UNIVERSITE DES ANTILLES
U.F.R Sciences de la Vie et de la Terre

THESE
Pour l'obtention du grade de
DOCTEUR DE L'UNIVERSITE DES ANTILLES
Discipline : Bio-informatique

Ecole Doctorale Pluridisciplinaire de l'Université des Antilles
Présentée et soutenue publiquement par

Christophe NOROY
Le 14 mai 2018

Analyse de la plasticité génomique des bactéries de la famille des *Anaplasmataceae* en lien avec les effecteurs du système de secretion de Type IV

Thése de Doctorat Présentée devant le jury composé de :

FICHANT Gwennaële – Rapporteur –
Professeur de l'Université Paul Sabatier, Toulouse
ARLAT Matthieu – Rapporteur –
Professeur de l'Université Paul Sabatier, Toulouse
GUYOMARD Stéphanie – Examinateur –
Chargée de recherche 1ère classe
TEYCHENEY Pierre-Yves – Examinateur –
Directeur de recherche 2ème classe
REYNAUD Yann – Examinateur –
Chargé de recherche de 1ère classe
GROS Olivier – Directeur –
Professeur de l'Université des Antilles, Guadeloupe
MEYER Damien – Co-Directeur –
Chargé de recherche 1ère classe, Guadeloupe

Recherches effectuées au sein de l'UMR CIRAD-INRA ASTRE,
- Animal, Santé, Territoires, Risque et Ecosystèmes –
Domaine de Duclos, 97170 Petit-Bourg, Guadeloupe

Financement de thèse : Union européenne (Projet Malin, Epigenesis), et CIRAD.

RÉSUMÉ

Identifier les effecteurs du système de sécrétion de type IV (SST4) des Anaplasmataceae, et élucider leurs fonctions à l'intérieur de la cellule hôte pour manipuler les voies de signalisation et l'immunité de l'hôte, est crucial pour concevoir des alternatives thérapeutiques contre ces bactéries zoonotiques. De même la plasticité génomique des bactéries intracellulaires obligatoires, via les transferts géniques horizontaux ou des évènements de recombinaison, semble jouer un rôle majeur dans l'évolution de cette bactérie, son adaptation à des conditions environnementales changeantes et la colonisation de nouvelles niches écologiques (hôtes). En utilisant les outils déjà disponibles et en développant de nouvelles méthodes analytiques, l'objectif de cette thèse a été de mieux comprendre quels sont les facteurs génomiques associés à la virulence bactérienne mais aussi à la spécificité d'hôte. Plusieurs approches menées en parallèle ont permis i) de prédire les effecteurs du systéme de sécrétion de type IV (SST4) chez les bactéries, ii) d'identifier les répertoires d'effecteurs en lien avec la virulence bactérienne au sein de l'espèce *Ehrlichia chaffeensis*, iii) d'identifier les répertoires d'effecteurs en lien avec la spécificité d'hôte au sein du genre *Ehrlichia* et iv) de développer de nouvelles méthodes d'analyse de l'évolution de l'architecture des génomes et mieux comprendre le concept de plasticité génomique.

Nous avons développé une seconde version (S4TE 2.0) du logiciel S4TE (Searching Algorithm for Type 4 secretion system Effectors) qui a consisté en la création d'une interface web et de nouveaux programmes, la construction de plusieurs bases de données et la gestion de leurs interactions entre elles, le site web, et le cluster de calcul. Ce logiciel permet des prédictions de qualité des effecteurs du SST4 mais comporte aussi des outils de génomique comparative. Le travail a ensuite consisté en l'identification des répertoires d'effecteurs chez Ehrlichia chaffeensis, dont les souches présentent des niveaux de virulence différents alors que ces souches sont très conservées au niveau génomique. Grâce au logiciel S4TE2.0, nous avons pu prédire les répertoires d'effecteurs pour chaque souche d'E. *chaffeensis* et mettre en évidence la présence d'un effecteur spécifique d'une souche (Liberty), qui pourrait avoir un lien avec la virulence accrue observée chez cette souche. En confrontant la liste des effecteurs d'E. *chaffeensis* à une base de données (HPIDB) pour prédire l'interaction avec certaines protéines de la cellule

hôte ainsi que leur localisation subcellulaire (CELLO2GO), nous avons mis en évidence que la plupart des effecteurs avec des domaines de localisation nucléaires (NLS) prédits par S4TE 2.0 présentaient des cibles ayant une localisation nucléaire. De nombreuses cibles d'effecteurs semblent avoir un lien direct avec la manipulation de la cellule hôte. Dans un troisième volet visant à étudier la spécificité d'hôte du genre Ehrlichia, nous avons comparé les effectomes des différentes espèces du genre *Ehrlichia*. Nous avons prédit les effecteurs en utilisant le logiciel S4TE2.0 puis utilisé de nouveaux outils d'analyse de données complexes pour explorer les relations synténiques entre gènes (logiciel Circos) et visualiser les réseaux de manière rationnelle (logiciel Hive plots). L'ensemble de ce travail a permis de comparer les répertoires d'effecteurs du SST4 entre différentes souches d'Ehrlichia et de suggérer quels effecteurs spécifiques de chaque souche pourraient être liés à la spécificité d'hôte. Nos résultats mettent en lumière l'impact de la plasticité génomique liée aux répertoires d'effecteurs conservés et accessoires dans le pouvoir pathogène et l'histoire évolutive des bactéries pathogènes intracellulaires de la famille des Anaplasmataceae.

Enfin, nous avons exploré l'utilisation de méthodes d'écologie spatiale pour l'étude de la plasticité génomique en prenant la famille de gènes « effecteurs » comme modèle. Nous avons mis en évidence que certaines familles de gènes semblent être associées entre elles et semblent avoir des préférences d'habitat. L'écologie spatiale des génomes pourrait constituer un nouveau champ de recherches et permettre une nouvelle définition de la plasticité génomique.

MOTS-CLES : *Ehrlichia*, *Anaplasma*, plasticité génomique, système de sécrétion de type IV, prédiction d'effecteurs, S4TE2.0, spécificité d'hôte, évolution

---

DISCIPLINE : Bio-informatique

SOMMAIRE

INTRODUCTION GENERALE

La plasticité génomique est une notion biologique très vaste. De façon intuitive, la plasticité génomique se définit comme la capacité du génome à être remodelé. D'un point de vue plus moléculaire, la plasticité génomique est souvent définie comme étant la résultante d'insertions, de délétions, de mutations et de réarrangements dans la séquence nucléique du génome (Bennett, 2004). La plasticité génomique s'observe au travers d'évènements évolutionnaires passés. L'observation de la présence d'anciens évènements de recombinaison génétique indique ainsi que le génome a été plastique au cours de l'évolution d'un individu donné. Au delà de cet aspect purement qualitatif, il est possible de quantifier la plasticité d'un génome, chez les eucaryotes et quelques bactéries, grâce aux microsatellites (Jansen et al., 2012). Les microsatellites sont de courtes séquences composées de moins de cinq paires de bases et répétées afin de former de longues chaines dans les génomes. Les microsatellites sont des sites de variation génomique favorisant les évènements de contraction (délétion de nucléotides dans la séquence d'un gène) et d'expansion (insertion et duplication d'un gène) (Metzgar et al., 2002). Il est donc possible, en quantifiant le taux de microsatellites, d'avoir une idée de la plasticité d'un génome. Cependant, ces séquences répétées étant très faiblement représentées chez les bactéries, il est difficile de se rendre compte de la plasticité effective des génomes bactériens.

La plasticité génomique peut être retracée grâce à des évènements de recombinaison génétique ayant eu lieu dans le passé. Les mécanismes liés à ces évènements sont bien identifiés chez les bactéries et se décomposent en trois grands types : (i) les mouvements d'éléments génétiques mobiles (EGM), (ii) les transferts horizontaux de gènes (THG) et (iii) les évènements de réarrangements du génome.

Parmi les EGM, on retrouve en particulier les éléments transposables qui sont de petits segments d'ADN capable de se déplacer et de se répliquer dans le génome (Mit'kina, 2003). L'insertion de ces séquences dans le génome peut entrainer des mutations ou des duplications (Peterson, 2013). Un autre élément mobile important dans l'évolution et l'adaptation des bactéries à leur milieu est le plasmide bactérien. Les plasmides bactériens sont des molécules d'ADN distinctes de l'ADN chromosomique et capables d'assurer leur réplication de façon autonome. Ces plasmides ne sont pas forcément essentiels à la survie de la cellule mais permettent cependant d'améliorer le fitness de la bactérie. Les plasmides sont transmis entre les bactéries partageant une même niche écologique via transfert de gènes horizontaux (Madsen et al., 2012).

Les transferts horizontaux de gènes sont des processus d'entrée de gènes d'autres espèces, se produisant par transformation, conjugaison ou transduction (Dorman, 2014). Les THG peuvent être aussi bien létaux que bénéfiques pour la bactérie. Des études ont montré que lorsque l'acquisition d'un gène par THG était néfaste, il était supprimé, alors que lorsque l'acquisition apportait un gain, il était incorporé au génome (Mozhayskiy and Tagkopoulos, 2012). Une des stratégies exercées par les bactéries pour limiter les perturbations associées aux THG est l'atténuation transcriptionnelle (Dorman, 2014). Malgré les risques encourus, les bactéries permettent l'acquisition de gènes par THG afin de mieux faire face aux changements environnementaux. Il a été montré que les THG conduisent à une diversification des souches et à la formation de nouvelles espèces (Papke et al., 2015; Polz et al., 2013).

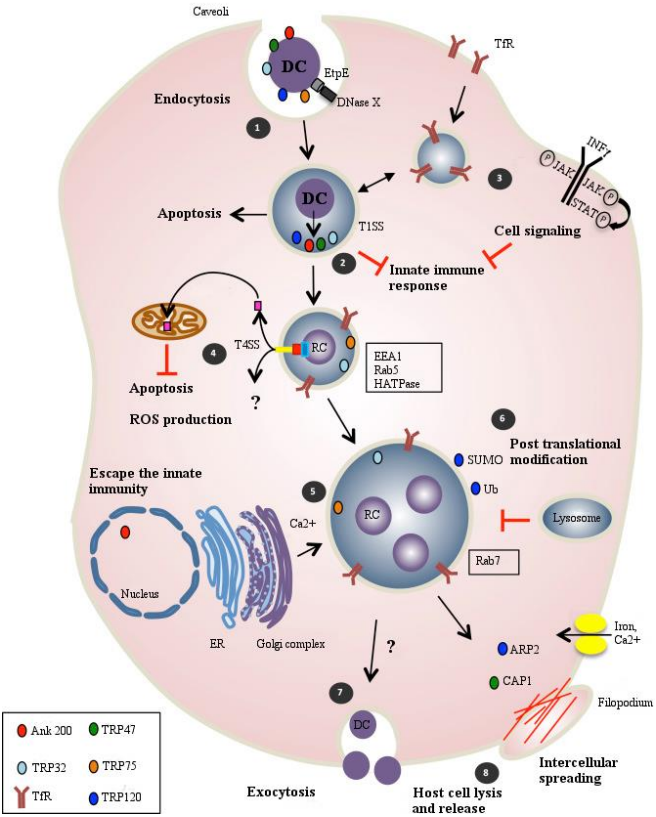


Figure 1. Cycle de développement intracellulaire d'*Ehrlichia* spp.

(1) Le corps élémentaire infectieux (DC) adhère et entre dans les cellules de mammifères en utilisant le récepteur EtpE qui se lie à la protéine DNase X. (2) Une fois entrée par endocytose, la bactérie se retrouve dans une vacuole bactérienne ressemblant aux endosomes précoces et sécrète des protéines via le système de sécrétion de type I (TRP 32, TRP47, TRP120 et Ank200) pour échapper aux réponses immunitaires innées de l'hôte. (3) Le corps élémentaire se différencie alors en corps réticulé (RC) réplicatif. A ce stade, les bactéries fusionnent avec l'endosome TfR pour acquérir du fer et perturber les voies de signalisation cellulaire comme JAK/STAT. (4) Dans le même temps, *Ehrlichia* échappe à la voie de dégradation lysosomale et sécrète des effecteurs du système de sécrétion de type IV, dont ECH_0825 pour inhiber l'apoptose et la production d'espèces réactives de l'oxygène. (5) Le corps réticulé se divise par fission binaire pour former des micro-colonies (*morulae*). (6) *Ehrlichia* exploite les voies de SUMOylation de l'hôte pour maintenir la cellule en vie grâce à l'interaction TRP120-hôte. (7) Le corps élémentaire est relâché dans le milieu extracellulaire par exocytose ou lyse de la cellule hôte. (8) *Ehrlichia* se propage aux cellules voisines. D'après la figure originale de Moumène et Meyer, 2016.

Les réarrangements génomiques constituent une grande famille de processus modifiant l'architecture du génome. L'architecture peut ainsi être réorganisée par des évènements tels que des délétions, des insertions, des duplications, des amplifications, des inversions ou des translocations (Darmon et Leach 2014). Ces variations structurelles, seules ou en combinaison, peuvent jouer un rôle dans la génération, l'activation, l'extinction ou la perturbation d'un gène (Periwal et Scaria, 2015). Ces modifications sont assez importantes pour générer de nouveaux phénotypes aboutissant finalement à l'évolution de nouvelles souches bactériennes (Cangi *et al*., 2016).

La famille *Anaplasmataceae* fait partie de la classe des α-protéobactéries et de l'ordre des *Rickettsiales*. Dans cette famille, on comptabilise quatre genres bactériens dont le genre *Ehrlichia* et le genre *Anaplasma*. Les espèces de ces deux genres sont principalement des pathogènes d'animaux causant diverses maladies souvent létales. Certaines de ces espèces sont également responsables de maladies observées chez l'humain comme l'anaplasmose granulocytaire humaine causée par *Anaplasma phagocytophilum* ou encore l'ehrlichiose monocytaire humaine pour *Ehrlichia chaffeensis* (Thomas *et al*., 2009). Les *Anaplasmataceae* sont des bactéries intracellulaires obligatoires capable de se développer au sein de deux hôtes différents, le vecteur (tique) et l'hôte mammifère (Moumène et Meyer, 2016). Ces bactéries se multiplient dans les cellules de la paroi intestinale puis dans les glandes salivaires de la tique, elles sont ensuite transmises à l'hôte par une morsure de tique lors des repas sanguins. Les bactéries vont alors infecter la cellule hôte cible (cellules endothéliales, monocytes, neutrophiles) afin de pouvoir se répliquer (Allsopp, 2010). Des études par microscopie électronique ont révélé que ces bactéries possèdent deux formes différentes. La première, la forme extracellulaire infectieuse que l'on nomme corps

élémentaire, va s'attacher à la surface de la cellule hôte avant d'entrer par endocytose. Une fois dans la cellule, la bactérie va se différencier en corps réticulé pour se répliquer jusqu'à former de petites colonies que l'on appelle *morulae*. Avant la lyse de la cellule hôte, la bactérie va se redifférencier en corps élémentaire afin d'infecter une nouvelle cellule hôte (Thomas *et al*., 2009, Figure 1).

Les bactéries appartenant à la famille des *Anaplasmataceae* présentent des spectres d'hôte variés. Pour chacun des genres *Anaplasma* et *Ehrlichia*, on observe des espèces avec un spectre d'hôte large (*E. chaffeensis* et *A. phagocytophilum*) et d'autres avec un spectre d'hôte étroit (par exemple *E. ruminantium* et *A. marginale*). *E. chaffeensis* et *A. phagocytophilum* présentent des spectres d'hôtes proches avec au moins quatre hôtes en commun comprenant l'homme, les cervidés, les canidés et les rongeurs. Cette forte similarité entre les différents spectres d'hôtes indique que ces deux bactéries pourraient avoir des niches écologiques se recoupant entre-elles. L'hôte deviendrait alors un point de « rencontre et d'échange » lors d'éventuelles co-infections. D'autre part, les autres espèces présentant des spectres d'hôte très restreints pourraient avoir développé au cours de leur évolution un répertoire de gènes associés à cette forte spécificité d'hôte.

