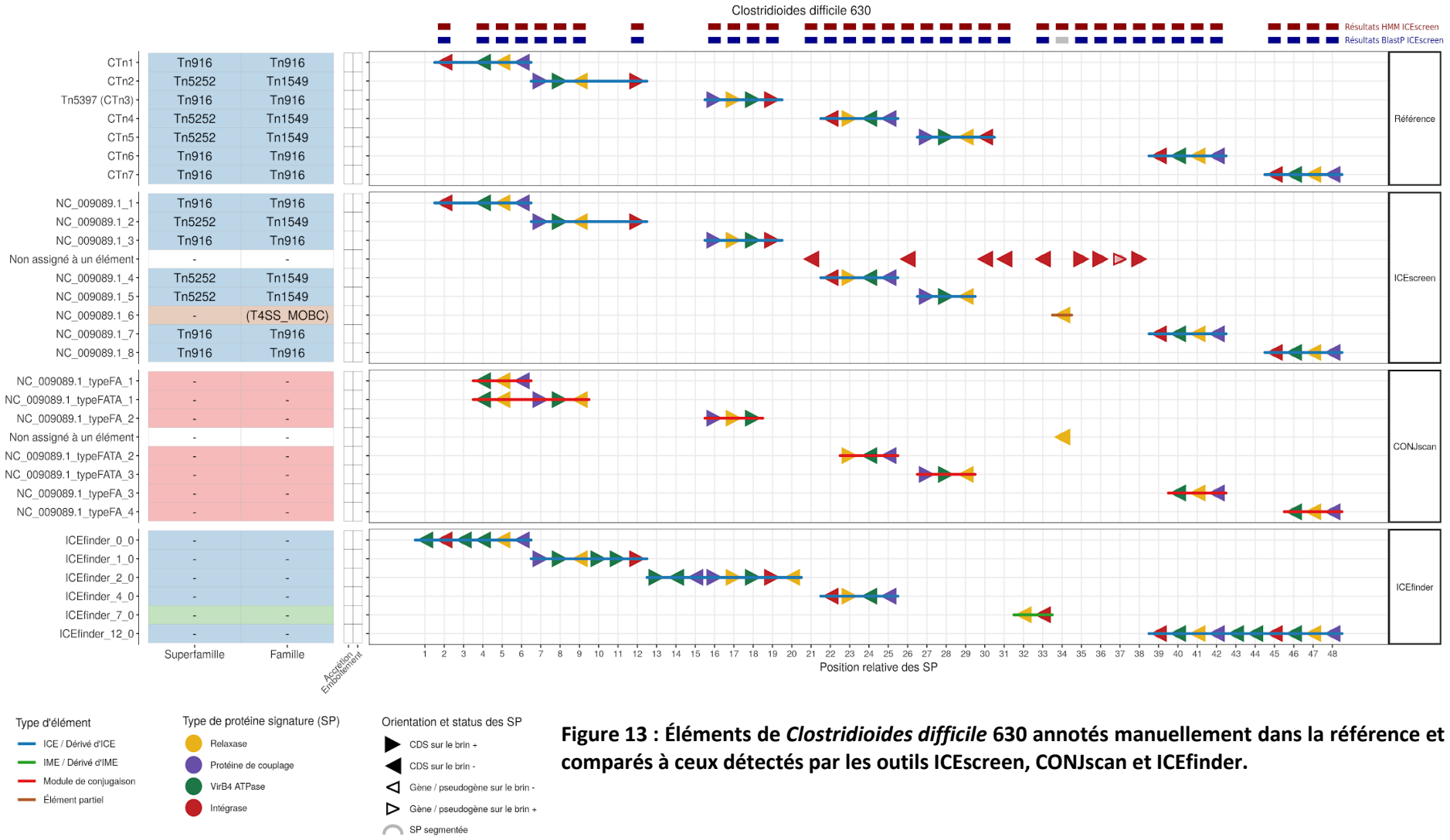

Figure 13 : Éléments de *Clostridioides difficile* 630 annotés manuellement dans la référence et comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder.

Effet de l'annotation structurale des gènes codant les protéines signatures

L'annotation des éléments ayant un gène de SP segmenté du fait de l'insertion d'un ou plusieurs IME peut être impactée. Un bon exemple est l'ICE Tn*6103* du génome *Clostridium difficile* R20291 (Figure 14) qui porte trois IME dont l'un, Tn*6104,* interrompt le gène de la CP. Malgré la segmentation du gène de la CP, ICEscreen résout correctement cette structure. Cependant, le second fragment du gène de la CP n'est pas assigné à l'ICE. Ainsi un cas de segmentation « simple » ne semble pas impacter la résolution de ces structures par ICEscreen. Les deux autres outils agrègent les quatre éléments.

L'outil n'est toutefois pas aussi performant pour des gènes de SP très segmentés c'est-à-dire lorsque plusieurs éléments mobiles sont intégrés dans le même gène de SP. L'ICE ICE_*LphP3177T_intser* du génome de *Lachnoclostridium phocaeense* Marseille-P3177 (voir figure 19a) illustre ce propos. Le gène codant la CP de cet élément est segmenté en trois par l'insertion de deux IME différents. À cause de ces segmentations, le premier fragment n'est détecté par aucun des trois outils. Le second fragment qui est de très petite taille est détecté correctement en tant que CP par CONJscan, mais comme intégrase par ICEscreen. Quant au dernier fragment, il est correctement détecté par ICEscreen et CONJscan. Ainsi, ICEscreen n'a pas pu résoudre correctement la structure composite du fait de la non détection du premier fragment de la CP.

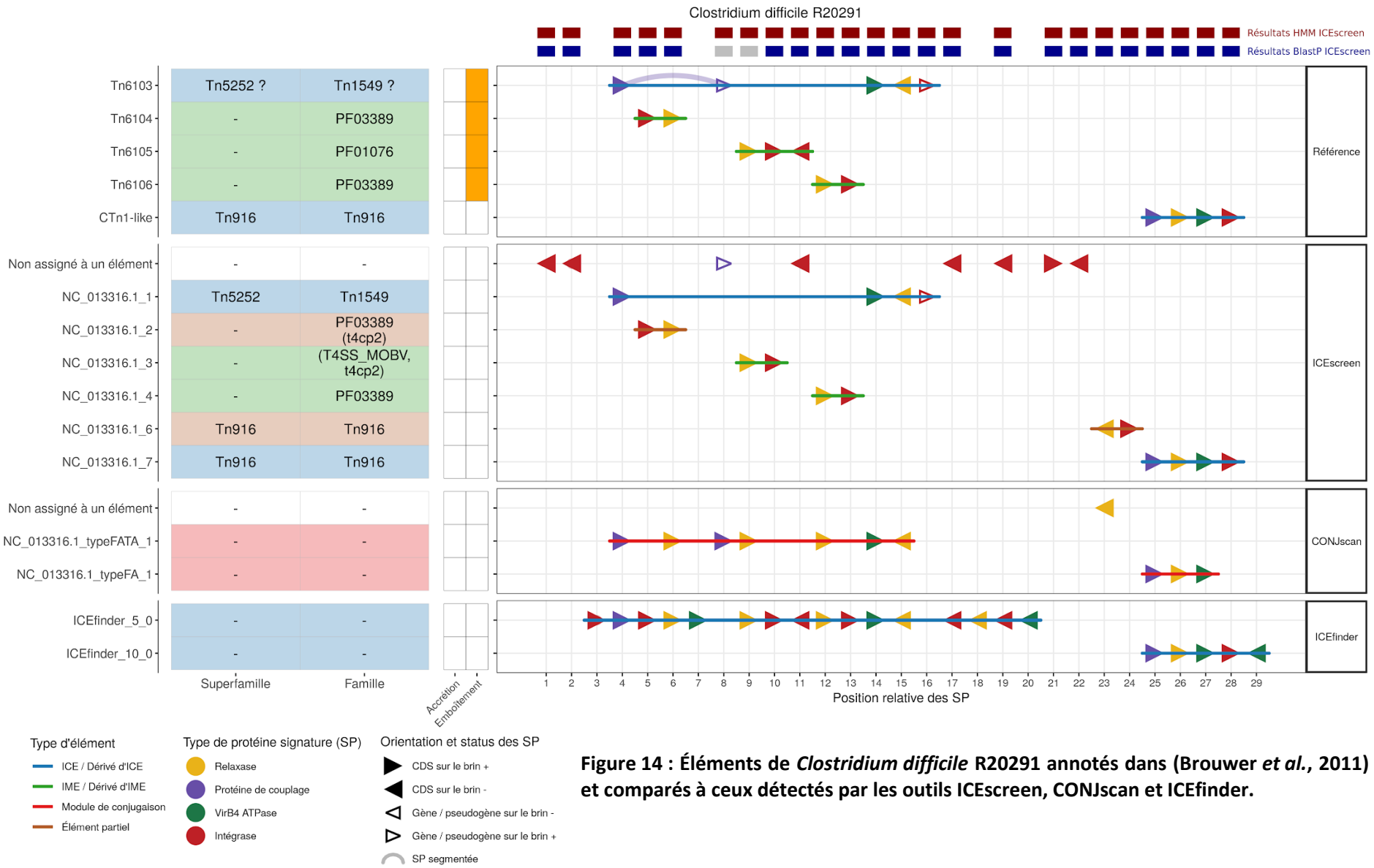

Figure 14 : Éléments de *Clostridium difficile* R20291 annotés dans (Brouwer *et al.*, 2011) et comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder.

3.4.4.2 Détection et attribution correcte des intégrases aux éléments

Une des difficultés pour l'attribution correcte des intégrases aux éléments est la présence dans les génomes d'intégrases d'autres éléments mobiles qui peuvent être assignés à un ICE ou un IME par erreur. C'est le cas par exemple du génome de *S. epidermitis* ATCC 12228 qui contient un ICE de superfamille Tn*GBS1* dont les SP du module de conjugaison ont tous été détectés par les trois outils ICEscreen, CONJscan et ICEfinder. En plus de l'intégrase de l'élément, ce génome contient également six intégrases supplémentaires. Alors que l'outil ICEscreen assigne la bonne intégrase à l'élément, ICEfinder lui assigne à tort une autre intégrase qui n'est ni dans la référence, ni détectée par ICEscreen. Il est ainsi fort probable qu'il s'agisse d'une erreur et que la protéine détectée par ICEfinder ne soit pas une intégrase d'ICE. Concernant les intégrases supplémentaires, elles appartiennent à des éléments intégrés qui ne sont ni des ICE ni des IME. ICEscreen ne les a pas assignées à un ICE ou IME et les considère comme des « SP isolées » et ne réalise donc pas d'erreurs d'assignation.

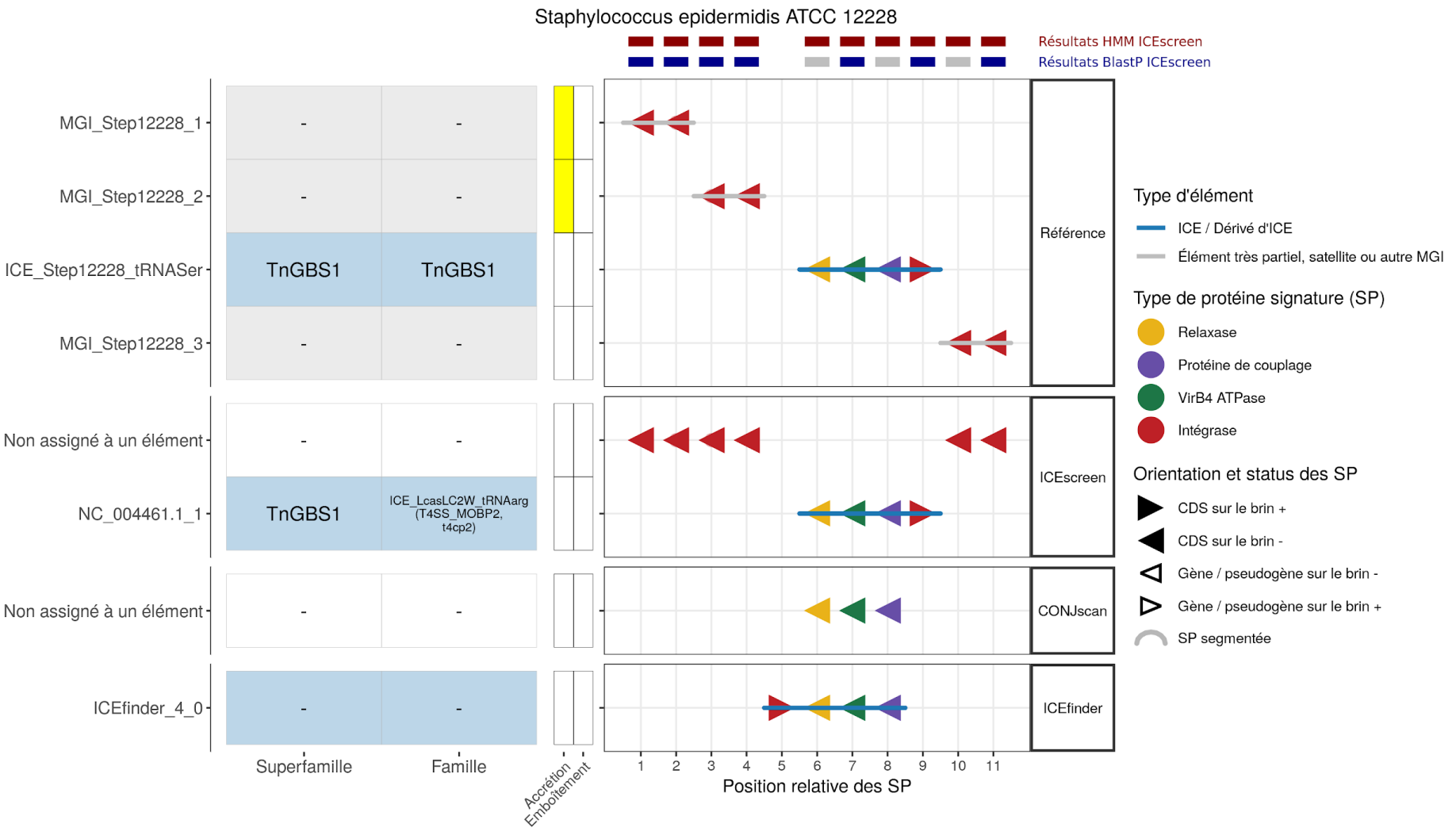

Figure 15 : Éléments de *Staphylococcus epidermitis* ATCC 12228 annotés manuellement dans la référence et comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder.

En plus de la détection d'intégrases d'autres éléments, des détections incorrectes de SP (faux positifs) compliquent l'annotation des éléments. Ceci est particulièrement vrai pour ICEfinder. Ainsi, dans le génome *Lactococcus lactis* subsp. *lactis* IO-1 (Figure 16), le problème de détection incorrecte des intégrases de l'outil ICEfinder est couplé à la détection incorrecte de CP et de VirB4. En effet, l'élément « ICEfinder_2_0 » annoté ICE possède trois intégrases et a une composition en SP très différente de l'ICE de la référence. L'élément détecté n'est clairement pas un ICE valide.

Pour conclure, malgré les bonnes performances de ICEscreen pour la détection et l'attribution des intégrases aux éléments, des améliorations sont possibles. Dans *L. lactis* subsp. *lactis* IO-1, l'intégrase de l'ICE remnant « ICE_remnant_*LlaO-1_dif* » n'est pas détectée par l'outil. Un alignement a toutefois été obtenu, cependant il a été exclu à cause de la longueur de la protéine. Celle-ci, a une longueur de 248 aa et est ainsi d'une taille inférieure aux 320 aa requis pour être considérée comme une intégrase à tyrosine par l'outil. Ce filtre a été mis en place à partir des connaissances sur les intégrases à tyrosine d'ICE et d'IME de streptocoques et pourrait ainsi être inadapté pour cette situation.

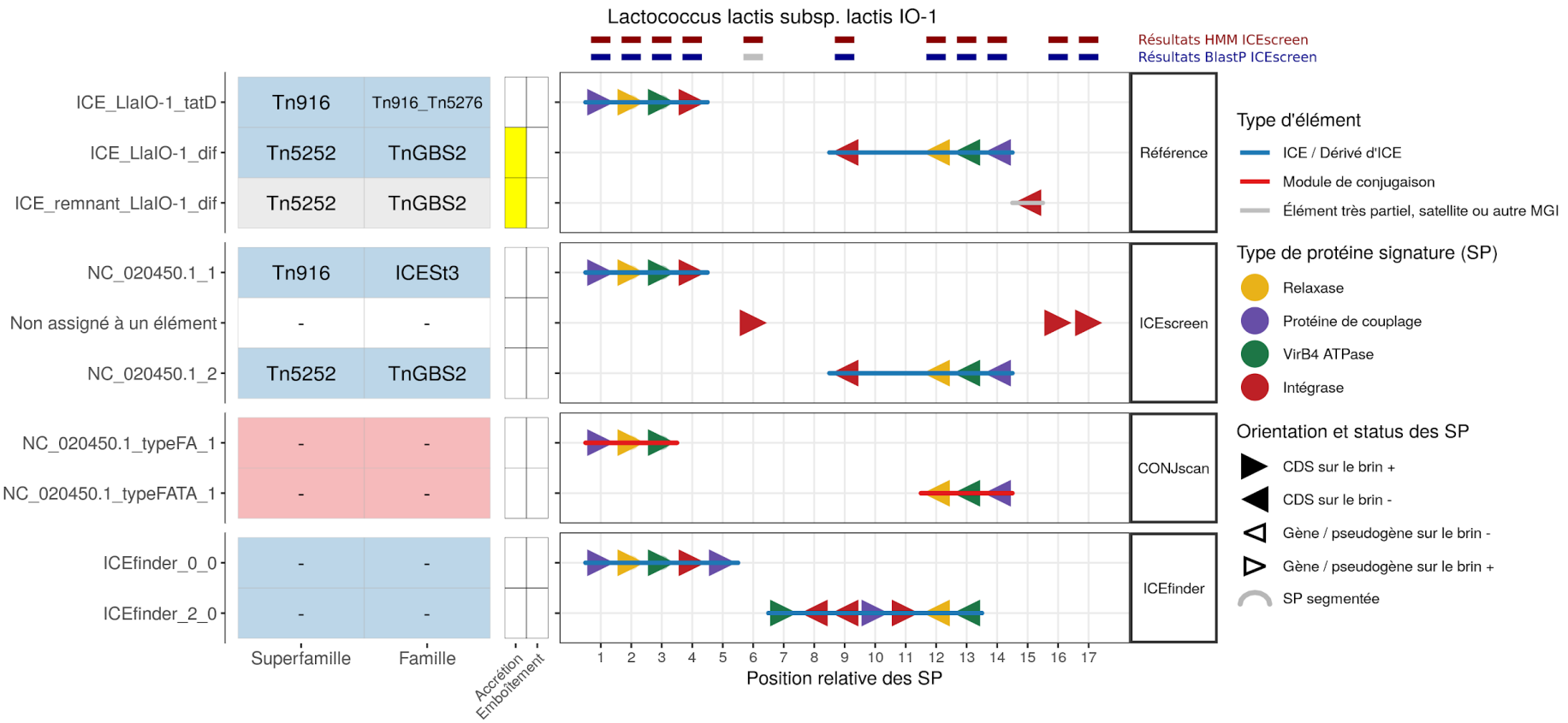

Figure 16 : Éléments de *Lactococcus lactis* subsp. *lactis* IO-1 annotés manuellement dans la référence et comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder.

3.4.4.3 Aide à la découverte de nouveaux éléments

Les éléments détectés par l'outil ICEscreen sont annotés à plusieurs niveaux de complétude (voir l'article 2.3.3 sur les règles de caractérisation des éléments de l'outil ICEscreen). Pour rappel, il y a trois niveaux d'annotation de l'élément en fonction de la complétude de l'annotation :

- Annotation complète : les éléments considérés par l'outil comme des ICE et des IME et dont les SP ont toutes été retrouvées sont caractérisés comme « ICE » et « IME » respectivement.
- Le module de transfert de l'élément est annoté : il s'agit d'éléments annotés en tant que « module de conjugaison » ou « élément mobilisable ». Ces éléments n'ont pas d'intégrase assignée par l'outil.
- Annotation partielle : il s'agit d'éléments codant au moins deux SP qui peuvent correspondre à des ICE ou IME partiels. Ces éléments sont classés comme « éléments partiels », cependant si une VirB4 est détectée, l'élément est caractérisé en tant que « ICE partiel ».

En plus d'annoter les éléments possibles, ICEscreen met à disposition l'annotation des SP détectées qui n'ont pas été assignées à un élément. Ces SP sont indiquées par l'outil en tant que « SP isolée ». Ces SP dites isolées peuvent correspondre à des ICE et IME très dégénérés ou très éloignés des éléments déjà connus.

Recherche d'éléments et SP atypiques :

- Parmi les cas d'annotation d'éléments complets nouveaux, les résultats du génome de *Lactobacillus paracasei* LOCK919 (Figure 17) sont à relever. Ce génome possède un ICE de superfamille Tn*GBS1* et un IME possédant une relaxase de la famille Rep_1 (PFAM PF01446). Cette famille de relaxase est retrouvée dans des petits plasmides de Firmicutes (Lee *et al.*, 2012). L'outil ICEscreen est le seul à avoir réussi à détecter ces deux éléments. La non détection par CONJscan s'explique par la nature des éléments du génome. En effet, les modules de conjugaison de type Tn*GBS1* ne sont pas

recherchés par CONJscan. Quant à l'IME, les relaxases de la famille Rep_1 ne sont pas recherchés ni par CONJscan, ni par ICEfinder.

- Le cas de IME_*LphP3177T_tRNACys* du génome *Lachnoclostridum phocaeense* Marseille-P3177 (voir figure 19b) qui n'est détecté que par ICEscreen est aussi intéressant. L'IME comporte une relaxase MOB_V, une CP VirD4 et une intégrase à tyrosine. L'association d'une relaxase MOB_V avec une CP VirD4 n'a jamais été observée, il s'agit donc possiblement d'une nouvelle famille de module de mobilisation.

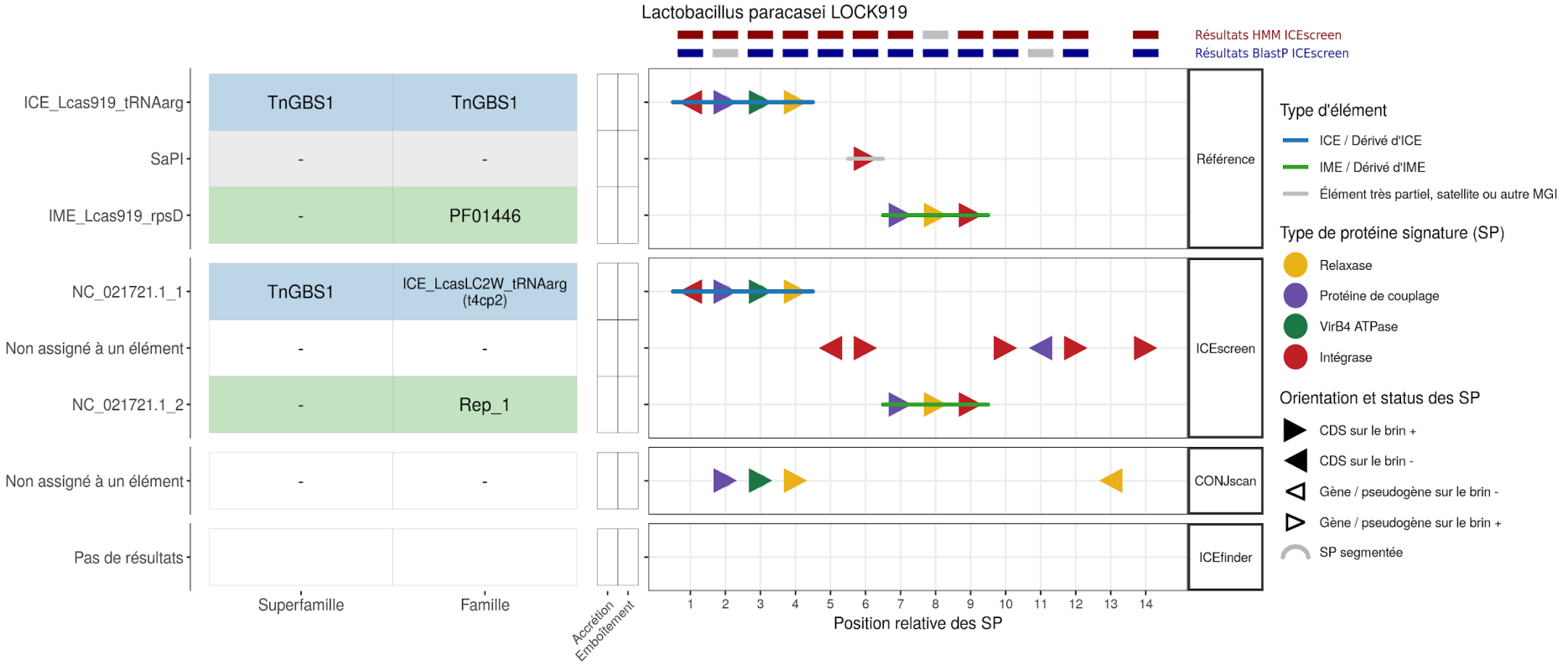

Figure 17 : Éléments de *Lactobacillus paracasei* LOCK919 annotés manuellement dans la référence et comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder.

Aide à la découverte par annotation des SP isolés :

- Dans le génome *L. phocaeense* Marseille-P3177, le dérivé d'IME dIME_*LphP3177T_unk* (voir figure 19c) n'est détecté par aucun des outils sauf l'outil ICEscreen qui le détecte en tant qu'élément partiel car seule la relaxase a été détectée.

- L'analyse des SP isolées peut aussi permettre la découverte de nouveaux éléments. Cela est illustré par les résultats du génome *Enterococcus faecalis* V583 (Figure 18) où le dérivé d'IME dIME_*EfalV583_rpsI* n'est détecté par aucun des trois outils. Cependant, ICEscreen annote la relaxase de l'élément en tant que SP isolée.

3.4.4.4 Les éléments de *Lachnoclostridium phocaeense* Marseille-P3177

Nous allons détailler l'exemple des résultats du génome de *L. phocaeense* Marseille-P3177 qui illustre la capacité d'ICEscreen à trouver correctement la plupart des ICE et d'IME d'une espèce très éloignée des streptocoques. Cet exemple permettra également de montrer les limites actuelles de notre outil sur des SP trop dégénérées et l'incapacité d'ICEfinder à détecter correctement et exhaustivement les éléments de ce génome.

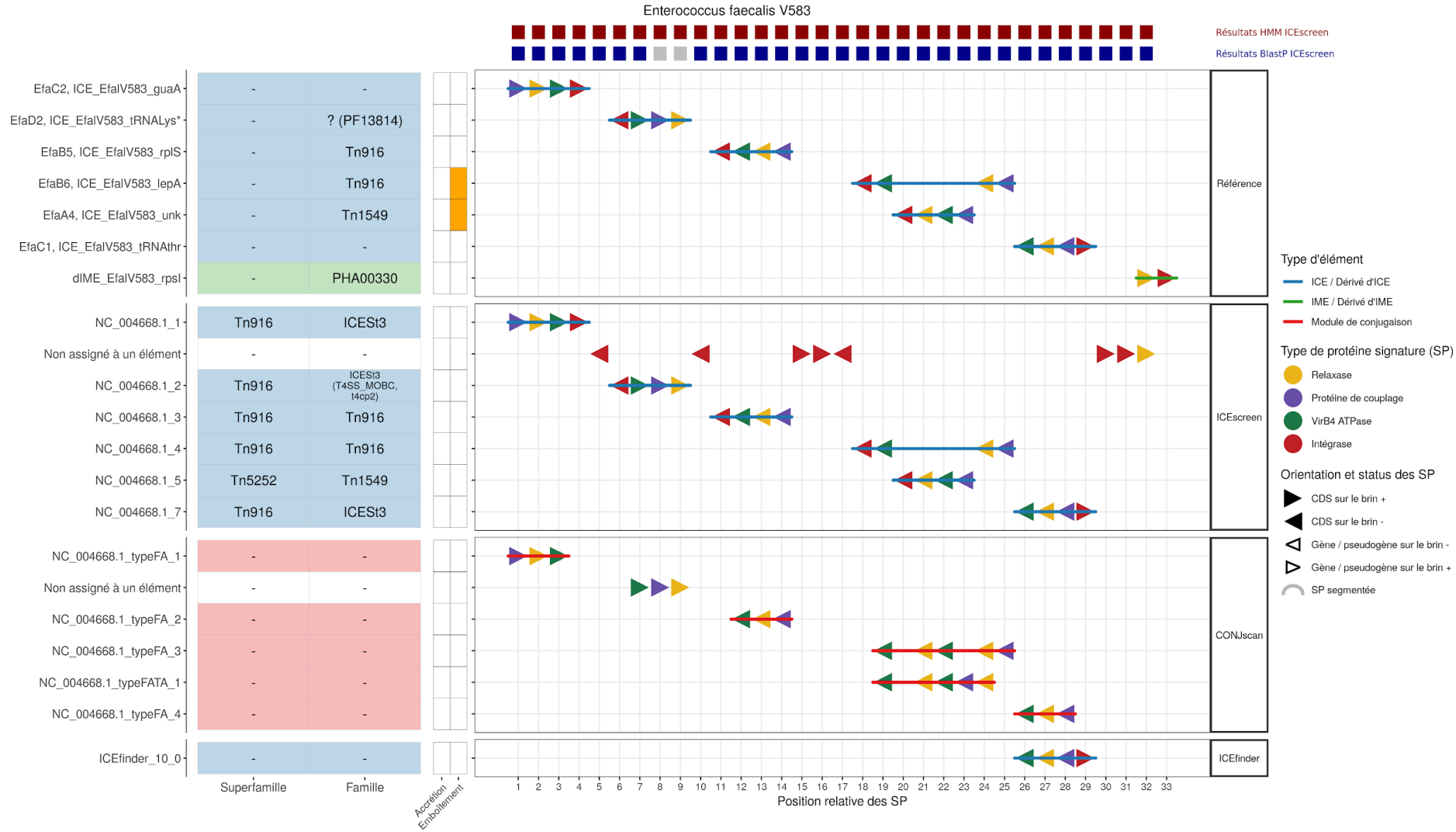

Figure 18 : Éléments de *Enterococcus faecalis* V583 annotés manuellement dans la référence et comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder. L'annotation des protéines signatures des six premiers éléments a été extraite de la publication de (Burrus *et al.*, 2002). L'annotation du 7ème élément a été effectuée par G. Guédon.

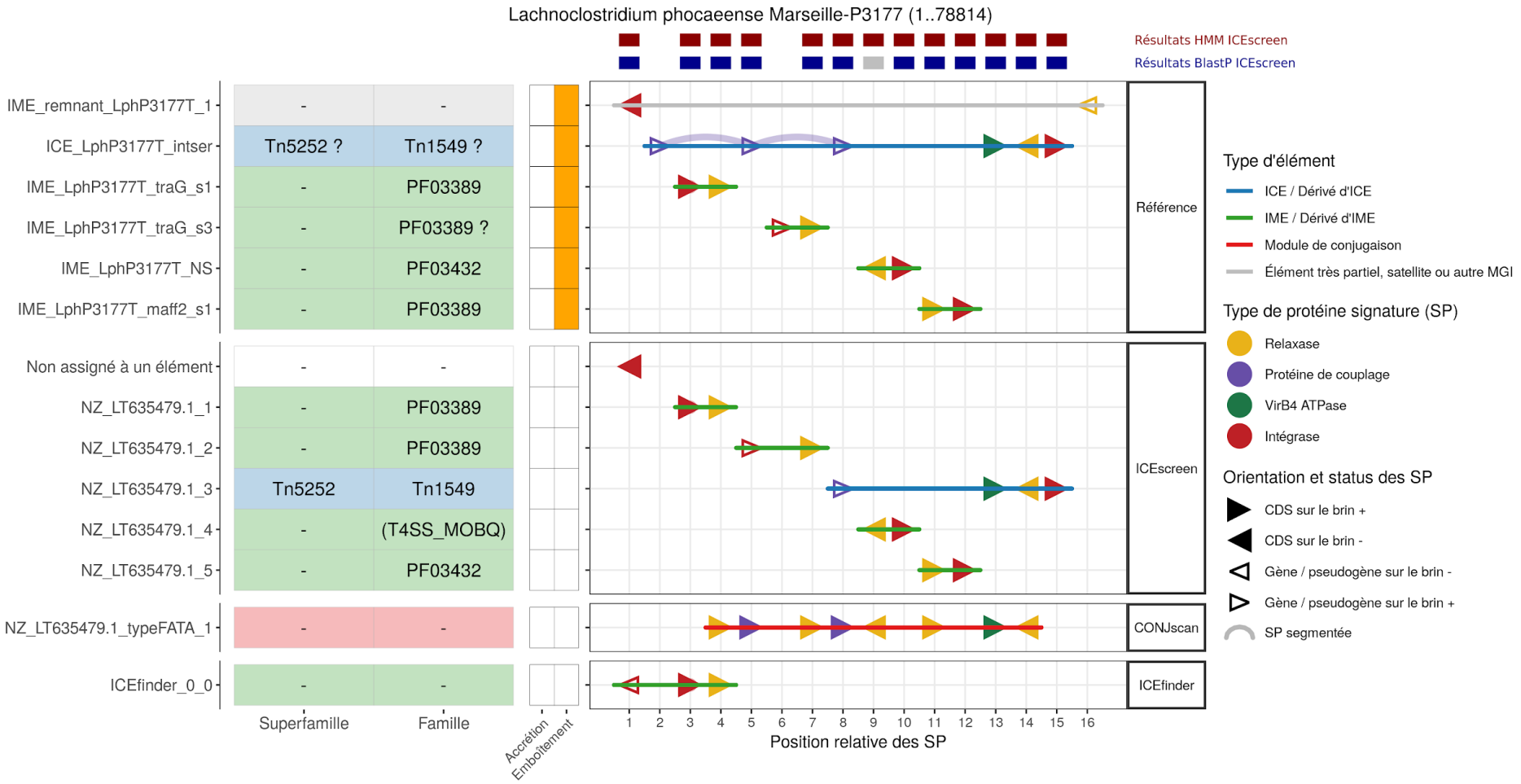

Figure 19a : Éléments de *Lachnoclostridium phocaeense* Marseille-P3177 (positions 1 à 78814 du génome) annotés manuellement dans la référence et comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder.

