**Bilan des modifications de la base de données des SP d’ICEScreen**

Ce bilan des modifications de la base des données inclut :

1. De très nombreuses propositions de modifications de la base de SP que j’avais proposées de faire à Julie avant le 7/07/2022 ou à Charles et que j’ai retrouvées, mais de date réelle inconnue, dans l’ordre des dates supposées. Comme la réalisation était incertaine a priori, j’ai vérifié que la modification avait bien été faite (vérification faite le 7/07/2022), ce qui était effectivement fait dans l’immense majorité des cas. Les modifications proposées mais apparemment non faites (à faire ?) sont en rouge.

2. Les modifications faites que j’ai finalisées moi-même le 7/07/2022 sur le fichier envoyé par Thomas. Les modifications futures seront incluses dans l’ordre des dates.

3. Les faux positifs signalés à l’occasion de ces demandes de modification dans l’ordre des dates. clairement non exhaustif.

4. Des propositions/pistes d’amélioration ultérieures non seulement pour la base de données mais aussi pour ICEScreen

NB : De nombreuse autres modifications ont été faites / demandées dans d’autres fichiers utilisés par ICEScreen, non répertoriées ici (site d’intégration, nom des éléments codant les SP…).

**1. Modifications de la base de SP avant le 7/07/2022 (dates imprécises, très probablement incomplet)**

On est parti initialement de fichiers dits Ref0, à partir de données publiées, données très limitées ? Ces séquences provenaient d’ICE, IME et pour faire nombreou certains éléments de nature indéterminée qui pouvaient en être (éléments SCCMec). Les données ont été compilées par Sophie et par moi, pour des organismes et/ou éléments différents (Sophie, si mes souvenirs sont bons pour les éléments les mieux connus comme SCCmec ou Tn916). Cependant notre approche était différente. Sophie avait vérifié le degré de ressemblance des séquences, et n’avait gardé qu’une séquence si elle avait des séquences très proches, ce qui avait limité le nombre de SP par exemple pour Tn916 ou les éléments SCCMec, tandis que je les vais tous pris sans vérifier le degré de ressemblance. A noter que ces données n’étaient finalement pas exhaustives, car les premiers IMEs de firmicutes connus, Tn4451 et apparentés n’avaient pas été inclus par inadvertance. Cette banque utilisait des numéros Uniprot.

**1.1 Du temps de Charles**

**- Ajout des protéines venant des ICE et IME des 124 génomes de streptocoques**

- **Elimination de toutes les redondances > 95% id,** en ne gardant qu’une protéine par groupe, aussi bien dans les séquences venant de streptocoques que dans Ref0, en essayant de garder les séquences venat des éléments les mieux connus ou venant d’éléments apparemment fonctionnels.

- **Retrait de toutes les Int DDE de Ref0 de famille IS30** qui détectaient essentiellement des transposases d’IS et ne permettaient pas de savoir si les séquences détectées étaient des gènes ou pseudogène du fait des annotations incorrectes des CDS (décalage du cadre de lecture non pris en compte dans GenBank, START imaginaires). Ces séquences pouvaient toutes correspondre de surcroît à des pseudos et/ou des séquences incomplètes ou partiellement fausses (décalage du cadre). Numéros Uniprot retirés B6V376, D2N3A9, D2N3B0, D2N8C5

- **Retrait des Int DDE de Ref0 de famille IS1595** (détection de MGE dont il est impossible de savoir si ce sont des IS ou des IME sans relaxase). Numéros Uniprot A9VV42, A6CNB6, C9WIW2, A9KPZ8, Q4KRH3

- **Remplacement des numéros Uniprot, par des numéros GenBank.**

**- Ajout de séquences venant d’Actinobactéries pour rechercher les AICE.**

**1.2. Modifications anciennes de date indéterminée sur les protéines de Ref0**

- **Attribution de CAR69098 à la famille Tn916 au lieu de Tn5252** (vient d’un Tn916 guest au lieu de l’ICE host ICESp23FST81

**- Retrait de CAL92505.1 relaxase d’un Tn916** présentant un START faux et possédant une fusion en phase avec **répresseur oméga**, ce qui pouvait fausser complètement les taux de couverture trouvés, ce domaine oméga n’ayant aucune fonction dans la protéine et pouvant être absent des protéines détectées.