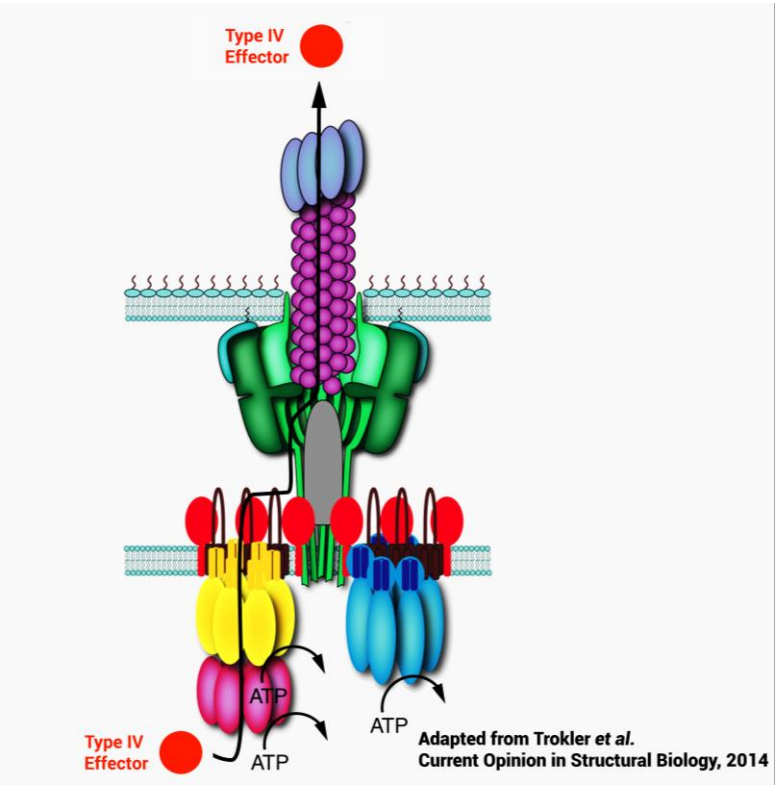


Figure 2. Le système de sécrétion de type IV (SST4), une seringue moléculaire permettant la translocation des effecteurs de la bactérie vers le cytoplasme de la cellule hôte.

Le SST4 permet la sécrétion de protéines nommées effecteurs (orange) du système de sécrétion de type IV (ET4), qui ont un rôle dans le détournement de l'immunité et des métaboliques de la cellule hôte. Le SST4 est composé d'un corps basal membranaire (vert), d'un pilus (violet) et de composants énergétiques (ATPases) participant à la sélection des substrats et fournissant l'énergie nécessaire à leur translocation dans la cellule hôte (jaune, rose et bleu).

Les bactéries intracellulaires utilisent les systèmes de sécrétions pour infecter, proliférer et persister à l'intérieur de leur cellule hôte. A ce jour, six principaux systèmes de sécrétions de protéines ont été mis en évidence chez les bactéries gram-négatives. Ces systèmes de sécrétions sont essentiels pour permettre aux protéines effectrices de franchir les différentes membranes plasmiques des bactéries (membrane interne et externe de la paroi bactérienne) et atteindre le milieu extracellulaire. La sécrétion de protéines est réalisée au moyen de différentes stratégies. Tout d'abord, la translocation peut se faire en deux étapes comprenant l'export à travers la membrane cytoplasmique puis une sécrétion à travers la membrane externe de la bactérie comme pour les systèmes de sécrétion de type II et V. La translocation peut aussi être directe à travers les deux membranes comme pour les systèmes de sécrétions de type I, III, IV et VI. Chez les *Anaplasmataceae,* le système de sécrétion majoritairement impliqué dans la pathogenèse est le système de sécrétion de type IV (SST4) (Rikihisa et al., 2010, Figure 2). Le SST4 présent dans le génome des bactéries des genres *Ehrlichia* et *Anaplasma* est le système VirB/D4 initialement décrit chez *Agrobacterium tumefaciens* (Middleton et al., 2005). Les gènes de la famille *virB/D4* codent pour des protéines membranaires internes, périplasmiques ou externes afin de former une véritable seringue moléculaire qui permet la translocation d'autres protéines – les effecteurs - dans le cytoplasme de la cellule cible (Low et al., 2014). Le SST4 et ses effecteurs (ET4) sont très étudiés chez les α-protéobactéries (*Anaplasma spp., Ehrlichia spp., Brucella spp., Bartonella spp., Agrobacterium spp.*) mais aussi chez les γ-protéobactéries (*Legionella spp., Coxiella spp.*). Les ET4 sont nécessaires au développement de la bactérie dans la cellule hôte (Niu et al, 2012) et agissent de différentes manières pour permettre à la bactérie de résister aux mécanismes de défenses, de contourner la

réponse immunitaire innée de la cellule hôte mais aussi de favoriser l'acquisition de nutriments. Un grand nombre d'ET4 sont déjà connus et décrits chez les γ-protéobactéries avec près de 300 ET4 connus chez *Legionella pneumophila* (10% du génome) et environ 100 effecteurs connus chez *Coxiella burnetii* (5% du génome). A l'inverse, seuls 19 ET4 ont été mis en évidence chez les α-protéobactéries (Schulein *et al.*, 2005 ; Niu *et al.*, 2010 ; Rikhisa et Lin, 2010 ; Lockwood *et al.*, 2011; Marchesini *et al.*, 2011 ; Liu *et al.*, 2012). A l'heure actuelle, seulement deux protéines effectrices ont été caractérisées chez *A. phagocytophilum* (*AnkA* et *Ats-1*) et *E. chaffeensis* (*Ets-1*, homologue d'*Ats-1*) comme étant sécrétées par le système de sécrétion de type IV. Le premier effecteur à avoir été caractérisé est donc la protéine *AnkA*, qui contient des répétitions en tandem de domaines ankyrine. Une fois sécrétée dans le cytoplasme, cette protéine est phosphorylée sur les tyrosines d'un motif EPIYA (motif protéique de phosphorylation) puis est adressée vers le noyau de la cellule hôte afin de diminuer l'expression du gène CYBB (Garcia-Garcia et al, 2009, zhu et al., 2009). Cet effecteur est un effecteur de la famille des nucléomodulines qui agit comme un facteur de transcription pour manipuler la machinerie cellulaire de l'hôte (Bierne and Cossart, 2012). Le second effecteur connu d'*A. phagocytophilum* est *Ats-1*. Cet ET4 peut jouer deux rôles distincts dans la cellule hôte. D'une part, il facilite le recrutement des autophagosomes afin d'apporter des nutriments à la vacuole bactérienne et permettre la réplication de la bactérie. D'autre part, une fois clivé dans le cytoplasme, *Ats-1* est adressé vers la mitochondrie afin d'inhiber l'apoptose de la cellule hôte infectée (Niu et al., 2010). Certains ET4 comme les AnkX chez *Legionella pneumophila* détournent et réorganisent les vésicules du réticulum endoplasmique en utilisant les microtubules de la cellule hôte afin de circonscrire la vacuole bactérienne et de la masquer vis-à-vis des réponses immunes

innées de la cellule hôte (Pan *et al*., 2008 ; Ge et Shao, 2011). D'autres ET4 vont agir sur certaines voies métaboliques, à l'instar de l'effecteur AnkG de *Coxiella burnetii* qui va interagir avec la protéine de la cellule hôte C1qR (p32) entrainant une inhibition de l'apoptose de la cellule hôte (Lürhrmann *et al.*, 2010). En ciblant ou en mimant certaines protéines de la cellule hôte, les ET4 sont ainsi capables d'interférer avec la machinerie cellulaire, de réguler l'expression des gènes, de détourner le trafic vésiculaire ou encore d'inhiber les réponses imunitaires innées de la cellule hôte.

Enfin, si l'on estime que pour une bactérie intracellulaire, le pool de gènes codant des effecteurs du SST4 représente entre 5 et 10% du génome (ce qui est le cas pour *Legionella* et *Coxiella* dont les répertoires d'effecteurs ont été caractérisés de manière extensive), il est raisonnable de penser qu'il reste une myriade d'effecteurs à découvrir et à caractériser chez les *Anaplasmataceae*.

Pourquoi prédire les effecteurs ? Identifier les effecteurs du système de sécrétion de type IV (SST4) des *Anaplasmataceae*, et élucider leurs fonctions dans la cellule hôte est crucial pour imaginer de nouvelles alternatives thérapeutiques aux antibiotiques ou aux vaccins contre ces bactéries zoonotiques. En effet, une meilleure compréhension du mode d'action des effecteurs du SST4 dans la cellule hôte permet d'identifier les cibles protéiques et les voies métaboliques nécessaires à l'infection. Il sera ainsi possible d'imaginer de nouvelles stratégies de lutte contre ces infections bactériennes en renforçant par exemple les voies dérégulées par la bactérie ou en en inactivant certains mécanismes-clés de reconnaissance ou d'interaction moléculaire. Les méthodes biologiques classiques permettant d'identifier les effecteurs – comme la génération de banque de mutants ou le criblage *in vivo* par double-hybride - sont difficiles à mettre en place chez les *Anaplasmataceae*

en raison de leur mode de vie particulier (intracellulaires stricts). La prédiction *in silico* des effecteurs est donc une alternative intéressante qui permet de réduire le nombre de candidats potentiels en éliminant les gènes les moins intéressants. Cependant, prédire les effecteurs du système de sécrétion de type IV reste un véritable défi. En effet, contrairement à ce que l'on peut retrouver chez les Bartonelles avec le domaine BID (Schulein et al. 2005) ou encore chez d'autres pathogènes pour les effecteurs du système de sécrétion de type III (Filloux, 2010), le signal de sécrétion des ET4 n'est pas clairement identifiable chez les *Anaplasmataceae*.

Comment prédire les effecteurs ? Afin d'automatiser la recherche d'effecteurs, nous avons donc préalablement développé un logiciel capable de prédire les protéines effectrices des autres protéines grâce à différentes caractéristiques liées aux effecteurs (Meyer et al., 2013).

En effet, à l'heure actuelle, il n'existe que deux approches pour prédire des effecteurs du SST4. La première est basée sur l'alignement des séquences protéiques N-terminale et C-terminale. Cette approche d'apprentissage automatique, appliquée chez *L. pneumophila*, analyse les séquences N-ter et C-ter des effecteurs connus afin de rechercher un modèle de prédiction cohérent (Wang et al., 2017). Elle est intéressante si le nombre d'effecteurs connus dans la bactérie pathogène est suffisamment grand, permettant de valider les modèles avec de vrais positifs comme pour le genre *Legionella* pour lequel l'effectome est considéré comme quasiment exhaustif. La deuxième méthode, qui est celle que nous avons choisie, propose de prédire les effecteurs du SST4 en recherchant d'une part les caractéristiques physiques qui semblent être liées au signal de sécrétion et d'autre part, les domaines protéiques présent dans des effecteurs connus. Ainsi, des signaux putatifs de sécrétion

portés sur la charge, la basicité et l'hydrophobicité dans les régions C-terminales des protéines ont pu être mis en évidence grâce à l'analyse de différents effecteurs connus dont la sécrétion par le SST4 a été validée fonctionnellement (Rikihisa et al., 2009; Lifshitz et al., 2013). De plus, notre approche vise donc à associer d'autres caractéristiques propres à certains effecteurs connus, comme par exemple la présence de domaines eucaryotes ou encore la présence de domaines de localisation subcellulaire tels que les NLS (signaux de localisation nucléaire) ou MLS (séquence de localisation mitochondriale) (Meyer et al., 2013; Noroy et al., 2018). L'association des résultats de recherche de ces différentes caractéristiques permet au logiciel S4TE d'identifier *in silico* une liste d'effecteurs putatifs tout en attribuant un score de prédiction à chaque protéine d'une bactérie donnée. De plus, S4TE nous apporte un grand nombre d'informations sur la fonction possible des effecteurs prédits (recherche de domaines d'intérêt) ou encore leur localisation subcellulaire. Ceci permet d'affiner le choix des candidats potentiellement intéressants pour une validation fonctionnelle ultérieure, accélérant ainsi l'étude des effecteurs.

Le logiciel S4TE permet donc d'identifier les répertoires d'effecteurs putatifs pour différentes souches bactériennes. La comparaison de ces répertoires putatifs permet de définir des effecteurs conservés entres différentes souches et des effecteurs spécifiques de chaque souche et pouvant avoir un rôle important dans la virulence bactérienne ou la spécificité d'hôte.

La génomique comparative est l'étude comparative de la structure et des fonctions des génomes au sein d'une espèce, d'un genre ou d'une famille. La génomique comparative est un champ de recherche qui a virtuellement débuté en 1995 lorsque les deux premiers génomes complets ont été séquencés. Elle permet de comparer différentes caractéristiques génomiques telles que les séquences ADN, le polymorphisme présence/absence des gènes, la synténie des gènes et d'autres repères structurels des génomes (Primrose et Twyman, 2003). Elle permet d'identifier et de comprendre les effets de la sélection sur l'organisation et l'histoire évolutive des génomes. En effet, les génomes bactériens peuvent être vus comme une association de gènes essentiels à la survie de la bactérie et donc très conservés (génome cœur) et de gènes accessoires, variables, qui permettent à la bactérie d'acquérir de nouvelles fonctions (McFall-Ngai et al., 2013). L'acquisition de nouvelles fonctions, portées par le génome accessoire dirigée par des mutations de gènes existants ou par insertion de nouveaux gènes, permet l'adaptation rapide de la bactérie à de nouvelles niches écologiques. L'étude des différents répertoires de gènes peut donc permettre d'identifier les gènes impliqués dans la diversité et la 'spéciation' bactérienne. C'est la raison pour laquelle nous avons privilégié les outils de génomique comparative pour étudier les différents répertoires d'effecteurs putatifs afin d'évaluer leur rôle potentiel dans la virulence et la spécificité d'hôte chez les *Anaplasmataceae*. En effet, la comparaison des répertoires d'effecteurs permet de mettre en avant l'effectome cœur (ensemble des effecteurs partagés entre toutes les souches bactériennes impliquées dans la comparaison) essentiel à l'infection bactérienne et les effectomes accessoires (ensemble des effecteurs présents dans une seule souche bactérienne) qui sont des gènes pouvant avoir un lien

avec les différents phénotypes différentiels observés (virulence, spécificité d'hôte).

En plus de l'étude classique des répertoires de gènes, avec l'identification et le dénombrement des gènes associés à chaque répertoire, il est possible de développer de nouvelles manières de comparer les génomes et de visualiser les différences portées par ces génomes. Ainsi, l'utilisation des logiciels Circos (Krzywinski et al., 2009) et hive plot (Krzywinski et al., 2012) peut permettre une meilleure compréhension des relations d'orthologie et de synténie entre les effecteurs, tout en visualisant plus facilement les réarrangements chromosomiques et le polymorphisme présence/absence des effecteurs prédits pour chaque génome.