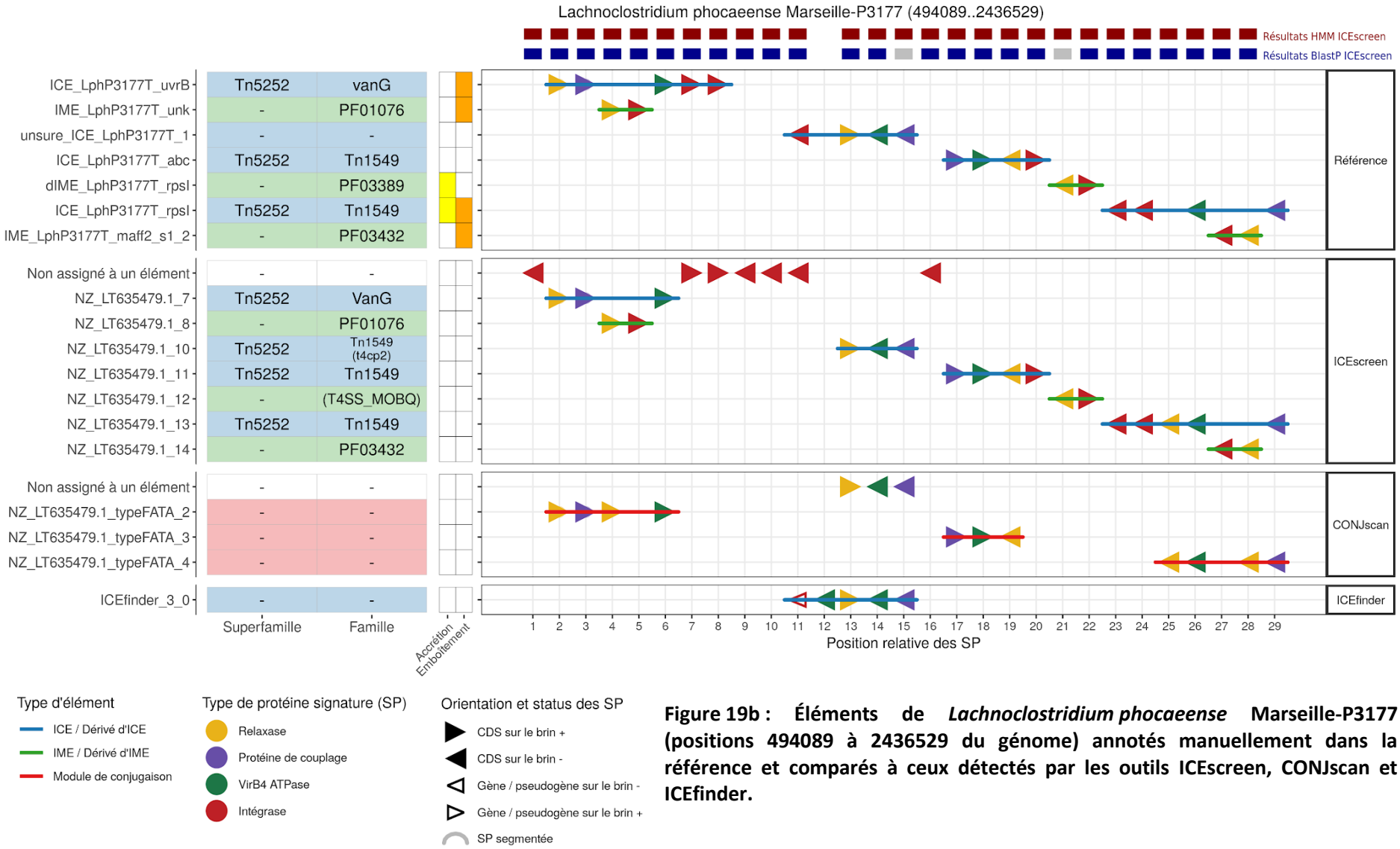

Figure 19b : Éléments de *Lachnoclostridium phocaeense* Marseille-P3177 (positions 494089 à 2436529 du génome) annotés manuellement dans la référence et comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder.

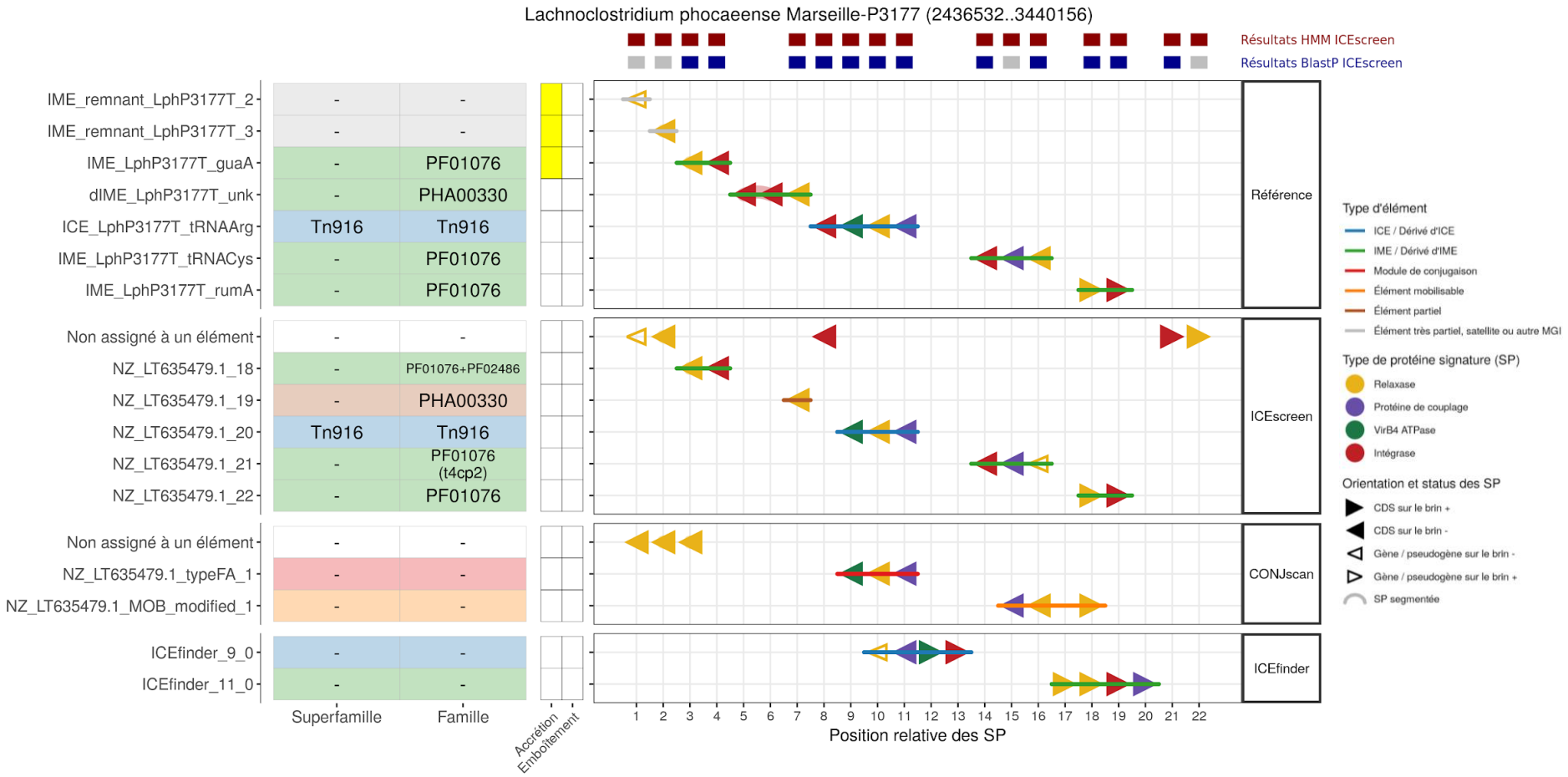

Figure 19c : Éléments de *Lachnoclostridium phocaeense* Marseille-P3177 (positions 2436532 à 3440156 du génome) annotés manuellement dans la référence et comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder.

Selon la référence, ce génome contient 20 éléments dont 11 IME, 6 ICE et 3 éléments très dégénérés. Parmi eux, quatre structures composites, pouvant être très complexes, sont détectées :

- Une première structure contenant un IME très dégénéré hôte d'un ICE qui est lui-même hôte de trois IME différents dont l'un est lui-même hôte d'un IME (IME_*Lph93177T_maff2_s1* est hôte de IME_*Lph3177T_NS*) (voir figure 19a) ;

- Une deuxième structure avec un ICE hôte d'un IME (voir figure 19b) ;

- Une troisième structure avec un ICE hôte d'un IME et en accrétion avec un autre IME (voir figure 19b) ;

- Une quatrième structure avec un IME en accrétion avec deux IME très dégénérés (voir figure 19c).

En plus de ces quatre structures complexes, l'annotation de référence inclut trois ICE et trois IME isolés : l'ICE ICE_*Lph3177T_tRNAArg* étant uniquement considéré comme un ICE potentiel (Figure 19c).

ICEscreen retrouve globalement tous les éléments sauf les trois très dégénérés. Par contre, il ne résout pas correctement la structure de l'ICE ICE_*Lph3177T_intser* (Figure 19a) car le gène codant pour la protéine de couplage est morcelé en trois fragments à cause de l'insertion de deux IME différents. De plus, ICEscreen détecte un ICE sans intégrase (« NZ_LT635479.1_7 » qui correspond au module de conjugaison de ICE_*Lph3177T_uvrB* dans la référence ; Figure 19b), des intégrases sont détectées mais ne sont pas assignées à l'élément. Des problèmes similaires sont retrouvés pour d'autres éléments de ce génome. Cela montre que les règles d'assignation des intégrases ne sont pas adaptées pour ces éléments. Il sera nécessaire de rajouter des règles spécifiques pour résoudre ce type de cas.

CONJscan détecte six de ces 20 éléments : 5 ICE et 1 IME ainsi que deux groupes de protéines signatures non assignées à un élément. Les 5 ICE détectés sont des éléments isolés ou hôtes

de structures complexes. L'élément annoté en tant qu'IME n'en est pas un car il s'agit, en fait, de deux IME en accrétion.

ICEfinder détecte seulement quatre de ces 20 éléments, 2 ICE et 2 IME, mais aucun des éléments ne correspond précisément à ceux identifiés dans la référence montrant, là encore, les limites de cet outil.

3.4.5 Bilan des éléments détectés au sein du jeu Firmidata

Le tableau 5 ci-dessous présente les pourcentages d'éléments détectés par chacun des outils sur le jeu de données FirmiData. Pour rappel, l'annotation de FirmiData est composée d'éléments de streptocoques et d'autres souches de Firmicutes validés manuellement par l'unité DynAMic et pour lesquels l'annotation est très fiable. Pour les autres génomes de Firmicutes, les annotations des éléments sont tirées de publications, comme les génomes de *C. difficile* et qui n'ont donc pas tous été validés par manque de temps. Au total, 38 génomes ont été annotés et leurs éléments sont recensés dans le tableau 5.

Tableau 5 : Bilan des ICE, IME et éléments dégénérés détectés par les outils ICEscreen, CONJscan et ICEfinder par rapport à l'annotation de référence.

	Annotation de référence	ICEscreen	CONJscan	ICEfinder
ICE	89	89 (100 %)	80 (89,9 %)	53 (59,6 %)
IME	109	108 (99 %)	38 (34,9 %)	54 (49,5 %)
Éléments dégénérés type "remnant"	11	4 (36,4 %)	2 (18 %)	5 (45,4 %)
Total	209	201 (96 %)	120 (57,4 %)	112 (53,6 %)

Globalement les résultats montrent qu'ICEscreen et CONJscan détectent la totalité ou la plupart des ICE de FirmiData alors qu'ICEfinder n'en détecte que 59 %. ICEscreen est le seul outil à détecter correctement les IME alors qu'ICEfinder et CONJscan en détectent moins de la moitié, respectivement 49 % et 35 %. On peut noter que CONJscan n'est pas un outil dédié à la recherche des IME, pourtant il détecte près de 35 % des IME du jeu FirmiData. Ces IME correspondent le plus souvent à des IME en accrétion ou en emboîtement avec des ICE. Ce qui justifie qu'ils aient été détectés par l'outil.

Compte tenu du fait que la méthode de recherche des éléments de ICEfinder n'est pas publiée, nous ne pouvons pas aller plus loin dans la compréhension des erreurs commises par cet outil. La suite de ce paragraphe se concentrera uniquement sur les outils ICEscreen et CONJscan pour une comparaison un peu plus poussée de ces outils.

3.4.5.1 Bilan de la détection des ICE

Concernant ICEscreen notre objectif a été atteint avec un bilan global de détection de 100 % des ICE par rapport à notre annotation de référence de FirmiData.

Parmi les ICE identifiés par ICEscreen et non détectés par CONJscan, il y a les ICE de la superfamille Tn*GBS1*. Leur détection par ICEscreen résulte d'une part de la recherche des relaxases de type MOB_L, relaxases caractéristiques de la superfamille Tn*GBS1* et d'autre part de la recherche spécifique des VirD4 des ICE de cette famille qui ont la particularité d'avoir une insertion d'environ 200 aa en leur sein. L'outil CONJscan ne possède pas de modèle dédié

pour la détection du système T4SS des ICE de cette superfamille. Il était donc attendu que ces éléments ne puissent être détectés en tant qu'ICE par cet outil.

3.4.5.2 Bilan sur la détection des IME

En plus des ICE, ICEscreen intègre également des fonctionnalités de recherche d'IME et détecte la quasi-totalité des IME de FirmiData (108/109). Pour permettre cela, les relaxases d'IME identifiés lors de l'analyse des génomes de streptocoques (PF01719-*like*, PHA00330-*like*, PF02407-*like*, etc.) (thèse de C. Coluzzi, 2017) ont été répertoriées et des profils HMM ciblant ces protéines ont été ajoutés à l'outil ICEscreen (voir l'a 2.2.2 sur la banque de profils HMM de l'outil ICEscreen). Ceci explique que ICEscreen soit plus performant que d'autres systèmes pour l'annotation des IME du jeu de données FirmiData.

3.4.5.3 Bilan sur la détection des éléments composites

En plus de rechercher les ICE et les IME isolés, l'outil ICEscreen intègre une stratégie permettant d'identifier des ICE et des IME en emboîtement. Les ICE et IME en accrétion sont identifiés par l'outil comme étant des ICE et IME isolés. Ainsi l'outil résout 21 des 37 structures composites du jeu FirmiData, ce qui correspond à 50 éléments du jeu soit près de la moitié (49,5 %) des éléments composites de FirmiData. Les structures composites peuvent être constituées d'un seul type d'éléments, par exemple un ICE hôte d'un autre ICE (7 sur 11 structures composites résolues) ou deux IME en accrétion (4 sur 6 structures composites résolues). Ainsi, les structures constituées d'un seul type d'éléments sont bien résolues par ICEscreen. Toutefois, la moitié des structures composites sont « mixtes », c'est-à-dire constituées d'ICE et d'IME qui peuvent être en emboîtement ou accrétion. ICEscreen résout la moitié d'entre elles (10 sur 20).

De plus, ICEscreen résout aussi les structures composites plus complexes (accrétion et/ou emboîtement de plus de trois éléments ainsi que les emboîtements semblables aux poupées « matryoshka »).

À notre connaissance, ICEscreen est à ce jour le seul outil bioinformatique capable de résoudre des structures composites complexes d'éléments conjugatifs intégrés.

Les quatre structures mal résolues par ICEscreen ont toutes la particularité de posséder un élément très dégénéré (« remnant »), qui peut dériver d'un ICE, d'un IME, ou d'un autre type d'éléments (voir les génomes *S. dysgalactiae* subsp. *equisimilis* RE378 et *S. equi* subsp. *zooepidemicus* ATCC 35246 de la figure 11a et *Lachnoclostridium phocaeense* Marseille-P3177 et *Roseburia hominis* A2-183 de la figure 11b).

Annotation des familles d'ICE

Un autre intérêt de l'outil ICEscreen est qu'il intègre une annotation des familles et des superfamilles d'ICE basée sur la nature des SP du module de conjugaison. Cela permet, pour tout ICE nouvellement identifié, de déterminer s'il s'agit déjà d'un élément appartenant à une famille ou superfamille connue ou non. L'utilisation de profils HMM permet de détecter de nouvelles relaxases et de nouvelles protéines de couplage et de nouvelles combinaisons de SP qui correspondent à de nouvelles familles d'ICE, comme l'illustre le cas de IME_*LphP3177T_tRNACys* (détaillé dans le paragraphe 3.4.4.3).

Au total trois éléments partiels incluant une protéine VirB4 ont été détectés. Ils correspondent tous à des ICE annotés dans la référence de FirmiData. Il s'agit probablement d'ICE nouveaux dont il faudra caractériser les protéines manquantes pour compléter leur annotation.

Annotation des IME

Concernant les IME, ICEscreen donne des informations sur la famille des relaxases et, lorsque cela se justifie, également sur la famille de la protéine de couplage. Certaines de ces SP ne sont détectées que par HMM et non par BlastP, ce qui suggère qu'elles sont éloignées en terme de conservation de séquence de celles déjà répertoriées chez les streptocoques.

Annotation des éléments partiels

Au total 28 éléments partiels, sans protéine VirB4, ont été détectés. Cinq correspondent à des IME et trois à des ICE dont le gène de la VirB4 est segmenté ou un pseudogène. Deux autres correspondent à des éléments dégénérés. Les 18 éléments partiels restants sont probablement de nouveaux IME pour lesquels des SP n'ont pas été détectés ou bien des ICE ayant un gène VirB4 segmenté.

Discussion et perspectives

1. Automatisation de l'annotation des SP

La détection des ICE et IME dans les génomes de *Streptococcus* avec ICEscreen montre des résultats extrêmement satisfaisants. Ceci est dû en grande partie à l'expertise acquise par l'équipe ICE-TeA de l'unité DynAMic sur les éléments de streptocoques qui a servi de base solide à la mise au point de nouveaux profils HMM et de règles de typage de ces éléments. Une plus-value de l'outil est d'intégrer une méthode automatique de détection des ICE mais également des IME complets ou partiels.

L'automatisation a été possible grâce à l'accumulation des connaissances des ICE et IME de streptocoques (Ambroset *et al.*, 2016; Coluzzi *et al.*, 2017; Coluzzi, 2017) et plus largement des éléments de Firmicutes (pour revues Bellanger *et al.*, 2014; Guédon *et al.*, 2017). L'outil ICEscreen détecte tous les éléments des streptocoques y compris les éléments incomplets et dégénérés et garde la trace de protéines signatures non assignées. La détection des éléments incomplets, dégénérés et des protéines signatures non assignées est un choix délibéré. D'une part, la détection d'éléments incomplets est nécessaire pour résoudre correctement la structure d'éléments composites. D'autre part, les SP non assignées pourraient éventuellement correspondre à des protéines signatures éloignées de celles connues et appartenir à de nouveaux types d'éléments. Ces SP feront l'objet d'une caractérisation afin de déterminer si tel est le cas. Le choix qui a été fait de garder les traces d'annotation de ces éléments incomplets et des protéines signatures non assignées à des éléments constitue donc une aide à la découverte possible de nouveaux éléments.

Importance de la détection des protéines signatures

Globalement, la combinaison de l'utilisation de BlastP et des profils HMM donnent des résultats très satisfaisants pour ICEscreen puisque 98 % des SP de FirmiData sont détectées. Les quelques cas de non détection des SP sont dûs à trois raisons principales :

- Annotation manquante dans les fichiers Genbank de la banque RefSeq : ICEscreen identifie les SP à partir des séquences codantes annotées "CDS" dans le fichier Genbank. Si cette annotation est manquante, la séquence du CDS n'est pas extraite et donc la SP ne peut pas être trouvée.

- Pseudogènes : Parmi les séquences codantes d'un fichier Genbank, certains sont des pseudogènes, ce qui conduit à une protéine partielle. Celle-ci est le plus souvent éliminée à l'étape de filtrage des hits sur les critères de taille et de taux de couverture insuffisants.

- SP trop distantes des SP connues : pour les identifier une analyse experte des éléments est nécessaire.

Concernant les annotations manquantes ou erronées, la réannotation des gènes du génome avec un logiciel de prédiction de gène comme Prokka (Hyatt *et al.*, 2010; Seemann, 2014) permettrait l'obtention d'une annotation plus homogène que celles des fichiers Genbank.

2. Automatisation de l'annotation des éléments

L'outil ICEScreen permet de détecter 96 % des éléments de FirmiData et dans 92,9 % des cas caractérise correctement le type de l'élément c'est à dire différentie les ICE des IME (87,6 % des cas si les éléments remnants sont pris en compte). Lorsque le type de l'élément est mal défini (7 % des cas), cela est dû à la non détection d'une SP Outre la non détection des SP évoquée dans le paragraphe précédent, la non détection des ICE et des IME peut résulter de la mauvaise résolution de leur structure. Celle-ci est généralement due à la présence d'éléments de structure complexe c'est à dire d'éléments de structure composite. En effet, 48,3 % des éléments de FirmiData appartiennent à des structures composites composées d'au moins deux éléments (ICE, IME ou dérivés codant au moins une SP. Une autre difficulté est le fait que la moitié des structures composites sont « mixtes » (contiennent des ICE et des IME).

L'algorithme développé répond à cette difficulté et permet de résoudre les structures complexes automatiquement. Il s'appuie sur une recherche récursive par un algorithme « glouton » qui explore toutes les combinaisons possibles d'ICE et d'IME complets ou partiels pour identifier celles qui sont les plus vraisemblables selon les connaissances que nous avons de ces éléments. Sur les 18 structures composites mixtes du jeu FirmiData, ICEscreen en résout dix. Les structures non résolues correspondent le plus généralement à des structures très complexes incluant au moins quatre éléments (ICE et IME) et des éléments très dégénérés. Dans quelques rares cas, les structures non résolues correspondent à des emboîtements simples de deux éléments. Pour les structures non ou mal résolues, le recours à un expert est nécessaire pour leur résolution. L'outil ICEscreen a pour vocation de repérer ces structures complexes et peut donner une idée du nombre d'éléments.

Aucune des structures complexes de FirmiData n'est résolue correctement par les autres outils. C'est parce qu'il prend en compte les ICE et les IME et qu'il analyse les cas d'accrétions et d'emboîtement, que l'outil ICEscreen est plus performant que les autres outils pour la résolution de structures complexes.

3. Délimitation des éléments

ICEscreen permet actuellement de détecter et de typer des éléments de type ICE et IME, partiels ou complets, fonctionnels ou dégénérés mais ne réalise pas une délimitation précise des extrémités des éléments. La délimitation précise des éléments serait un atout à la fois pour caractériser les éléments simples mais aussi pour améliorer la résolution des structures complexes.

Pour réaliser cette tâche deux approches sont possibles : la recherche de répétitions directes qui bornent les éléments (approche de l'outil ICEfinder) ou l'utilisation de méthodes comparatives (par exemple l'approche utilisée par Cury *et al.* en complément de l'outil CONJscan).

Actuellement l'approche utilisée par ICEfinder ne recherche que les DR des éléments intégrés dans l'extrémité 3' des gènes codant des ARNt (Liu *et al.*, 2019). Chez les streptocoques (Ambroset *et al.*, 2016; Coluzzi *et al.*, 2017) et dans le jeu FirmiData, ceux-ci ne représentent qu'une petite minorité des éléments. Une difficulté de la recherche des DR est qu'il n'est pas toujours possible de trouver des répétitions directes notamment lorsqu'elles sont très courtes (par exemple 2 nt pour de nombreux ICE ou IME codant une intégrase à tyrosine), très dégénérées (par exemple pour de nombreux ICE ou IME intégrés dans le gène *rumA*), voire absentes (par exemple pour les ICE de famille Tn*916*). Localiser la position de ces DR est particulièrement difficile lorsque la spécificité d'intégration est inconnue et/ou lorsque les éléments sont de très grande taille.

Les méthodes comparatives, telles que celle proposée par (Cury *et al.*, 2020) utilisant l'outil CONJscan, permettent généralement de borner les éléments au gène près dans les cas les plus simples. En effet, elles exigent une comparaison du génome analysé avec des génomes de même espèce complètement séquencés et assemblés dont certains ne présentent aucun élément apparenté intégré au même site. Les défauts de cette approche sont :

1) de ne pas pouvoir délimiter les éléments de structures composites,
2) d'exiger un échantillon de souches proches mais au contenu différent en ICE et IME, ce qui est difficile à estimer à priori.

Nous envisageons à terme d'utiliser une méthode basée à la fois sur la détection de répétitions et l'analyse du degré de conservation des gènes pour proposer un nouveau module de délimitation des éléments dans ICEscreen.

Concernant la recherche des DR, nous disposons actuellement de toutes les informations concernant les intégrases, les gènes cibles et les DR des ICE et des IME de streptocoques (Ambroset *et al.*, 2016; Coluzzi *et al.*, 2017). Ces données seront utiles pour automatiser la délimitation de tous les éléments codant ces intégrases. L'analyse des génomes de FirmiData hors streptocoques, a permis de mettre en évidence des intégrases présentant de nouvelle spécificité d'intégration. Ces données seront incluses dans le module dédié à la délimitation des éléments. Cette procédure s'inscrit dans une démarche itérative basée sur l'acquisition de connaissances et leur intégration dans l'outil.

Concernant l'analyse du degré de conservation, des études préliminaires ont été effectuées avec Thomas Lacroix de l'unité MaIAGE sur le degré de conservation des protéines qui composent et bornent huit ICE et treize IME de streptocoques à quatre niveaux phylogénétique différents (au niveau de l'espèce, du genre *streptococcus*, du phylum Firmicutes et en dehors des Firmicutes) avec le logiciel Insyght (Lacroix *et al.*, 2016). Cette analyse a mis en évidence un fort degré de conservation des gènes cibles de l'élément qui pourrait aider à localiser les DR et/ou le site d'intégration.

4. Élargissement à l'ensemble des Firmicutes

Le premier objectif de ma thèse était d'automatiser l'annotation des éléments de streptocoques. Les résultats de ICEscreen sur les 26 génomes de streptocoques de FirmiData montrent que l'ensemble des ICE et la très grande majorité des IME des éléments sont retrouvés. Seul un IME appartenant à une structure composite complexe n'a pas été annoté par ICEscreen. Ainsi, nous pouvons considérer que l'objectif initial est atteint.

Le deuxième objectif était d'élargir l'annotation à l'ensemble des ICE et IME de Firmicutes. C'est dans cette perspective que 14 génomes d'autres Firmicutes ont été analysés. Les résultats les plus saillants sont :

- la présence d'éléments apparentés aux éléments de streptocoques y compris chez des bactéries éloignées de streptocoques. Notamment les ICE de la famille Tn*1549* sont très abondants chez les bactéries du tractus intestinal.

- l'identification de nouvelles familles d'ICE et d'IME, y compris chez des bactéries proches de streptocoques. Par exemple ICE*_LlaIO-1_tatD* de la superfamille Tn*916* présente chez *Lactococcus lactis* subsp. *lactis* IO-1.

- une nouvelle superfamille d'ICE, codant une relaxase de type MOB_C, est probablement présente chez *E. faecalis*. Une nouvelle superfamille d'IME codant une relaxase Rep_1 (PF01446) est présente chez *Lactobacillus paracasei* (IME*_Lcas919_rpsD*).

- des structures différentes en particulier des éléments composites de très grande complexité. Par exemple, le génome de *Lachnochlostridium phocaeense* Marseille-P3177 contient une structure composée d'un dérivé d'IME hébergeant un ICE qui contient lui-même trois IME dont l'un héberge un autre IME.

- de nouveaux sites d'intégration dont certains présentent des caractéristiques radicalement nouvelles par rapport aux éléments de streptocoques. Par exemple, un ICE de lactocoque est intégré dans le site *dif* de partition des chromosomes et code une intégrase à tyrosine apparentée à la résolvase reconnaissant ce site *dif*.

- des éléments mobiles de type RIT (Ricker *et al.*, 2013), ne codant aucun gène de conjugaison ou de mobilisation mais possédant un triplet d'intégrase à tyrosine, sont intégrés de manière site spécifique dans des gènes *ardC* (codant une protéine d'anti-restriction) de trois ICE de *Lachnochlostridium* sp. YL32.

Ce travail permet d'ores et déjà d'affirmer que l'outil ICEscreen permettra d'identifier non seulement des éléments proches de ceux des streptocoques mais aussi des ICE et IME très éloignés. Toutefois, les résultats obtenus sur le jeu FirmiData n'est probablement qu'un avant-goût de l'immense diversité et prévalence des ICE et des IME de Firmicutes, de la diversité de leur structure et de la diversité de leurs sites d'intégration.

5. Perspectives

Nos perspectives sont à ce jour de trois types :

- **implémentation d'un module de délimitation des éléments à l'outil ICEscreen.** Ce travail permettra de caractériser l'ensemble des gènes et donc des fonctions codées par ces éléments et notamment les fonctions d'adaptation. Il permettra également de faciliter la résolution des structures complexes.

- **généralisation de l'outil ICEscreen à la détection d'éléments éloignés de ceux de FirmiData.** Il faudra pour cela continuer le développement de ICEscreen :

    - Pour élargir la recherche à des éléments très éloignés de ceux déjà identifiés dans ce travail, il pourra être utile d'analyser les éléments partiels détectés par ICEscreen, voire les protéines signatures isolées. Cela permettra éventuellement d'identifier des SP manquantes que l'outil ne détecte pas automatiquement. Il faudra ensuite inclure ces nouvelles protéines dans l'outil, soit par ajout dans la banque BlastP soit par création de profil HMM si nécessaire.

    - La caractérisation manuelle de ces éléments permettra de définir leurs caractéristiques et d'identifier de nouvelles familles et superfamilles d'éléments. Ces informations pourront conduire à l'établissement de nouvelles règles qui seront incluses dans ICEscreen pour la détection automatique de ces nouvelles familles ou superfamilles.

  Notre ambition n'est pas de caractériser finement chacune des nouvelles familles/superfamilles d'éléments sur l'ensemble des Firmicutes mais plutôt de proposer une première annotation des éléments qui pourrait ensuite être améliorée par une analyse experte. Des collaborations pourraient être établies pour des éléments étudiées par d'autres équipes. Un bon exemple chez les Firmicutes sont les ICE de mycoplasmes (Citti *et al.*, 2020).

- **distribution de l'outil ICEscreen à l'ensemble de la communauté scientifique** via une interface conviviale type Galaxy et via un accès programmatique pour les bioinformaticiens. Ce travail est d'autant plus utile qu'actuellement les ICE et les IME ne sont pas ou peu annotés dans les génomes de Firmicutes. Bien que puissante, la méthode développée par Cury *et al.* n'est pas accessible via une interface. Le seul outil accessible pour les biologistes actuellement est l'outil ICEfinder qui est peu performant pour l'annotation des ICE et IME de Firmicutes.