- **Retrait d’AAC98428.1, intégrase de Tn5252 originel. Erreur de séquençage** des auteurs pour cet ICE. Cette erreur provoque un décalage du cadre de lecture en 5' (dans le deuxième codon décrit) et une troncature du début de la séquence protéique.

**1.3 Modification en 2019**

**\* Retrait de la relaxase AGY40713 d’IME\_SspIG2\_tRNAleu qui a un problème de START** **probable (trop court)**, ce qui a causé des problèmes à répétition. J’avais proposé une protéine de remplacement (WP\_038675027.1 d’IME\_SsaJF\_tRNALeu dont le START est correct (famille PF02486\_5, 97 % id). Retrait effectué, mais apparemment ajout non fait. Cependant, Un blastP d’**AGY40713 vs** cette banque montre qu’elle contient 2 protéines présentant 96% id (AGY37395.1 et CCC18877.1), ce qui implique i) **que le toilettage à 95 % est incomplet**, ii) que l’ajout de WP\_038675027.1 est inutile.

**\* Retrait de la relaxase CCC20892 de dIME\_SthJIM8232\_tRNAlys (pseudo) et ajout pour remplacer, d’une relaxase 59% id AAV61548** d’IME\_Sth18311\_tRNAlys (PF02486\_2),

**\* Retrait de ZP\_06474984, YP\_003378314, YP\_004082809 et YP\_003381205 (soi-disant relaxases) d'actinomycètes.** ZP\_06474984 « relaxase » de Frankia n’était en rien une relaxase ou un initiateur RCR mais une DNA polymérase I. de même, YP\_003378314, YP\_004082809 de Micromonospora et YP\_003381205 de Kribbella flavida ne sont pas ni des initiateurs de réplication RCR ni des relaxases mais des primase/polymérases. Provoquent la détection de primase/polymérase de phages de firmicutes, ce qui fait passer des prophages pour des IME du fait de la présence d’une intégrase. En fait, tout ce qui a pour meilleur query une protéine d’actinomycète chez S salivarius est une MERDE (avec malheureusement des E corrects) et non une « relaxase » (heureusement nombre de cas totaux faibles, exclusivement des « relaxases ». Aucune détection utile. **A la suite de cela les autres protéines d'actinobactéries introduites par Charles ont été virées de la banque utilisée pour les Firmicutes (mais gardés pour les actinobactéries) car aucune n'avait été vérifiée quant à la nature et/ou la qualité des résultats chez les firmicutes**. Toute incorporation de protéine étrangère aux Firmicutes pour une analyse chez les firmicutes est dangereuse et doit être faite avec précaution.

\* **Ajout de la relaxase d’IME\_oriT** (EAO62073 d’IME\_Sag18RS21\_oriT), type de relaxase mobT non présente dans les 124 génomes de streptocoques.

\* **Ajout de la relaxase WP\_118058175.1 d’IME\_SsaAF10-23\_rpsI**, S. salivarius AF10-23 PF01719-PF00910 (nouvelle famille).

**\* Ajout de la CP WP\_175063099 identique à celle de CDS\_1741 d’IME\_SsaT93\_rpsI**, S. salivarius T93, nouvelle famille de TcpA, TcpA\_13)

**\* Ajout de la CP WP\_004183002.1 d’IME\_SaM18\_rplL**, S. salivarius M18, family PF01719-PF00910 (TcpA, 27% id avec meilleure query, nouvelle famille de TcpA TcpA\_14,)

\* **Ajout de la CP WP\_038675054.1 d’IME\_SsaJF\_ND**, (nouvelle famille de VirD4 42% id avec meilleure query)

\* **Remplacement de family PF13814 par PF02486 dans l’annotation de la Relaxase AGK71980**.