Ce manuscrit de thèse est organisé en trois parties. La première présente l'outil de prédiction des effecteurs du système de sécrétion de type IV (S4TE 2.0). Ce logiciel, que nous avons développé, nous a permis d'une part, d'identifier le super répertoire d'ET4 dans le pan-génome de la famille des *Anaplasmataceae* et d'autre part, de comparer les différents effectomes des *Anaplasmataceae* grâce au module de génomique comparative (S4TE-CG) associé à S4TE 2.0. La seconde partie concerne le travail effectué sur l'analyse du pan-effectome des souches d'*E. chaffeensis* présentant des virulences différentes. La troisième l'analyse des répertoires d'effecteurs au niveau du genre *Ehrlichia*, et l'évaluation de l'implication de l'effectome dans la spécificité d'hôte. Enfin, une discussion générale propose de replacer l'étude de la plasticité génomique dans un contexte plus intégratif à différentes échelles et les perspectives nouvelles s'ouvrant à la suite de ce travail.

RESULTATS

Partie 1

Prédiction des effecteurs du système de sécrétion de type IV

1. Préambule

Les bactéries pathogènes ont développé de nombreuses stratégies pour corrompre, détourner et imiter les processus cellulaires afin de contourner les défenses innées de la cellule hôte pour survivre et se répliquer. Parmi ces stratégies, les effecteurs du système de sécrétion de type IV (ET4) sont des protéines secrétées par la bactérie pathogène pour manipuler les processus cellulaires durant l'infection. Elles sont injectées dans les cellules hôte eucaryote par un complexe multi protéique ATP-dépendant, le système de sécrétion de type IV (SST4).

Les ET4 présentent un grand nombre de caractéristiques comme la présence de domaines eucaryotes, de domaines de localisation subcellulaire ou de signaux de sécrétions dans la partie C-terminale de la protéine (Meyer et al., 2013).

Cette partie présente la mise à jour effectuée sur le logiciel S4TE 1.0 (Searching Algorithm for Type IV secretion system Effectors) avec le développement de nouvelles fonctionalités, celui de l'outil de génomique comparative et de l'environnement convivial axé utilisateur, qui manquait cruellement à la version précédente. Cette nouvelle version de S4TE propose de nouvelles fonctionnalités :

(i) un module permettant de comparer les effectomes de plusieurs bactéries (jusqu'à 4) en seulement quelques clics,

(ii) un mode expert permettant aux utilisateurs de moduler la recherche ou bien de faire une recherche selon une ou plus caractéristiques des effecteurs,

(iii) un outil simple permettant de faire des recherches directement dans la base de données afin de retrouver l'effecteur d'une bactérie en fonction de son nom, de son « locus_tag » ou de son numéro NCBI,

(iv) une base de données qui a été créé afin d'une part, de référencer et calculer les données pour un grand nombre de génomes bactériens et de plasmides, et d'autre part, de permettre aux utilisateurs d'importer leurs propres données génomiques,

(v) deux graphiques interactifs permettant l'analyse des génomes en fonction de leur composition en GC et de la densité locale de gènes,

Enfin S4TE 2.0, et en particulier sa base de données, a été pensé pour être évolutif grâce à l'ajout des nouveaux effecteurs expérimentalement validés permettant de renforcer le pouvoir prédictif de l'algorithme.

Il est important de noter que c'est l'ensemble du développement effectué pour la mise en ligne de S4TE2.0 (développement des nouvelles fonctionnalités de S4TE, création des parties Front-End et Back-End du site web, développement, remplissage et interaction des bases de données) qui a été effectué dans le cadre cette thèse.

2. Publication : Searching Algorithm for Type IV Effector proteins (S4TE) 2.0: improved tools for type IV effector prediction, analysis and comparison

Searching Algorithm for Type IV Effector proteins (S4TE) 2.0: improved tools for type IV effectors prediction, analysis and comparison.

Christophe Noroy[1,2,3], Thierry Lefrançois[2] and Damien F. Meyer[1,2*]

[1] CIRAD, UMR ASTRE, F-97170 Petit-Bourg, Guadeloupe, France
[2] ASTRE, Univ Montpellier, CIRAD, INRA, Montpellier, France
[3] Université des Antilles, 97159 Pointe-à-Pitre, Guadeloupe, France
* To whom correspondence should be addressed. Tel: +590 (0)590 25 59 47; Email: damien.meyer@cirad.fr

ABSTRACT

Bacterial pathogens have evolved numerous strategies to corrupt, hijack or mimic cellular processes in order to survive and proliferate. Among those strategies, Type IV effectors (T4Es) are proteins secreted by pathogenic bacteria to manipulate host cell processes during infection. They are delivered into eukaryotic cells in an ATP-dependent manner via the type IV secretion system, a specialized multiprotein complex. T4Es contain a wide spectrum of features including eukaryotic-like domains, localization signals or a C-terminal translocation signal. A combination of these features enables prediction of T4Es in a given bacterial genome. In this study, we developed a web-based comprehensive suite of tools with a user-friendly graphical interface. This version 2.0 of S4TE (Searching Algorithm for Type IV Effector Proteins; http://sate.cirad.fr) enables accurate prediction and comparison of T4Es. Search parameters and threshold can be customized by the user to work with any genome sequence, whether publicly available or not. Applications range from

characterizing effector features and identifying potential T4Es to analyzing the effectors based on the genome G+C composition and local gene density. S4TE 2.0 allows the comparison of putative T4E repertoires of up to four bacterial strains at the same time. The software identifies T4E orthologs among strains and provides a Venn diagram and lists of genes for each intersection. New interactive features offer the best visualization of the location of candidate T4Es and hyperlinks to NCBI and Pfam databases. S4TE 2.0 is designed to evolve rapidly with the publication of new experimentally validated T4Es, which will reinforce the predictive power of the algorithm. The computational methodology can be used to identify a wide spectrum of candidate bacterial effectors that lack sequence conservation but have similar amino acid characteristics. This approach will provide very valuable information about bacterial host-specificity and virulence factors, and help identify host targets for the development of new anti-bacterial molecules.

INTRODUCTION

Proteobacteria have evolved specific effector proteins to manipulate host cell gene expression and processes, hijack immune responses and exploit host cell machinery during infection. These proteins are secreted by ATP-dependent protein complexes named type IV secretion systems (T4SS). Some T4Es have been identified and shown to be crucial for pathogenicity. To facilitate the identification of putative T4Es, we previously developed a bioinformatics tool called S4TE 1.0 (Searching Algorithm for Type IV secretion system effector proteins) [1].

In the present article, we present the second version of 'S4TE'. S4TE 2.0 is a tool for *in silico* screening of proteobacteria genomes and T4E prediction based on the combined use of 14 distinctive features. In this updated version, modules searching for promoter motifs, homology, NLS, MLS and E-block are more efficient. A new module has been added in the workflow to locate phosphorylation (EPIYA-like) domains.

S4TE 2.0 consists of the S4TE 1.4 tool and a web interface available to non-commercial users at http://sate.cirad.fr. The web interface is designed to make S4TE 2.0 easy to use for biologists and more time efficient. Most of the genomes and plasmids available in the NCBI database of pathogenic bacteria that have a type IV secretion system have been loaded into the S4TE 2.0 database so effectors can be predicted in only a few clicks.

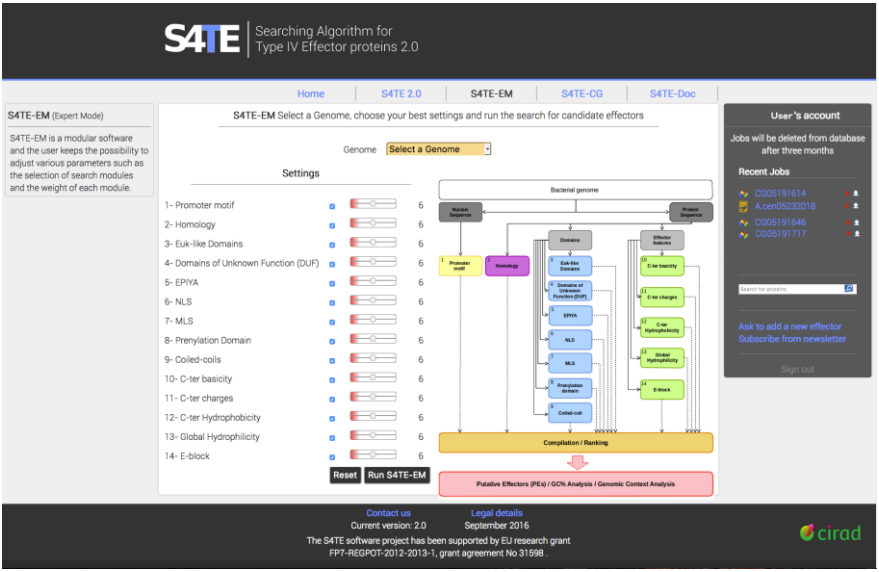


Figure 1. The new front page of the S4TE-EM tool. The right side provides some information about the page. The right side matches the user account. The user account shows all the jobs previously ran in S4TE 2.0 and S4TE-CG. This account makes it possible to search a protein with the search bar and to ask to add a proven T4 effector in the database. In the central part of the work space, the user can select a genome in the drop-down menu. In S4TE-EM, the user can change the weighting or disable one or more modules (on the left) shown in the S4TE diagram (on the right), and run S4TE-EM by clicking on the 'Run S4TE-EM' button.

S4TE 2.0 offers advanced users an expert mode (S4TE-EM) they can use to customize S4TE 2.0 search parameters (*e.g.* exclude modules, modify module weightings). In this mode, S4TE 2.0 can be used as 14 independent programs to search for particular features in a given bacterial genome (*e.g.* NLS, C-ter charges). A new function for comparative genomics (S4TE-CG) has been added to compare up to four predicted effectomes in just a few seconds.

All S4TE 2.0 results are interactive and linked to NCBI and Pfam databases.

SOFTWARE AND ALGORITHM

Programming

S4TE 2.0 software consists in a graphical interface (website) to use the S4TE 1.4 algorithm for genome analysis, Type IV

effectors (T4Es) prediction and comparison of effectomes. S4TE 1.4 is an update of S4TE 1.0[1]. It is written in Perl programming language and uses NCBI, Pfam, EMBOSS, BioPerl and MitoFates libraries and its own proper programs and database. It was developed to improve the prediction performances of S4TE 1.0 and to provide new functionalities to search for new features, enable interactivity and comparative genomics. The 10 S4TE search modules in S4TE 1.0 were kept in S4TE 1.4. However, some modules have been modified (promoter motif search, homology, MLS, NLS, E-block and Pfam database) to improve their predictive power. A supplementary module (EPIYA search) has been added to the workflow. In this paper, only the EPIYA search module and the revised modules are described.

Name	Organism	Length	Threshold	Effector [1]	Non-effector [2]	Logo[3]
PmrA	Legionella	20	0.748	18.6	4.1	
Cpm	Coxiella	20	0.87	18.2	0.02	
Cpm2	Coxiella	7	0.875	13.8	3.9	
Apm	Anaplasma	14	0.7	77.8	9.9	
Apm2	Anaplasma	15	0.86	66.7	0.41	
Bapm	Bartonella	19	0.75	62.5	2.2	
Hpm	Helicobacter	20	0.68	1	0.39	
Bopm	Bordetella	20	0.8	52.6	0.51	

[1]Frequency of motif in effector promoters
[2]Frequency of motif in non-effector promoters
[3]Logo and motif were established using MEME software (Bailey TL *et al.*, 2009).

Table 1. Enriched DNA motifs found in several bacteria in the 100 nucleotides upstream of known type IV effectors and implemented in S4TE 2.0 searches

Promoter motif search

As several T4Es in a given bacterium can be subjected to coordinated regulation with the same protein, *e.g.* PmrA[2], we used S4TE 2.0 to conduct a search for conserved motifs (potential regulatory motifs) in the short promoter regions of the genes. The aim was to improve S4TE 2.0 prediction of possible regulons of T4Es. Enriched DNA motifs were searched in a window of 100 nucleotides (nt) placed upstream of the start codon, using MEME[3]. Eight consensus motifs were identified in different bacteria (table 1). The corresponding motif search module of S4TE 2.0 extracts the 5' Flanking intergenic regions (5' FIRs) and searches for all these motifs thanks to a position-specific scoring matrix generated from multiple sequence alignments with the promoters of known T4Es. Only alignments with a score above the chosen threshold are selected. The threshold that yielded the highest sensitivity and specificity for each motif in the corresponding bacterium was chosen (Table 1).

Homology

BLAST 2.2 was used to compare proteins to search for homologies with known T4Es [4]. The cut-off of the S4TE 1.0 homology module was changed. S4TE 2.0 compares the database containing all known T4Es with the query proteome and returns all homologs with a cut-off of the expected value (E) <10-4. This E-value cut-off was selected to find real homologs between phylogenetically distant bacterial species. Databases containing proven effectors have also been updated (Table S1).

Nuclear localization signals (NLS)

NLS are protein sequences that target proteins in the nucleus of eukaryotic cells[5]. We assume that the occurrence of NLS in a

bacterial protein sequence would be a good indicator of secretion. There are two classes of NLS,

monopartite and Bipartite. In S4TE 2.0, the search for monopartite NLS has been improved according to Ruhanen *et al*. [6]. We rewrote this module to add more known NLS motifs in the search. Monopartite NLS consist of [KR]-[KR]-[KR-][KR]-[KR], X-K-[KR]-[KRP]-[KR]-X, X-R-K-[KRP]-[KR]-X, X-R-K-X-[KR]-[KRP], X-K-[KR]-[KR]-X-[KRP], X-R-K-[KR]-X-[KRP], X-K-[KR]-X-[KR]-X-X, X-R-K-X-[KR]-X-X, X-K-[KR]-[KR]-X-X-X and X-R-K-[KR]-X-X-X motifs. Bipartite NLS were also searched with S4TE 1.0 motif (K-[KR]-X(6,20)-[KR]-[KR]-X-[KR]). The new module was tested with a dataset of 32 NLS and 32 no-NLS containing proteins (dataset 1). The module selected 24 true positives (TP) and only three false positives (FP). This represents a sensitivity (Se) of 75% and a specificity (Sp) of 91%.

Mitochondrial Localization Signals (MLS)

MLS are signal sequences located in the N-terminus of proteins that are targeted to mitochondria. This sequence is cleaved after translocation of the protein inside the mitochondria[5,7]. To predict MLS in S4TE 2.0, we used the MitoFates tool[8]. MitoFates predicts mitochondrial presequences, a cleavable localization signal located in the N-terminal, and its cleaved position.

E-block

The E-block domain consists of a glutamate sequence rich in C-terminal 30 amino acids and is associated with T4Es translocation in *L. pneumophila*. Huang *et al*. showed that an E-block motif is also important for the translocation of T4SS substrates[9]. In S4TE 2.0,

the E-block module was modified according to Lifshitz *et al.* [10]. The E-block was searched

in a window of 22 amino acids between position -4 C-terminal and -26 C-terminal. The motif that is searched for is a motif of 10 amino acids containing three or more glutamate (E) residues. The module was tested on 98 E-block and 98 no-E-block containing proteins (dataset 2). This module selected 60 TP and only 6 FP (Sensitivity of 61%, Specificity of 94%).

Pfam database

The local Pfam database has been updated to find more eukaryotic domains of known effectors of *Legionella pneumophila*[10]. Eukaryotic domains were extracted from the whole Pfam database and added to the S4TE 2.0 workflow. All eukaryotic domains used for this search are listed in Table S2.

EPIYA search

EPIYA search is a new module implemented in S4TE 2.0. The EPIYA domain is an eukaryotic phosphorylation motif[11]. In *H. pylori*, EPIYA has been shown to contribute to the secretion of a CagA effector[12]. We searched for conserved EPIYA motifs (EPIYA, ENIYE, NPLYE, EHLYA, TPLYA, EPLYA, ESIYE, EDLYA, EPIYG, EPVYA, VPNYA, EHIYD) in different bacteria that have a type IV secretion system and we searched for hypothetical EPIYA motifs using the motif E-X-X-Y-X.