Références bibliographiques

Abby, S.S., Cury, J., Guglielmini, J., Néron, B., Touchon, M., Rocha, E.P.C., 2016. Identification of protein secretion systems in bacterial genomes. Sci. Rep. 6, 23080. https://doi.org/10.1038/srep23080

Abby, S.S., Néron, B., Ménager, H., Touchon, M., Rocha, E.P.C., 2014. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. PLOS ONE 9, e110726. https://doi.org/10.1371/journal.pone.0110726

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. https://doi.org/10.1093/nar/25.17.3389

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

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

Bellanger, X., Morel, C., Gonot, F., Puymege, 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: Accretion of an ICE and a genomic island. Mol. Microbiol. 81, 912–925. https://doi.org/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. https://doi.org/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. https://doi.org/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, e800. https://doi.org/10.1371/journal.pone.0000800

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. https://doi.org/10.1093/nar/gkr846

Bjørkeng, E.K., Hjerde, E., Pedersen, T., Sundsfjord, A., Hegstad, K., 2013. ICESluvan, 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. https://doi.org/10.1128/JB.02165-12

Bordeleau, E., Ghinet, M.G., Burrus, V., 2012. Diversity of integrating conjugative elements in actinobacteria. Mob. Genet. Elem. 2, 119–124. https://doi.org/10.4161/mge.20498

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. https://doi.org/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. https://doi.org/10.1371/journal.pone.0023014

Burrus, V., Bontemps, C., Decaris, B., Guédon, G., 2001. Characterization of a novel type II restriction-modification system, Sth368I, encoded by the integrative element ICESt1 of Streptococcus thermophilus CNRZ368. Appl. Environ. Microbiol. 67, 1522–1528. https://doi.org/10.1128/AEM.67.4.1522-1528.2001

Burrus, V., Pavlovic, G., Decaris, B., Guédon, 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. https://doi.org/10.1016/s0147-619x(02)00102-6

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

Camilli, R., Bonnal, R.J.P., Del Grosso, M., Iacono, M., Corti, G., Rizzi, E., Marchetti, M., Mulas, L., Iannelli, F., Superti, F., Oggioni, M.R., De Bellis, G., Pantosti, A., 2011. Complete genome sequence of a serotype 11A, ST62 Streptococcus pneumoniae invasive isolate. BMC Microbiol. 11, 25. https://doi.org/10.1186/1471-2180-11-25

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. https://doi.org/10.1128/JB.00041-10

Citti, C., Baranowski, E., Dordet-Frisoni, E., Faucher, M., Nouvel, L.-X., 2020. Genomic Islands in Mycoplasmas. Genes 11, 836. https://doi.org/10.3390/genes11080836

Coluzzi, C., 2017. L'exploration des génomes par l'outil ICEFinder révèle la forte prévalence et l'extrême diversité des ICE et des IME de streptocoques (Theses). Université de Lorraine. https://tel.archives-ouvertes.fr/tel-01743816

Coluzzi, C., Guédon, G., Devignes, M.-D., Ambroset, C., Loux, V., Lacroix, T., Payot, S., 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. https://doi.org/10.3389/fmicb.2017.00443

Criscuolo, A., Gribaldo, S., 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 10, 210. https://doi.org/10.1186/1471-2148-10-210

Cury, J., Abby, S.S., Doppelt-Azeroual, O., Néron, B., Rocha, E.P.C., 2020. Identifying Conjugative Plasmids and Integrative Conjugative Elements with CONJscan, in: de la Cruz, F. (Ed.), Horizontal Gene Transfer: Methods and Protocols, Methods in Molecular Biology. Springer US, New York, NY, pp. 265–283. https://doi.org/10.1007/978-1-4939-9877-7_19

Cury, J., Touchon, M., Rocha, E.P.C., 2017. Integrative and conjugative elements and their hosts: composition, distribution and organization. Nucleic Acids Res. 45, 8943–8956. https://doi.org/10.1093/nar/gkx607

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, e00337-17, e00337-17. https://doi.org/10.1128/AEM.00337-17

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. https://doi.org/10.1111/j.1574-6976.2009.00195.x

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

Delorme, C., Poyart, C., Ehrlich, S.D., Renault, P., 2007. Extent of Horizontal Gene Transfer in Evolution of Streptococci of the Salivarius Group. J. Bacteriol. 189, 1330–1341. https://doi.org/10.1128/JB.01058-06

Eddy, S.R., 2011. Accelerated Profile HMM Searches. PLOS Comput. Biol. 7, e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L.L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S.C.E., Finn, R.D., 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47, D427–D432. https://doi.org/10.1093/nar/gky995

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

Franke, A.E., 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. https://doi.org/10.1128/jb.145.1.494-502.1981

Frisoni, E.D., Marenda, M.S., Sagné, E., Nouvel, L.X., Guérillot, R., Glaser, P., Blanchard, A., Tardy, F., Sirand-Pugnet, P., Baranowski, E., Citti, C., 2013. ICEA of Mycoplasma agalactiae: a new family of self-transmissible integrative elements that confers conjugative properties to the recipient strain. Mol. Microbiol. 89, 1226–1239. https://doi.org/10.1111/mmi.12341

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. https://doi.org/10.1038/nrmicro1235

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinforma. Oxf. Engl. 28, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., Koonin, E.V., 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 43, D261-269. https://doi.org/10.1093/nar/gku1223

Garcillán-Barcia, M.P., Alvarado, A., de la Cruz, F., 2011. Identification of bacterial plasmids based on mobility and plasmid population biology. FEMS Microbiol. Rev. 35, 936–956. https://doi.org/10.1111/j.1574-6976.2011.00291.x

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. https://doi.org/10.1111/j.1574-6976.2009.00168.x

Garcillán-Barcia, M.P., Redondo-Salvo, S., Vielva, L., de la Cruz, F., 2020. MOBscan: Automated Annotation of MOB Relaxases, in: de la Cruz, F. (Ed.), Horizontal Gene Transfer: Methods and Protocols, Methods in Molecular Biology. Springer US, New York, NY, pp. 295–308. https://doi.org/10.1007/978-1-4939-9877-7_21

Ghinet, M.G., Bordeleau, E., Beaudin, J., Brzezinski, R., Roy, S., Burrus, V., 2011. Uncovering the Prevalence and Diversity of Integrating Conjugative Elements in Actinobacteria. PLOS ONE 6, e27846. https://doi.org/10.1371/journal.pone.0027846

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol. Biol. Evol. 27, 221–224. https://doi.org/10.1093/molbev/msp259

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

Guédon, G., Libante, V., Coluzzi, C., Payot, S., Leblond-Bourget, N., 2017. The Obscure World of Integrative and Mobilizable Elements, Highly Widespread Elements that Pirate Bacterial Conjugative Systems. Genes 8, 337. https://doi.org/10.3390/genes8110337

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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1371/journal.pgen.1002222

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

Haft, D.H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F., Derbyshire, M.K., Gonzales, N.R., Gwadz, M., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Zheng, C., Thibaud-Nissen, F., Geer, L.Y., Marchler-Bauer, A., Pruitt, K.D., 2018. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res. 46, D851–D860. https://doi.org/10.1093/nar/gkx1068

Haren, L., Ton-Hoang, B., Chandler, M., 1999. Integrating DNA: Transposases and Retroviral Integrases. Annu. Rev. Microbiol. 53, 245–281. https://doi.org/10.1146/annurev.micro.53.1.245

Heuer, H., Smalla, K., 2012. Plasmids foster diversification and adaptation of bacterial populations in soil. FEMS Microbiol. Rev. 36, 1083–1104. https://doi.org/10.1111/j.1574-6976.2012.00337.x

Holden, M.T.G., Hauser, H., Sanders, M., Ngo, T.H., Cherevach, I., Cronin, A., Goodhead, I., Mungall, K., Quail, M.A., Price, C., Rabbinowitsch, E., Sharp, S., Croucher, N.J., Chieu, T.B., Mai, N.T.H., Diep, T.S., Chinh, N.T., Kehoe, M., Leigh, J.A., Ward, P.N., Dowson, C.G., Whatmore, A.M., Chanter, N., Iversen, P., Gottschalk, M., Slater, J.D., Smith, H.E., Spratt, B.G., Xu, J., Ye, C., Bentley, S., Barrell, B.G., Schultsz, C., Maskell, D.J., Parkhill, J., 2009. Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen Streptococcus suis. PloS One 4, e6072. https://doi.org/10.1371/journal.pone.0006072

Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinforma. Oxf. Engl. 26, 680–682. https://doi.org/10.1093/bioinformatics/btq003

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. https://doi.org/10.1186/1471-2105-11-119

Iannelli, F., Santoro, F., Oggioni, M.R., Pozzi, G., 2014. Nucleotide sequence analysis of integrative conjugative element Tn5253 of Streptococcus pneumoniae. Antimicrob. Agents Chemother. 58, 1235–1239. https://doi.org/10.1128/AAC.01764-13

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. https://doi.org/10.1111/j.1574-6976.2008.00136.x

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066. https://doi.org/10.1093/nar/gkf436

Köster, J., Rahmann, S., 2012. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28, 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kuenne, C., Billion, A., Mraheil, M.A., Strittmatter, A., Daniel, R., Goesmann, A., Barbuddhe, S., Hain, T., Chakraborty, T., 2013. Reassessment of the Listeria monocytogenes pan-genome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome. BMC Genomics 14, 47. https://doi.org/10.1186/1471-2164-14-47

Lacroix, T., Thérond, S., Rugeri, M., Nicolas, P., Gendrault, A., Loux, V., Gibrat, J.-F., 2016. Synchronized navigation and comparative analyses across Ensembl complete bacterial genomes with INSYGHT. Bioinformatics 32, 1083–1084. https://doi.org/10.1093/bioinformatics/btv689

Lang, A.S., Beatty, J.T., 2007. Importance of widespread gene transfer agent genes in alpha-proteobacteria. Trends Microbiol. 15, 54–62. https://doi.org/10.1016/j.tim.2006.12.001

Lao, J., Guédon, G., Lacroix, T., Charron-Bourgoin, F., Libante, V., Loux, V., Chiapello, H., Payot, S., Leblond-Bourget, N., 2020. Abundance, Diversity and Role of ICEs and IMEs in the Adaptation of Streptococcus salivarius to the Environment. Genes 11, 999. https://doi.org/10.3390/genes11090999

Laslett, D., Canback, B., 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32, 11–16. https://doi.org/10.1093/nar/gkh152

Lee, C.A., Thomas, J., Grossman, A.D., 2012. The Bacillus subtilis Conjugative Transposon ICEBs1 Mobilizes Plasmids Lacking Dedicated Mobilization Functions. J. Bacteriol. 194, 3165–3172. https://doi.org/10.1128/JB.00301-12

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinforma. Oxf. Engl. 22, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Li, X., Xie, Y., Liu, M., Tai, C., Sun, J., Deng, Z., Ou, H.-Y., 2018. oriTfinder: a web-based tool for the identification of origin of transfers in DNA sequences of bacterial mobile genetic elements. Nucleic Acids Res. 46, W229–W234. https://doi.org/10.1093/nar/gky352

Libante, V., Nombre, Y., Coluzzi, C., Staub, J., Guédon, G., Gottschalk, M., Teatero, S., Fittipaldi, N., Leblond-Bourget, N., Payot, S., 2020. Chromosomal Conjugative and Mobilizable Elements in Streptococcus suis: Major Actors in the Spreading of Antimicrobial Resistance and Bacteriocin Synthesis Genes. Pathogens 9, 22. https://doi.org/10.3390/pathogens9010022

Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z., Ou, H.-Y., 2019. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. Nucleic Acids Res. 47, D660–D665. https://doi.org/10.1093/nar/gky1123

Lorenz, M.G., Wackernagel, W., 1994. Bacterial gene transfer by natural genetic transformation in the environment. Microbiol. Rev. 58, 563–602. Sur https://mmbr.asm.org/content/mmbr/58/3/563.full.pdf

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Geer, L.Y., Bryant, S.H., 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45, D200–D203. https://doi.org/10.1093/nar/gkw1129

Marchler-Bauer, A., Bryant, S.H., 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 32, W327–W331. https://doi.org/10.1093/nar/gkh454

Mell, J.C., Redfield, R.J., 2014. Natural competence and the evolution of DNA uptake specificity. J. Bacteriol. 196, 1471–1483. https://doi.org/10.1128/JB.01293-13

Mendes Oliveira, V.R., Paiva, M.C., Lima, W.G., 2019. Plasmid-mediated colistin resistance in Latin America and Caribbean: A systematic review. Travel Med. Infect. Dis. 31, 101459. https://doi.org/10.1016/j.tmaid.2019.07.015

Miele, V., Penel, S., Duret, L., 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12, 116. https://doi.org/10.1186/1471-2105-12-116

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

Nang, S.C., Li, J., Velkov, T., 2019. The rise and spread of *mcr* plasmid-mediated polymyxin resistance. Crit. Rev. Microbiol. 45, 131–161. https://doi.org/10.1080/1040841X.2018.1492902

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinforma. Oxf. Engl. 31, 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Pavlovic, G., Burrus, V., Gintz, B., Decaris, B., Guédon, 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. https://doi.org/10.1099/mic.0.26883-0

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, Shaochuan, Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, Shengting, Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, Songgang, Qin, N., Yang, H., Wang, Jian, Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S.D., Wang, Jun, 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59–65. https://doi.org/10.1038/nature08821

Ramsay, J.P., Firth, N., 2017. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. Curr. Opin. Microbiol., Mobile genetic elements and HGT in prokaryotes * Microbiota 38, 1–9. https://doi.org/10.1016/j.mib.2017.03.003

Rauch, P.J., Vos, W.M.D., 1992. Characterization of the novel nisin-sucrose conjugative transposon Tn5276 and its insertion in Lactococcus lactis. J. Bacteriol. 174, 1280–1287. https://doi.org/10.1128/jb.174.4.1280-1287.1992

Redfield, R.J., Schrag, M.R., Dean, A.M., 1997. The evolution of bacterial transformation: sex with poor relations. Genetics 146, 27–38. Sur https://www.genetics.org/content/genetics/146/1/27.full.pdf

Ricker, N., Qian, H., Fulthorpe, R.R., 2013. Phylogeny and organization of recombinase in trio (RIT) elements. Plasmid 70, 226–239. https://doi.org/10.1016/j.plasmid.2013.04.003

Roberts, A.P., Mullany, P., 2009. A modular master on the move: the Tn916 family of mobile genetic elements. Trends Microbiol. 17, 251–258. https://doi.org/10.1016/j.tim.2009.03.002

Rozwandowicz, M., Brouwer, M.S.M., Fischer, J., Wagenaar, J.A., Gonzalez-Zorn, B., Guerra, B., Mevius, D.J., Hordijk, J., 2018. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. J. Antimicrob. Chemother. 73, 1121–1137. https://doi.org/10.1093/jac/dkx488

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Shimizu-Kadota, M., Kato, H., Shiwa, Y., Oshima, K., Machii, M., Araya-Kojima, T., Zendo, T., Hattori, M., Sonomoto, K., Yoshikawa, H., 2013. Genomic Features of Lactococcus lactis IO-1, a Lactic Acid Bacterium That Utilizes Xylose and Produces High Levels of L-Lactic Acid. Biosci. Biotechnol. Biochem. 77, 1804–1808. https://doi.org/10.1271/bbb.130080

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539. https://doi.org/10.1038/msb.2011.75

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. https://doi.org/10.1128/microbiolspec.MDNA3-0030-2014

Smalla, K., Jechalke, S., Top, E.M., 2015. Plasmid Detection, Characterization, and Ecology. Microbiol. Spectr. 3. https://doi.org/10.1128/microbiolspec.PLAS-0038-2014

Smillie, C., Garcillán-Barcia, M.P., Francia, M.V., Rocha, E.P.C., Cruz, F. de la, 2010. Mobility of Plasmids. Microbiol. Mol. Biol. Rev. 74, 434–452. https://doi.org/10.1128/MMBR.00020-10

Smyth, D.S., Robinson, D.A., 2009. Integrative and sequence characteristics of a novel genetic element, ICE6013, in Staphylococcus aureus. J. Bacteriol. 191, 5964–5975. https://doi.org/10.1128/JB.00352-09

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. https://doi.org/10.1073/pnas.92.19.8985

Sullivan, J.T., Ronson, C.W., 1998. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. Proc. Natl. Acad. Sci. U. S. A. 95, 5145–5149. https://doi.org/10.1073/pnas.95.9.5145

Szöllosi, G.J., Derényi, I., Vellai, T., 2006. The maintenance of sex in bacteria is ensured by its potential to reload genes. Genetics 174, 2173–2180. https://doi.org/10.1534/genetics.106.063412

Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A genomic perspective on protein families. Science 278, 631–637. https://doi.org/10.1126/science.278.5338.631

te Poele, E.M., Bolhuis, H., Dijkhuizen, L., 2008. Actinomycete integrative and conjugative elements. Antonie Van Leeuwenhoek 94, 127–143. https://doi.org/10.1007/s10482-008-9255-x

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

Treangen, T.J., Rocha, E.P.C., 2011. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLOS Genet. 7, e1001284. https://doi.org/10.1371/journal.pgen.1001284

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. https://doi.org/10.1128/jb.182.23.6577-6583.2000

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

Annexes

Tableau 1 : Génomes du jeu FirmiData.

N	Organisme	N° accession	Référence des annotations
	Génomes sans éléments en accrétions ni éléments en emboîtements		
1	*Streptococcus dysgalactiae* subsp. *equisimilis* D166B, ATCC 12394	NC_017567.1	(Suzuki *et al.*, 2011) ; (Coluzzi, 2017)
2	*Streptococcus equi* subsp. *zooepidemicus* MGCS10565	NC_011134.1	(Coluzzi, 2017)
3	*Streptococcus agalactiae* GD201008-001	NC_018646.1	(Coluzzi, 2017)
4	*Streptococcus salivarius* JF	NZ_CP014144.1	Données non publiées
5	*Streptococcus salivarius* FDAARGOS_259	NZ_CP020451.2	(Lao *et al.*, 2020)
6	*Streptococcus pneumoniae* SPN034183	NC_021028.1	(Coluzzi, 2017)
7	*Streptococcus anginosus* C1051	NC_022244.1	(Coluzzi, 2017)
8	*Streptococcus suis* ST1	NC_017950.1	(Coluzzi, 2017)
9	*Lactobacillus paracasei* LOCK919	NC_021721.1	(Coluzzi, 2017)
10	*Listeria monocytogenes* sv 4e SLCC 2378	NC_018585.1	(Kuenne *et al.*, 2013)
11	*Clostridioides difficile* 630	NC_009089.1	(Mullany *et al.*, 1990) ; (Brouwer *et al*., 2011)
	Génomes avec éléments en accrétions ou éléments en emboîtements		
12	*Streptococcus dysgalactiae* subsp. *equisimilis* RE378	NC_018712.1	(Coluzzi, 2017)
13	*Streptococcus pyogenes* HKU QMH11M0907901	NZ_AFRY01000001.1	Données non publiées
14	*Streptococcus pyogenes* sv. M12 MGAS2096	NC_008023.1	(Beres and Musser, 2007) ; (Coluzzi, 2017)
15	*Streptococcus equi* subsp. *zooepidemicus* ATCC 35246	NC_017582.1	(Guérillot *et al.*, 2013) ; (Ma *et al.*, 2013) ; (Coluzzi, 2017)
16	*Streptococcus agalactiae* 09mas018883	NC_021485.1	(Coluzzi, 2017)

Tableau 1 (suite)

N	Organisme	N° accession	Référence des annotations
17	*Streptococcus agalactiae* sv. III/III NEM316	NC_004368.1	(Glaser *et al*., 2002) ; (Brochet *et al.*, 2008) ; (Brochet *et al*., 2009) ; (Coluzzi, 2017)
18	*Streptococcus gallolyticus* subsp. *gallolyticus* ATCC BAA-2069	NC_015215.1	(Guérillot *et al*., 2013) ; (Coluzzi, 2017)
19	*Streptococcus salivarius* ATCC 25975	NZ_CP015283.1	(Lao *et al*., 2020)
20	*Streptococcus thermophilus* JIM 8232	NC_017581.1	(Coluzzi, 2017)
21	*Streptococcus pneumoniae* P1031	NC_012467.1	(Coluzzi, 2017)
22	*Streptococcus parasanguinis* ATCC 15912	NC_015678.1	(Coluzzi, 2017)
23	*Streptococcus constellatus* subsp. *pharyngis* C1050	NC_022238.1	(Olson *et al*., 2013) ; (Coluzzi, 2017)
24	*Streptococcus constellatus* subsp. *pharyngis* C232	NC_022236.1	(Olson *et al.*, 2013) ; Données non publiées
25	*Streptococcus suis* BM407	NC_012926.1	(Holden *et al.*, 2009) ; (Coluzzi, 2017)
26	*Streptococcus suis* SC84	NC_012924.1	(Holden *et al*., 2009) ; (Coluzzi, 2017)
27	*Streptococcus suis* sv. SS2 05ZYH33	NC_009442.1	(Li *et al.*, 2011) ; (Coluzzi, 2017)
28	*Streptococcus suis* NSUI002	NZ_CP011419.1	(Libante *et al*., 2020)
29	*Streptococcus suis* T15	NC_022665.1	(Coluzzi, 2017)
30	*Lactococcus lactis* subsp. *lactis* IO-1	NC_020450.1	(Shimizu-Kadota *et al*., 2013) ; G. Guédon
31	*Enterococcus faecalis* V583	NC_004668.1	(Burrus *et al.*, 2002) ; Données non publiées
32	*Staphylococcus pseudintermedius* HKU10-03	NC_014925.1	Données non publiées
33	*Staphylococcus epidermidis* ATCC 12228	NC_004461.1	Données non publiées

Tableau 1 (suite)

N	Organisme	N° accession	Référence des annotations
34	*Lachnoclostridium phocaeense* Marseille-P3177	NZ_LT635479.1	Données non publiées
35	*Roseburia hominis* A2-183	NC_015977.1	Données non publiées
36	*Clostridioides difficile* QCD-63q42	NZ_CM000637.1	(Brouwer *et al.*, 2011) ; Données non publiées
37	*Clostridium difficile* R20291	NC_013316.1	(Brouwer *et al.*, 2011)
38	*Dehalobacterium formicoaceticum* DMC	NZ_CP022121.1	Données non publiées
	Génomes sans élément annotés		
39	*Lachnoclostridium* sp. YL32	NZ_CP015399.2	-
40	*Faecalibacterium prausnitzii* A2-165	NZ_CP022479.1	-

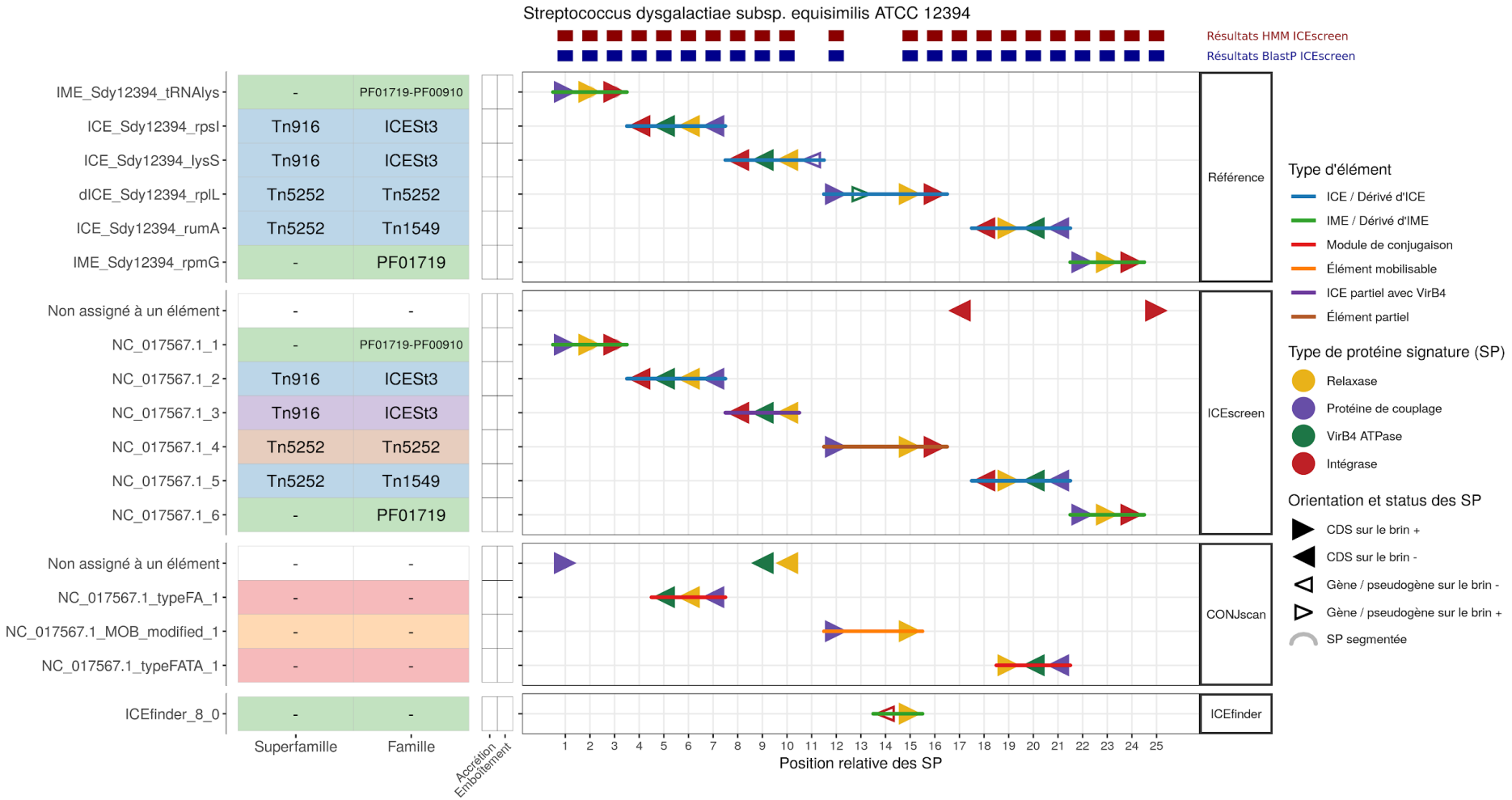

Figure 1 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus dysgalactiae* subsp. *equisimilis* D166B, ATCC 12394 (NC_017567.1).

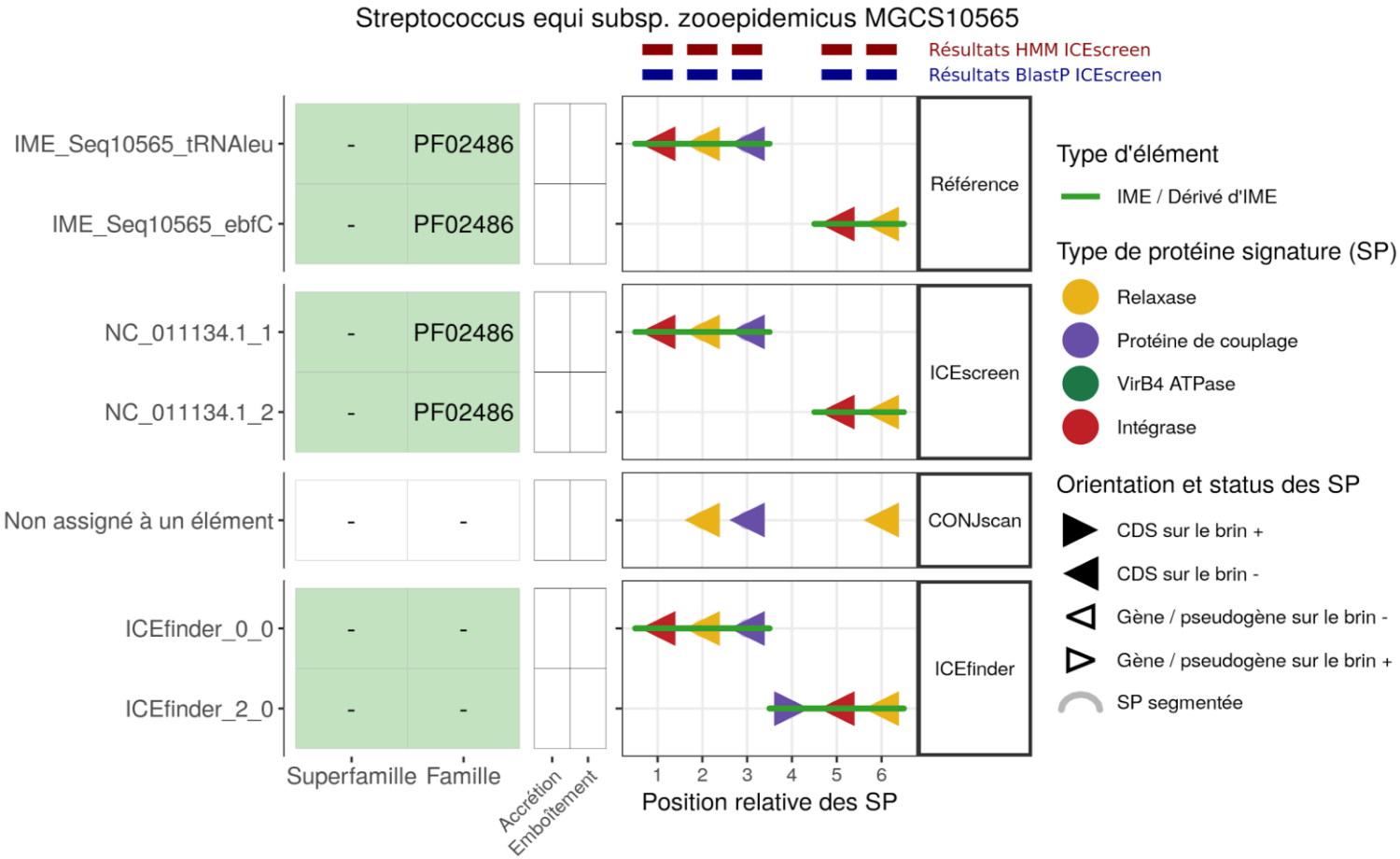

Figure 2 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus equi* subsp. *zooepidemicus* MGCS10565 (NC_011134.1).

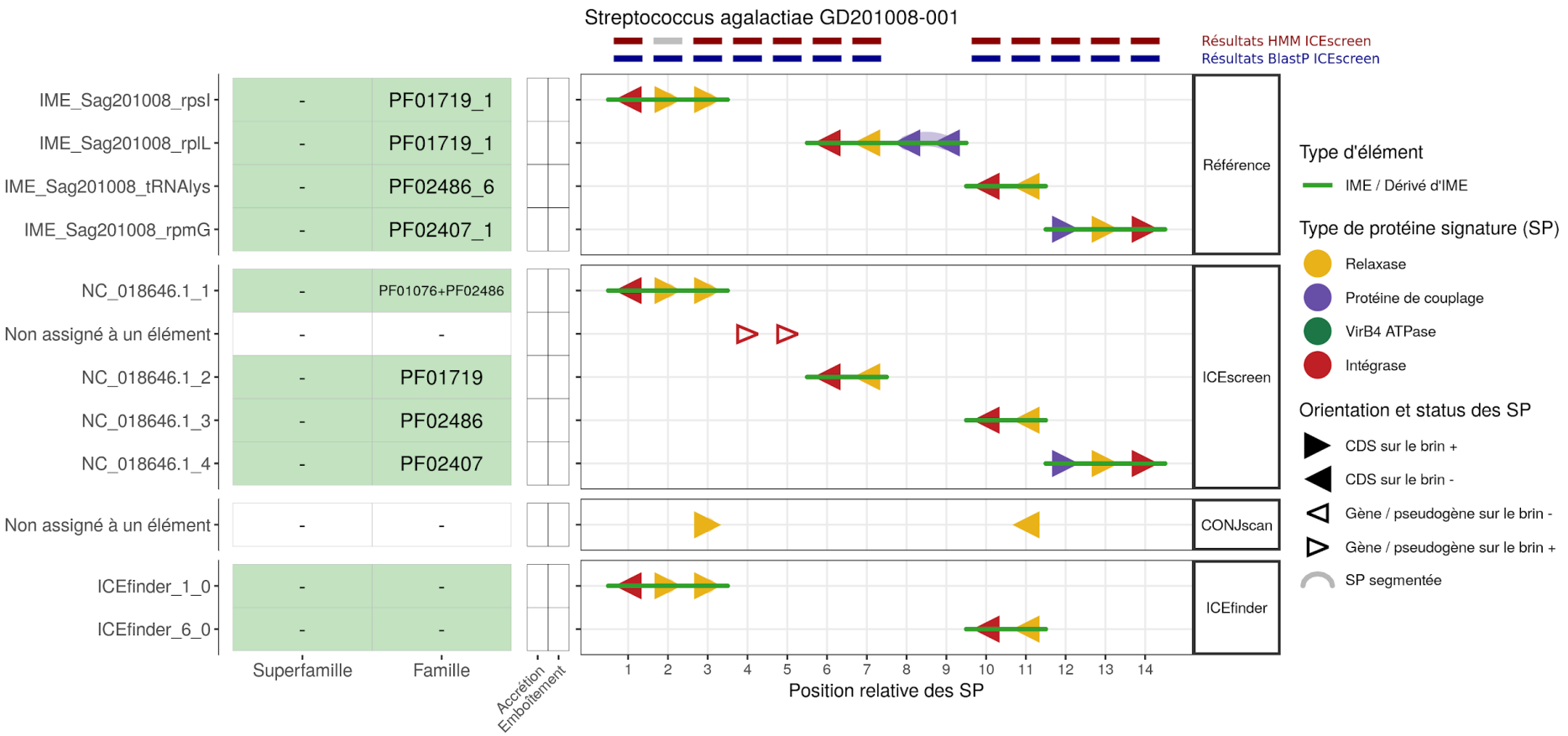

Figure 3 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus agalactiae* GD201008-001 (NC_018646.1).