**1.4. Modifications apparemment demandées à Julie avril-juin 2021, ou confirmation de modification demandées en juin 2021**

- **Complément d’annotation des protéines de Ref0.** Au 7/07/2022, il ne semble plus y avoir de différences de qualité par rapport aux protéines ajoutées ultérieurement.

- **Retrait d’intégrases valables mais sans rapport avec ICE et IME**

- int ser CAJ68758 de Ref0 n’appartient pas à un ICE (introduite comme celle de CTn1). Adjacente à l’int réelle qui n’avait pas été annotée de façon correcte dans le génome (extérieure à CTn1, nature de l’élément porteur indéterminé) , ce qui avait conduit à l’erreur. L’intégrase réelle fait 1612 aa du fait qu’att est interne au gène et provoque la fusion avec le gène « ciblé » (site secondaire).

- Int tyr AGP68019 de Lactobacillus paracasei Lock 919 présente dans ref n’est pas une protéine d’ICE ou IME

- **Vérification du retrait de protéines ref tronquées ou de pseudogènes**

- Int Ser BAN60853, tronquée pour la partie C terminale

- Int Ser CAP20376 de dICE\_Seq4047\_traG.

- CP VirD4 CAP20376.1 venant en fait de la traduction d’un pseudogène.

**2 . Modifications de la base de SP du 22/07/2022 et ultérieures**

**2 .1. Modifications faites au 22/07/2022**

-Précédemment pour les **IME ayant 2 relaxases potentielles**, les familles (ou plutôt superfamilles de relaxases) affichées pour la SP étaient non celle de la SP mais celles des deux relaxases, ce qui pouvait conduire à des aberrations donnant comme famille PF01076+PF02486 et MobV+MobP comme famille pour une SP de superfamille PF01076 (MobV) ou de superfamille PF02486 (MobT), et pire PF01076+PF02486 pour un module de mobilisation provenant d’un résidu de plasmide intégré possédant une relaxase PF01076 (MobV) et une PF01446 !

**\* remplacement de famille PF01076+PF02486 par PF02486 et de famille MobV+MobT par** **mobT** pour AAM99131, ABA45702, AFS45036, AGU82204, AGU83894, BAN61292, CAD45862, CBZ47880, CCW37084

**\* remplacement de famille PF01076+PF02486 par PF01076 et de famille MobV+MobT par** **mobV** pour **ABA44955 (et non)** **ABA44681**, AFS45038, AGU82202, AGU83896, BAH80804, CAD45864, CBZ47878, CCW37086

- **Suppression de l’attribution indue de famille de relaxase PF01441 à deux SP qui ne sont pas des relaxases** mais qui appartenaient à un IME codant une relaxase PF01446 : protéine de couplage AGP68114 et intégrase AGP68118 et

**- Remplacement de rep\_1 par PF01446 dans les familles de l’élément codant la protéine de couplage AGP68114, l’intégrase AGP68118 et la relaxase AGP68116 pour harmoniser la nomenclature.**

- **Correction de l’erreur de numéro de domaine** **PF01441** **par le bon numéro PF01446** pour la relaxase AGP68116

**- Retrait des intégrases à sérine des MGI SCCmec** (BAA86648, BAC67561, BAG06212) dont les mécanismes de transfert sont toujours inconnus, qui avaient été inclus dans Ref0, et n’ont jamais permis de détecter des intégrases d’ICE ou IME, mais uniquement celles d’autres SCC ou d’éléments de nature inconnue. Ce sont peut-être des IME sans relaxase, mais on en connaît désormais beaucoup d’autres dont on n’a jamais inclus les intégrases dans la banque de SP. Donc, pas de vraie raison de les maintenir. A noter que ces intégrases à sérine sont toujours en duo, présentent une cible jamais trouvée pour les ICE et IME et peuvent provoquer des accrétions de SCC.