Validation

S4TE 2.0 is a software program with 14 independent modules. We tested all the modules independently. The 14 modules were weighted to make S4TE 2.0 efficient. The weighting of each module was calculated according to its performance in finding effectors in *L. pneumophila* Philadelphia I which has been shown to have the most extensive repertoire of T4Es ever identified, with 286 confirmed effectors [10].

S4TE 2.0 Features	1	2	3	4	5	6	7	8	9	10	11	12	13	14
True Positives	108	285	13	2	30	105	6	1	100	262	62	41	114	98
False Positives	434	34	27	106	101	783	79	6	231	2376	863	339	156	232
PPV(%)	20	89	32	2	23	12	7	14	30	10	7	10	42	30

Table 2. Calculation of S4TE 2.0 weighting according to *Legionella pneumophila* Philadelphia 1 Positive Predictive Values (PPV) of each module

Software	S4TE 1.0		S4TE 2.0	
Homology	With	Without	With	Without
Sensitivity	0.86	0.16	1	0.41
Specificity	0.97	0.97	0.93	0.93
Positive Predictive Value	0.74	0.44	0.60	0.43
Negative Predictive Value	0.98	0.91	1	0.94

Table 3. Comparison between S4TE 1.0 and S4TE 2.0

Each module has its own weighting in S4TE 2.0 searches. The weightings were calculated for each module based on their Positive Predictive Value (PPV [PPV=TP/(TP+FP)]) for *L. pneumophila* (Table 2).

The S4TE 2.0 prediction threshold was then defined to enable the best prediction by disregarding homology with known effectors. The threshold was chosen by examining the Sensitivity (Se), Specificity (Sp), Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Accuracy (Acc) for thresholds ranging from 40 to 120 on the test dataset (Figure 2). The threshold was set at a score of 72 to obtain the global PPV possible with the least possible impact on sensitivity.

This threshold combined with weightings led to the correct prediction (true positives) of 282 of the 286 effectors of *L. pneumophila* (Se=98%, PPV=60%) and 96 incorrect predictions (false positives) (Sp = 96%, NPV = 99%).

With this update, S4TE 2.0 prediction is more powerful than that of S4TE 1.0 whose sensitivity was 14% lower. Without

homology, sensitivity increased by 25% (data not shown). Other characteristics including specificity, accuracy and negative predictive value did not change significantly (table 3). S4TE 2.0 allows flexible, highly sensitive and specific detection of new putative T4SS effectors.

SATE-CG

S4TE-CG is a new tool designed to compare different repertoires of putative T4Es identified by S4TE 2.0. The corresponding S4TE-CG algorithm is described in Figure 3. The user can compare up to four effectomes simultaneously. S4TE 2.0 results from selected genomes (effectomes) are compared with Blastp 2.2

with an expected value (E) cut-off of <10-4 to find homologous proteins in each effectome. S4TE-CG successively compares all effectomes in a pairwise manner, the overlaps between the effectomes of each genome are calculated and the final results are plotted on a Venn diagram and listed in an interactive table. All effectors are clickable and the user is redirected to the S4TE 2.0 results on the effector concerned. The table can be easily copied and pasted for export.

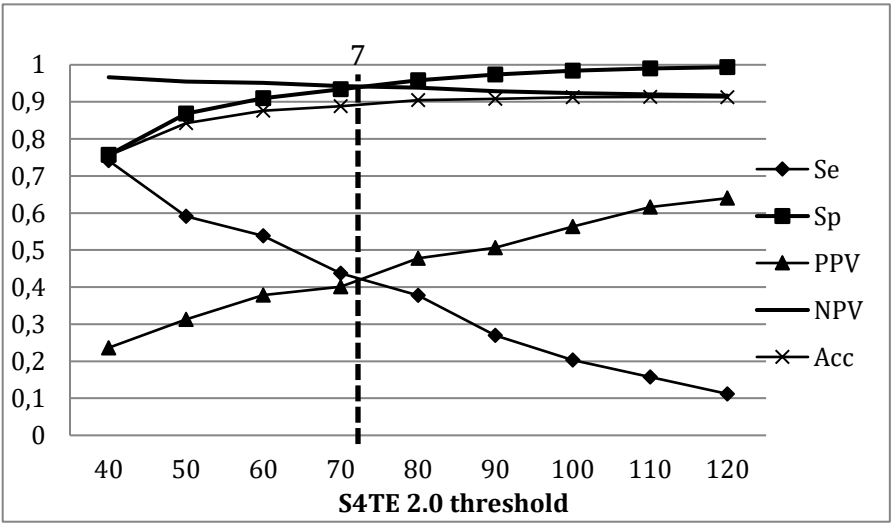


Figure 2. Distribution of S4TE 2.0 performances according to the threshold. Plot of the sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV) and accuracy (Acc) of S4TE 2.0 with no homology module on *L. pneumophila* genome as a function of the S4TE 2.0 threshold. A threshold of 72 proved to be the best combination of these characteristics.

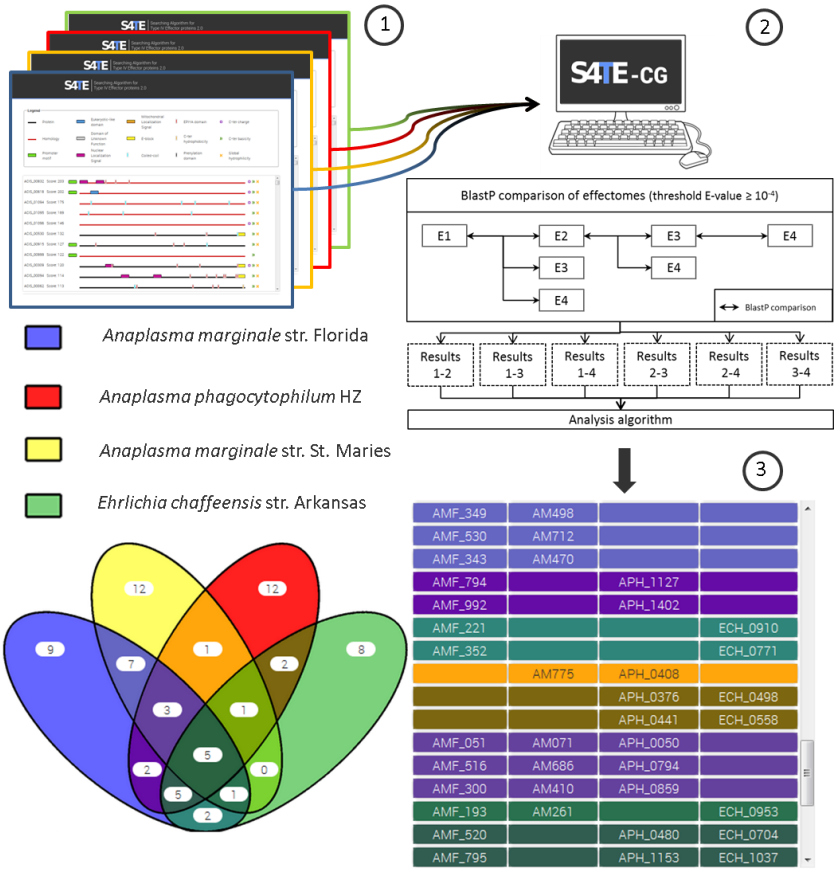


Figure 3. Flow chart of the comparison of 4 effectomes using S4TE-CG. Users can compare up to four genomes simultaneously. **1**. S4TE 2.0 results from selected genomes (effectomes) are compared with Blastp 2.2 to find homologous proteins in each effectome. **2**. S4TE-CG successively compares all effectomes in a pairwise manner, and calculates any overlaps between the effectomes of each genome. **3**. The final results are plotted on a Venn diagram and listed in an interactive table.

Software availability

S4TE 2.0 is a web interface and the S4TE 1.4 package is freely available to non-commercial users at http://sate.cirad.fr/S4TE-Doc.php. All programming was done using Perl 5.18 and BioPerl 1.6.1. The software runs on Linux platforms (Ubuntu 14.04 and Mac OS X). All required packages and the installation process are described in the user guide included in the package. The user guide also details S4TE options for running S4TE. By default, the command line to launch S4TE is **./S4TE.pl -f "Genbank_file"** in the S4TE folder (**cd way_to_S4TE/S4TE/**). Some options are available for the user to launch S4TE: **-c**, suppression of a module in the pipeline; **-w**, modification of the weight of each module in the pipeline; **-t**, imposition of a threshold for effector selection. Each S4TE module creates a .txt file in the folder **way_to_S4TE/S4TE/Jobs/ job<Name_of_genome_folder><year><month><day><hour><min>**

All the results are compiled in the *CompilationFile.txt* and *Results.txt* in the same folder.

WEB INTERFACE

Design and general features

The S4TE 2.0 website is powered from scratch on the 'CIRAD web server'. All the features of the web site were tested on common web browsers. S4TE 2.0 found T4Es in large genome databases (Table S3) available to all users. A user account is available and necessary to keep your jobs up to three months, to import your own genome in a S4TE 2.0 temporary database and to ask to add a new proved effector in the database. The addition of an effector to the

database must be accompanied by a reference (scientific article) and will be checked manually before the effector is added to the database. Those who subscribe to the newsletter will be notified by email about the addition of new effectors to the database and the effector will be visible in the S4TE 2.0 tab strip. This free account allows users to search for proteins in the S4TE 2.0 database using the name, the locus tag or NCBI number of a protein in the search bar. The account also allows the user to subscribe to the S4TE newsletter that summarizes any changes made to the software, and provide updates on the latest research on Type IV Effectors.

S4TE 2.0 is a simple and user-friendly tool

S4TE 2.0 is a web-based user-friendly tool that gets results in only a few clicks. The user can locate a chromosome in more than 340 bacterial genomes and plasmids available in the database and the results can be viewed by clicking on run S4TE 2.0 (Table S3).

If the desired genome is not available in the databases, the user can import it with a GenBank file (.gbk). S4TE 2.0 will import the file to a temporary database for three months. All S4TE tools (S4TE-EM and S4TE-CG) can then be used on the genome by the owner.

The S4TE 2.0 web page allows users to read some of the news published in the newsletter. Five news items are visible on the S4TE2.0 web page, but all the news can be found by clicking on the bottom right link.

Figure 4 presents some results obtained with S4TE 2.0. All the proteins in the selected genome are represented on the S4TE 2.0 web results page. A score was calculated for each protein based on the weighting of each module. Proteins were ranked according to the same score. All proteins whose scores are above the threshold are considered as belonging to the S4TE 2.0 effectome. An iconography was created to help read the list (Figure 4A). Users can find all the

details concerning each characteristic of a given protein by clicking on the protein concerned on the web results page.

When a user runs S4TE 2.0, in addition to the results page, two graphs are automatically drawn. The first shows the distribution of predicted effectors according to local gene density (Figure 4B). The second one displays the distribution of predicted T4Es according to the G+C content along the genome (Figure 4C).

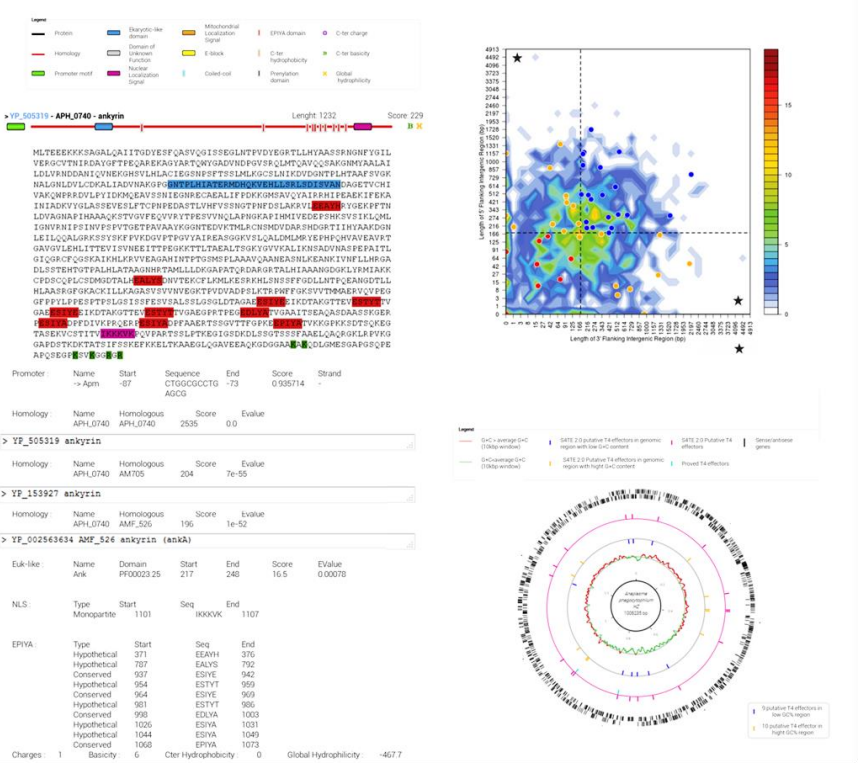


Figure 4. Example of S4TE 2.0 results for *Anaplasma phagocytophilum* HZ. APH-0740. A. Schematic representations of proteins with different characteristics present in the sequence are shown. Characteristics are easy to find by highlighting the corresponding sequence in the effector sequence. These characteristics are detailed below the sequence. **B**. Distribution of S4TE 2.0 predicted type IV effectors (T4Es) according to local gene density. The predicted T4Es are plotted according to the length of their flanking intergenic regions (FIRs). All *A. phagocytophilum* genes were sorted into 2-dimensional bins according to the length of their 5′ (y-axis) and 3′ (x-axis) FIRs. The number of genes in the bins is represented by a color-coded density graph. Genes whose FIRs are both longer than the median FIR length were considered as gene-sparse region (GSR) genes. Genes whose FIRs are both below the median value were considered as gene-dense region (GDR) genes. In-between region (IBR) genes are genes with a long 5′FIR and short 3′FIR, or inversely. Candidate T4Es predicted using the S4TE2.0 algorithm were *s* plotted on this distribution according to their own 3′ and 5′ FIRs. A color is assigned to each of the three following groups: Red to GDRs, orange to IBRs, and blue to GSRs. **C**. Genome-wide distribution of predicted effectome according to the G+C content. From outer track to inner track, sense and antisense genes (black), S4TE 2.0 putative T4Es (pink), proved T4Es (turquoise), S4TE 2.0 putative T4Es in genomic region with low G+C content (yellow), S4TE 2.0 putative T4Es in genomic region with high G+C content (blue), G+C ≥ average G+C (red), G+C < average G+C (green).

S4TE-EM Expert mode for accurate searching

S4TE-EM is the expert mode of S4TE 2.0. S4TE-EM allows the user to modify the weights of each module and to deactivate one or more modules in the search (Figure 1). The weight of a module can be changed by moving the position of the cursor next to the name of each module. Weightings can be changed between the lowest weighting available for the module and the threshold of S4TE 2.0 (**t**=72). The lowest weight is calculated independently for each module as a function of the positive predictive value and corresponds to a value equal to 0.5. Users can also cancel one or more modules in the pipeline by unchecking the box next to the name of the module (Figure 1).

All the modules are independent and users can use S4TE-EM to locate the same characteristic throughout the genome. For example, if the user disables all the modules except NLS, S4TE-EM will find all proteins with an NLS in the genome, meaning users can use S4TE-EM as a new genome analysis tool.