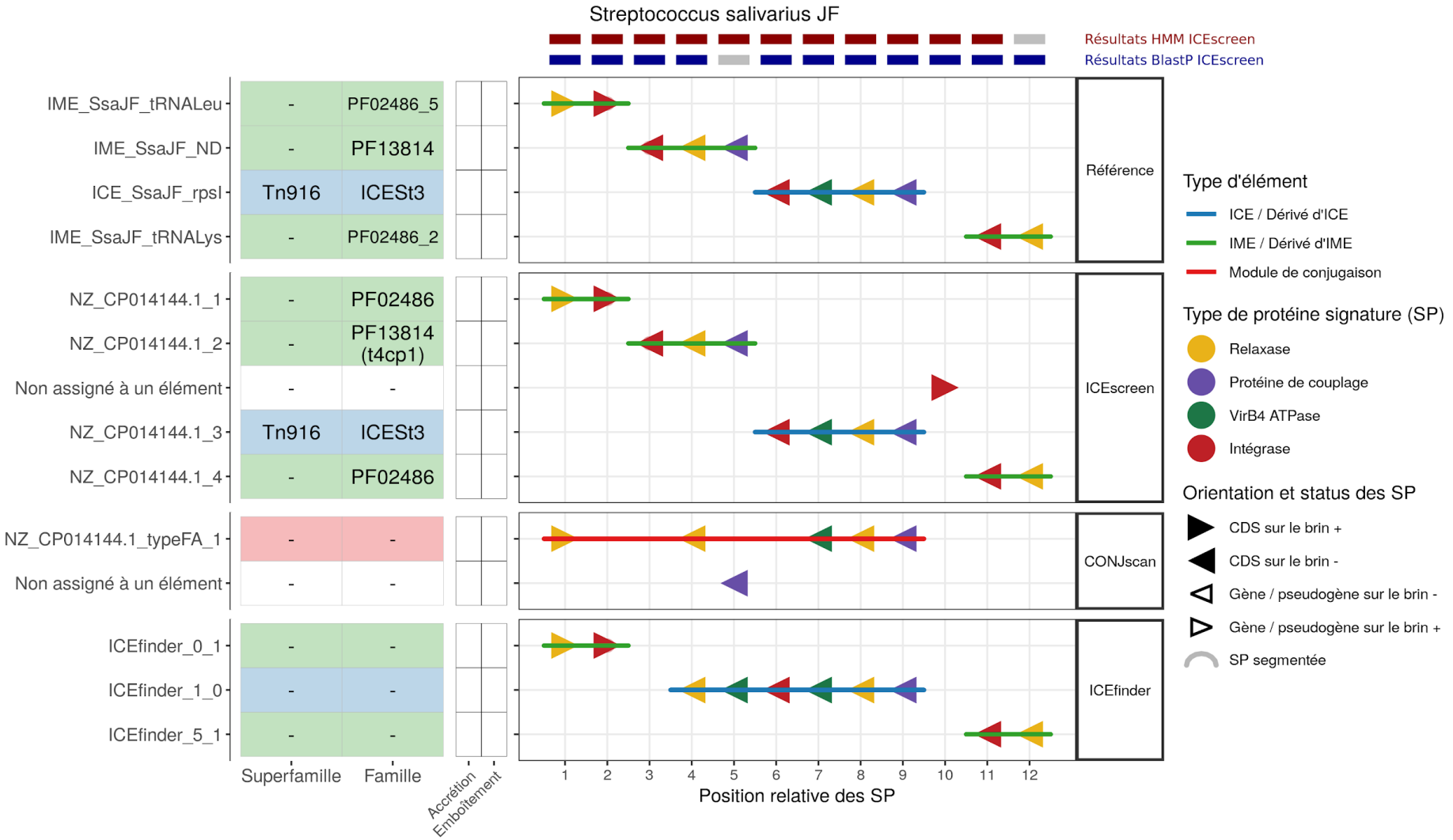

Figure 4 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus salivarius* JF (NZ_CP014144.1).

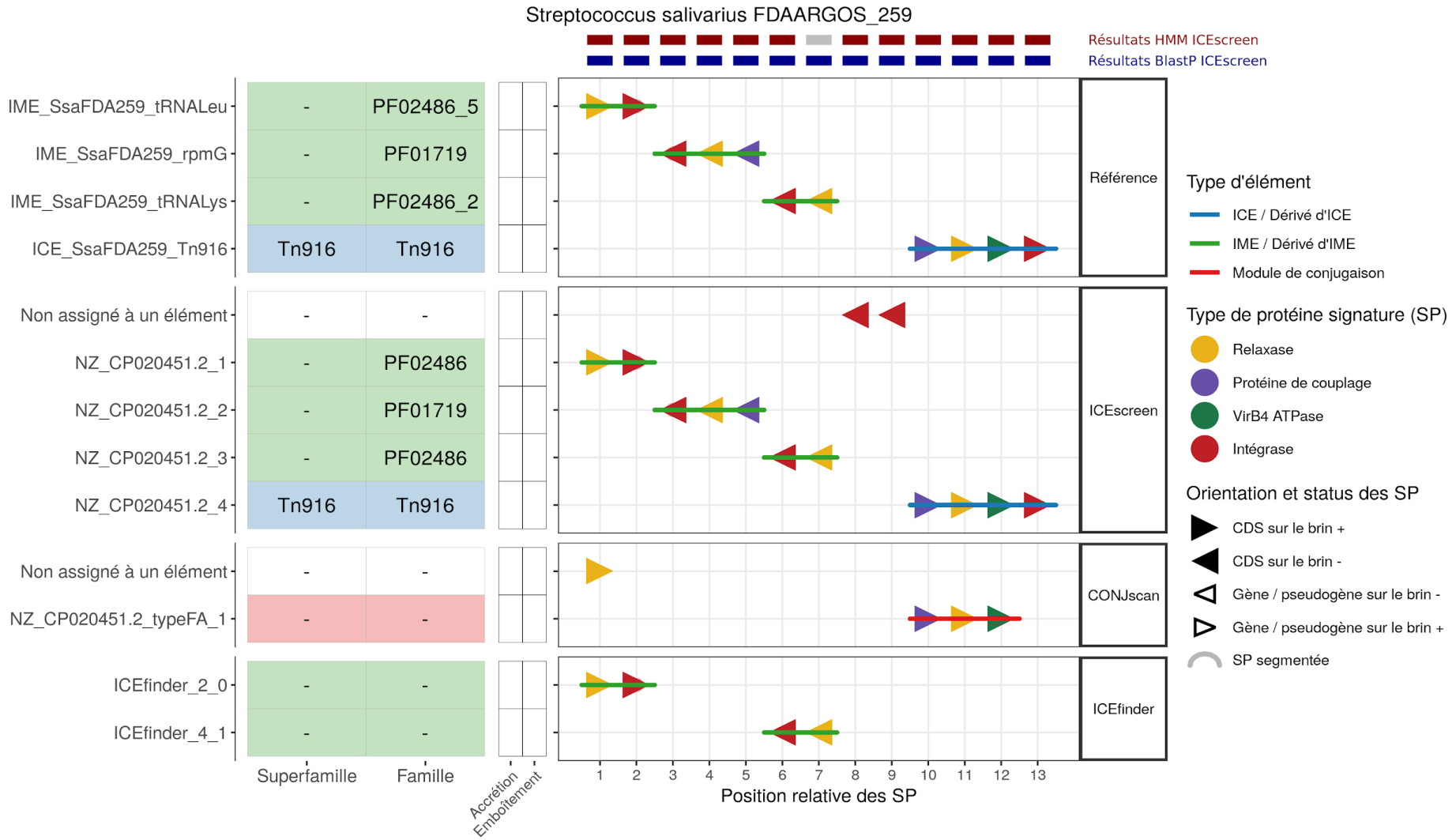

Figure 5 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus salivarius* FDAARGOS_259 (NZ_CP020451.2).

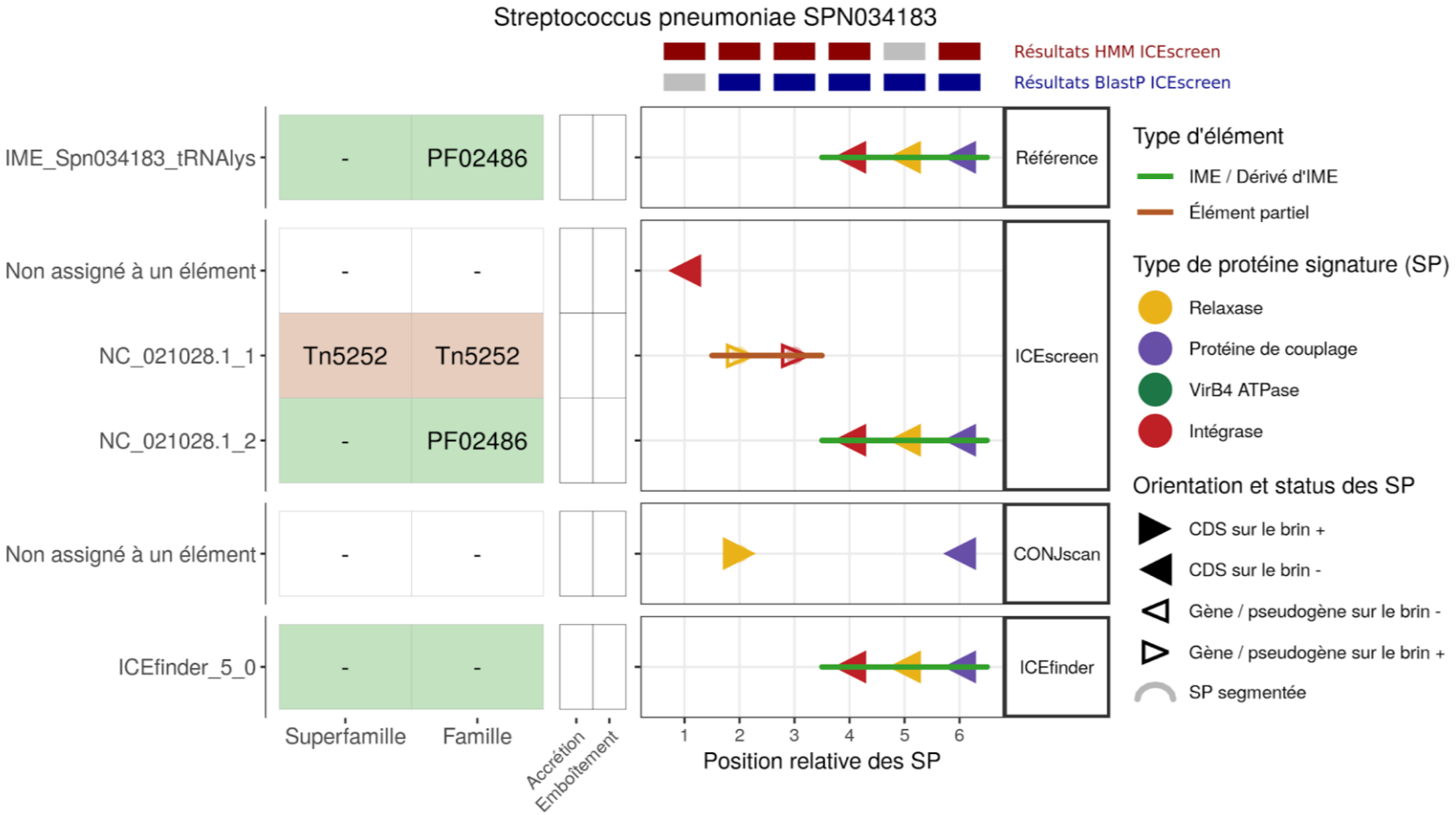

Figure 6 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus pneumoniae* SPN034183 (NC_021028.1).

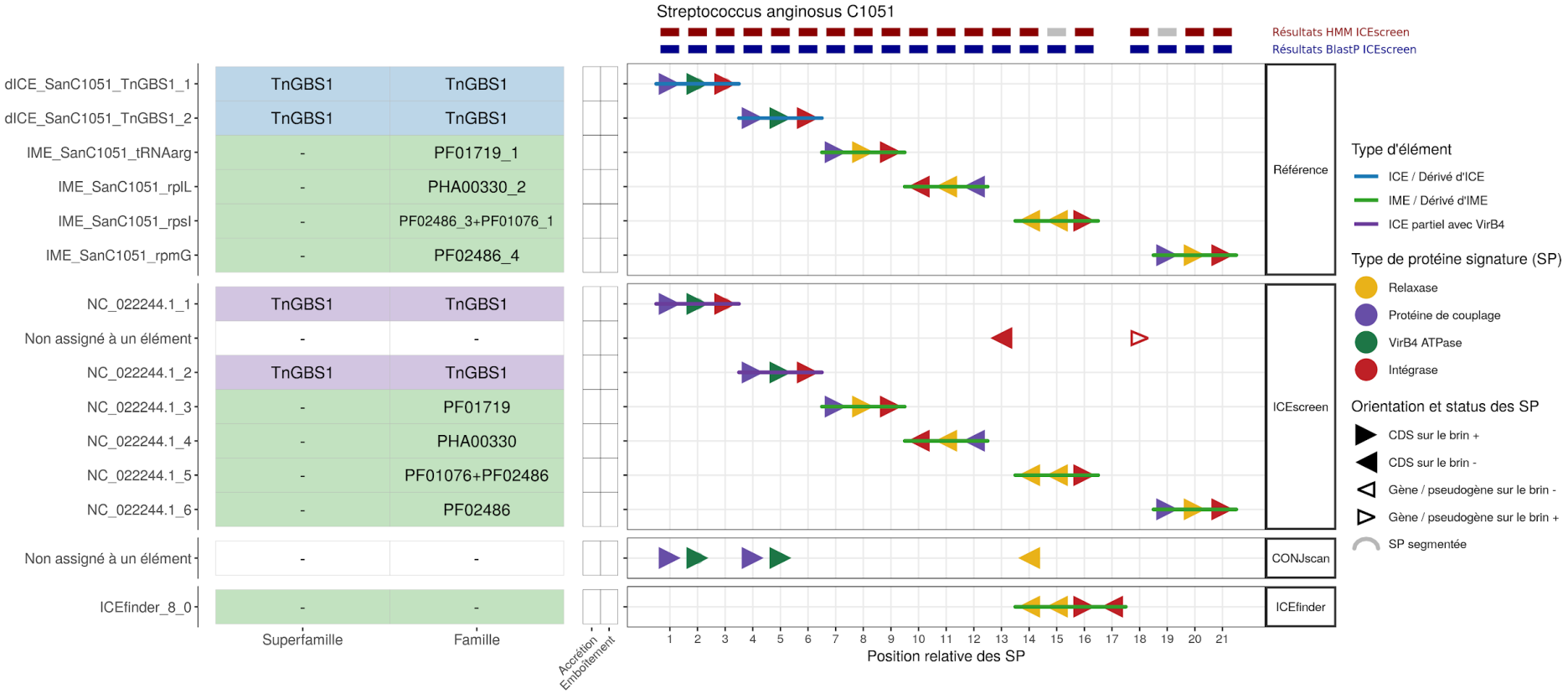

Figure 7 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus anginosus* C1051 (NC_022244.1).

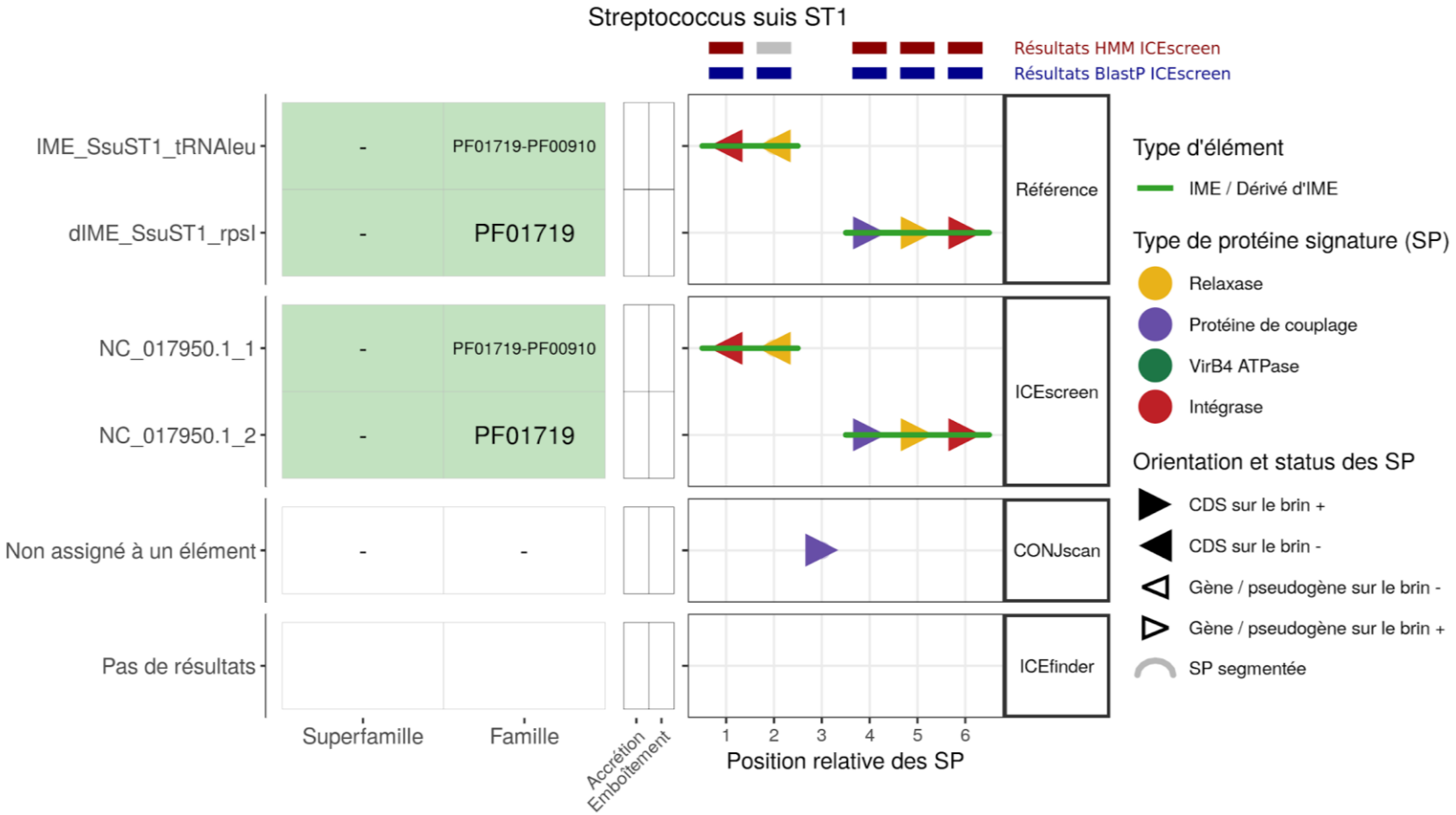

Figure 8 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus suis* ST1 (NC_017950.1).

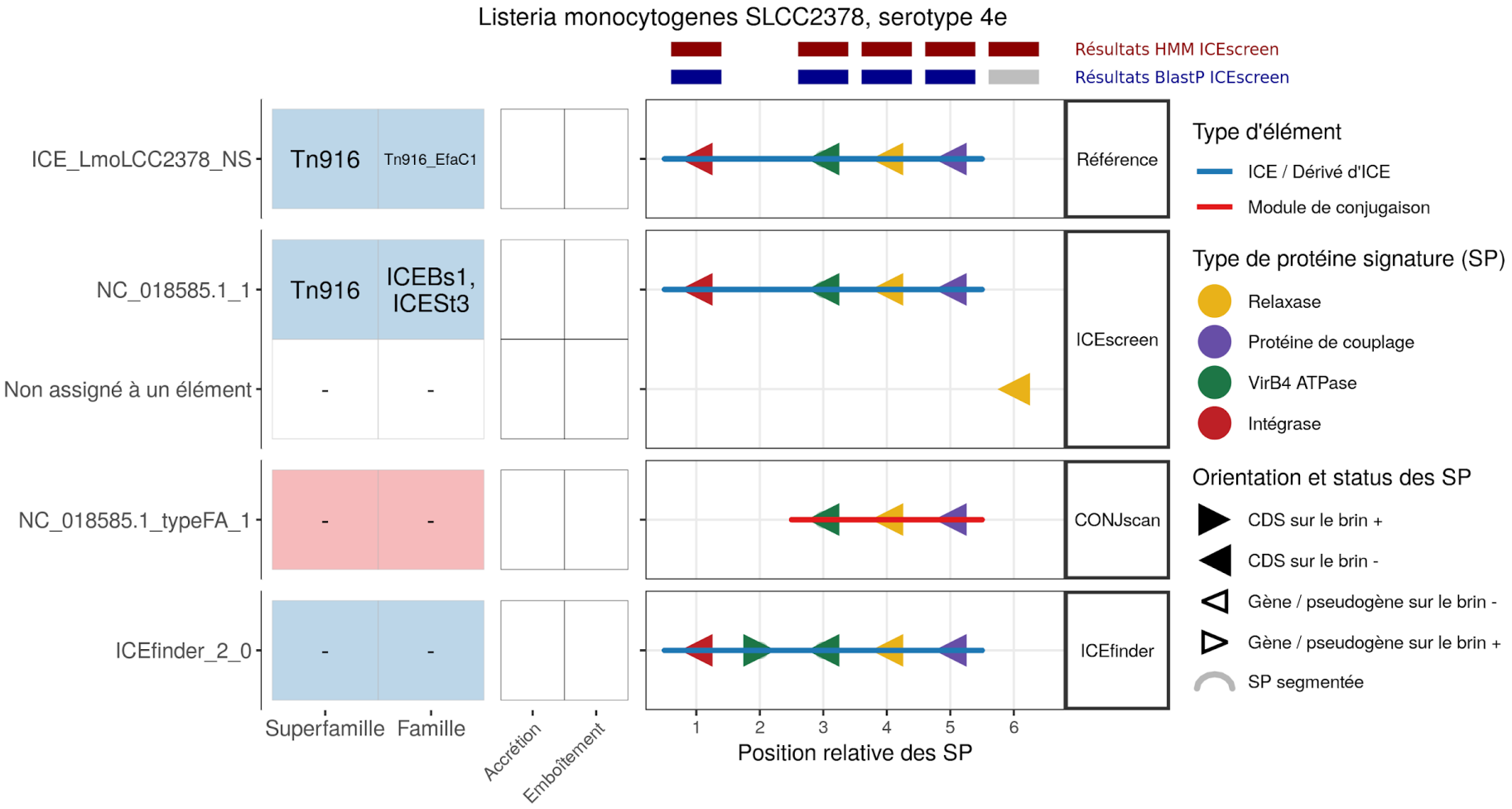

Figure 9 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Listeria monocytogenes* sv. 4e SLCC 2378 (NC_018585.1). ICE_*LmoLCC2378_NS* correspond en partie à l'îlot génomique annoté dans (Kuenne *et al.,* 2013) sous le nom de « LGI2 ».

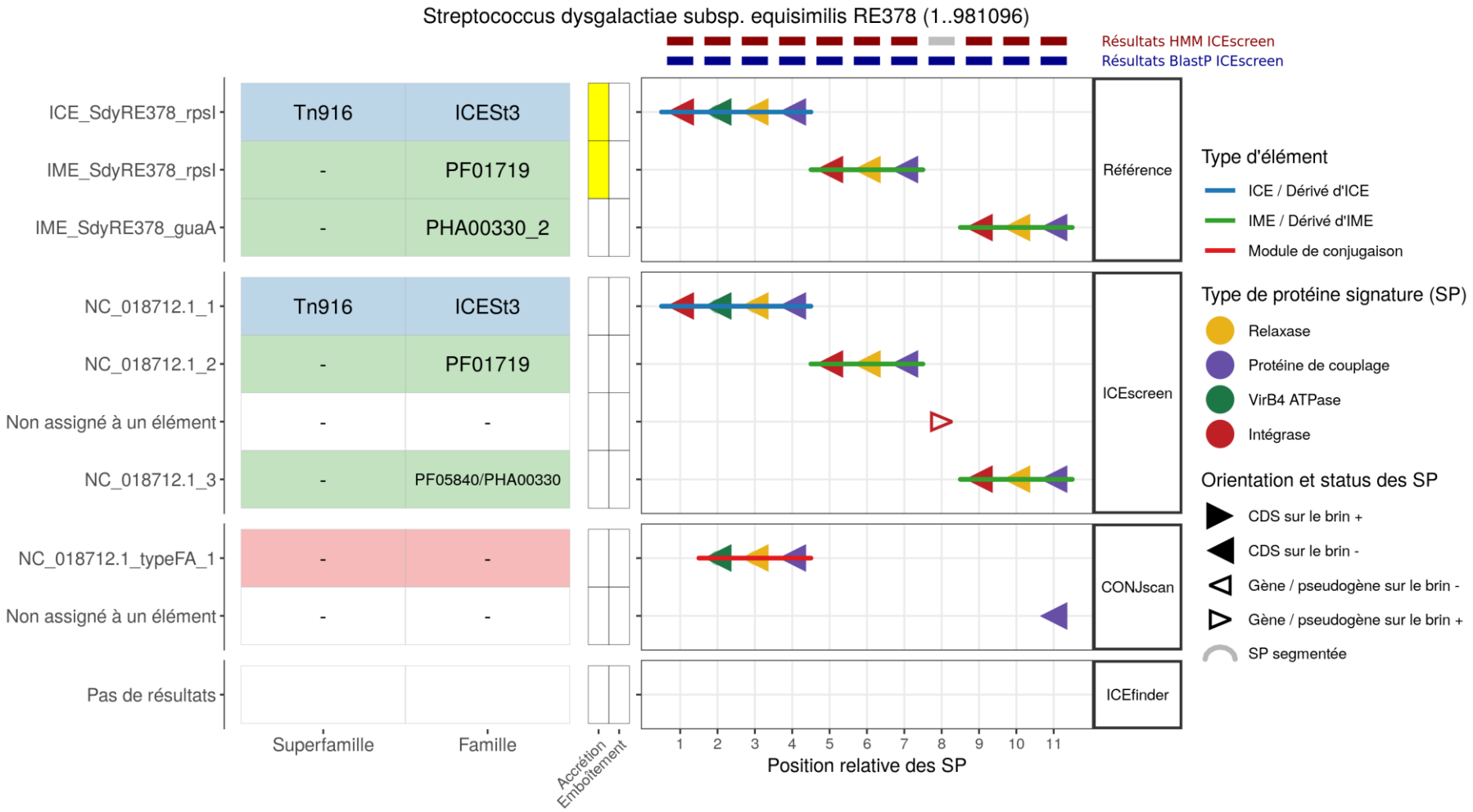

Figure 10a : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus dysgalactiae* subsp. *equisimilis* RE378 (positions 1 à 981096) (NC_018712.1).

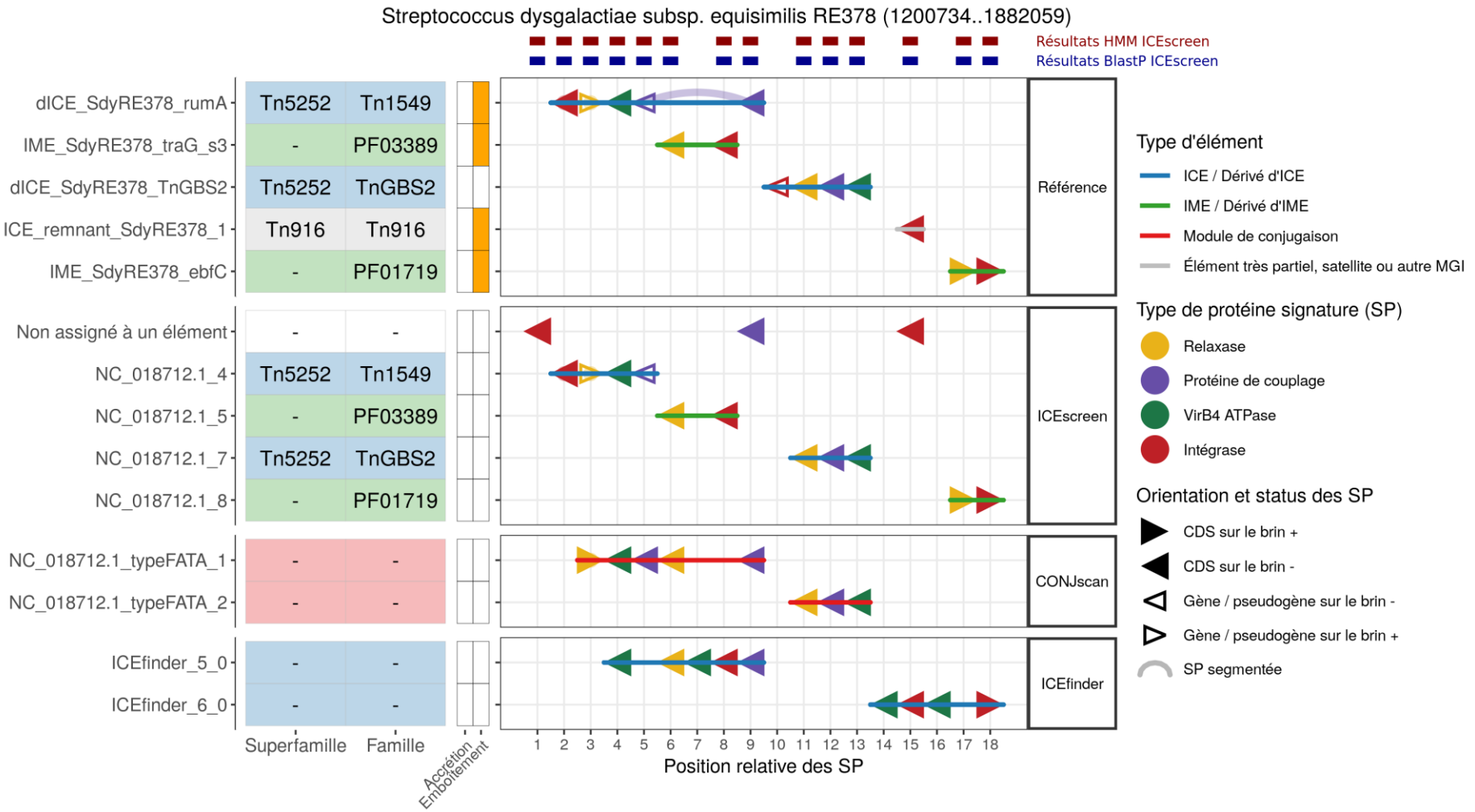

Figure 10b : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus dysgalactiae* subsp. *equisimilis* RE378 (positions 1200734 à 1882059) (NC_018712.1).

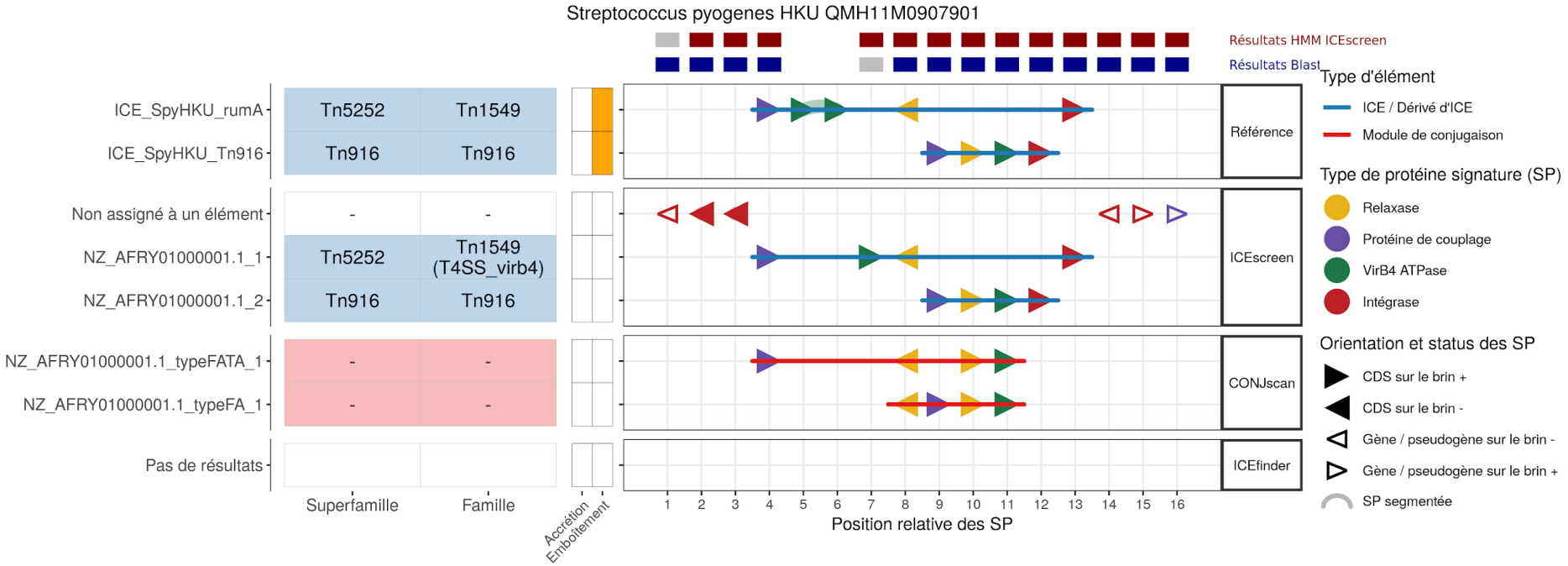

Figure 11 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus pyogenes* HKU QMH11M0907901 (NZ_AFRY01000001.1).