- **Attribution de CP BAN75392, Relaxase BAN75394, VirB4 BAN7540, Intégrase BAN75415 à un IME de famille Tn916** au lieu de la famille ICE\_Lcas393\_rpsI, les CP, relaxase et VirB4 ayant respectivement 59% id, 46% id et 66% id avec celle de Tn916.

- **Remplacement de la superfamille PF05840/PHA00330 de l’IME** codant l’intégrase CCW3779, Relaxase CCW37800, CP CCW37801 ainsi que de **la superfamille PF05840/PHA00330 de la relaxase CCW37800** **par PHA00330.** Selon la banque et un CDSearch, CCW37800 possède un domaine PF05840 et non PHA00330. Cependant les analyses effectuées lors des travaux pour l’article sur les 124 génomes de streptocoques ont démontré que les relaxases ayant ce domaine sont apparentées sur la région correspondant à ce domaine à la région des relaxases ayant le domaine PHA00330. Donc considéré comme une famille particulière des relaxases PHA00330 dans la publi, nomenclature à reprendre dans ICEScreen.

- **remplacement d’AGZ23064 qui a un mauvais START** (il manque les 26 premiers aa, le vrai START est un UUG en amont de l’AUG qui avait été retenu) **par WP\_041179472.1** provenant de l’annotation RefSeq qui est correcte. Cette intégrase ciblant SNF2 ne présentait aucune séquence proche dans la base.

>WP\_041179472.1 recombinase family protein [Streptococcus suis]

MNRTKKYIAALYCRLSKDDGSTNESMSIYSQKAMLKQYAEQNNIAVYDYYVDDGYSGTNFERPAFKKMIT

DIENGKINCVITKDLSRLGRNYLESGAYIEMYFPQKNVRYIAITDGIDTINSYEFDIMPFKNILNEMYAK

DTSKKVKSALKSRMKEGTYIGSKAPFGFKKDPDDKHRLVIDERVKPIIELVYELCLEGKGTQLISQEMMK

RKIPRPSSFLENADKLYGLTEENKYKWTHRMVLSILRDPVYCGNMERNKRPTLSFKNRKRLYVSKEDRIV

VKDTHEGIVSEEVWTQVQQMLDKRKNTNKSGITYDNIFKGLVKCPTCGYALTPKADYRLNKKETIDFVHF

SCSGYKKYGVNACTSHRINARDLYNVVLEDIQYHGQMALSSREDFVMKIAEKIDKDKVDERKDKTEKLNS

SKIKLKDLDKAFEKLYEDRLSESISERNFNLMNEKLSRQQEKLIEEIDLLEEEIKEIADTEENCEQFVEN

ISKFAKIKELNRYILNQVIDKIYVYDKEEVDGEIKQKVEIHYKFIGKLD

**3. Faux positifs**

**3.1 2019**

**\* XerS** détectée dans la plupart des génomes de streptocoques et homologue de XerS dans les mégaplasmides, Eliminée actuellement par comparaison avec une XerS. Attention, des homologues de XerS sont utilisées comme Int ciblant dif dans des ICe de L. lactis et pourraient être rejetées lors de l’analyse du génome (mais Xers avait été incluse dans l’ICE en plus de l’int lors de l’analyse de L. lactis IO-1)

\* Détection inattendue mais normale : En plus de la présence d’intégrases chez les prophages et MGI, **initiateur de réplication PF01719-PF00910 trouvé dans un prophage** de L11 (25% id avec query)

\* plusieurs faux positifs dans une première analyse des génomes de S salivarius puis non dans une 2ème sans raison claire dans le bilan considéré (paramétrages différents ?)