S4TE-CG Comparative genomics to compare effectomes

S4TE-CG is a new tool designed to compare different effectomes predicted by S4TE 2.0. Users can choose up to four effectomes in S4TE 2.0 databases or upload a genome present in the temporary database. S4TE-CG displays results in a Venn diagram and in an interactive table. Users can easily find different subsets of information in the appropriate table by referring to the different colors in the Venn diagram (Figure 3). Information about each effector can easily be found by clicking on the name of the effector in the table. Or users can simply copy and paste the table in a .csv file.

CONCLUSION

This paper presents updated S4TE software. The computational tool is designed to predict the presence of T4SS effector proteins in bacteria. The identification of T4Es and some characteristics are improved in this update. Compared with a machine learning approach, using S4TE 2.0 to predict T4Es in *Legionella* and *Coxiella* species[10,13,14] improved sensitivity (98% for S4TE 2.0 and 89% for Wang *et al.*) and equivalent specificity (97% for Wang *et al.* and 93% for S4TE 2.0). S4TE 2.0 is easy to use. Only an internet connection and a few clicks are needed to search for T4Es in more than 340 bacterial genomes and plasmids. The results are displayed instantaneously for easy reading. An automated pipeline is also provided to analyze and visualize effector distribution in the genome according to G+C content and local gene density. S4TE 2.0 results are linked to bioinformatics databases like NCBI and Pfam. The S4TE 2.0 database is designed to evolve and will be updated by adding new proven effectors and new bacterial genomes. S4TE 2.0 not only predicts the T4Es but also their subcellular localization (NLS, MLS, prenylation) and the function of these proteins (Coiled coils, EPIYA, Euk-like, etc.). All these features make S4TE 2.0 a powerful software for studies of T4Es.

S4TE 2.0 also offers an expert mode, which allows users to make manual adjustments to the weight of the modules. Each module that searches for a feature or a characteristic can be used independently. S4TE EM can be viewed and use as 14 independent programs. This could facilitate the annotation of new genomes by looking for specific features such as NLS, prenylation domains, etc.

Finally, S4TE-CG makes it possible for users to compare effectomes to highlight core T4 effectomes and/or accessory T4

effectomes to understand how effectomes evolved, and may provide clues to the specificity of different strains.

ACKNOWLEDGEMENTS

This study was partly conducted in the framework of the project MALIN “Surveillance, diagnosis, control and impact of infectious diseases of humans, animals and plants in tropical islands” supported by the European Union in the framework of the European Regional Development Fund (ERDF) and the Regional Council of Guadeloupe. We thank D. Goodfellow for reading of this manuscript.

REFERENCES

1. Meyer DF, Noroy C, Moumène A, Raffaele S, Albina E, Vachiéry N: **Searching algorithm for type IV secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context.** *Nucleic acids research* 2013, **41**:9218–2910.1093/nar/gkt718.
2. Zusman T, Aloni G, Halperin E, Kotzer H, Degtyar E, Feldman M, Segal G: **The response regulator PmrA is a major regulator of the icm/dot type IV secretion system in Legionella pneumophila and Coxiella burnetii.** *Molecular microbiology* 2007, **63**:1508–2310.1111/j.1365-2958.2007.05604.x.
3. Bailey T, Boden M, Buske F, Frith M, Grant C, Clementi L, Ren J, Li W, Noble W: **MEME SUITE: tools for motif discovery and searching.** *Nucleic acids research* 2009, **37**:W202–810.1093/nar/gkp335.
4. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden T: **NCBI BLAST: a better web interface.** *Nucleic acids research* 2008, **36**:W5–910.1093/nar/gkn201.

5. Hicks SW, Galán JE: **Exploitation of eukaryotic subcellular targeting mechanisms by bacterial effectors**. *Nature Reviews Microbiology* 2013, **11**:316–32610.1038/nrmicro3009.
6. Ruhanen H, Hurley D, Ghosh A, O'Brien KT, Johnston CRR, Shields DC: **Potential of known and short prokaryotic protein motifs as a basis for novel peptide-based antibacterial therapeutics: a computational survey.** *Front Microbiol* 2014, **5**:410.3389/fmicb.2014.00004.
7. Niu H, Kozjak-Pavlovic V, Rudel T, Rikihisa Y: **Anaplasma phagocytophilum Ats-1 Is Imported into Host Cell Mitochondria and Interferes with Apoptosis Induction**. *PLoS Pathogens* 2010, **6**:e100077410.1371/journal.ppat.1000774.
8. Fukasawa Y, Tsuji J, Fu S-CC, Tomii K, Horton P, Imai K: **MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites.** *Mol. Cell Proteomics* 2015, **14**:1113–2610.1074/mcp.M114.043083.
9. Huang L, Boyd D, Amyot W, Hempstead A, Luo Z-Q, O'Connor T, Chen C, Machner M, Montminy T, Isberg R: **The E Block motif is associated with Legionella pneumophila translocated substrates.** *Cellular Microbiology*, **13**:227–24510.1111/j.1462-5822.2010.01531.x.
10. Lifshitz Z, Burstein D, Peeri M, Zusman T, Schwartz K, Shuman H, Pupko T, Segal G: **Computational modeling and experimental validation of the Legionella and Coxiella virulence-related type-IVB secretion signal**. *Proc National Acad Sci* 2013, **110**:E707–E71510.1073/pnas.1215278110.
11. Safari F, Murata-Kamiya N, Saito Y, Hatakeyama M: **Mammalian Pragmin regulates Src family kinases via the Glu-Pro-Ile-Tyr-Ala (EPIYA) motif that is exploited by bacterial effectors.** *Proc. Natl. Acad. Sci. U.S.A.* 2011, **108**:14938–4310.1073/pnas.1107740108.

12. Papadakos KS, Sougleri IS, Mentis AF, Hatziloukas E, Sgouras DN: **Presence of terminal EPIYA phosphorylation motifs in Helicobacter pylori CagA contributes to IL-8 secretion, irrespective of the number of repeats.** *PLoS ONE* 2013, **8**:e5629110.1371/journal.pone.0056291.
13. Lifshitz Z, Burstein D, Schwartz K, Shuman H, Pupko T, Segal G: **Identification of Novel Coxiella burnetii Icm/Dot Effectors and Genetic Analysis of Their Involvement in Modulating a Mitogen-Activated Protein Kinase Pathway**. *Infect Immun* 2014, **82**:3740–375210.1128/IAI.01729-14.
14. Wang Y, Wei X, Bao H, Liu S-L: **Prediction of bacterial type IV secreted effectors by C-terminal features**. *Bmc Genomics* 2014, **15**:1–1410.1186/1471-2164-15-50

Partie 2

Etude de la plasticité génomique des effecteurs du système de sécrétion de type IV au sein de l'espèce *E. chaffeensis*

1. Préambule

Dans cette étude, nous avons émis l'hypothèse que la variation de virulence entre certaines souches d'*E. chaffeensis* pouvait être dirigée par la plasticité génomique et l'acquisition de différents répertoires d'effecteurs du SST4 , à l'instar de ce qui avait déjà été observé chez les pathogènes de plante *Xanthomonas* spp. (Cesbron et al. 2015 10.3389/fpls.2015.01126). Grâce à l'analyse comparative de génomes entiers, nous avons caractérisé les relations entre les huit génomes disponibles d'*E. chaffeensis* isolé chez l'homme et montré que ces génomes sont hautement conservés. Nous avons identifié l'effectome cœur putatif d'*E. chaffeensis* et certaines stratégies intracellulaires adaptatives. Nous avons prédit les effecteurs candidats du SST4 ainsi que certaines cibles dans la cellule hôte en utilisant une base de donnée (HPIDB) référençant les interactions protéiques hôte-pathogène. Nous avons aussi identifié deux effecteurs n'appartenant pas à l'effectome conservé et pouvant être impliqués dans les différences de virulence observées entre les souches. Enfin, la prédiction des effecteurs et de leurs cibles potentielles dans la cellule hôte nous a permis d'identifier d'importantes voies cellulaires de l'hôte pouvant être la cible de ces bactéries pathogènes.

2. Publication : Comparative genomics of the zoonotic pathogen *Ehrlichia chaffeensis* reveals candidate type IV effectors and putative host cell targets.

frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 25 January 2017
doi: 10.3389/fcimb.2016.00204

Comparative Genomics of the Zoonotic Pathogen *Ehrlichia chaffeensis* Reveals Candidate Type IV Effectors and Putative Host Cell Targets

Christophe Noroy[1,2,3] *and Damien F. Meyer*[1,2]*

[1] CIRAD, UMR ASTRE, Guadeloupe, France, [2] INRA, UMR 1309 ASTRE, Montpellier, France, [3] Université des Antilles, Guadeloupe, France

During infection, some intracellular pathogenic bacteria use a dedicated multiprotein complex known as the type IV secretion system to deliver type IV effector (T4E) proteins inside the host cell. These T4Es allow the bacteria to evade host defenses and to subvert host cell processes to their own advantage. *Ehrlichia chaffeensis* is a tick-transmitted obligate intracellular pathogenic bacterium, which causes human monocytic ehrlichiosis. Using comparative whole genome analysis, we identified the relationship between eight available *E. chaffeensis* genomes isolated from humans and show that these genomes are highly conserved. We identified the candidate core type IV effectome of *E. chaffeensis* and some conserved intracellular adaptive strategies. We assigned the West Paces strain to genetic group II and predicted the repertoires of T4Es encoded by *E. chaffeensis* genomes, as well as some putative host cell targets. We demonstrated that predicted T4Es are preferentially distributed in gene sparse regions of the genome. In addition to the identification of the two known type IV effectors of *Anaplasmataceae*, we identified two novel candidates T4Es, ECHLIB_RS02720 and ECHLIB_RS04640, which are not present in all *E. chaffeensis* strains and could explain some variations in inter-strain virulence. We also identified another novel candidate T4E, ECHLIB_RS02720, a hypothetical protein exhibiting EPIYA, and NLS domains as well as a classical type IV secretion signal, suggesting an important role inside the host cell. Overall, our results agree with current knowledge of *Ehrlichia* molecular pathogenesis, and reveal novel candidate T4Es that require experimental validation. This work demonstrates that comparative effectomics enables identification of important host pathways targeted by the bacterial pathogen. Our study, which focuses on the type IV effector repertoires among several strains of *E. chaffeensis* species, is an original approach and provides rational putative targets for the design of alternative therapeutics against intracellular pathogens. The collection of putative effectors of *E. chaffeensis* described in our paper could serve as a roadmap for future studies of the function and evolution of effectors.

Keywords: type IV effectors, *Ehrlichia chaffeensis*, comparative genomics, host-pathogen interactions, genome plasticity

OPEN ACCESS

Edited by:
Rey Carabeo,
Washington State University, USA

Reviewed by:
Jere W. McBride,
University of Texas Medical Branch, USA
Matteo Bonazzi,
Centre National de la Recherche Scientifique, France

***Correspondence:**
Damien F. Meyer
damien.meyer@cirad.fr

Received: *16 June 2016*
Accepted: *21 December 2016*
Published: *25 January 2017*

Citation:
Noroy C and Meyer DF (2017) Comparative Genomics of the Zoonotic Pathogen Ehrlichia chaffeensis Reveals Candidate Type IV Effectors and Putative Host Cell Targets. Front. Cell. Infect. Microbiol. 6:204. doi: 10.3389/fcimb.2016.00204

INTRODUCTION

Ehrlichia chaffeensis is an intracellular rickettsial pathogen transmitted by *Amblyomma americanum* ticks, which is the etiologic agent of human monocytic ehrlichiosis (HME) (Dumler et al., 1993). This pathogen also causes disease in several other vertebrates, including dogs and deer (Paddock and Childs, 2003). The white-tailed deer is the reservoir host for *E. chaffeensis*, while humans, dogs and other vertebrate hosts, such as coyotes and goats, are regarded as incidental hosts (Paddock and Childs, 2003). This bacterium is able to replicate within two hosts, a mammalian host and a tick vector, and is capable of orchestrating highly sophisticated strategies to persist and infect their natural hosts (Rikihisa, 2010). Thus, studying *E. chaffeensis* provide a wealth of information about bacterial adaptation to various environments.

E. chaffeensis has a biphasic developmental cycle involving two morphologically distinct forms (Zhang et al., 2007). The infectious extracellular forms (dense core cells) first attach to the surface of host target cells before entering by endocytosis. Inside the host cells, the bacteria differentiate into reticulate cells within a membrane-bound vacuole where they create a safe niche for survival and replication by binary fission to form large colonies, called morulae. After a few days, the bacteria redifferentiate into infectious forms to be released outside the cell and start a new cycle of infection (Zhang et al., 2007).

In *E. chaffeensis*, the genome sequences of eight human isolates with variable pathogenicity, are available (**Table 1**). The first strain was discovered in 1991 in a 21-year old man and was named *Arkansas* for its geographic origin (Dawson et al., 1991). The most recently identified strain, called West Paces, was found in Tennessee in 2000 (Cheng et al., 2003). The other strains, Heartland, Jax, Liberty, Osceola, Saint Vincent, and Wakulla have also been isolated in humans (**Table 1**) and show different pathogenesis. In severe combined immunodeficiency (SCID) mice, Miura et al. observed differences in virulence in three of the strains, the Arkansas strain causing mild, the Liberty strain causing acute severe pathogenesis, and the Wakulla strain causing acute lethal pathogenesis (Miura and Rikihisa, 2007). The eight strains of *E. chaffeensis* used in this study were separated into three genetic groups based on the sequence polymorphisms of the *p28* outer membrane protein genes (Yu et al., 1999). The Arkansas and Osceola strains were classified in group I, the Heartland, Saint Vincent, and Wakulla strains in group II, and the Jax and Liberty strains in group III. The West Paces strain had not yet been isolated when the genetic groups were defined. Other genetic classifications were based on genes encoding immunoreactive proteins. The gene encoding tandem-repeat proteins (TRP) 32 (formerly VLPT, the variable length PCR target gene) contains the region specifying three to six nearly identical, highly hydrophilic 90–amino acid tandem repeats (Sumner et al., 1999). Similarly, in TRP120 (formerly gp120), there are two to four imperfect, direct, tandem 80 bp repeats (Sumner et al., 1999). The number of repeats varies depending on the isolate, resulting in variations in size in the encoded protein. The TRP32 gene shows great inter-strain diversity and is characterized by a series of direct tandem repeats whose number varies among isolates (Paddock and Childs, 2003). The DNA of TRP32 genes amplified from cultured isolates of *E. chaffeensis*, or from ticks, or from samples of patients' blood infected with this pathogen, has shown two to six repeats (summarized in **Table 1**). TRP120gene plays an important role in *E. chaffeensis* infection as it is a type I secretion system effector which is sumoylated on lysine residues and mediates interactions with host protein targets such as actin and myosin cytoskeleton components (Myo10) or GGA1 involved in vesicular trafficking (Wakeel et al., 2009) (**Table 1**).

Like other mammalian pathogenic bacteria, *E. chaffeensis* uses specific molecular mechanisms to evade host immune responses and to modulate host cell processes to its own advantage. Among these pathogenicity determinants, the type IV secretion system (T4SS) is a specialized protein complex involved in the injection of type IV effector (T4E) proteins into eukaryotic cells in order to subvert host cell processes during infection (Cascales and Christie, 2003). Rapid progress has been made toward identifying the proteins that form different parts of the T4SS, the translocated effectors and how these effectors subvert eukaryotic cellular processes during infection (Voth et al., 2012). However, to date, only two T4Es have been identified in the *Anaplasmataceae* family and shown to be critical for pathogenicity. After being injected in the host cells, AnkA (*Anaplasma phagocytophilum*), is tyrosine-phosphorylated in the cytoplasm at EPIYA motifs and binds to SHP-1 phosphatase (Lin et al., 2007; Garcia-Garcia et al., 2009). AnkA is then translocated to the nucleus of the infected cell and interacts with gene promoter regions, leading to the downregulation of the CYBB and other key host defense genes (IJdo et al., 2007). In *E. chaffeensis*, the only known T4E is ECH_0825, homologous to *A. phagocytophilum* Ats-1 (Liu et al., 2012). This effector is translocated to host mitochondria where it restrains ROS and apoptosis for more efficient infection.