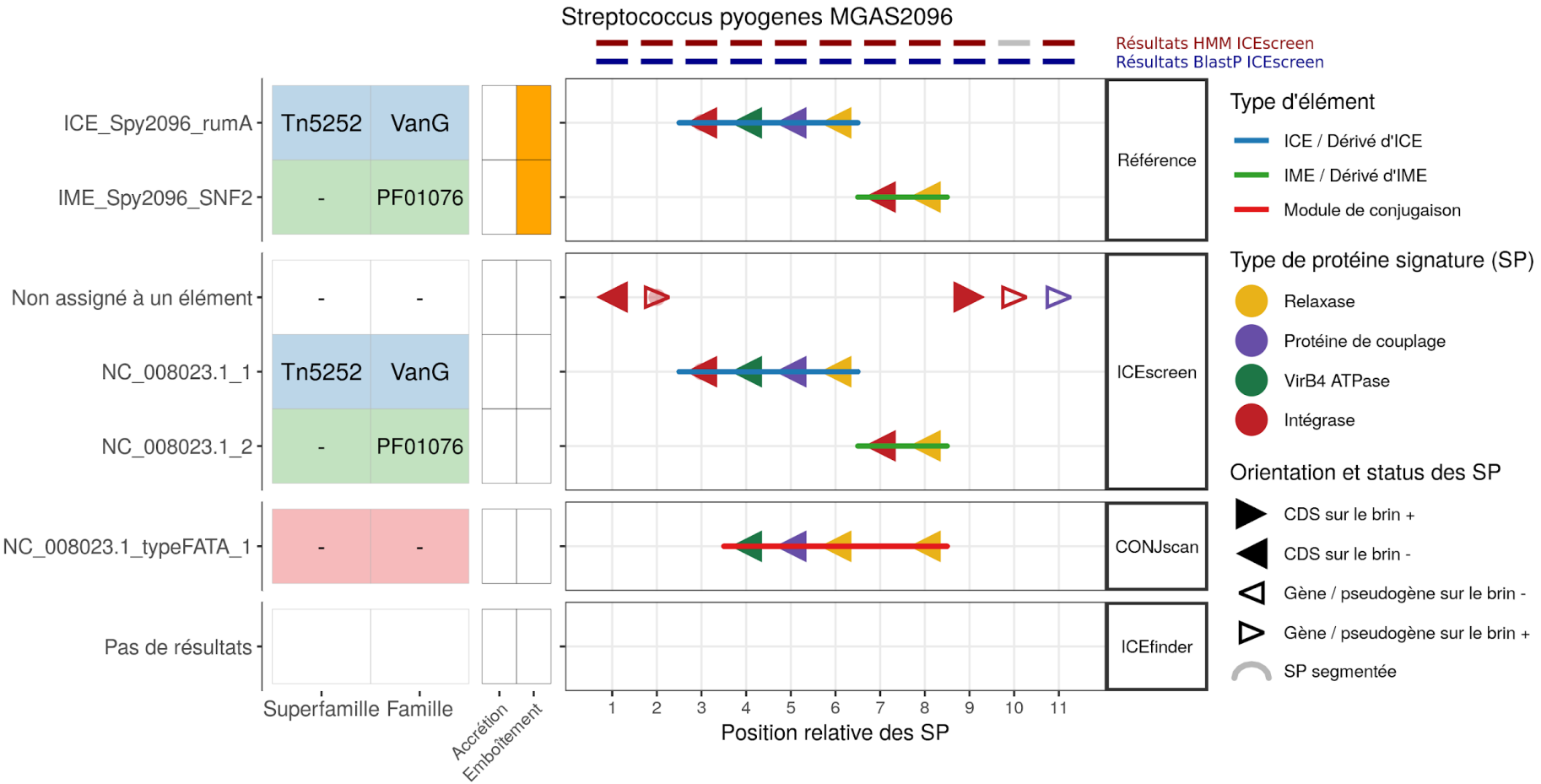

Figure 12 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus pyogenes* sv. M12 MGAS2096 (NC_008023.1).

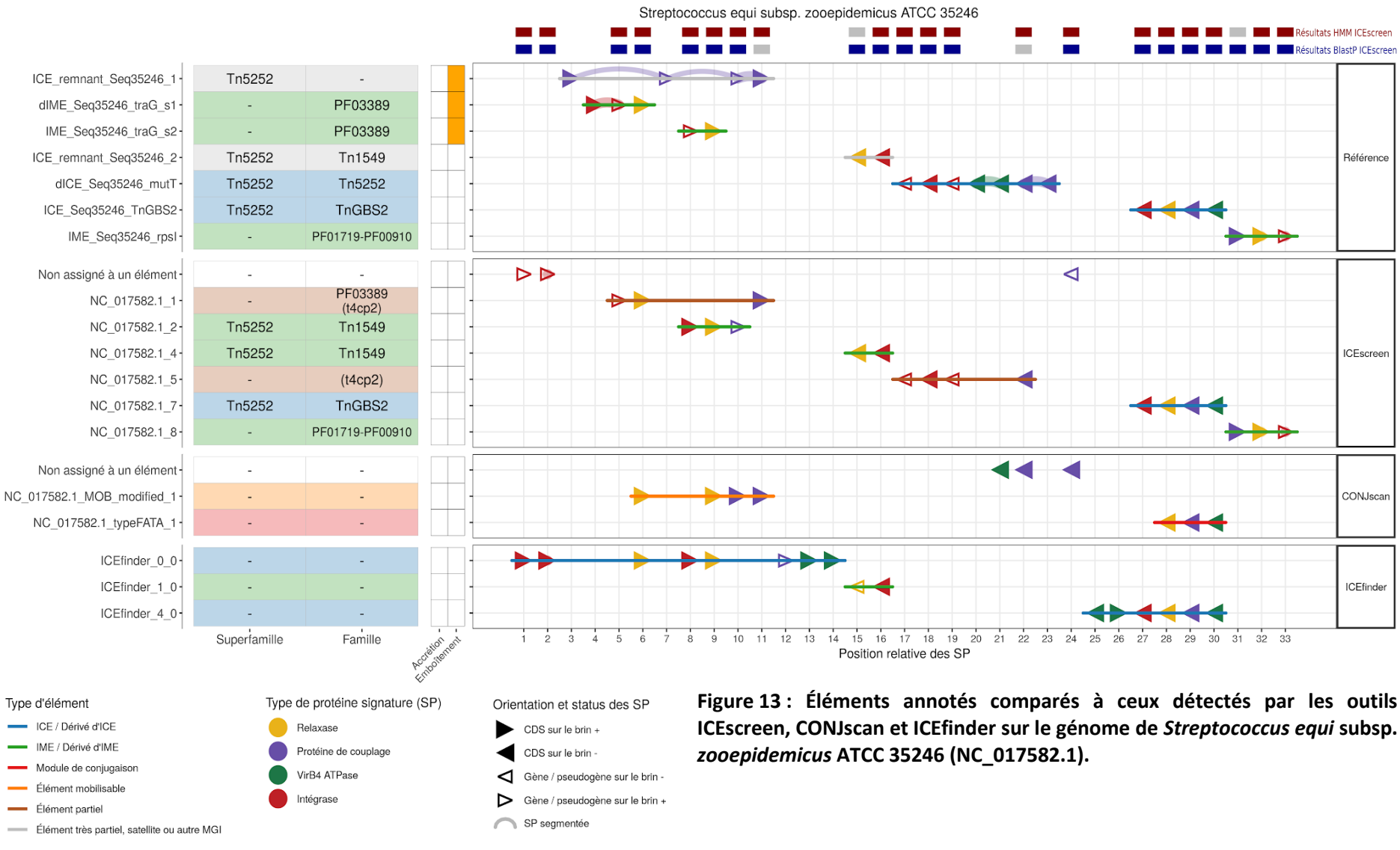

Figure 13 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus equi* subsp. *zooepidemicus* ATCC 35246 (NC_017582.1).

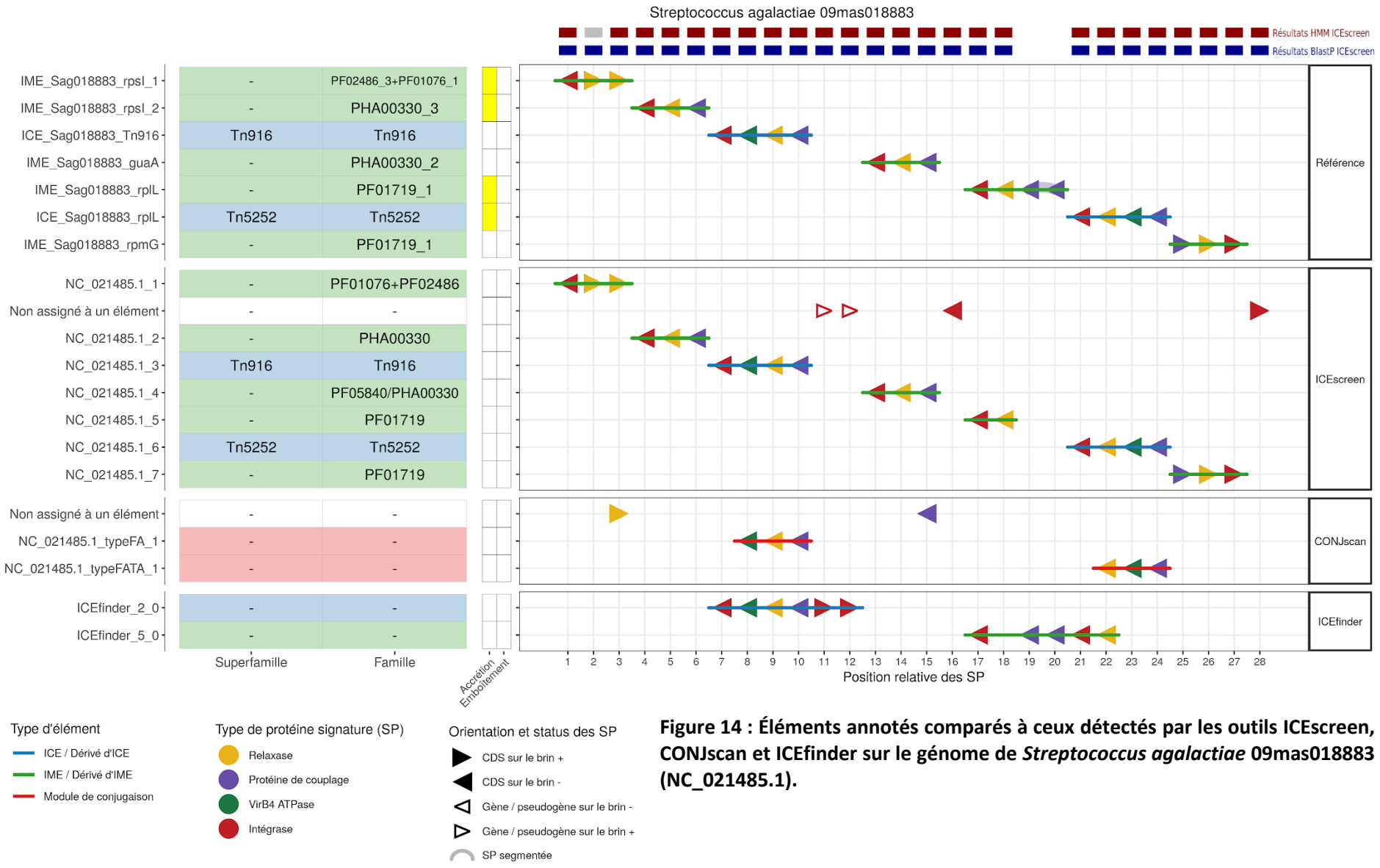

Figure 14 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus agalactiae* 09mas018883 (NC_021485.1).

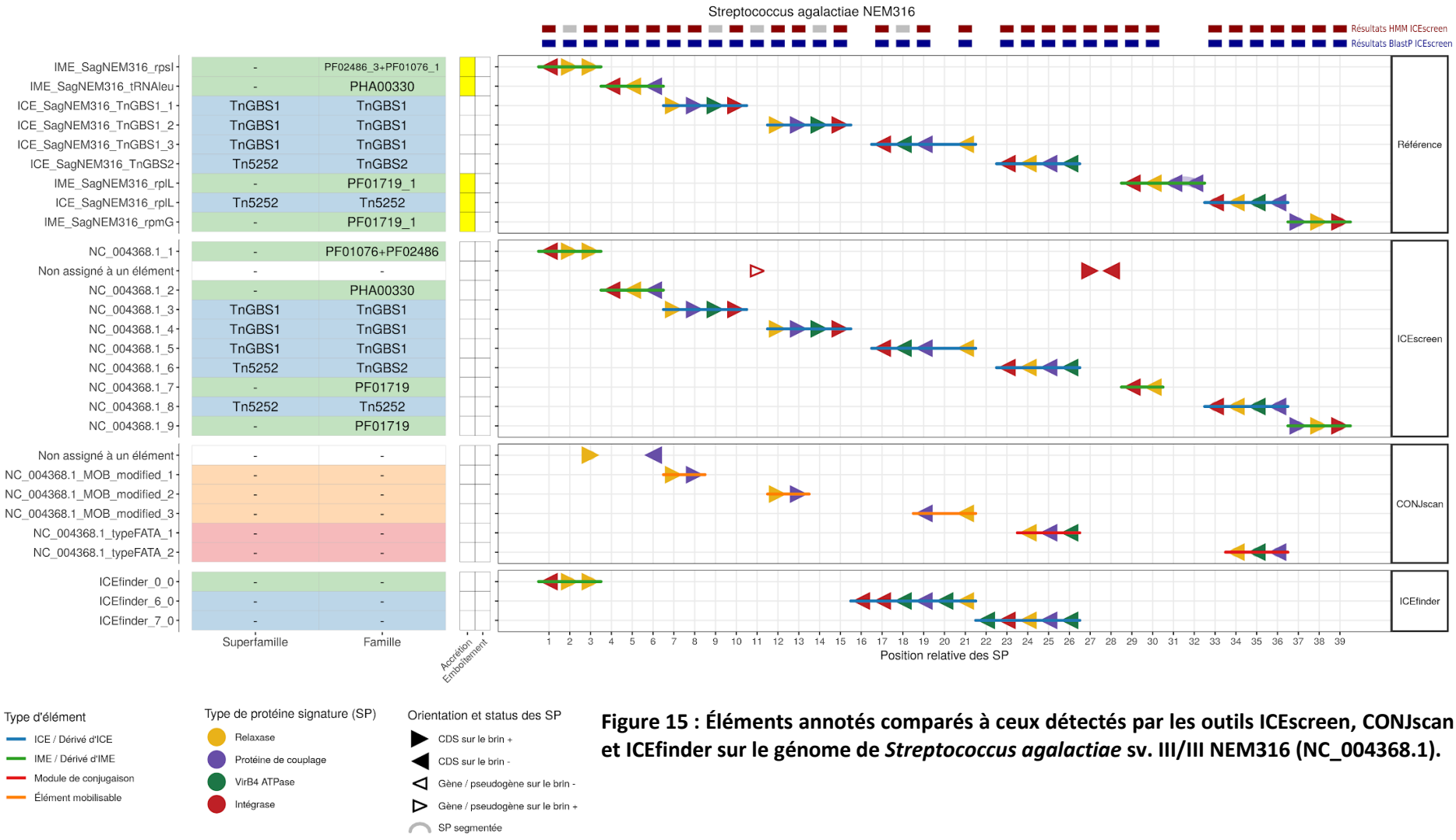

Figure 15 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus agalactiae* sv. III/III NEM316 (NC_004368.1).

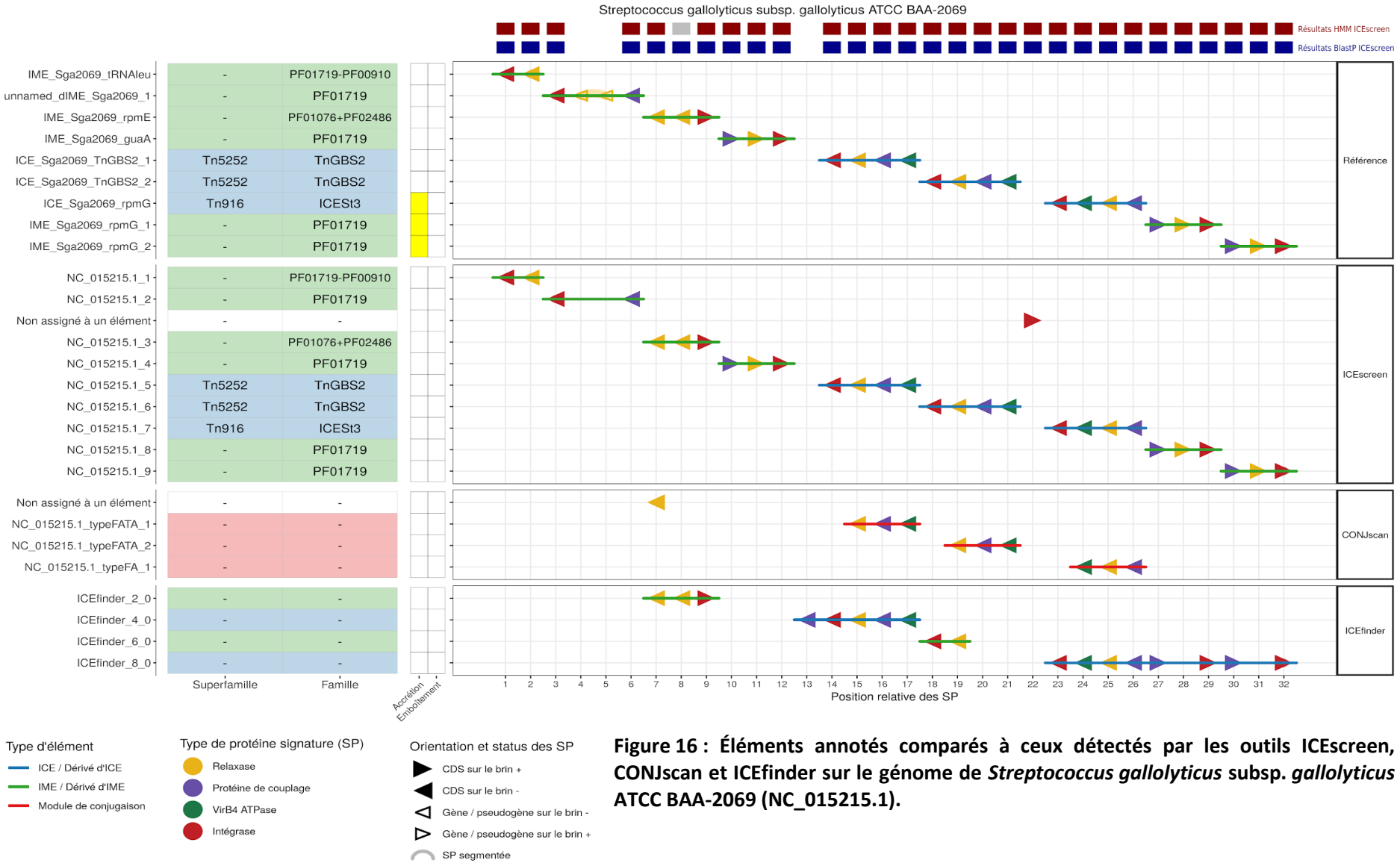

Figure 16 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus gallolyticus* subsp. *gallolyticus* ATCC BAA-2069 (NC_015215.1).

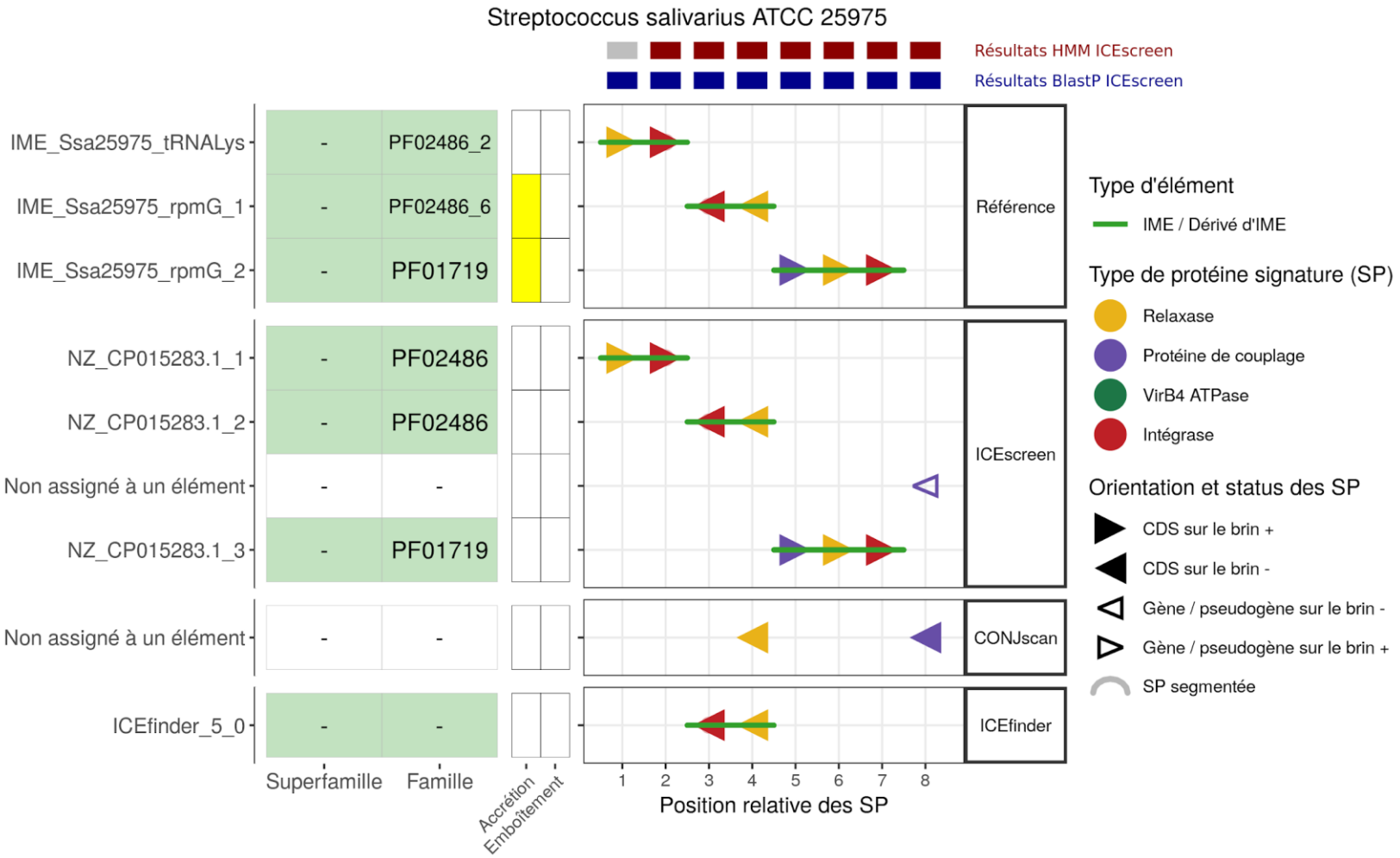

Figure 17 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus salivarius* ATCC 25975 (NZ_CP015283.1).

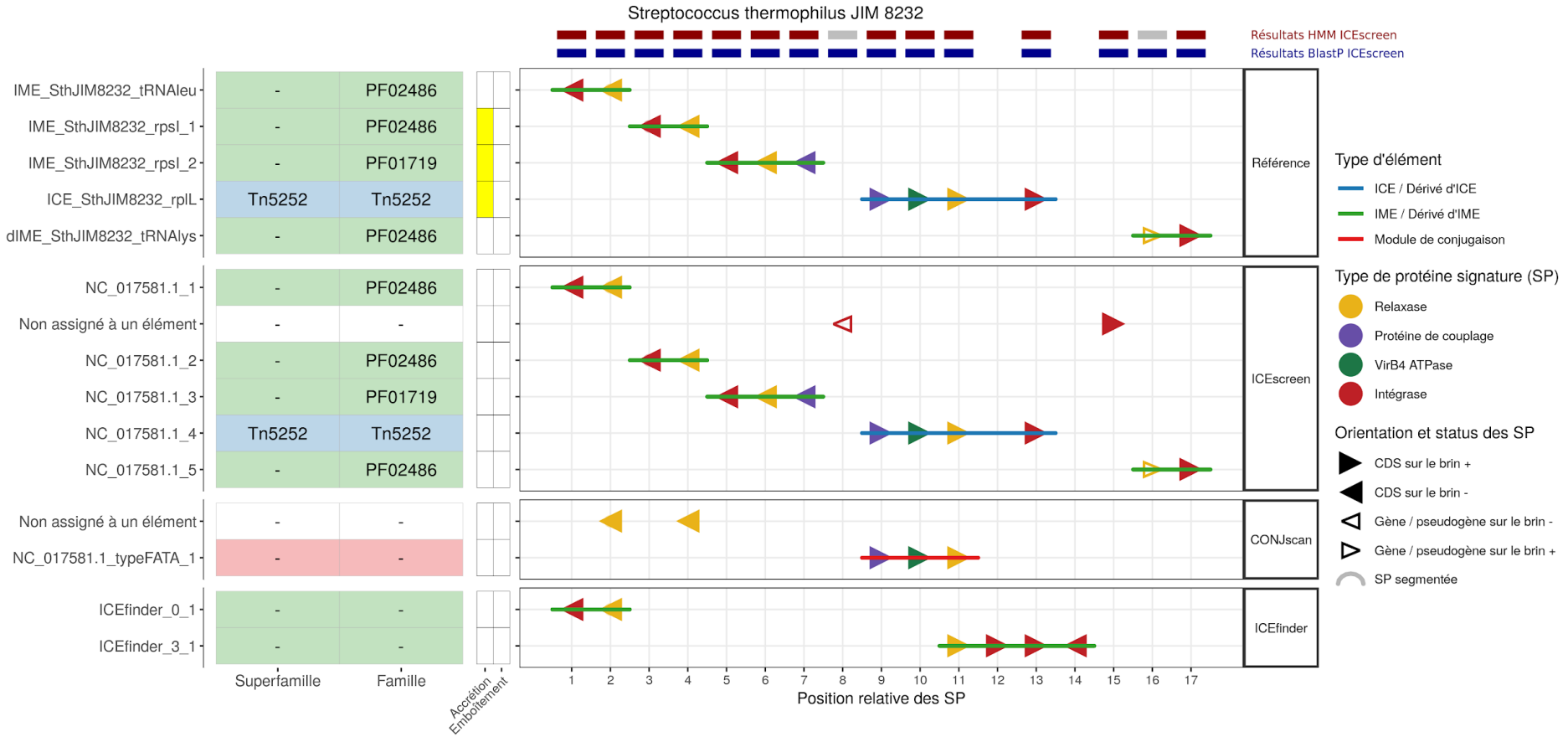

Figure 18 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus thermophilus* JIM 8232 (NC_017581.1).

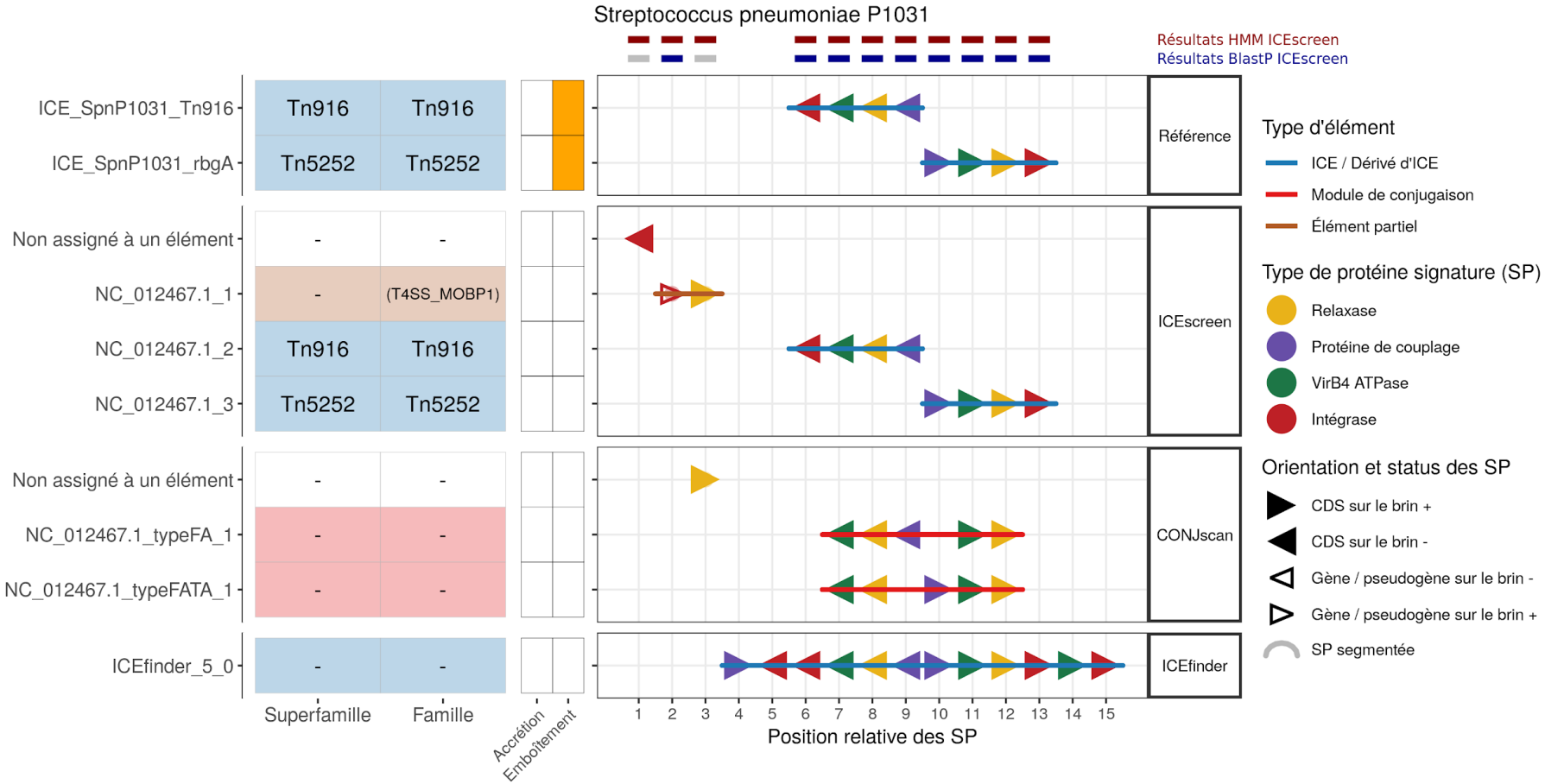

Figure 19 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus pneumoniae* P1031 (NC_012467.1).

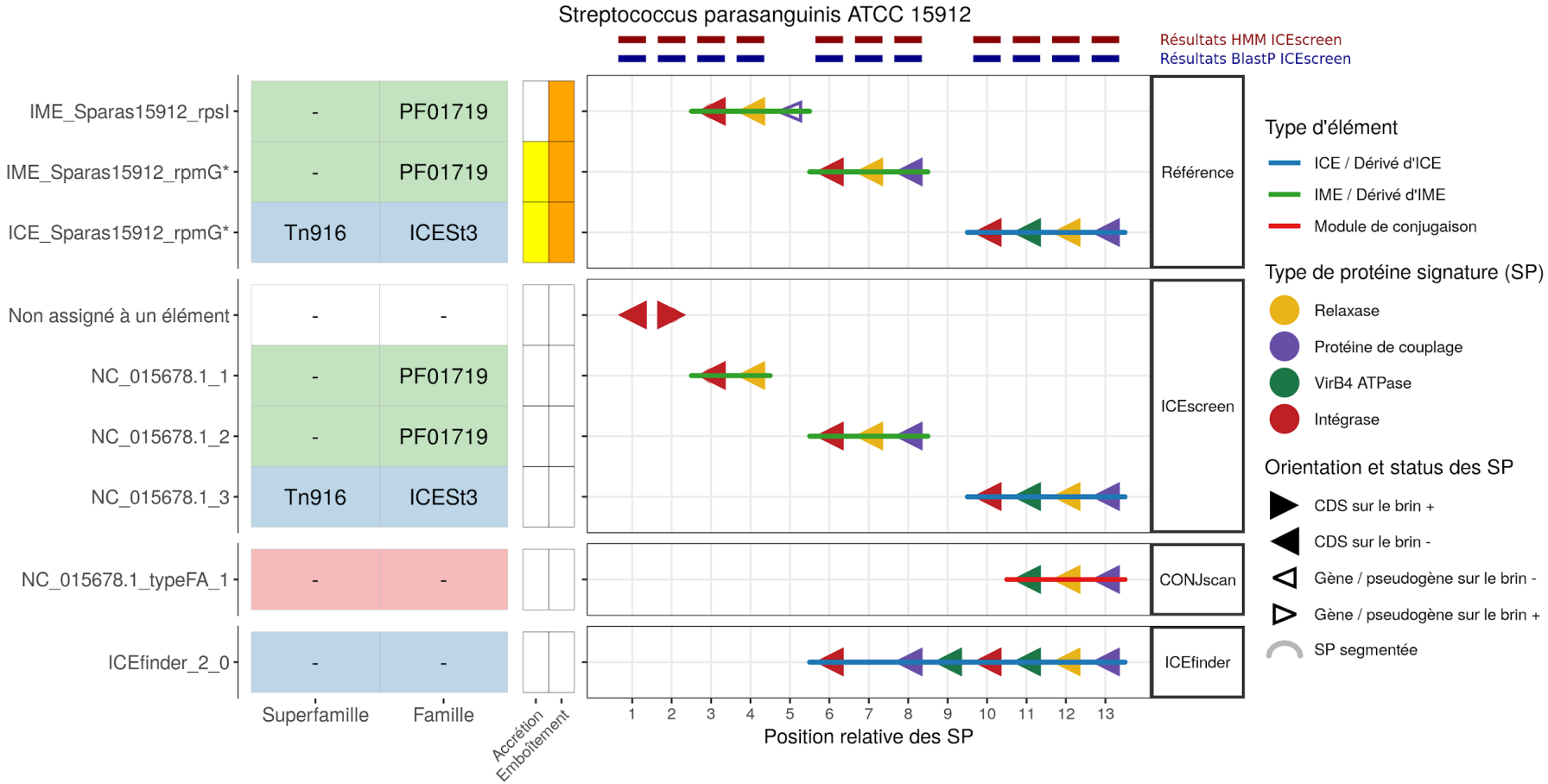

Figure 20 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus parasanguinis* ATCC 15912 (NC_015678.1).