**3.2. 2020 – 2021 (date pas claire, plutôt 2020)**

\* WP\_011861425.1 de difficile 630 détecté comme relaxase mobC par HMM, alors que ce n’en est clairement pas une relaxase i) E un peu marginale 2E-10, ii) Aucun domaine mobC selon la banque, iii) domaine central HTH. Yn domaine HTH\_MARR de liaison à l'ADN décrit dans les régulateurs transcriptionnels de famille MarR (2.2e-19). Cependant, coïncidence des plus étrange : une protéine de Tn5398, élément mobilisable, codée par le gène adjacent de la seule séquence homologue de l'ICE mobilisateur probable, détectée par HMM comme MobC

**\* WP\_013245814.1** de Lactobacillus paracasei Lock 919 détectée comme CP VirD4 (E 10-6). Non associé au moindre système de conjugaison ou intégration. Possède un domaine apparenté aux domaine recherchés de CP et pourrait correspondre à FtsK elle-même. Meilleure E : domaine tronqué FtsK COG1674 =2.2x10-150 avec un domaine C terminal inclus Ftsk\_gamma 2.5x10-22 que je n'avais jamais vu dans les TcpA ni VirD4 mais uniquement dans les FtsK. Dans le même génome une autre protéine protéine "FtsK" non détectée, WP\_003569507, non associée à des modules de conjugaison ou intégration. Possède une longueur et architecture similaire avec un domaine tronqué FtsK COG1674 =2x10-167 et un domaine C terminal Ftsk\_gamma 1x10-33. Ces 2 protéines possèdent des domaines chevauchants de CP TrwB\_TraG\_TraD\_VirD4 (7x10-10 et 4.1x10-07) et FtsK/SpoIIIE (8.3e-66 et 7.8e-60).

**\* WP\_032800909.1** de Lactobacillus paracasei Lock 919 détecté comme Int par une XerS comme meilleure query.

**4. Pourrait être fait pour compléter la base de données ou la recherche, ce qui pourrait améliorer les performances d’ICEScreen ou faciliter son utilisation :**

- **Ajout du nom de l’élément d’origine de la SP.** Je crois que cette information est disponible dans un autre fichier utilisé par ICEScreen, puisqu’elle apparaissait dans la version privée d’ICEScreen, mais elle **manquait pour les Ref0**. Est-ce encore le cas ?

- **Spécificité d’intégration** (gène ciblé, position dans le gène ciblé longueur du DR), **positions du gène de l’intégrase** par rapport au site d’intégration ou extrémité de l’élément. Je crois que cette information est ailleurs puisqu’elle apparaissait dans la version privée d’ICEScree, mais elle manquait pour les Ref0. **Elle n’est cependant pas complète non seulement pour tous les éléments ref0 mais aussi pour la plupart des autres.**

- **Protéines à ajouter en provenance d’éléments trouvés plus** **récemment** chez S salivarius (apparemment fait du moins pour les nouvelles familles, donc un cut à 40%), S suis (demander à Sophie), de l’analyse des génomes de FirmiData, de l’analyse des génomes contenants des éléments publiés, des éléments publiés, y compris ceux de Mycoplasma.

- **Toilettage diminuant le nombre de SP pour ne retenir qu’une par groupe de % identité de plus de 90 %** (si ce n’est déjà fait) voire 80 % au lieu de 95%

- **Ajout de la famille pour les relaxases et CP de d’IME, comme nous l’avons fait pour les ICE**. A priori peu utile pour le fonctionnement d’ICEscreen car pas ou peu de corrélation entre famille de CP et d’IME, mais uniquement entre superfamilles déjà indiquées. Serait utile dans le cas des IME à 2 relaxases, ceci correspondant à une famille particulière de PF02486 et une famille particulière de PF01076.

- Distinction de familles ou superfamilles distinctes de relaxases pour MobT et **HTH-MobT** ?

- inclure les **petites recombinases à sérine des trios ou quatuors** et les rechercher en HMM et Blast.