Our laboratory developed a searching algorithm for type IV effector proteins (S4TE), which identifies candidate T4Es in genome sequences based on a combinatorial approach with 14 different parameters (Meyer et al., 2013).

To better understand the evolution and pathogenicity of *E. chaffeensis*, we analyzed the eight available *E. chaffeensis* genomes of distinct geographical origin and of varying virulence isolated from humans (**Table 1**). We identified the relationship between *E. chaffeensis* strains using comparative whole genome analysis based on phylogenetic analysis, alignment of locally collinear blocks (LCB), and analysis of shared and specific genetic content. We provide evidence that the West Paces strain belongs to genetic group II and that *E. chaffeensis* is a highly conserved species. We describe likely virulence traits (candidate type IV effectors) encoded by their genomes and some putative host cell targets. Most notably some strains lack one or two candidate T4Es, but show conserved intracellular adaptive strategies.

Our results show that using our S4TE software and approach even for strains which are really close at the intraspecies level, enables the prediction of candidate type IV effectors that could be relevant for the study of bacterial pathogenesis.

MATERIALS AND METHODS

Retrieval of Genome Sequences and Comparison of Genomes

Complete genome sequences of *E. chaffeensis* strains were obtained from the National Center for Biotechnology Information (NCBI) database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Eight complete genomes were used in this study. Orthologous groups of all *E. chaffeensis* genes were identified using the PanOCT program (Fouts et al., 2012) with the following parameters: E-value 10^{-5}, percent identity ≥ 30, and length of match ≥ 65.

Prediction of *E. chaffeensis* Type IV Effectomes

The repertoires of T4Es were predicted using a S4TE algorithm with default parameters (Meyer et al., 2013) . S4TE 1.4 predicts and ranks candidate T4Es by using a combination of 11 independent modules to explore 14 characteristics of type IV effectors. One module searches for consensus motifs in promoter regions; three modules search for the five features of the type IV secretion signal (C-terminal basicity, C-terminal charges, C-terminal hydrophobicity, overall hydrophilicity, and E-blocks); six modules search for several domains (eukaryotic–like domains, the DUF domain, EPIYA motifs, the nuclear localization signal, the mitochondrial localization signal, the prenylation domain, coiled-coil domains); and one module searches for homology with known T4Es (Meyer et al., 2013)

Analysis of Type IV Effectome Distribution According to Local Gene Density

To visualize in a single representation the distance between each gene and its closest neighbors on the five prime and three prime borders, we sorted genes into two-dimensional bins defined by the length of their 5′ and 3′ flanking intergenic regions (hereafter denoted 5′ and 3′ FIRs) (Raffaele et al., 2010). The gene density distribution is represented in R by a heat map. We used the median length of FIRs to distinguish between gene-dense regions (GDRs); in-between regions (IBRs); and gene-sparse regions (GSRs). Putative type IV effectors identified by S4TE software were plotted on this graph according to their 5′ and 3′ FIRs (**Figure 2A**). The distribution of putative T4Es in each region was calculated for each strain (**Figure 2B**).

Prediction of *E. chaffeensis* Type IV Effectors and Host Protein–Protein Interaction Networks

Protein-protein interactions between human genomes and predicted type IV effectome of *E. chaffeensis* were predicted using the Host-Pathogen Interaction Database (HPIDB) (Kumar and Nanduri, 2010) with the identity and percentage query coverage set at 30%. Based on the homology approach, the HPIDB predicts protein-protein interactions from a plentiful template of eukaryotic-prokaryotic inter-species interactions available among 68 hosts and 602 pathogens. Subcellular locations of the host proteins interacting with putative T4Es of *E. chaffeensis* were predicted using the CELLO2GO algorithm (Yu et al., 2014). S4TE 1.4 results were used to predict the location of T4Es (Meyer et al., 2013)

Phylogenetic Reconstruction and Genomic Plasticity Analysis

For phylogenetic reconstruction, whole-genome nucleotide sequences of the eight *E. chaffeensis* strains were aligned using the progressiveMauve algorithm (Darling et al., 2010, http://gel.ahabs.wisc.edu/mauve/). FastTree was used with default parameters to build the unrooted tree (Price et al., 2009). Mauve software was also used to characterize the genomic rearrangements between the eight genomes of *E. chaffeensis* by showing LCBs. In order to accurately align conserved regions in the genomes, the progressiveMauve algorithm was parameterized with a match seed weight of 15 and a minimum LCB score of 70. The seed size parameter sets the minimum weight of the seed pattern used to generate local multiple alignments (matches) during the first pass of anchoring the alignment. The LCB weight sets the minimum number of matching nucleotides identified in a collinear region in order for the region to be considered a true homology rather than a random similarity (Darling et al., 2010).

RESULTS

The *Ehrlichia chaffeensis* Genomes Are Highly Conserved

In order to establish a whole genome-based phylogeny of these eight *E. chaffeensis* strains, we used the Mauve progressive alignment and FastTree to build the tree. Our results are in agreement with those of previous studies, with the eight strains being separated into three genetic groups. The Arkansas and Osceola strains were assigned to group I, and the Wakulla, Saint Vincent, West Paces, and Heartland strains were assigned to group II. We also assigned the West Paces strain to group II due to its phylogenetically close proximity to the Heartland strain (**Figure 1A**). The Jax and Liberty strains were assigned to group III (**Figure 1A**). With an average size of 1.2 Mb, the genomic features of the eight strains used in this study are similar. The GC (guanine-cytosine) content was seen to be highly homogenous (30.1%) and genome sequences relatively well-conserved (**Figure 1B**). The number of genes ranges from 871 to 883. Whole genome alignments revealed seven LCBs with some inversions and with rearrangements in the genomes with respect to one another (**Figure 1B**). In the Arkansas strain, we found a rearrangement between three LCBs with green and orange blocks switched with yellow LCB. The strains Arkansas, Osceola, Heartland, and West Paces showed an inversion of blue and red LCBs compared to other genomes. The Saint Vincent and Wakulla strains showed inversion of one small LCB (purple, **Figure 1B**). The structural variation among these genomes suggests a low degree of inter-species genome plasticity for *E. chaffeensis*.

We then analyzed the pan-genome of *E. chaffeensis*. We used PanOCT software to cluster the ortholog and compared the core and accessory genomes of the eight strains of *E. chaffeensis*.

TABLE 1 | Main biological and genetic characteristics of the eight *Ehrlichia chaffeensis* strains analyzed.

Strain	Arkansas	Heartland	Jax	Liberty	Osceola	St. Vincent	Wakulla	West Paces
Year	1991	1999	1997	1998	1997	1996	1997	2000
Origin	Arkansas	Nebraska	Florida	Florida	Florida	Georgia	Florida	Tennessee
Source	21-year old male	Human	51-year old Woman	Human	Human	52-year old, HIV+	Human	Human
Human, clinical and laboratory observations	Fever, headache, pharyngitis, nausea, vomiting, and dehydration. Cervical lymphadenopathy, splenomegaly, thrombocytopenia, leukopenia with left shift, elevations in serum aspartate transaminase concentration.	HME, no clinical description available.	Fever, non-productive cough, nausea, vomiting, and diarrhea, profoundly lethargic. Thrombocytopenia, leukopenia, elevations in serum aspartate transaminase concentration, doxycycline therapy, cerebrospinal fluid mononuclear pleocytosis, pulmonary oedema, hypotension, and anuria. The patient died in hospital on day 6.	Acute HME, no clinical description available.	Acute HME, no clinical description available.	Fever, headache, myalgia, nausea, and vomiting, orthostatic hypotension, thrombocytopenia, leukopenia with left shift, elevations in serum aspartate transaminase concentration, doxycycline therapy. Lobar pneumonia and acute renal failure. The patient died.	Acute HME, no clinical description available.	Acute HME, no clinical description available.
SCID mice pathogenesis	Mild, chronic	UN	UN	Acute, severe	UN	UN	Acute, lethal	UN
Genetic Group/ TRP32/TRP120 repeats	I/4/4	II/3/3	III/4/4	III/4/4	I/4/3	II/3/3	II/6/4	II/3/3
Literature	Dawson et al., 1991	Sumner et al., 1999	Paddock et al., 1997	Sumner et al., 1999	Sumner et al., 1999	Paddock et al., 1997	Sumner et al., 1999	Cheng et al., 2003

Human isolates of E. chaffeensis were classified in three genetic groups according to the 28-kDa major outer membrane gene cluster, the number of TRP32 repeats (variable-length PCR target gene, NCBI accession version # WP_011452439.1) and the number of TRP120 repeats (120-kDa immunodominant surface protein, NCBI accession version # WP_011452362.1). Data for pathogenesis in SCID mice for E. chaffeensis Arkansas, Liberty and Wakulla strains were obtained from Miura and Rikihisa (2007) and indicated. UN, unknown.

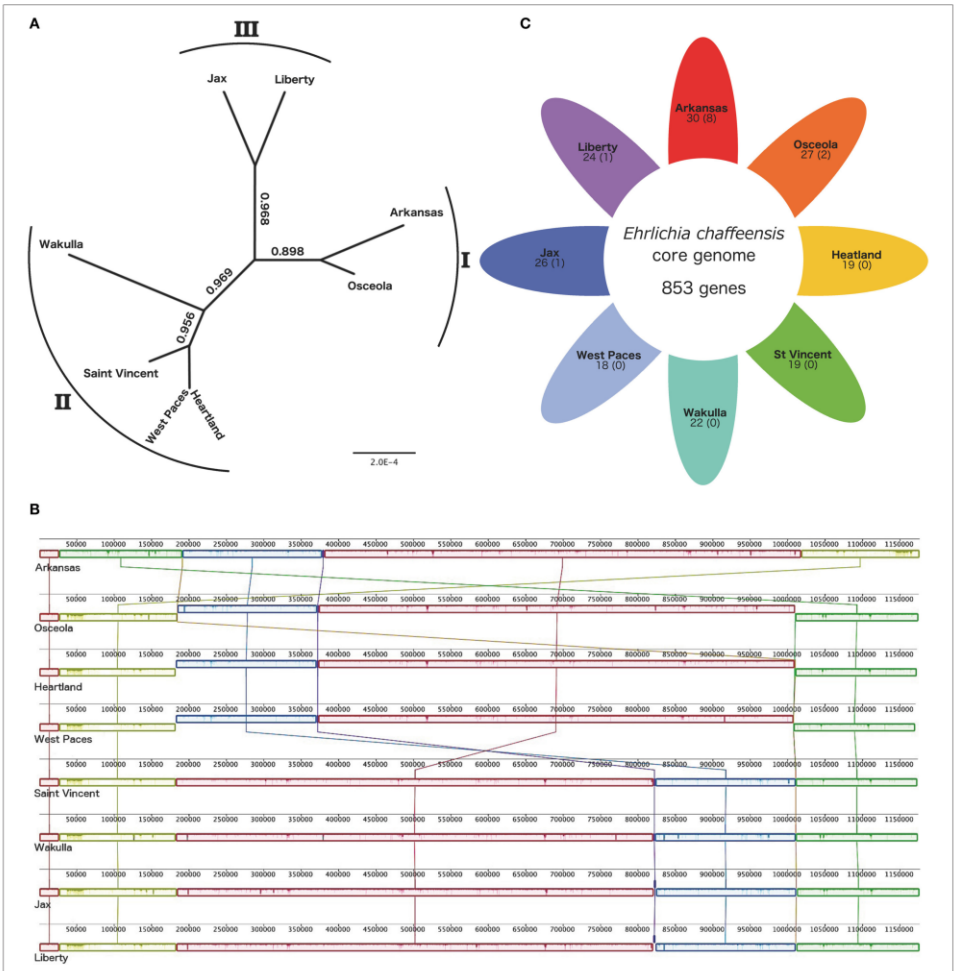


FIGURE 1 | Comparative genomics of 8 *Ehrlichia chaffeensis* strains. (A) Phylogenetic tree of 8 *E. chaffeensis* strains. FastTree based on the Mauve alignment of the whole genomes of 8 *E. chaffeensis* strains. The node values indicate the local support values of the Shimodaira-Hasegawa test. The number outside the tree shows the genetic group of each strain, the West Paces strain was assigned to genetic group II due to the high level of conservation with the Heartland strain. **(B)** Alignments of 8 *E. chaffeensis* genomes generated using Mauve software (Darling et al., 2010) (http://gel.ahabs.wisc.edu/mauve/). Locally collinear blocks (LCBs), shown as rounded rectangles, represent regions with no rearrangement of homologous sequences across genomes. The forward or reverse orientation of the LCBs is indicated by their position, respectively above or below the line. Lines between the genomes trace orthologous LCBs. Using default parameters resulting in a minimum LCB weight of 70, there are 7 LCBs across all the genomes. The LCB weight defines the minimum number of matching nucleotides in a collinear region for it to be considered homologous across genomes and not the result of a spurious match. Regions outside LCBs were too divergent in at least one genome to be aligned successfully. Inside each LCB, vertical bars represent the similarity profile of the genome sequence. The height of each bar corresponds to the average level of conservation in that region of the genome sequence. **(C)** Shared and specific gene content between 8 *E. chaffeensis* strains. Each colored petal represents a different *E. chaffeensis* genome. The number in the center of the diagram represents the number of orthologous genes shared by all the genomes, thus defining the *E. chaffeensis* core genome. The number inside each individual petal corresponds to the number of genes that are absent from the core genome, and the numbers in brackets correspond to the number of genes specific to this strain. The number outside each petal shows the genetic group of each strain.

TABLE 2 | Putative type IV effectors (T4Es) identified by the S4TE algorithm.