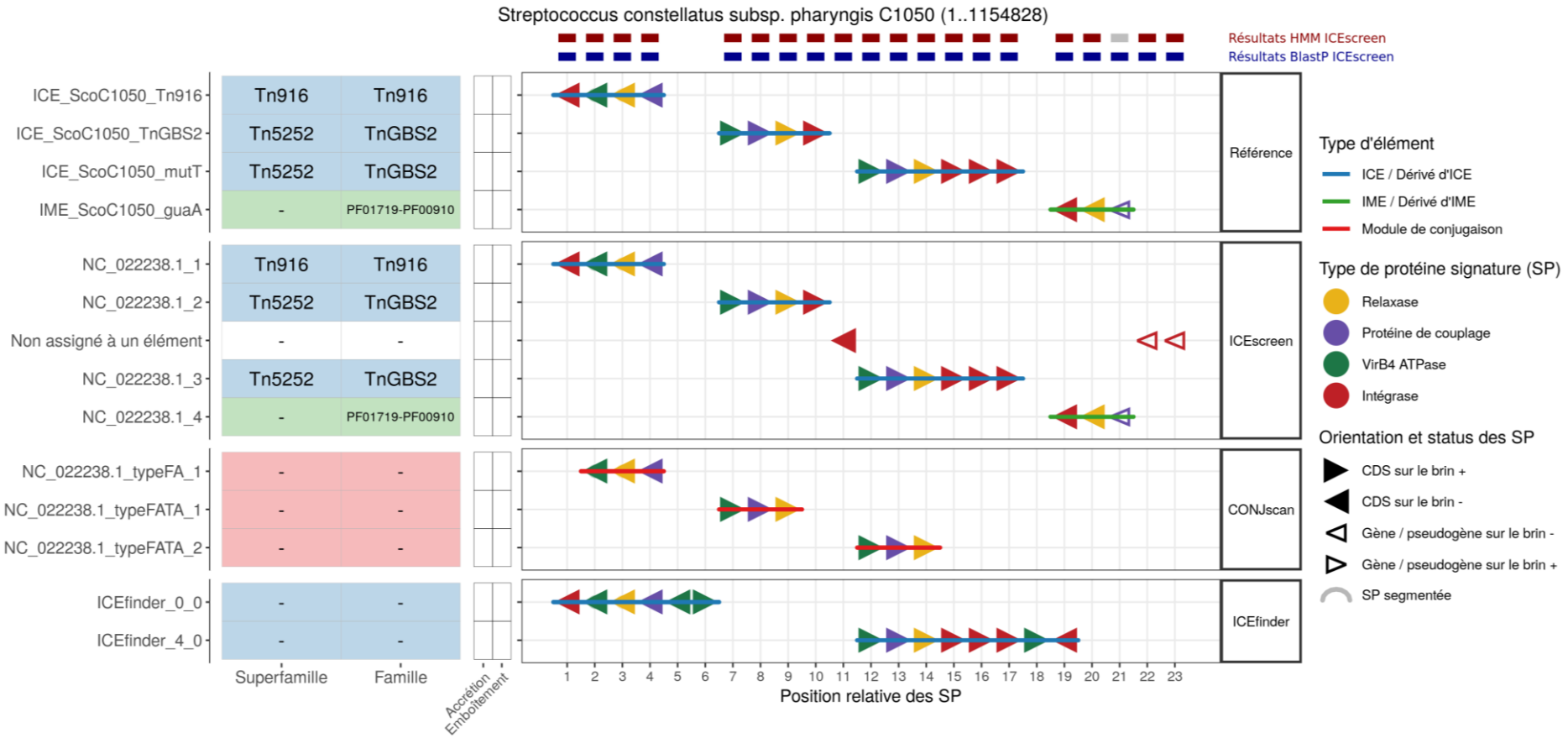

Figure 21a : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus constellatus* subsp. *pharyngis* C1050 (positions 1 à 1154828) (NC_022238.1).

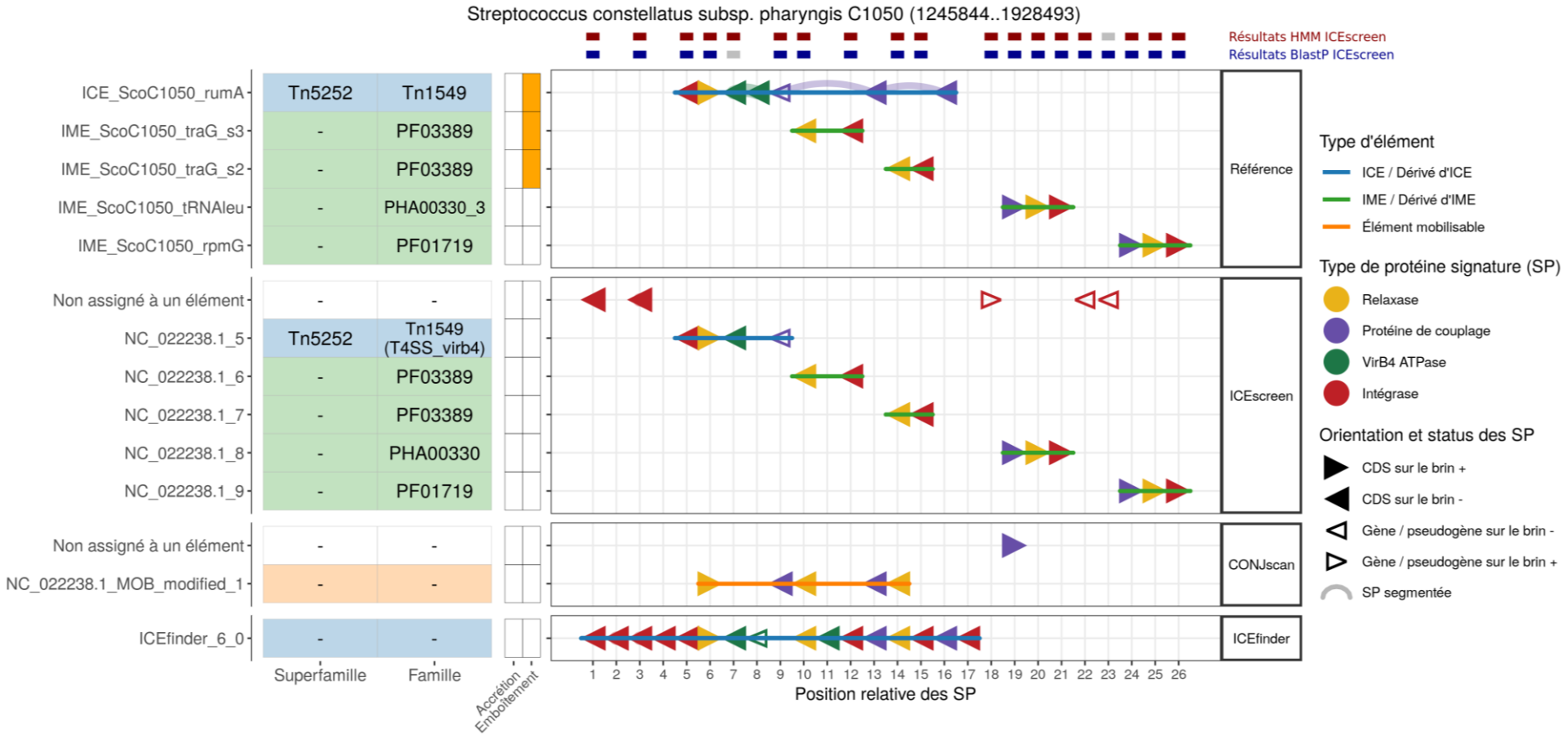

Figure 21b : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus constellatus* subsp. *pharyngis* C1050 (positions 1245844 à 1928493) (NC_022238.1).

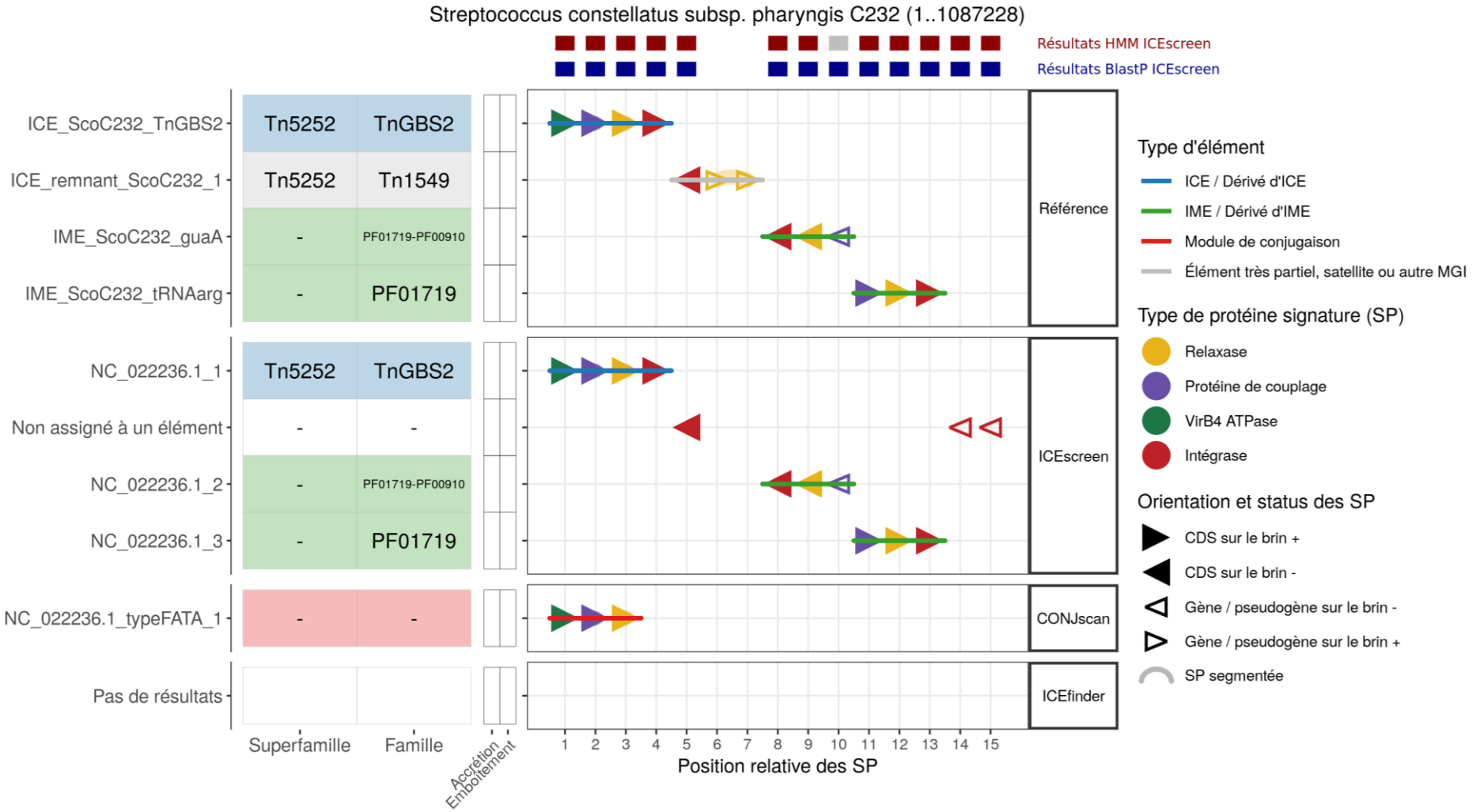

Figure 22a : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus constellatus* subsp. *pharyngis* C232 (positions 1 à 1087228) (NC_022236.1).

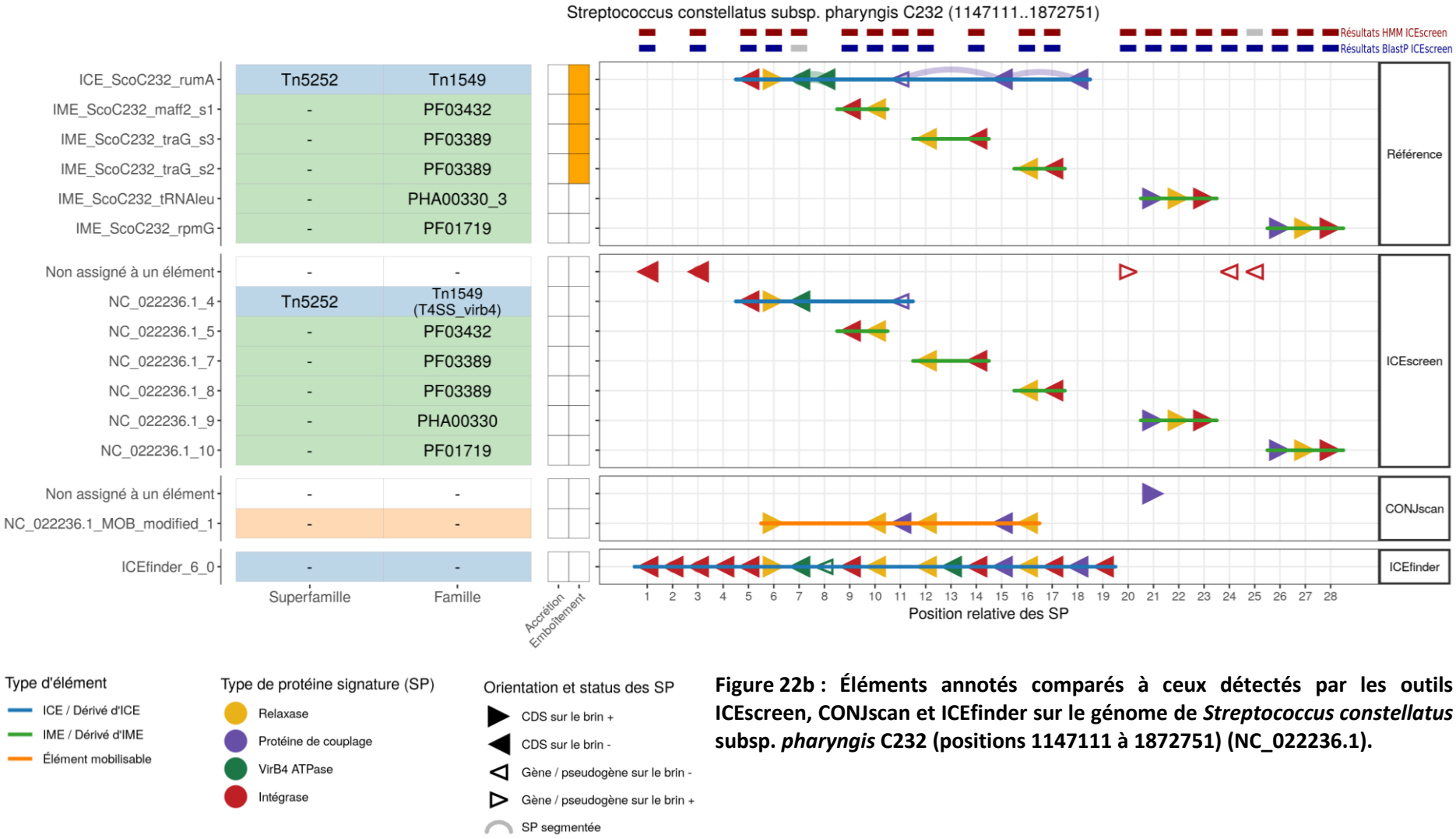

Figure 22b : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus constellatus* subsp. *pharyngis* C232 (positions 1147111 à 1872751) (NC_022236.1).

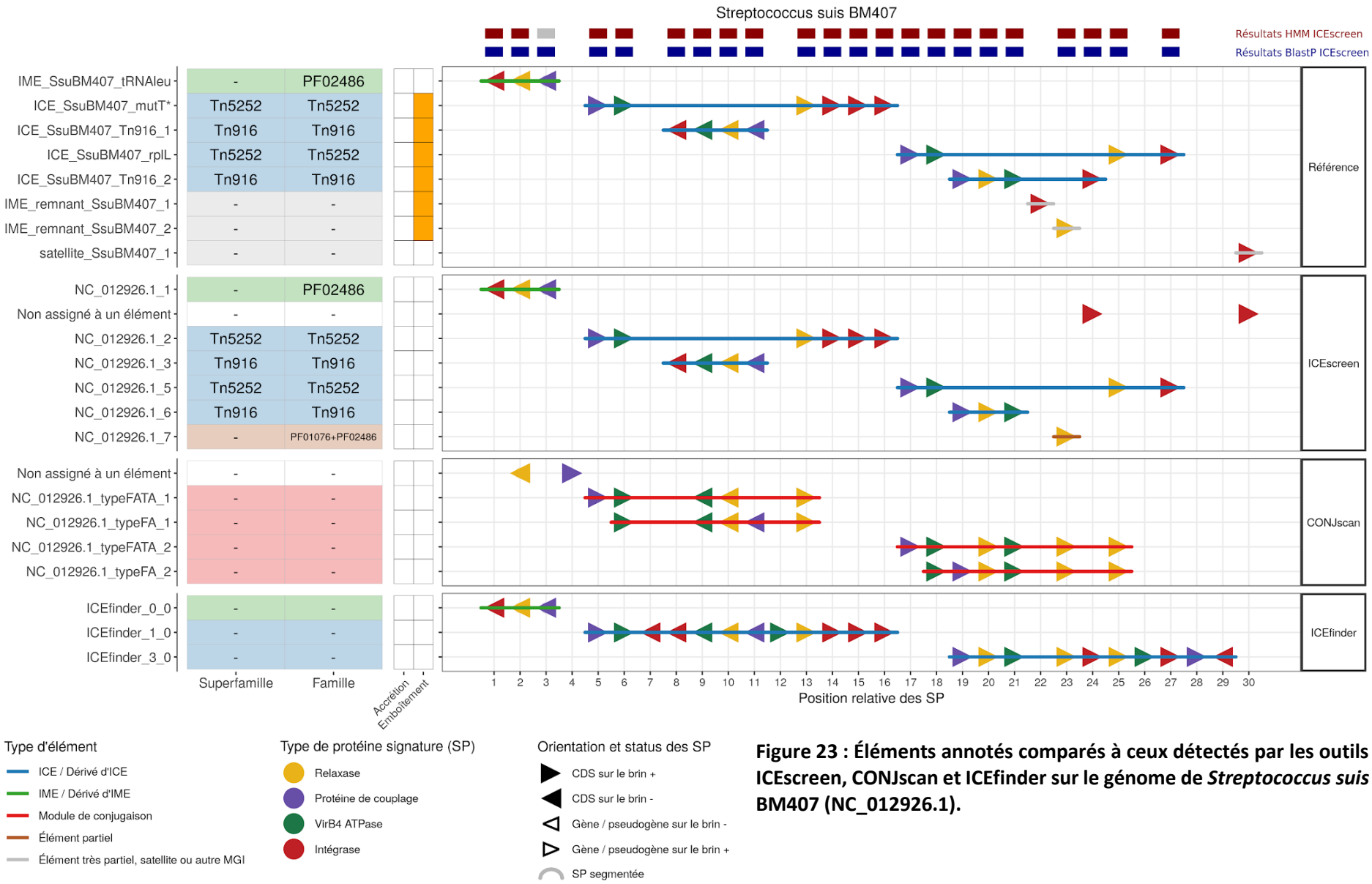

Figure 23 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus suis* BM407 (NC_012926.1).

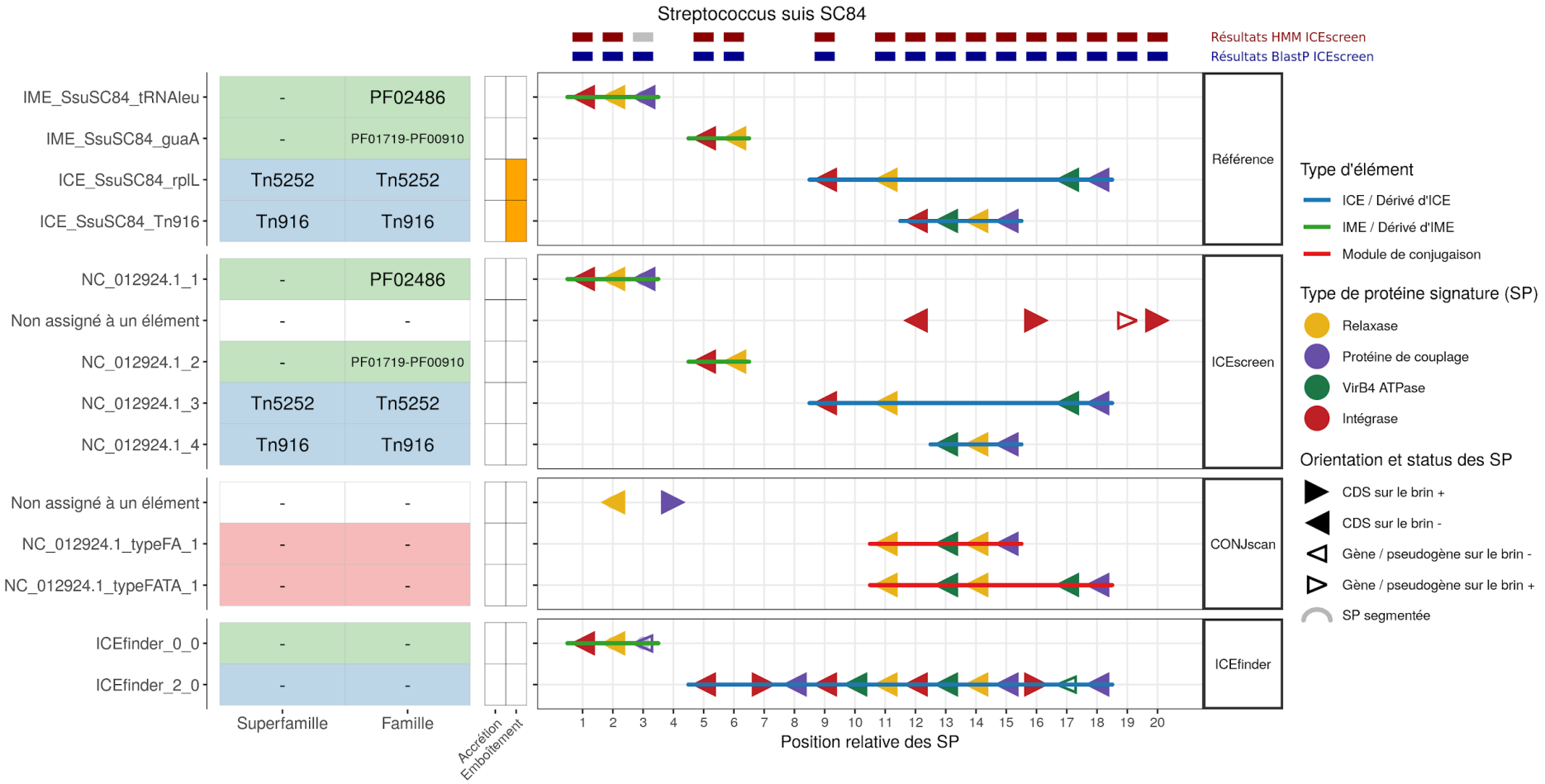

Figure 24 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de Streptococcus suis SC84 (NC_012924.1).

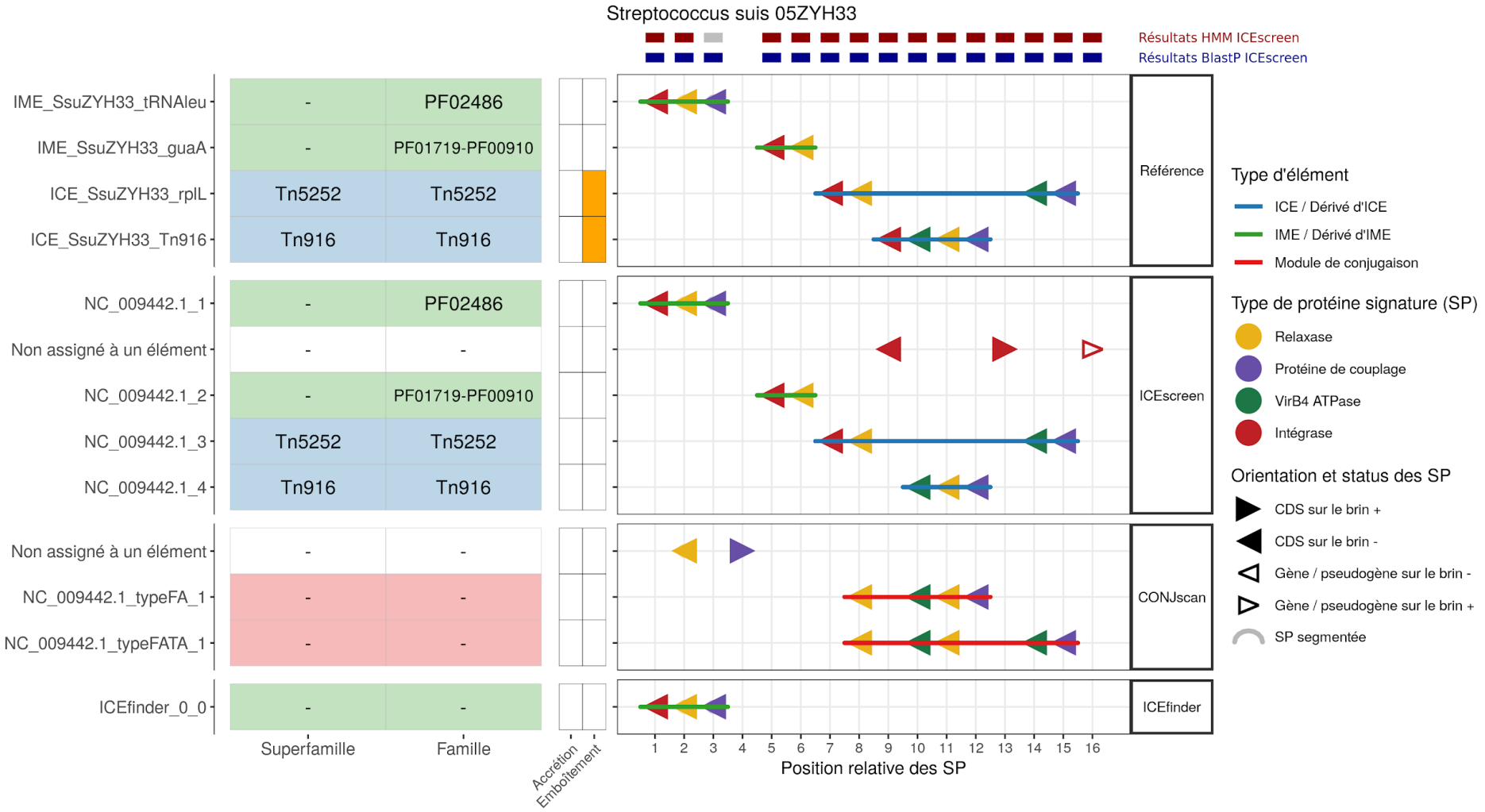

Figure 25 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus suis* sv. SS2 05ZYH33 (NC_009442.1).

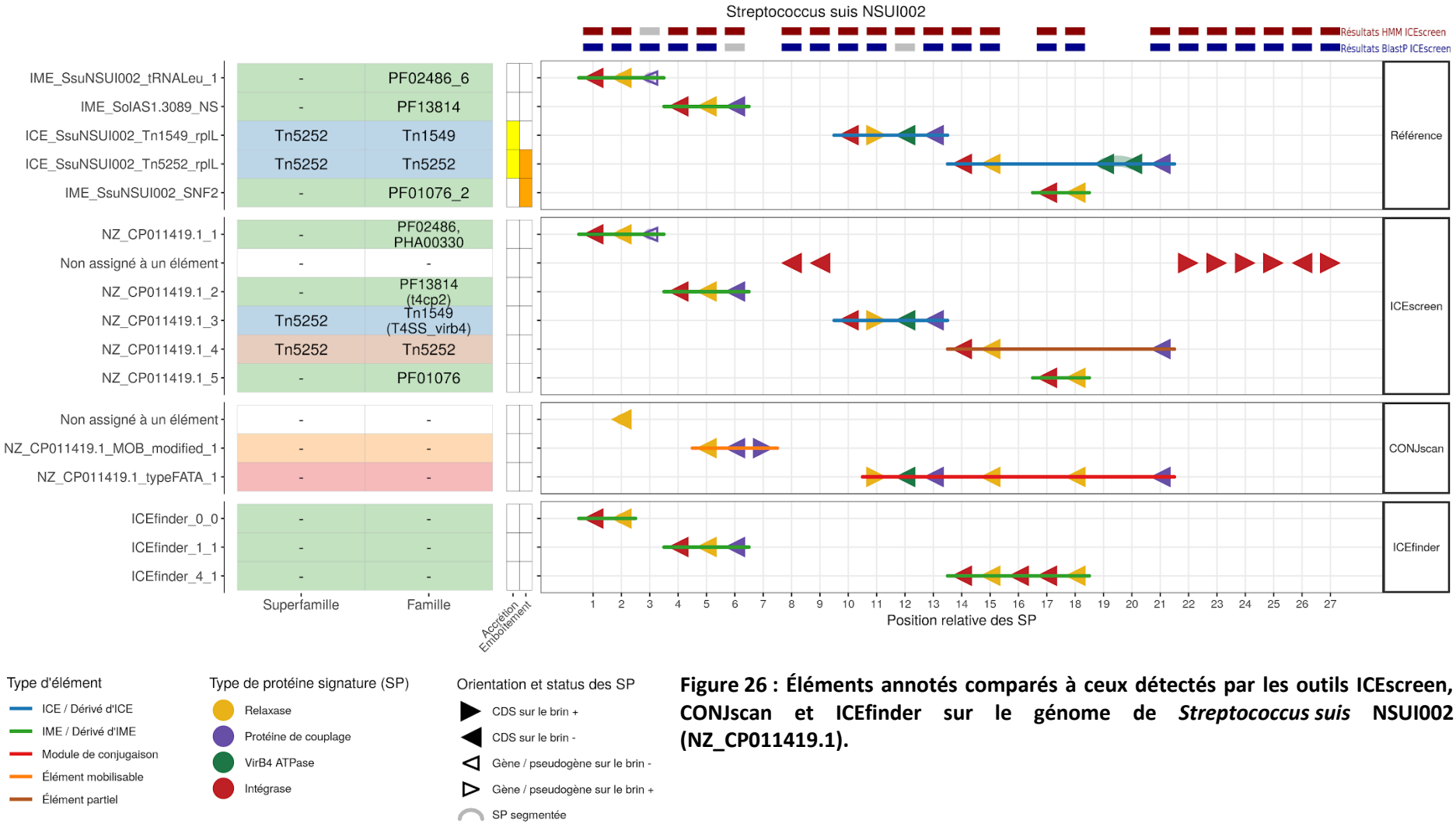

Figure 26 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus suis* NSUI002 (NZ_CP011419.1).

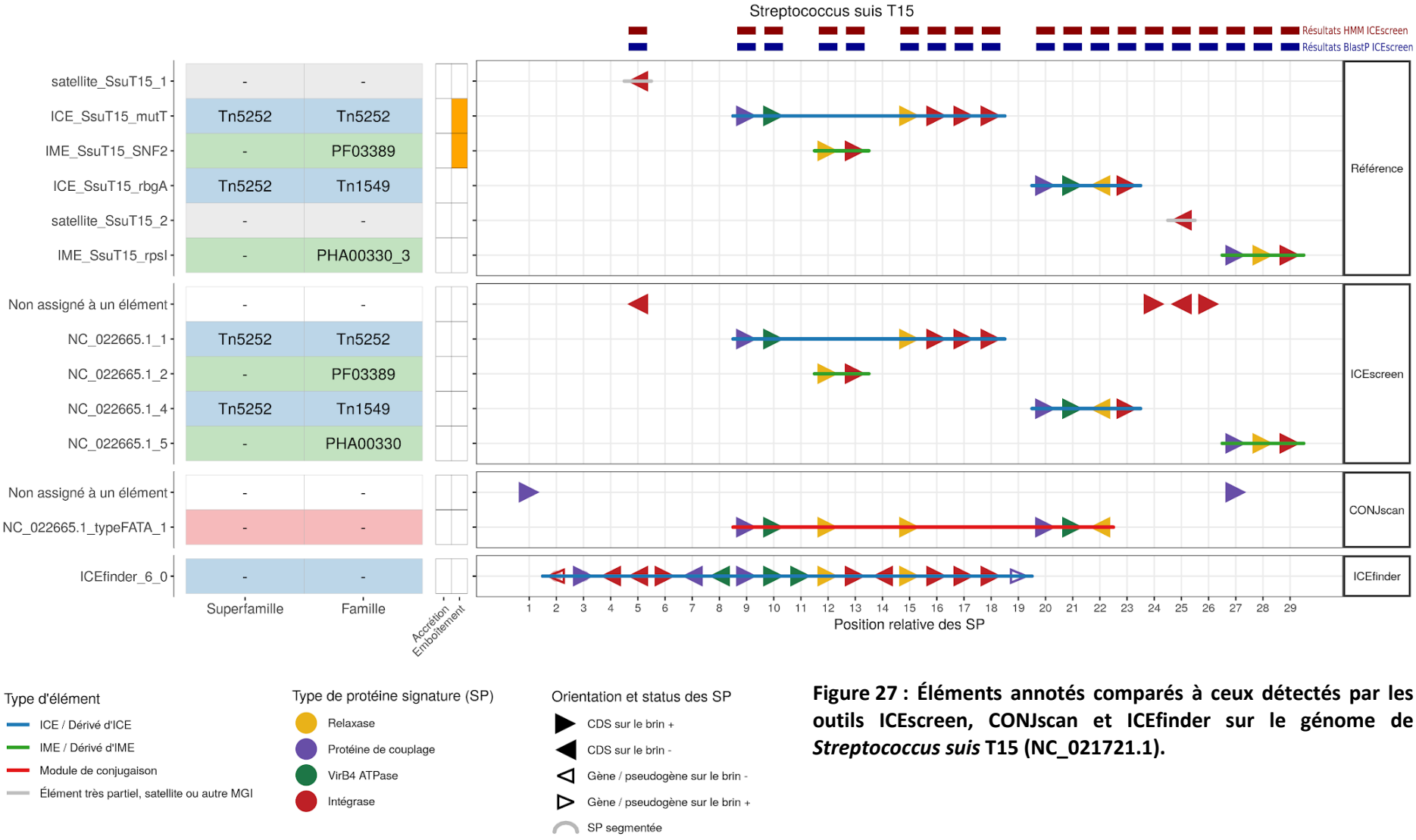

Figure 27 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Streptococcus suis* T15 (NC_021721.1).