- **Inclure des profils HMM supplémentaires** pour les relaxases qui ont le domaine « MobM » et non PF01076 (Faecalibacterium, MobM a une définition plus large que PF01076, mais les HMM PF01076/MobV les trouvent quand même); PRK13878 et non PF03432 (Faecalibacterium, PRK13878 a une définition plus large que PF03432/MobP, et **ICEScreen HMM et Blast ne les trouve pas**), relaxases mobT sans domaines PF02486 (Streptocoques, peut-être déjà fait), et PF01446 (Lactobacillus, **HMM ne les trouve pas**). Inclure les « MobM » de plasmides pCW3 ou d’ICE qui les auraient si on en trouve (**seraient peut-être trouvées comme des intégrases à tyrosine par ICEScreen HMM** !)  ou des relaxases proposées pour pXO16 ?… HMM avec pour base **VirD4 étrange de dIME/dICE\_Fpr94230-2\_tRNAlys\_1\_1** ?

- Rechercher **d’autres protéines du TIVSS ? de protéines de réplication autres ?**

- **Inclure et rechercher les transposases à DDE IS30 en Blast et HMM**, mais avec des règles très particulières, par exemple ne les rechercher qu’après tout le reste ou ne les associer qu’à des positions fixées ou qu’avec certains types d’éléments, sous peine de graves erreurs et de tomber dans l’un des travers d’ICEFinder. Faire attention que le HMM ne trouve pas les autres familles de Tnp DDE sous peine de catastrophe.

- attention, **certaines int Tyr d’IME sont internes et/ou dirigées vers l’intérieur**

**- inclure la recherche de gènes cibles de localisation appropriée** quand l’intégrase présente > 50% id avec une intégrase déterminée (gène codant une protéine), plus de 70% si int Tyr ciblant un gène de tRNA avec un anticodon particulier, 50% IntSer ciblant un tRNA (n’importe lequel pourvu que le gène soit en position) et des **DR associés.**

- **modifier en fonction des ICE et IME caractéristiques nouvelles, provenant des autres organismes de FirmiData, en particulier de de Faecalibacterium (détaillés ici)**

\* 1) Nombreux IME codant une **intégrase à sérine du côté opposé du gène de tRNA** (du jamais vu). Presque tous ces IME sont mal détectés (les isolés sont détectés comme éléments partiels) ou non détectés par ICEScreen (faux élément constitué d'un gène de l'IME et d'un gène d'élément en accrétion, ou intégré dans l'IME, en particulier des MGI...). A noter, qu'une **grande partie de ces gènes d'intégrase sont considérés par refseq comme des pseudogènes** en raison de la recombinaison interne. Heureusement, ils sont quand même détectés par ICEScreen, mais malheureusement de ce fait comme pseudogènes.

2) quelques éléments portant une **intégrase à sérine adjacente et convergeant avec le gène de tRNA** (du jamais vu).

3) la majorité des éléments codant des intégrases à sérine possèdent des DR courts (6 nt). La comparaison des intégrases de ces 4 souches, confirmée par l'analyse du contexte des intégrases quasi-identiques de GenBank, montre clairement que des **intégrases présentant > 99% id sont codées par des éléments intégrés dans des tRNA très différent**s (acide aminé reconnu différent, ou même acide aminé reconnu mais anticodon différent de l'ARNt et séquences éloignées des ARNt). Si certains de ces cas peuvent être interprétés par des intégrations dans un site secondaire (?), dans d'autres 2 sites d’intégration distincts (voire plus ?) sont répandus. Cette situation rappellerait un peu l'intégration spécifique d'IME présentant des intégrases à sérine quasi-identique dans SNF2 ou PPI, 2 gènes adjacents non apparentés des ICE de famille de Tn1549, que Sophie avait vu dans les données de S. suis.

4) quelques éléments codant des **intégrases à tyrosine du côté opposé du gène de tRNA** (du jamais vu)

**\* rechercher et détecter les SP interrompus des IME, Introns ou ICE… (impliquerait de faire des tBlastN).** Ceci inclut surtout les VirD4 Tn5252 et Tn1549 (IME et Introns), mais inclut aussi d’autres SP.