Ehrlichia chaffeensis strains				NCBI protein names																	
Arkansas	Liberty	Wakulla	West Paces		S4TE Score	Promoter motif	Homology	Euk-like domain	EPIYA	NLS	MLS	Prenylation	Coiled coils	Cter basicity	Cter charges	Cter basicity	Cter charges	Cterhydrophobicity	Eblock	Location
ECH_RS02870	ECHLIB_RS01940	ECHWAK_RS01950	ECHWP_RS02750	Hypothetical protein	229	1	1	1	0	1	1	0	0	0	1	0	0	1	0	GSRs
ECH_RS03425	ECHLIB_RS01385	ECHWAK_RS01390	ECHWP RS03295	Hypothetical protein	164	1	1	0	0	0	0	0	0	0	0	1	0	0	0	GDRs
ECH_RS04335	ECHLIB_RS00490	ECHWAK_RS00485	ECHWP_RS00485	Gamma carbonic anhydrase family protein	151	1	1	0	0	0	0	0	0	0	1	0	0	0	1	GSRs
ECH_RS02750	ECHLIB_RS02060	ECHWAK_R502070	ECHWP R502630	Hypothetical protein	141	1	0	1	0	1	1	0	0	0	1	0	0	0	0	GDRs
ECH_RS02620	ECHLIB_RS02190	ECHWAK_RS02200	ECHWP_RS02500	Alpha/beta hydrolase	139	1	1	0	0	0	1	0	0	0	1	1	0	0	0	GSRs
ECH_RS03745	ECHLIB_RS01065	ECHWAK_RS01070	ECHWP RS03615	AI-2E family transporter	122	1	1	0	0	0	0	0	0	0	1	0	0	0	0	GSRs
ECH_RS00450	ECHLIB_RS04345	ECHWAK_RS04360	ECHWP_RS04320	Hypothetica l protein	118	1	0	0	1	0	1	0	0	0	1	0	0	0	1	GSRs
ECH_RS01210	ECHLIB_RS03585	ECHWAK_RS03600	ECHWP RS01105	DNA ligase (NAD(+)) LlgA	117	0	0	1	0	1	1	0	0	0	1	0	0	1	0	BRs
ECH RS04225	ECHLIB RS00595	ECHWAK RS00600	ECHWP RS00590	Hypothetica l protein	115	1	0	0	0	1	1	0	0	0	1	1	0	1	0	GSRs
	ECHLIB_RS02720			Hypothetical protein	114	0	0	0	0	1	1	0	0	0	1	0	0	1	1	GDRs
ECH_RS02365	ECHLIB_RS02435	ECHWAK_RS02455	ECHWP_RS02250	Translation initiation factor IF-2	114	0	0	0	0	1	1	0	0	0	0	1	0	1	1	GDRs
ECH_RS03205	ECHLIB_RS01605	ECHWAK_RS01615	ECHWP RS03075	Diguanylate cyclase response regulator0	109	0	1	0	0	0	1	0	0	0	1	0	0	0	0	IBRs
ECH_RS03195	ECHLIB_RS01615	ECHWAK_RS01625	ECHWP_RS03065	NAD-glutamate dehydrogenase	108	1	0	0	0	1	0	0	1	0	1	0	1	1	0	IBRs
ECH_RS02495	ECHLIB_RS02315	ECHWAK_R502330	ECHWP R502375	Peptide chain release factor 1	105	1	0	0	0	0	0	0	0	0	1	0	0	1	1	IBRs
ECH_RS04685	ECHLIB_RS04640	ECHWAK_RS04650		Hypothetica l protein	103	0	0	0	1	1	1	0	0	0	1	1	1	1	0	GSRs
ECH_RS01570	ECHLIB_RS03225	ECHWAK_RS03240	ECHWP RS01465	Hypothetical protein	101	0	0	0	0	1	1	0	0	0	1	1	1	1	0	GDRs
ECH_RS04230	ECHLIB_RS00590	ECHWAK_RS00595	ECHWP_RS00585	Hypothetica l protein	99	1	0	0	0	1	1	0	0	0	0	0	0	1	0	GSRs
ECH RS02945	ECHLIB RS01860	ECHWAK RS01875	ECHWP RS02825	Transcriptional regulator	98	0	1	0	0	0	0	0	0	0	1	0	0	0	0	IBRs
ECH_RS02385	ECHLIB_RS02415	ECHWAK_RS02435	ECHWP_RS02270	Hypothetica l protein	98	1	0	0	0	1	0	0	0	0	1	0	0	1	0	GSRs
ECH_RS03860	ECHLIB_RS00950	ECHWAK_RS00955	ECHWP RS03725	Hypothetical protein	97	1	0	0	0	0	1	0	0	0	1	0	1	1	0	IBRs
ECH_RS04650	ECHLIB_RS00175	ECHWAK_RS00175	ECHWP_RS00175	Protein translocase subunit SecA	93	1	0	0	0	0	1	0	0	0	1	1	0	1	0	GSRs
ECH_RS02080	ECHLIB_R502725	ECHWAK_R502740	ECHWP R501965	Hypothetical protein	93	0	0	0	0	1	0	0	0	0	0	0	0	0	1	GDRs
ECH_RS02075	ECHLIB_RS02730	ECHWAK_RS02745	ECHWP_RS01960	Conjugal transfer protein TrbI	92	0	0	0	0	0	1	0	0	0	1	0	0	1	1	GDRs
ECH_RS03040	ECHLIB_RS01765	ECHWAK_R501780	ECHWP R502910	Peptidylprolyl isomerase	91	0	0	0	0	1	1	0	0	0	1	1	0	1	0	GSRs
ECH_RS02255	ECHLIB_RS02545	ECHWAK_RS02565	ECHWP_RS02140	165 rRNA (uracii(1498)-N(3))-methyltransferase	88	0	1	0	0	0	0	0	0	0	0	0	0	0	0	GDRs
ECH_RS03605	ECHLIB_RS01205	ECHWAK_R501210	ECHWP R503475	Hypothetical protein	87	1	0	1	0	0	1	0	0	0	1	0	1	0	0	IBRs
ECH RS03515	ECHLIB RS01295	ECHWAK RS01300	ECHWP RS03385	Hypothetical protein	87	1	0	0	0	0	1	0	0	0	1	0	0	1	0	GSRs
ECH_RS02340	ECHLIB_RS02460	ECHWAK_RS02480	ECHWP RS02225	Hypothetical protein	87	1	0	0	0	0	1	0	0	0	1	0	0	1	0	IBRs
ECH_RS01565	ECHLIB_RS03230	ECHWAK_RS03245	ECHWP_RS01460	Exodeoxyribonuclease V subunit beta	87	1	0	0	0	0	1	0	0	0	1	0	0	1	0	IBRs

(Continued)

TABLE 2 | Continued

Ehrlichia chaffeensis strains				NCBI protein names																
Arkansas	**Liberty**	**Wakulla**	**West Paces**		**S4TE Score**	**Promoter motif**	**Homology**	**Euk-like domain**	**EPIYA**	**NLS**	**MLS**	**Prenylation**	**Coiled coils**	**Cter basicity**	**Cter charges**	**Cter basicity**	**Cter charges**	**Cterhydrophobicity**	**Eblock**	**Location**
ECH_RS01140	ECHLIB_RS03660	ECHWAK_RS03675	ECHWP RS01035	Hypothetical protein	87	1	0	0	0	0	1	0	0	0	1	0	0	1	0	GSRs
ECH_RS03890	ECHLIB_RS00920	ECHWAK_RS00925	ECHWP_RS03755	DNA-directed RNA polymerase subunit beta	85	0	0	0	0	1	1	0	0	0	1	0	0	1	0	GDRs
ECH_RS00205	ECHLIB_RS04580	ECHWAK_RS04590	ECHWP RS04555	Type IV secretion system protein VirD4	85	0	0	0	0	1	1	0	0	0	1	0	0	1	0	IBRs
ECH_RS03630	ECHLIB_RS01180	ECHWAK_RS01185	ECHWP_RS03500	DNA processing protein DprA	84	1	0	0	0	0	1	0	0	0	1	0	1	0	1	GSRs
ECH_RS03440	ECHLIB_RS01370	ECHWAK_RS01375	ECHWP RS03310	Phage capsid protein	82	1	0	0	0	0	0	0	0	0	1	1	0	1	0	GSRs
ECH_RS03530	ECHLIB_RS01280	ECHWAK_RS01285	ECHWP_RS03400	Hypothetica l protein	81	0	0	0	0	0	0	0	0	0	1	0	0	1	1	IBRs
~RS01950	ECHLIB_RS02855	ECHWAK_RS02870	ECHWP RS01835	Molecular chaperone DnaK	81	0	0	0	0	0	0	0	0	0	1	0	0	1	1	IBRs
ECH_RS03610	ECHLIB_RS01200	ECHWAK_RS01205	ECHWP_RS03480	Hypothetica l protein	80	0	0	0	0	1	0	0	0	0	1	1	0	1	0	GSRs
ECH_RS03260	ECHLIB_RS01550	ECHWAK_RS01560	ECHWP RS03130	abc-ATPase UvrA	77	1	0	0	0	0	1	0	0	0	0	0	0	1	0	GSRs
ECH_RS03895	ECHLIB_RS00915	ECHWAK_RS00920	ECHWP_RS03760	505 ribosomal protein L7 /L12	74	1	0	0	0	0	1	0	0	0	1	0	0	0	1	GDRs
ECH_RS02525	ECHLIB_RS02285	ECHWAK_RS02300	ECHWP RS02405	Glutamate-tRNA ligase	74	1	0	0	0	0	1	0	0	0	1	0	0	0	1	GSRs
ECH RS00785	ECHLIB_RS04010	ECHWAK_RS04025	ECHWP_RS03985	Hypothetica l protein	74	0	0	0	0	1	0	0	0	0	1	0	0	1	0	GSRs
'Eei(RS00330	ECHLIB_RS04455	ECHWAK_RS04465	ECHWP RS04430	1-acyl-sn-glycerol-3-phosphate acyltransferase	74	1	0	0	0	0	1	0	0	0	1	0	0	0	1	GSRs
ECH_RS00255	ECHLIB_RS04530	ECHWAK_RS04540	ECHWP_RS04505	Hypothetical protein	74	1	0	0	0	0	1	0	0	0	1	0	0	0	1	BRs
ECH_RS03415	ECHLIB_RS01395	ECHWAK_RS01400	ECHWP RS03285	NAD+ synthetase	73	1	0	0	0	1	1	0	0	0	1	1	0	0	0	GSRs
ECH_RS02490	ECHLIB_RS02320	ECHWAK_RS02335	ECHWP_RS02370	GTP-binding protein	73	1	0	0	0	0	0	0	0	0	1	0	1	0	1	IBRs
ECH_RS00505	ECHLIB_RS04290	ECHWAK_RS04305	ECHWP_RS04265	Citrate (Si)-synthase	73	1	0	0	0	0	0	0	0	0	0	1	0	1	1	GDRs
ECH RS03555	ECHLIB RS01255	ECHWAK RS01260	ECHWP RS03425	Hypothetica l protein	72	0	0	0	0	0	0	0	1	0	0	1	1	0	1	GSRs

This table shows the candidate T4Es identified by S4TE software in four E. chaffeensis strains. The Liberty strain is used as a reference to sort predicted effectors, and the homolog candidate effectors are ranked by S4TE scores. Each T4E is defined by the gene ID, Name, and S4TE features.

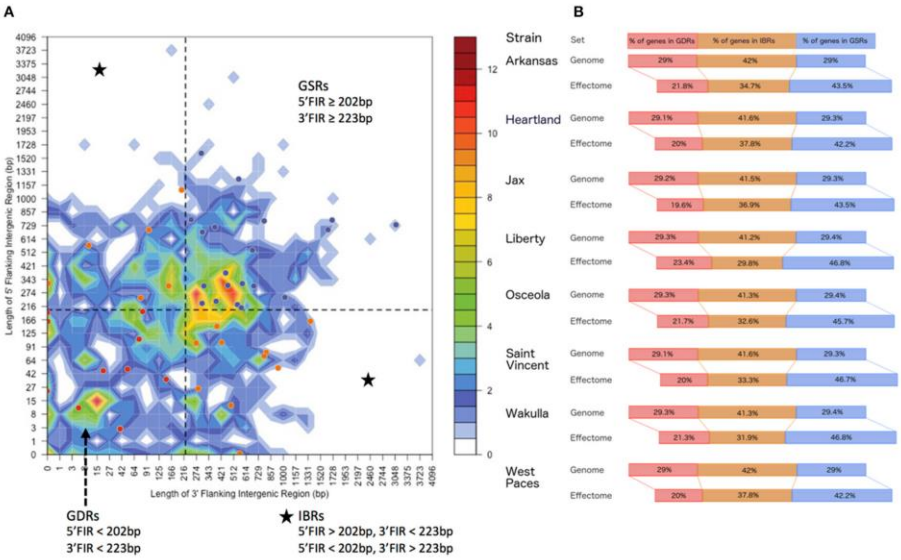


FIGURE 2 | Distribution of *Ehrlichia chaffeensis* effectomes according to local gene density. (A) Distribution of *E. chaffeensis* str. Arkansas genes according to the length of their flanking intergenic regions (FIRs). All *E. chaffeensis* genes were sorted into 2-dimensional bins according to the length of their 5′ (y-axis) and 3′ (x-axis) FIRs. The number of genes in the bins is represented by a color-coded density graph. Genes whose FIRs are both longer than the median FIR length were considered as gene-sparse region (GSR) genes. Genes whose FIRs are both below the median value were considered as gene-dense region (GDR) genes. In-between region (IBR) genes are genes with a long 5′FIR and short 3′FIR, or inversely. Candidate effectors predicted using the S4TE algorithm were s plotted on this distribution according to their own 3′ and 5′ FIRs. A color is assigned to each of the three following groups: Red to GDRs, orange to IBRs, and blue to GSRs. **(B)** Distribution of genes in GDRs, IBRs, and GSRs of *E. chaffeensis* strains. The proportion of the genome and the effectome that occurs in GDRs (red), IBRs (orange), and in GSRs (blue) is indicated.

(**Figure 1C**). The *E. chaffeensis* core-genome contained 853 orthologous genes, corresponding to ∼96% of the pan-genome and indicating that the *E. chaffeensis* accessory genome is narrow. Thus, four percent of *E. chaffeensis* genes are not in the core genome and only a few genes are specific to four out of these eight strains. The Arkansas strain harbored eight specific genes, the Osceola strain two specific genes and Liberty and Jax strains only one specific gene (**Figure 1C**).

To test if the genome plasticity and effector repertoires can explain the differential intra-species pathogenesis of *E. chaffeensis*, we decided to focus our study on four different representative strains. When data were available, we chose strains belonging to different genetic groups showing variations in virulence. From genetic group I, we chose the Arkansas strain, which is the most widely studied and best-described strain in the literature. This strain shows mild virulence in immunodeficient mice (Miura and Rikihisa, 2007). From genetic group II, we chose the West Paces and Wakulla strains, the latter causing acute lethal pathogenesis in SCID mice (Miura and Rikihisa, 2007). Finally, from genetic group III, we chose the Liberty strain, which causes acute pathogenesis in immunodeficient mice (Miura and Rikihisa, 2007).

Prediction of Type IV Effectors for *E. chaffeensis* Identifies the Core Type IV Effectome among Four Human Isolates

We used the S4TE algorithm to predict and compare the type IV effector repertoires in four human isolates (Arkansas, Liberty, Wakulla, and West Paces) of *E. chaffeensis* in order to determine how these repertoires differed between strains with respect to the presence or absence of whole candidate T4Es. We identified a conserved repertoire of 45 candidate T4Es, defining the core type IV effectome of *E. chaffeensis*.

Based on orthology analysis, we found few differences in T4E content between the four selected *E. chaffeensis* isolates. *E. chaffeensis* str. Liberty was the only strain to own all 47 predicted T4Es (**Table 2**). One candidate T4E, ECHLIB_RS02720, is specific to *E. chaffeensis* str. Liberty, whereas ECHLIB_RS04640 was only absent in *E. chaffeensis* str.

FIGURE 3 | Protein-protein interaction network between the *E. chaffeensis* str. Liberty effectome and the human genome. A sub-cellular location was predicted with the S4TE algorithm (http://sate.cirad.fr) for *Ehrlichia* candidate effectors (left) and with CELLO2GO software (http://cello.life.nctu.edu.tw/cello2go/) for human proteins (right). Blue and red circles represent predicted T4Es located in the cytoplasm and in the nucleus of the host cell, respectively. Blue, red, pink, green, purple, yellow, and turquoise hexagons represent the different locations of targeted human proteins in the host cell. Hexagons harbor several colors when CELLO2GO predicts several probable subcellular locations.

West Paces. All the other predicted T4Es (94% of predicted effectors) are common to the four strains, revealing the low diversity of effector repertoires in *E. chaffeensis* species. We did not discover any relation between the presence or absence of an effector and the variations in virulence exhibited by the Wakulla, Liberty, and Arkansas strains.