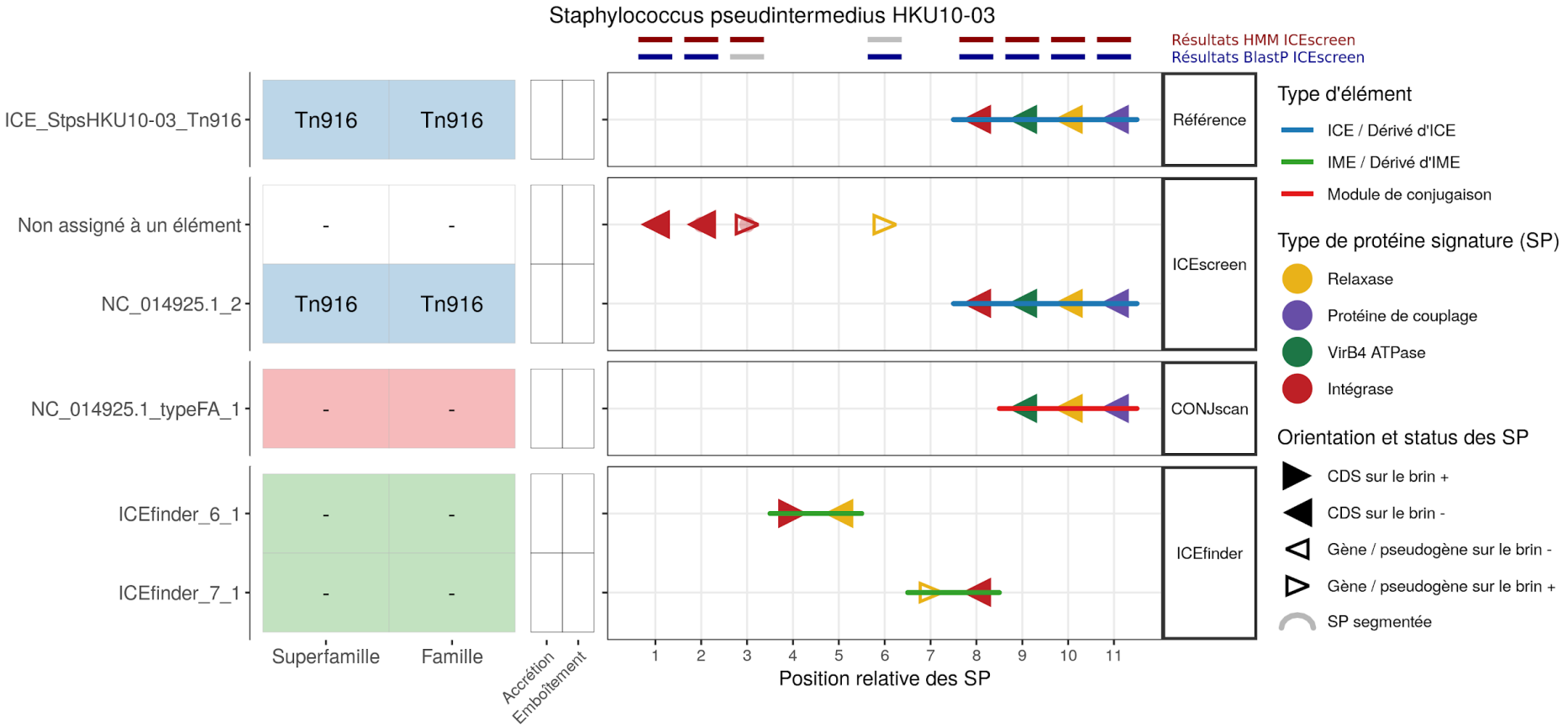

Figure 28 : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Staphylococcus pseudintermedius* HKU10-03 (NC_014925.1).

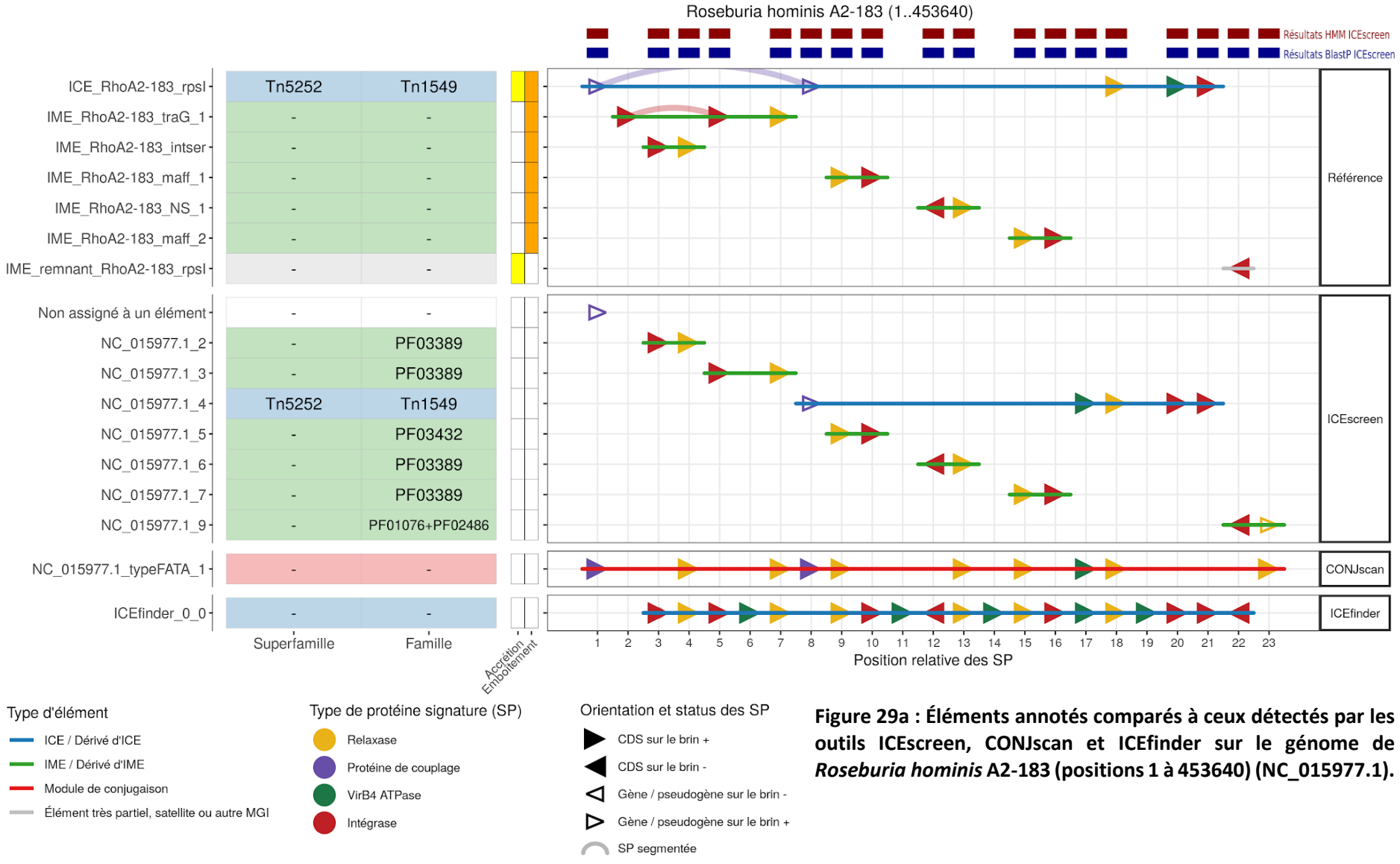

Figure 29a : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Roseburia hominis* A2-183 (positions 1 à 453640) (NC_015977.1).

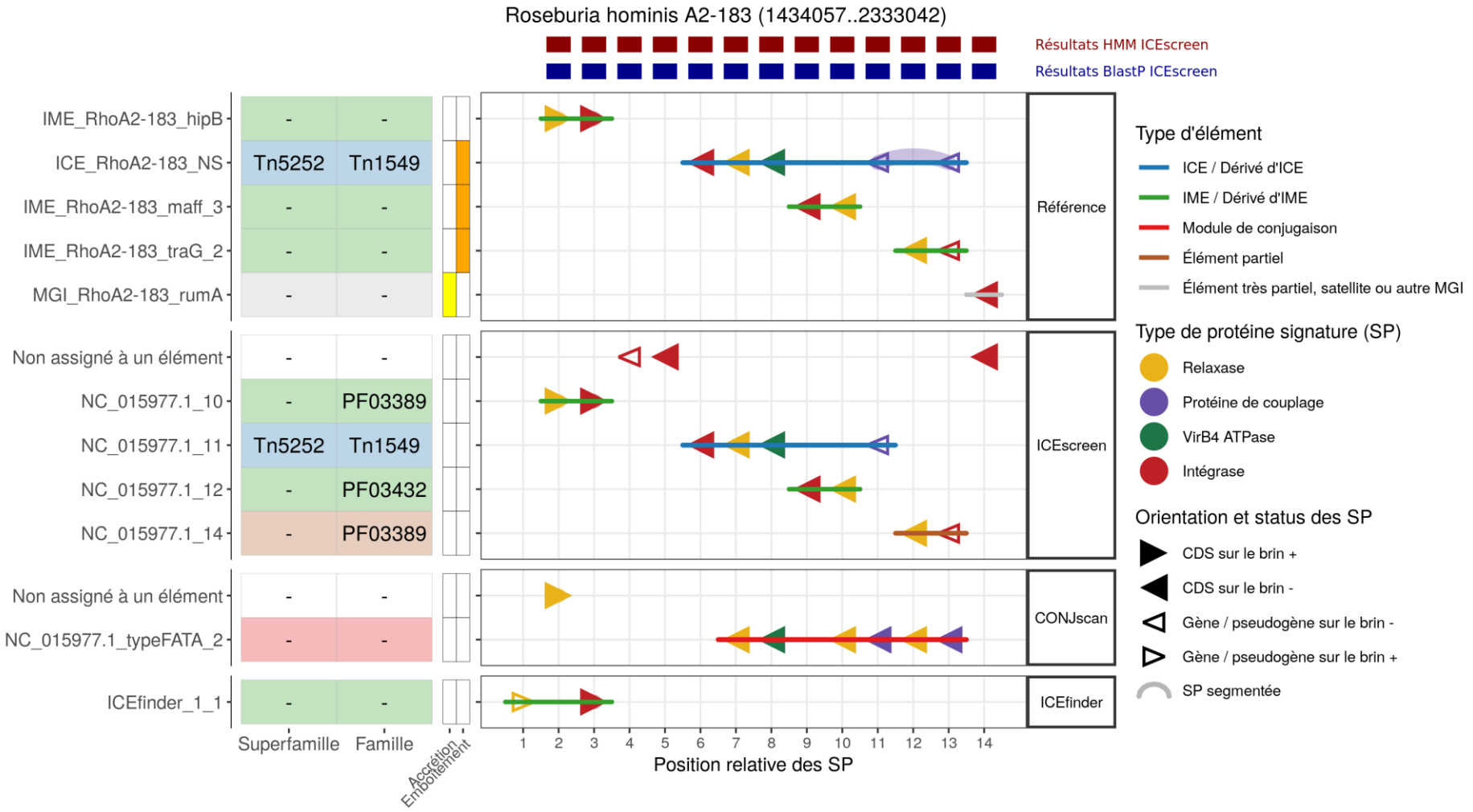

Figure 29b : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Roseburia hominis* A2-183 (positions 1434057 à 2333042) (NC_015977.1).

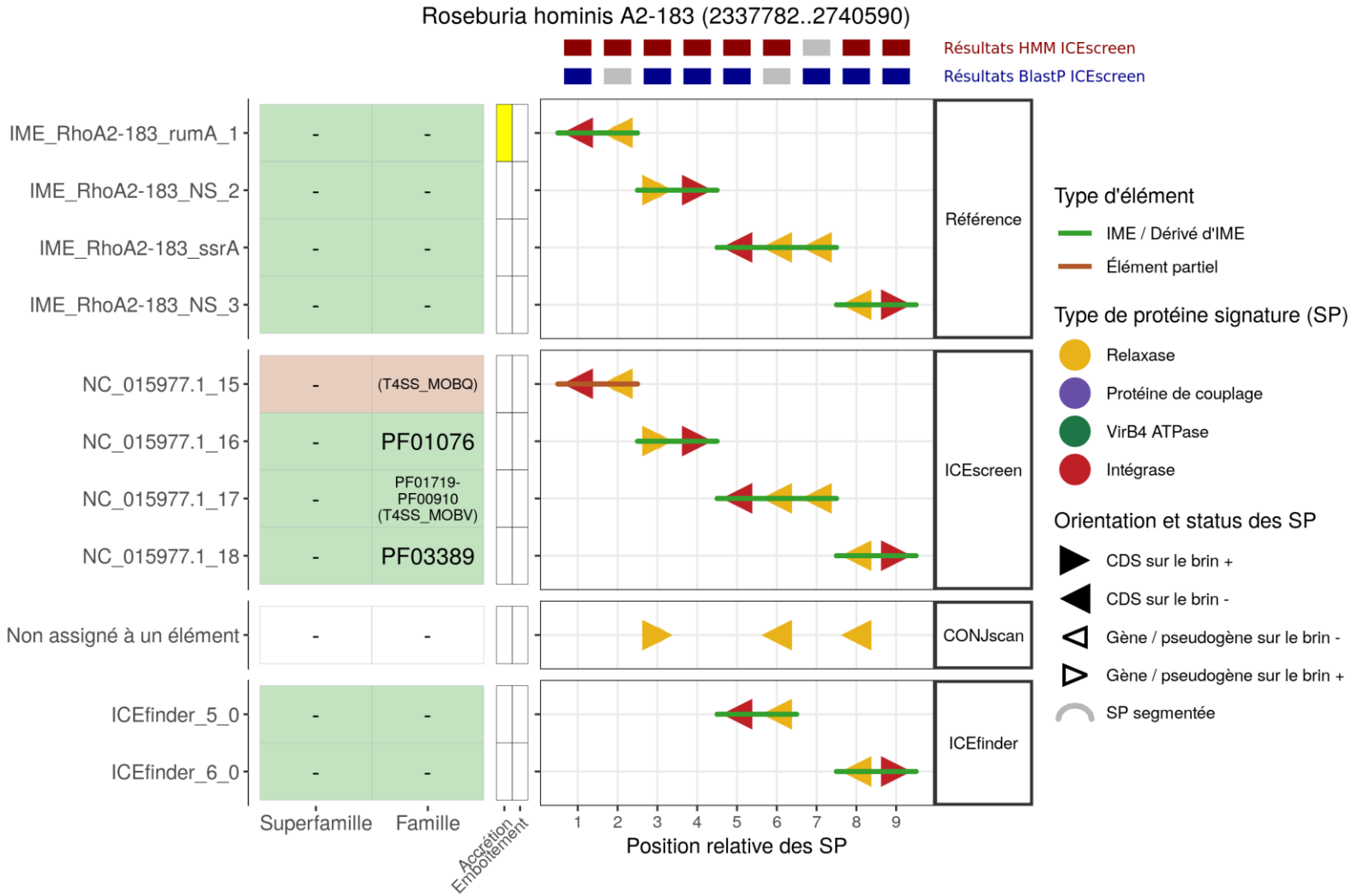

Figure 29c : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Roseburia hominis* A2-183 (positions 2337782 à 2740590) (NC_015977.1).

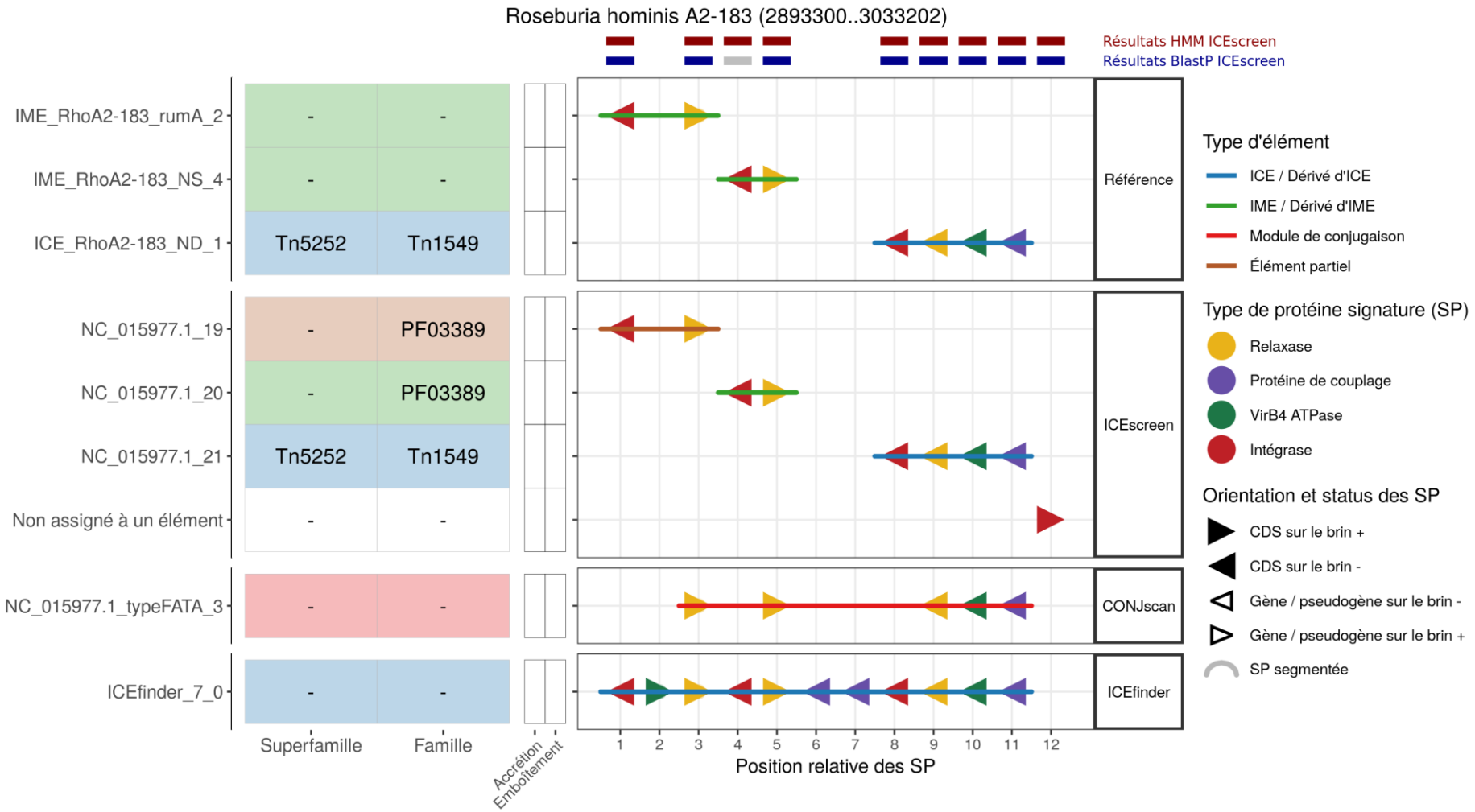

Figure 29d : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Roseburia hominis* A2-183 (positions 2893300 à 3033202) (NC_015977.1).

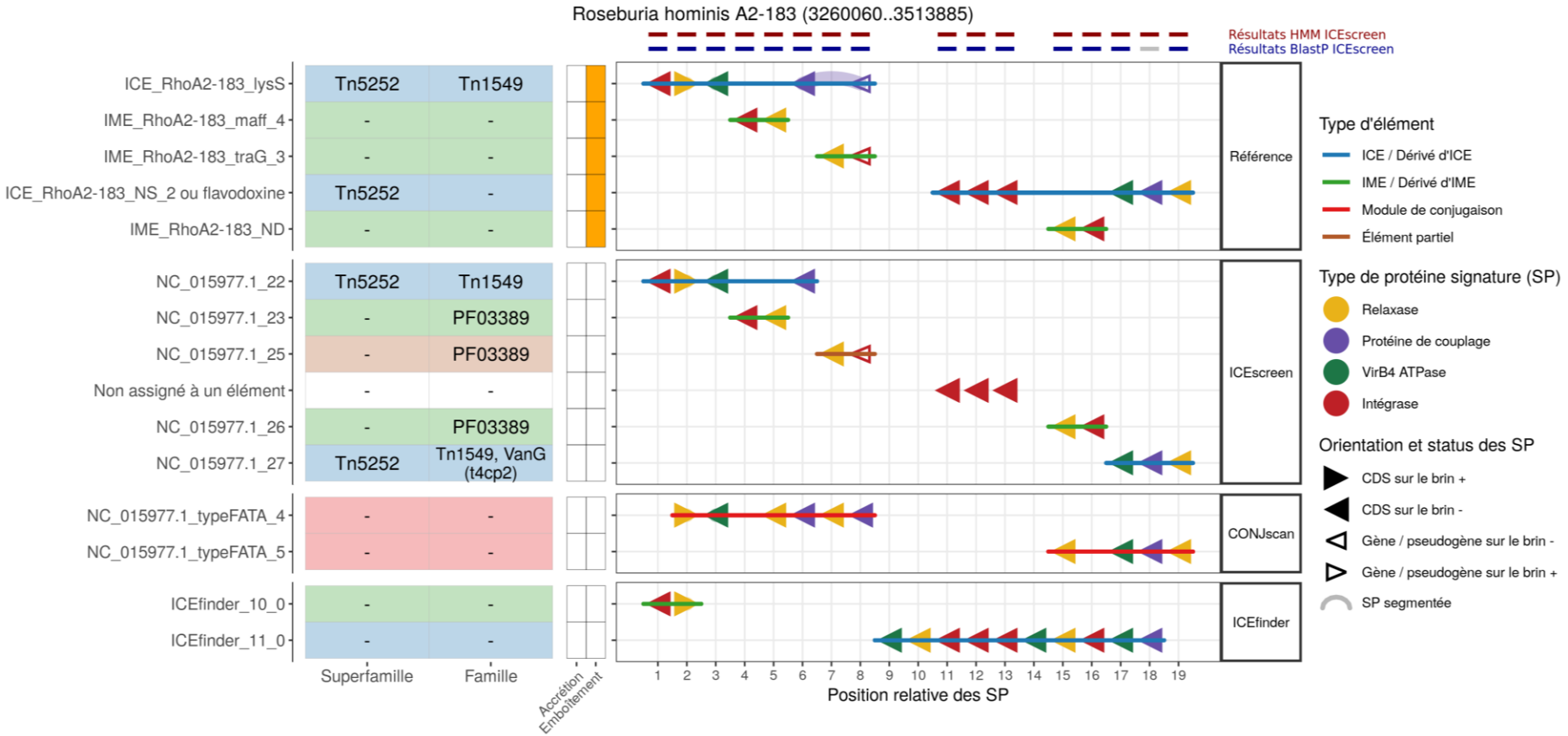

Figure 29e : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Roseburia hominis* A2-183 (positions 2893300 à 3033202) (NC_015977.1).

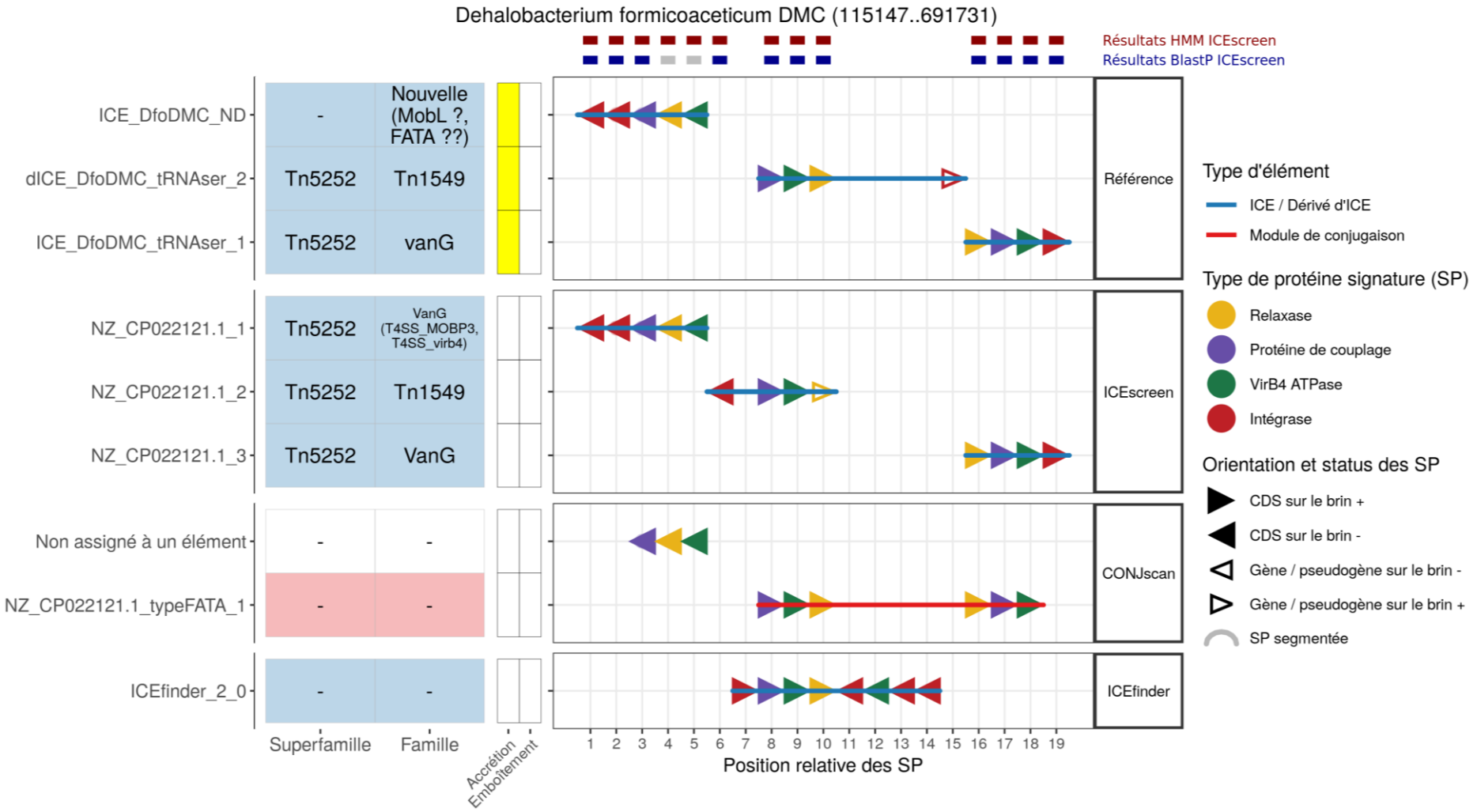

Figure 30a : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Dehalobacterium formicoaceticum* DMC (positions 115147 à 691731) (NZ_CP022121.1).

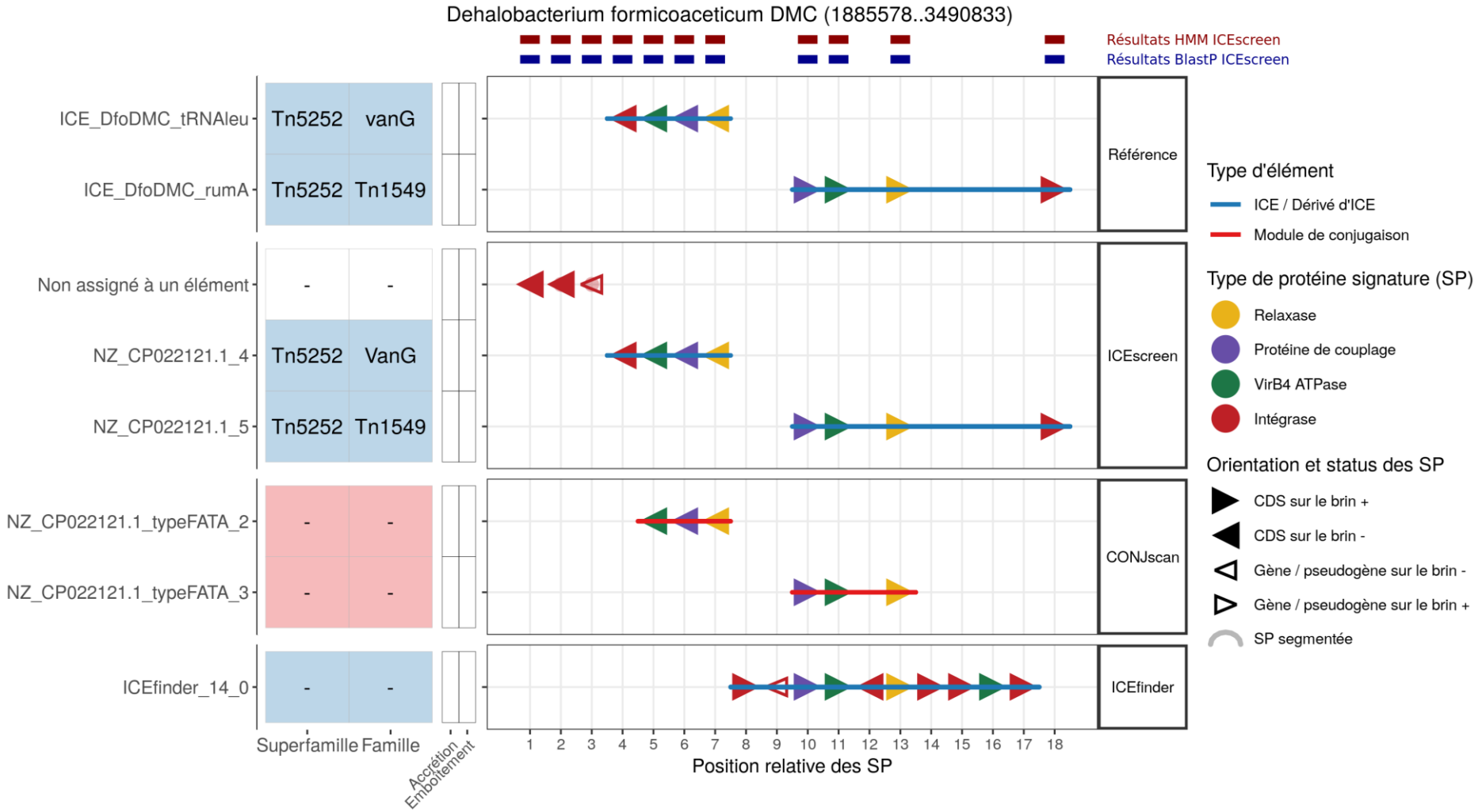

Figure 30b : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Dehalobacterium formicoaceticum* DMC (positions 1885578 à 3490833) (NZ_CP022121.1).

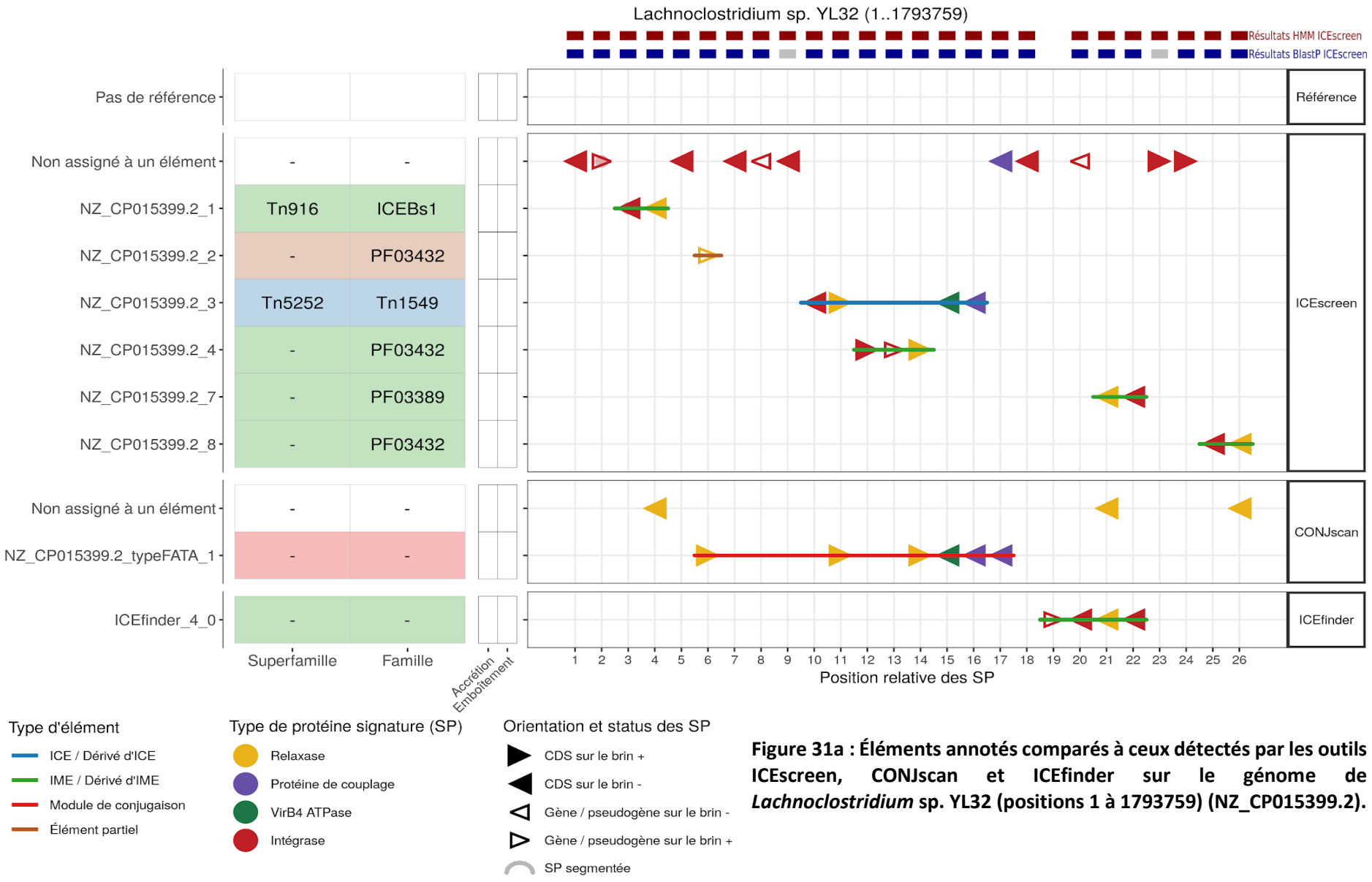

Figure 31a : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Lachnoclostridium* sp. YL32 (positions 1 à 1793759) (NZ_CP015399.2).

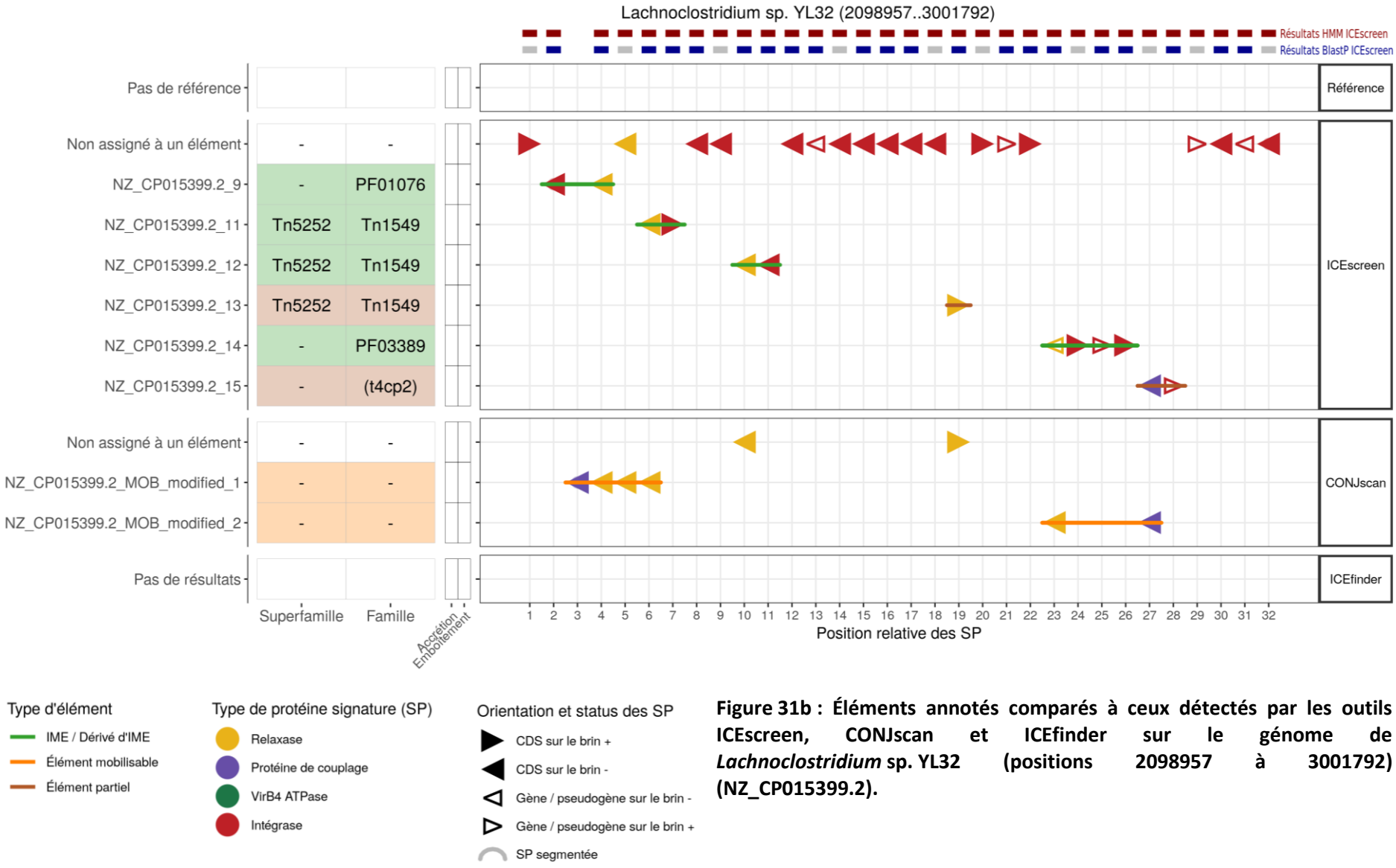

Figure 31b : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Lachnoclostridium* sp. YL32 (positions 2098957 à 3001792) (NZ_CP015399.2).

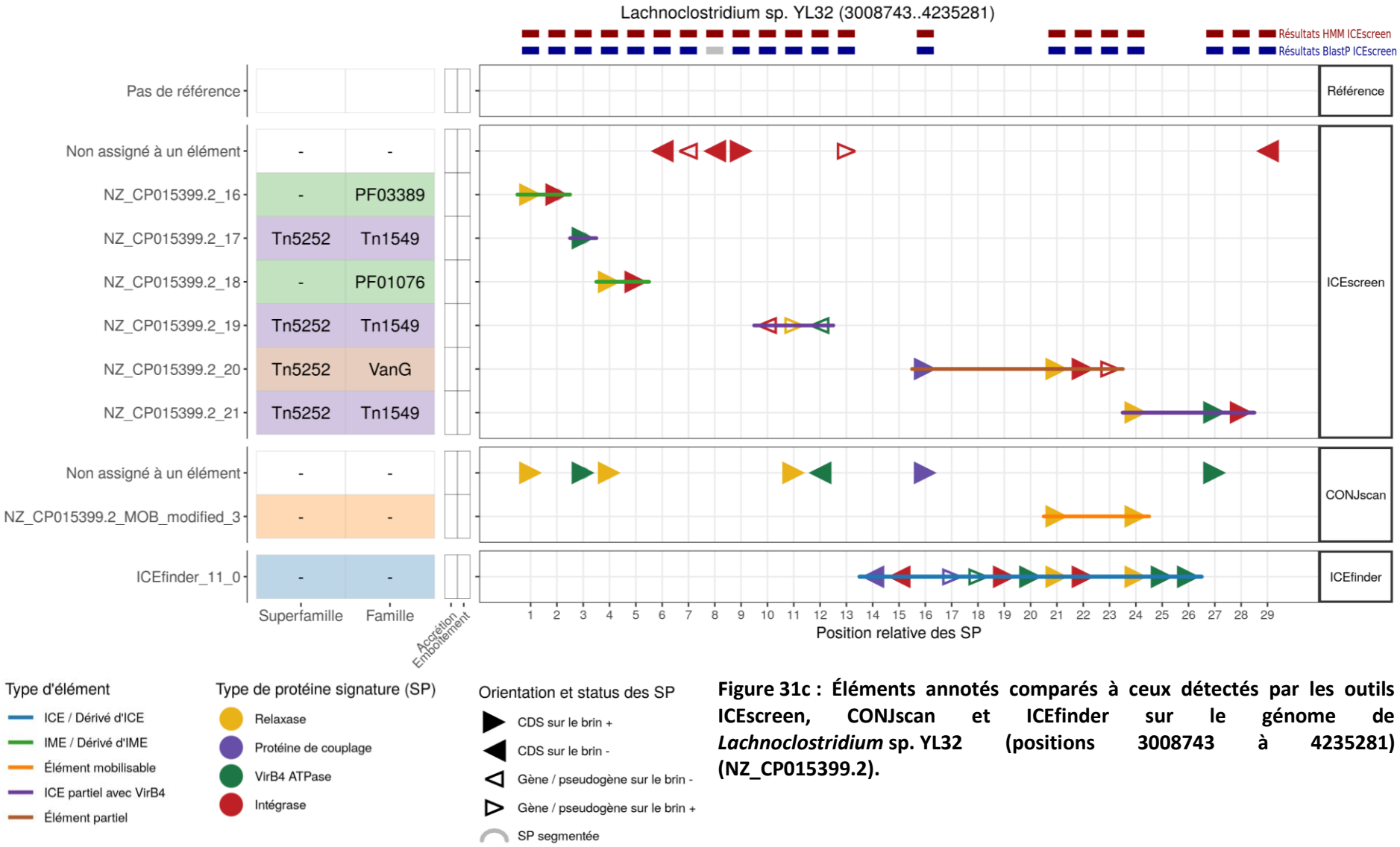

Figure 31c : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Lachnoclostridium* sp. YL32 (positions 3008743 à 4235281) (NZ_CP015399.2).

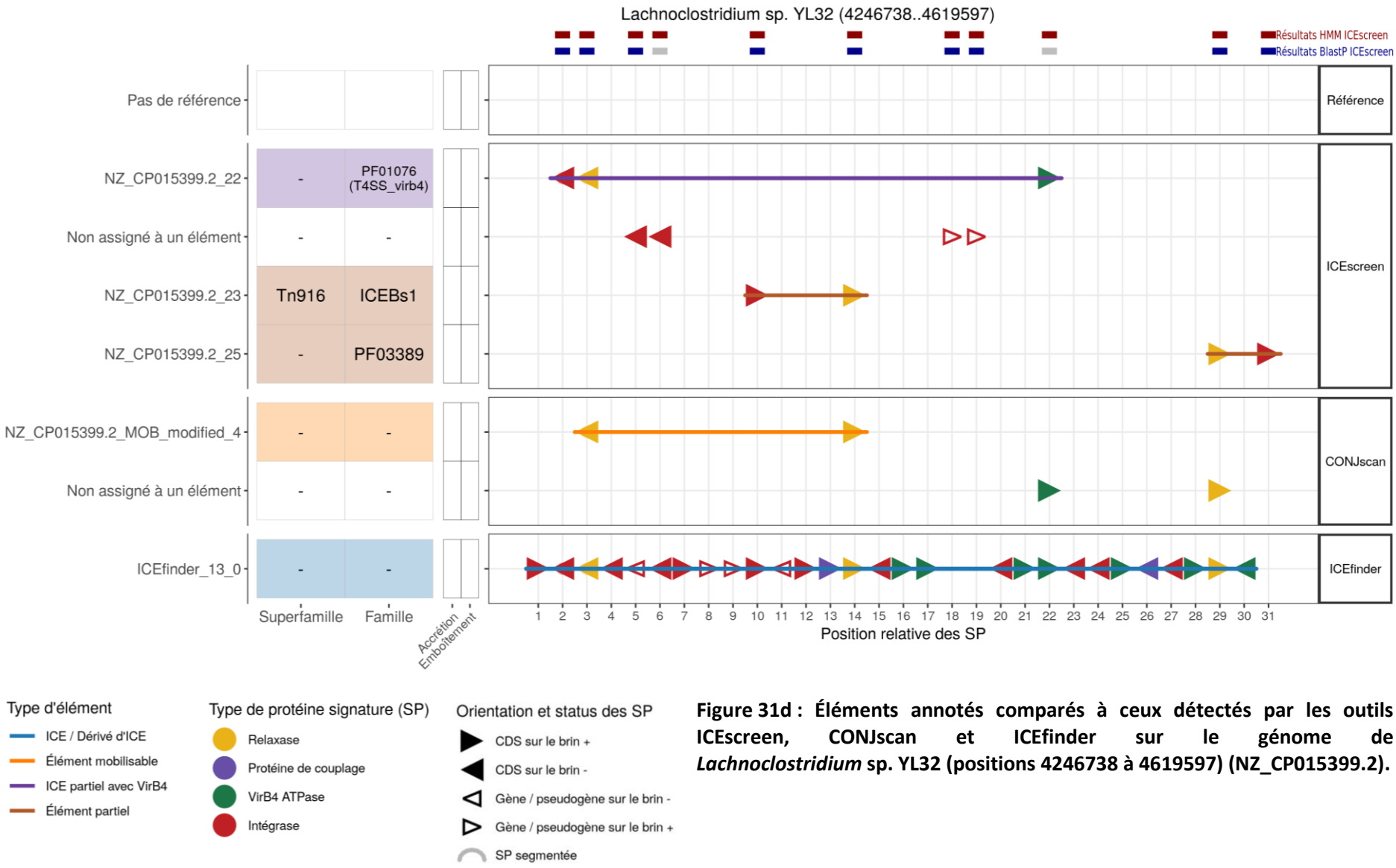

Figure 31d : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Lachnoclostridium* sp. YL32 (positions 4246738 à 4619597) (NZ_CP015399.2).

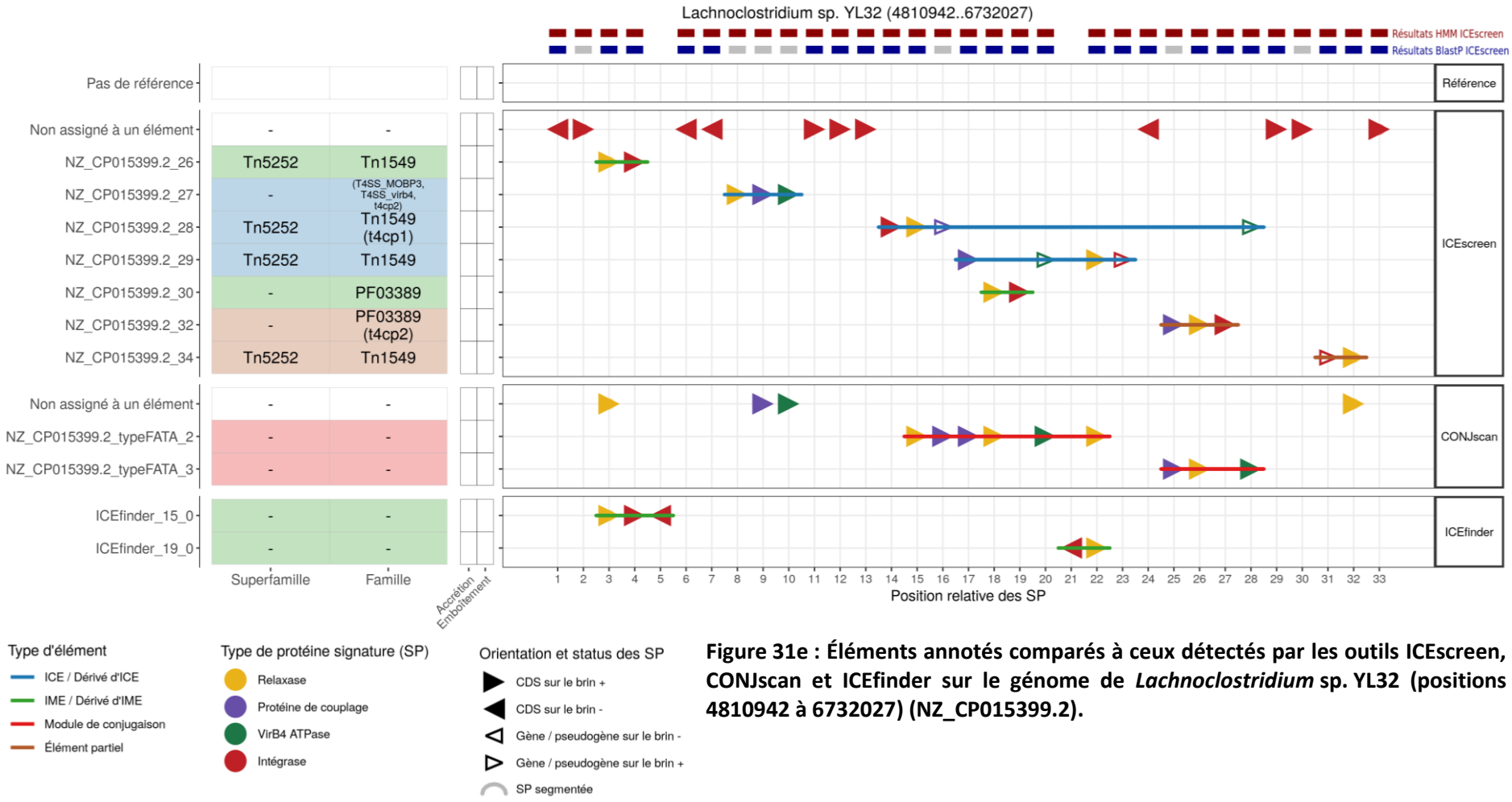

Figure 31e : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Lachnoclostridium* sp. YL32 (positions 4810942 à 6732027) (NZ_CP015399.2).

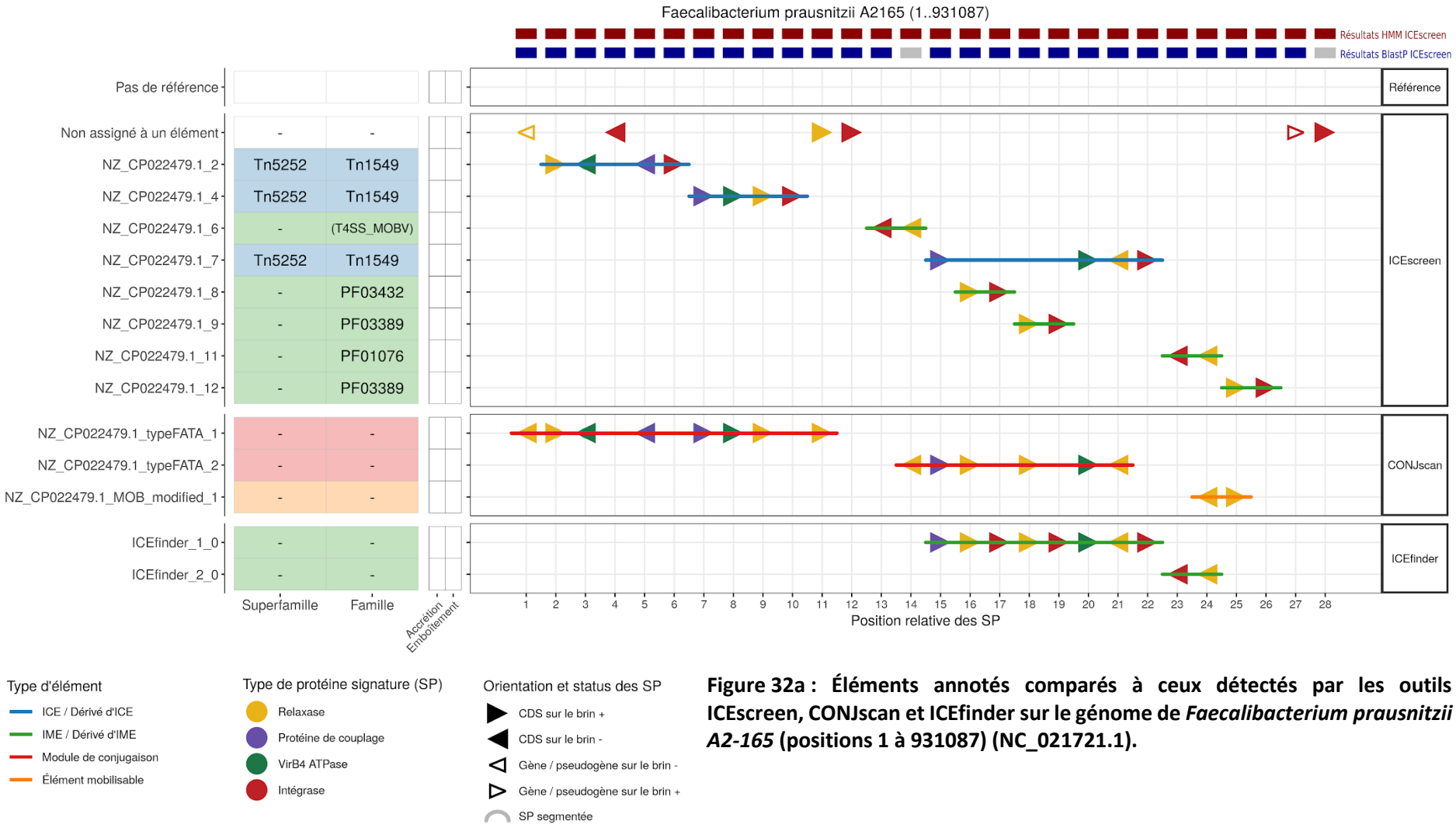

Figure 32a : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Faecalibacterium prausnitzii A2-165* (positions 1 à 931087) (NC_021721.1).

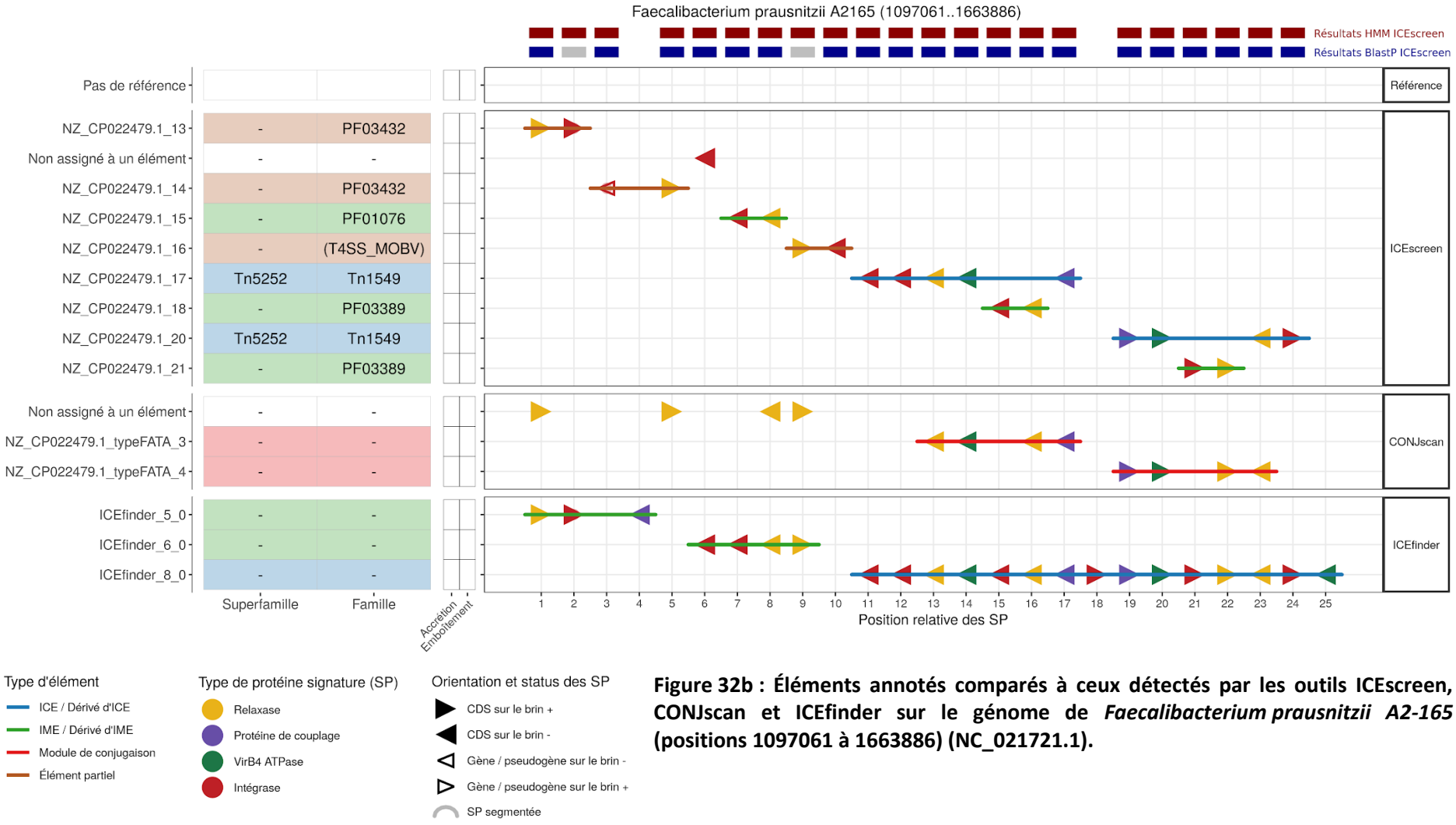

Figure 32b : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Faecalibacterium prausnitzii A2-165* (positions 1097061 à 1663886) (NC_021721.1).

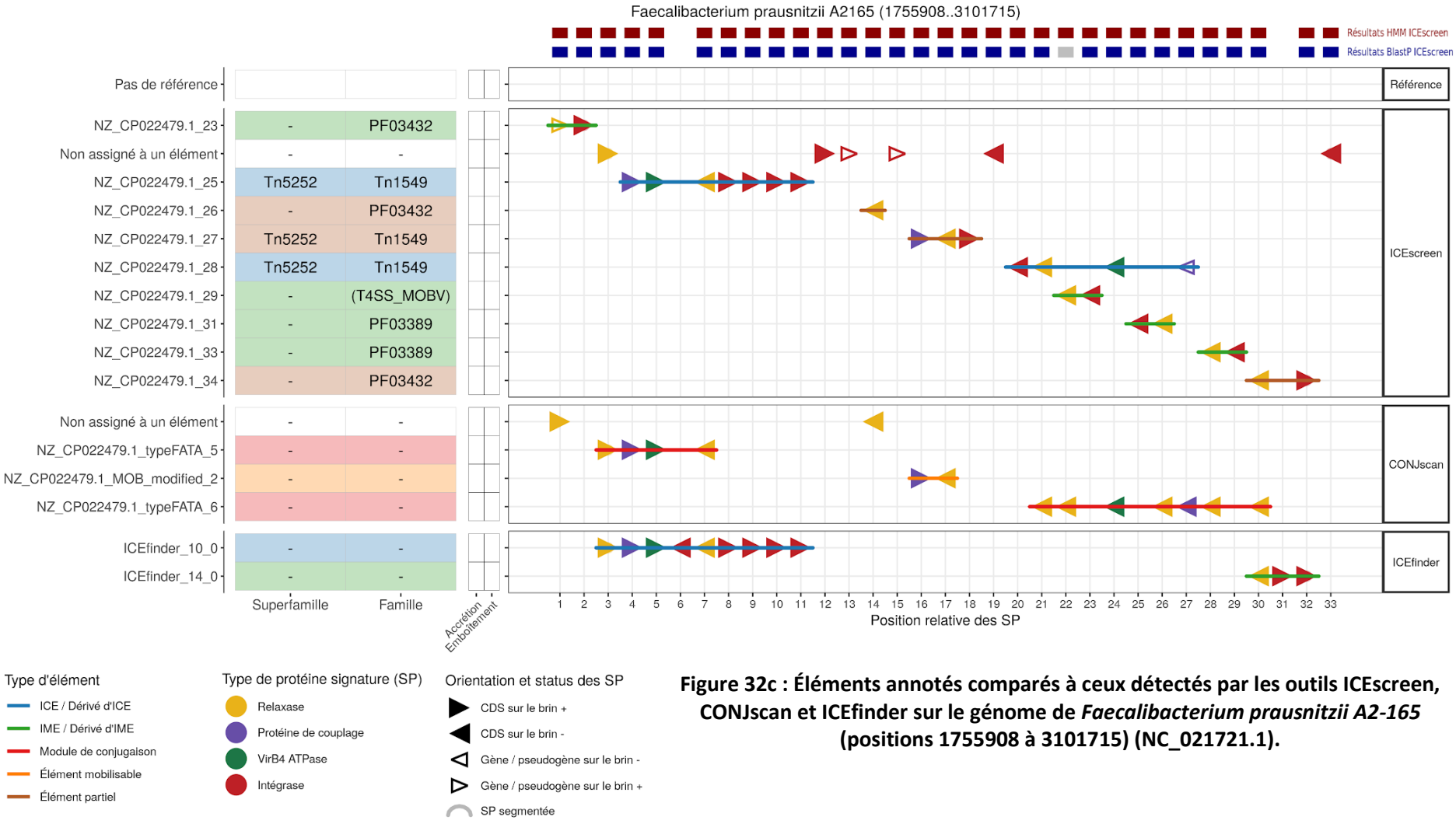

Figure 32c : Éléments annotés comparés à ceux détectés par les outils ICEscreen, CONJscan et ICEfinder sur le génome de *Faecalibacterium prausnitzii A2-165* (positions 1755908 à 3101715) (NC_021721.1).

Conception et la mise en œuvre d'une méthodologie bioinformatique dédiée à l'identification des ICE, IME et éléments composites dans les génomes de Firmicutes

Les ICE (Éléments intégratifs conjugatifs) et les IME (Éléments intégratifs mobilisables) sont des éléments mobiles bactériens qui jouent un rôle clé dans les transferts horizontaux. Ils ont la capacité de s'intégrer et de se transférer par conjugaison d'une bactérie à une autre. Ces éléments sont très répandus dans les génomes bactériens mais sont encore mal connus. Leur identification automatique est un défi et ils ne sont en général pas annotés dans les génomes. Jusqu'à présent, seules deux approches bioinformatiques permettent la détection des ICE et la détection des IME, mais leur fiabilité reste très variable, en particulier chez les Firmicutes. De plus, aucune de ces approches ne permet de détecter avec précision les éléments composites constitués d'ICE et d'IME emboîtés ou en accrétions, qui sont fréquemment observés dans des génomes bactériens. Nous avons développé une stratégie et un outil nommé ICEscreen permettant d'identifier les ICE et IME dans les génomes des Firmicutes, y compris les éléments emboîtés ou en accrétions. Notre outil, commence par la détection de quatre protéines signatures (SP) indispensables au fonctionnement de ces éléments puis effectue la détection et le typage des éléments à partir de la colocalisation des SP et de la caractérisation de leur contenu. Notre outil utilise un algorithme dédié permettant de résoudre la structure des éléments qu'ils soient composites ou non. Pour réaliser ces étapes nous avons construit une banque de protéines signatures d'ICE et d'IME de référence à partir d'une liste de gènes connus pour être impliqués dans la dynamique de ces éléments chez les streptocoques ainsi que de profils HMM publics ou construits pour cette étude. Pour valider les résultats d'ICEscreen nous avons construit un jeu de données, FirmiData, constitué de 40 génomes de Firmicutes pour lesquels les ICE et IME ont été annotés semi-manuellement et expertisés. Nous avons ensuite comparé les résultats de ICEscreen avec ceux de deux outils de référence : CONJscan et ICEfinder. ICEscreen détecte la quasi-totalité des éléments de la référence (96 %) ce qui en fait un outil plus performant que CONJscan (58 %) et surtout ICEfinder (53 %) sur notre jeu de données. ICEscreen est ainsi un outil d'aide à l'annotation et à la découverte d'ICE et d'IME dans les génomes de Firmicutes, ce qui peut aider à mieux caractériser leur contribution aux transferts horizontaux de gènes, notamment lors de la transmission de la résistance aux antibiotiques, auxquels ils sont fréquemment associés.

Mots-clés : Éléments intégratifs conjugatifs (ICE) ; Éléments intégratifs mobilisables (IME) ; Éléments génétiques mobiles ; Firmicutes ; Bioinformatique ; Annotation

Design and implementation of a bioinformatics methodology dedicated to the identification of ICEs, IMEs, and composite elements in the genomes of Firmicutes

ICEs (Integrative Conjugative Elements) and IMEs (Integrative Mobilizable Elements) are bacterial mobile elements that play a key role in horizontal transfers. They have the capacity to integrate and transfer by conjugation from one bacterium to another. These elements are widespread in bacterial genomes but are still poorly understood. Their automatic identification is a challenge and they are generally not annotated in genomes. So far, only two bioinformatic approaches allow the detection of ICEs and IMEs, but their reliability remains highly variable, particularly among Firmicutes. Moreover, neither of these approaches can accurately detect composite elements consisting of nested or accreted ICEs and IsMEs, which are frequently observed in bacterial genomes. We have developed a strategy and a tool called ICEscreen to identify ICEs and IMEs in the genomes of Firmicutes, including nested or accreted elements. Our tool starts with the detection of four signature proteins (SPs) that are essential to the functioning of these elements and then carries out the detection and typing of the elements based on the colocalization of the SPs and the characterisation of their content. Our tool uses a dedicated algorithm to solve the structure of the elements whether they are composite or not. To perform these steps, we have built a bank of ICEs and IMEs signature proteins from a list of genes known to be involved in the dynamics of these elements in streptococci and also public HMM profiles and HMM profiles constructed especially for this study. To validate the ICEscreen results, we built a dataset, FirmiData, consisting of 40 genomes of Firmicutes for which the ICEs and IMEs were annotated semi-manually and curated. We then compared the results of ICEscreen with those of two reference tools: CONJscan and ICEfinder. ICEscreen detects almost all the elements of the reference (96%) making it a more powerful tool than CONJscan (58%) and especially ICEfinder (53%) on our dataset. ICEscreen is thus a tool for the annotation and discovery of ICE and IME in the genomes of Firmicutes, which can help to better characterize their contribution to horizontal gene transfers, particularly during the transmission of antibiotic resistance, with which they are frequently associated.

Keywords: Integrative Conjugative Elements (ICEs); Integrative Mobilizable Elements (IMEs); Mobile Genetic Elements; Firmicutes; Bioinformatics; Annotation