Identified candidate T4Es were sorted according to their S4TE score, which ranged from 72 (corresponding to the S4TE algorithm threshold) to 229 (**Table 2**). Eight candidate T4Es showed homology with known T4Es (17% of predicted T4Es) as indicated by the number "1" in the Homology column in **Table 2**. Among these candidates, one is ECH_RS01385 previously called ECH0825 (old NCBI locus_tag) (Liu et al., 2012; Table S1). This effector was predicted with the second highest S4TE score of 164. The first predicted T4E, ECH_RS01940, matched the homologous gene of *A. phagocytophilum* AnkA (IJdo et al., 2007; Lin et al., 2007; Garcia-Garcia et al., 2009; Table S1). ECHLIB_RS02190, ECHLIB_RS01065, ECHLIB_RS01605, and ECHLIB_RS01860 are four candidate T4Es presenting homologies with known *Coxiella burnetii* effectors (Table S1). ECHLIB_RS022545 shows homology with a known *Legionella pneumophila* T4E (lpg2936, 16S ribosomal RNA methyltransferase RsmE) while ECHLIB_RS00490 shows homology with a *Brucella* effector (Table S1).

Besides homology with known effectors, several other candidate T4Es had interesting features (**Table 2**). Indeed, 59.6% of predicted T4Es had a promoter motif such as PmrA upstream of the effector genes of *Coxiella* spp. and *Legionella* spp. Furthermore, 8.5% of putative T4Es harbored eukaryotic-like domains such as AnkA (Ankyrin repeat-containing domain) and BRCT (phospho-protein binding domain) domains. Only two putative T4Es contained domains of unknown function (DUF). It is interesting to note that 38.3 and 61.7% of candidate T4Es had a tyrosine phosphorylation domain (EPIYA) and a nuclear localization signal (NLS), respectively. Moreover, nearly 38% of the proteins harboring an NLS also had an EPIYA phosphorylation domain. None of the predicted T4Es had a prenylation domain or a coiled-coil domain. Thirty-four percent of the candidate T4Es harbored the canonical *L. pneumophila* secretion domain (E-block).

Concerning other features related to the type IV secretion signal, 17% of the predicted T4Es showed C-terminal hydrophobicity, 68% showed global hydropathy < -200 (on the Kyte-Doolittle scale), 21.3% had a C-ter charge ≥ 2 and 89.4% had at least three alkaline amino acids in C-terminal 25 amino acids.

Putative Type IV Effectors of *E. chaffeensis* are Overrepresented in Gene Sparse Regions of the Genome

In order to understand how genomic plasticity influences the distribution of predicted T4Es, we first analyzed the genome architecture of *E. chaffeensis* by looking at local gene density (**Figure 2**). The median length of 3′ and 5′ flanking intergenic regions (FIRs) delimits four coherent gene pools when combined with the 2-variable binning representations (**Figure 2A**).

The gene dense region (GDR, genes with 5′ FIR $<$ 202 bp and 3′ FIR $<$ 223 bp) contains 254 genes, which account for 28.9% of the *E. chaffeensis* str. Arkansas genome (**Figure 2A**). The gene sparse region (GSR, genes with 5′ FIR $\geq$ 202 bp and 3′ FIR $\geq$ 203 bp) includes 255 genes, which account for 29% of the genome (**Figure 2A**).

The other two quadrants define in-between regions (IBRs) grouping genes with a 5′ FIR shorter than the median length and a longer 5′ FIR, and inversely. In the *E. chaffeensis* str. Arkansas genome, 370 genes, which account for 42% of the genome, fall into IBRs (**Figure 2A**). This genome architecture of *E. chaffeensis* str. Arkansas is representative of other strains of *E. chaffeensis* (**Figure 2B**).

We then performed a detailed analysis of the distribution of predicted *E. chaffeensis* T4Es according to local gene density. We found that the predicted T4Es of all isolates of *E. chaffeensis* frequently had both FIRs above the genome median value. Although 29% of *E. chaffeensis* genes belong to GSRs, 42.2% to 46.8% of predicted type IV effector genes fall in GSRs (**Figures 2A,B**). Thus, compared to the whole genome, the GSRs showed a 1.5-fold enrichment in candidate type IV effector genes. Consequently, the proportion of candidate T4Es in the GDRs and IBRs is lower than the proportion of genes of the whole genome (**Figure 2B**). These results suggest that plastic regions with low gene density harbor pathogenicity genes and could play a role in host-bacteria interactions.

Prediction of the Host-Pathogen Protein–Protein Interaction Network

We predicted the interactions of *E. chaffeensis* T4Es with human proteome and identified 57 protein-protein interactions with the involvement of 13 putative T4Es of *E. chaffeensis* str. Liberty (which harbors all predicted T4Es) and 56 human proteins (**Figure 3**).

The targeted host proteins are located in cellular compartments relevant to the pathogenesis mechanisms. The predicted cellular localizations of human interacting proteins were confirmed in cytoplasm, nucleus, extracellular, mitochondrial, plasma membrane, endoplasmic reticulum, and cytoskeleton (**Figure 3**). As described above, we predicted the subcellular localization in human host cell of *E. chaffeensis* T4Es using the S4TE algorithm (**Table 2**, Table S1). Out of the 13 predicted T4Es of *E. chaffeensis* that interact with human proteins, eight (~60%) harbor at least one nuclear location signal (NLS). Interestingly, most of these proteins had putative human targets located in the nucleus (**Figure 3**).

Thus, the putative targets of the ABC-ATPase UvrA (ECHLIB_RS01550) are involved in different processes including innate immunity, response to stress, the cell cycle, cell signaling, and cell death (Table S2). Another candidate nuclear effector (ECHLIB_RS02285) interacts with 11 human proteins, most of which are involved in metabolic processes such as amino acid synthesis (IDHC), carbohydrate metabolic process (DPM1 and IDHC), lipid metabolism (DPM1), and nitrogen compound metabolism (ZN627, IDHC, DPM1, WBS22, CNTRL, and PUF60) (Table S2).

The nuclear effectors ECHLIB_RS02315 and the DNA ligase ECHLIB_RS03585 interact with several putative targets involved in immune and stress responses, cell organization, and cell death. Most of the proteins targeted by ECHLIB_RS00175 are located in the nucleus and are involved in nuclear organization (chromosomal protein HMG2) or biosynthetic process (proline-rich and zinc finger proteins) (Table S2). The nuclear effector ECHLIB_RS02435 interacts with kinases and with the nuclear transcriptional repressor BCLF1 suggesting an important role in signal transduction and stress response, particularly activation of response to DNA damage. It is of note that this effector also harbors a tyrosine phosphorylation domain that could play an important role in the ATM/MAP kinases signaling pathway.

A dihydropyrimidine dehydrogenase is the only target of ECHLIB_RS00920 involved in catabolic and metabolic processes. The last putative nuclear effector with a target is a ligase (ECHLIB_RS01395), which interacted with a protein associated with cytochrome c oxidase and had one putative target on the human genome. This protein plays a role in the organization of mitochondria, the assembly of cell components and in the generation of precursor metabolite and energy.

Among the other *E. chaffeensis* T4Es whose interaction with human proteins was predicted, the transcriptional regulator ECHLIB_RS01860 interacts with two putative targets. The first is the nuclear factor NF-kappa-B, which plays a prominent role in immune responses, responses to stress, and cell death. The second target is SDCB1, which is involved in cytoskeleton organization, cell-cell signaling, locomotion, cell adhesion, and growth (Table S2).

Finally, four other cytoplasmic *E. chaffeensis* T4Es (ECHLIB_RS01605, ECHLIB_RS02855, ECHLIB_RS04290, ECHLIB_RS01280) were predicted to interact with one or two proteins involved in reticulum catabolic processes (HERP2), protein folding (SACS), cytoskeleton organization (INF2) and transcriptional repression (ARI4B), or in immune response (antigen processing by HG2A), respectively.

DISCUSSION

Motivated by the availability of eight genome sequences, we explored the world of pathogenicity determinants in the species *E. chaffeensis*. We hypothesized that variations in virulence between some strains could be driven by genome plasticity and the acquisition of different repertoires of type IV effectors (T4Es). Such mechanisms of evolution have already been observed in plant pathogenic and non-pathogenic *Xanthomonas* (Cesbron et al., 2015). The aim of our work was to show that computational methods to identify and categorize putative T4Es, prior to their functional characterization, could be a valuable approach to better understand *E. chaffeensis*-host interactions. We also aimed to identify novel candidate T4Es and their interactions with host cell proteins to advance our current understanding of *E. chaffeensis* pathogenesis.

We showed that *E. chaffeensis* genomes had low plasticity and with few intra-species genomic rearrangements. We also showed that the eight genomes of *E. chaffeensis* are highly conserved with 96% genes present in the core genome. Hence, the observed differences in pathogenesis and symptoms between the Arkansas, Liberty and Wakulla strains (**Table 1**) could be due to the absence of certain genes in the core genome.

The core type IV effectome of a bacterial species is defined by the minimum set of type IV effectors conserved in all strains within a species, which make it necessary for the bacterium to develop inside the host cell. Using our comparative genomics approach, we showed that the core type IV effectome of *E. chaffeensis* contains 45 candidate T4Es. In addition, we showed that the Liberty isolate of *E. chaffeensis* contains all the 47 predicted T4Es. Although, S4TE software was designed for optimal sensitivity (Meyer et al., 2013), the prediction of false positives can occur and is inherent to any predictive computational approach.

However in our study, the S4TE algorithm correctly predicted the two known type IV effectors in *Anaplasmataceae* family with *E. chaffeensis* mitochondrial effector ECH0825 (ECHLIB_RS01385) and the homolog of *A. phagocytophilum* nucleomodulin AnkA (ECHLIB_RS01940) (Table S1). In addition, S4TE predicted effectors that are homologous to known effectors in other bacteria, including *C. burnetii*, *L. pneumophila*, and *Brucella* spp. S4TE also predicted some new candidate T4Es that were not easy to identify *ab initio,* based solely on the poor quality of automated genome annotations, especially for bacteria harboring 30% or more unknown hypothetical proteins like *Anaplasmataceae*. For example, S4TE identified some bacterial enzymes as candidate effectors, including the annotated acyltransferase ECHLIB_RS04455, which is in agreement with current knowledge on bacterial effectors (Anderson et al., 2015).

Most of the predicted T4Es in *E. chaffeensis* belong to the core type IV effectome, showing that effector repertoires are highly conserved in this species. Thus, for bacteria with compact genomes, the type IV effector repertoires may not reflect the genetic diversity and the variations in pathogenesis observed within a species. However, two candidate T4Es, ECHLIB_RS02720 and ECHLIB_RS04640, are not present in all *E. chaffeensis* strains and could explain some within-strain variations in virulence. Indeed, in pathogens with bigger genomes and more complex lifestyles, some authors demonstrated that diversity in effector repertoires is linked to host specificity (Cooke et al., 2012; Guyon et al., 2014; Schwartz et al., 2015). The 45 T4Es predicted by S4TE in *E. chaffeensis* account for about 5% of the genome. In comparison, in the facultative intracellular *L. pneumophila* str. Philadelphia I, which contains a well-characterized type IV effectome, 286 T4Es account for about 9% of the genome (Lifshitz et al., 2013). Thus, in relation to the number of genes, the predicted type IV effectome of *E. chaffeensis* is significantly smaller than that of *L. pneumophila*. This could be explained by the reduced size of the *E. chaffeensis* genome, linked to its obligate intracellular lifestyle, thus leading to less functional redundancy in the type IV effectome.

Interestingly, the *E. chaffeensis* Liberty strain contained one specific candidate T4E, ECHLIB_RS02720, a hypothetical protein exhibiting EPIYA and NLS domains as well as a classical type IV secretion signal. These features strongly suggest this effector could be phosphorylated in the cytoplasm, addressed to the

nucleus, and play an important role inside the host cell, like the AnkA effector of *A. phagocytophilum* (IJdo et al., 2007; Garcia-Garcia et al., 2009). This effector could also be involved in the differential virulence phenotypes described between the Arkansas and Liberty strains in SCID mice (Miura and Rikihisa, 2007). Conversely, the identical putative type IV effectomes of the Arkansas and Wakulla strains cannot explain their differential pathogenesis in SCID mice. We cannot exclude the possibility that the homologous T4Es repertoires of these two strains contain point mutations in some effectors, which would alter the pathogenesis of the strain, as shown in *L. pneumophila* with the mutant protein kinase LegK2 (Hervet et al., 2011). Another explanation could be differences in the metabolisms or the kinetics of infection of the Arkansas and Wakulla strains. Indeed, Marcelino et al. showed that virulent and attenuated Gardel strains of *E. ruminantium*, which have the same gene content, only differ in their proteome expression, yet have different life cycles (Marcelino et al., 2015). At the whole genome level, some horizontal gene transfer (HGT) of genes that control advantageous phenotypic differences, might also have occurred during evolution to explain the differing degrees of virulence between Wakulla, Liberty and Arkansas isolates of *E. chaffeensis* (Dorman et al., 2016).

We demonstrated that predicted T4Es are preferentially distributed in gene sparse regions of the genome. In addition, some putative effectors harbor typical eukaryotic features such as Ank or BRCT domains. These results suggest that some effectors could be acquired via HGT from other bacterial species (McAdam et al., 2014) or from the host cell (Lurie-Weinberger et al., 2010).

To guide the functional characterization of the candidate T4Es of interest with respect to *E. chaffeensis* pathogenesis, we tried to predict some putative host targets. Among the 47 candidate T4Es in *E. chaffeensis* str. Liberty, most of the proteins with predicted NLSs were predicted to interact with human proteins located in the nucleus. Moreover, several putative targets of candidate T4Es affect human immunity-related proteins. Two predicted T4Es (ECHLIB_RS01550 and ECHLIB_RS01860) could interact with the nuclear factor NF-kappa-B1. This is a pleiotropic transcription factor induced by a vast array of stimuli and which is linked to many biological processes, including immunity, inflammation, and apoptosis. Another predicted T4E (ECHLIB_RS01280) may play a role in controlling innate immune responses by interacting with two human proteins in particular, ARIA4B and HG2A. The first is a transcriptional repressor, and the second plays a critical role in MHC class II antigen processing by stabilizing peptide-free class II alpha/beta heterodimers in a complex. Suppressing innate immunity of the host cells is one of the necessary actions for the proper development of this intracellular bacterium (Luo, 2012).

Other putative T4Es could affect host cell transcription like ECHLIB_RS01605, which targets two transcriptional repressors: N42L2 and HERP2. On the other hand, some putative targets involve the global organization of cell membranes. Thus, COA6 is involved in the maturation of the mitochondrial respiratory chain complex IV; CO1A2 and CO4A1 are involved in the extracellular membrane by forming fibrillar collagen, with SDCB1 playing a role in vesicular trafficking (Zimmermann et al., 2001). This modification of global membrane organization could be related to the lysosome-like vacuole recruitment in intracellular bacteria, as shown in *C. burnetii* (Moffatt et al., 2015).

Our analysis of the protein-protein interaction network also revealed that certain candidate T4Es could alter the phosphorylation cascades by putatively interacting with protein kinases (FYN, PTK7, TIE2, KAP2, ATM, MK14), enzymes which catalyze phosphorylation reactions (Dhanasekaran and Premkumar Reddy, 1998). Phosphorylation-dephosphorylation mechanisms are extremely common in signaling pathways where they regulate cell activity (Dhanasekaran and Premkumar Reddy, 1998). For example, PTK7 is a catalytically inactive receptor tyrosine kinase which is upregulated in many common human cancers. Knocking down this protein was shown to inhibit cell proliferation and induce apoptosis (Meng et al., 2010). MK14 is a serine/threonine kinase, which is an essential component of the MAP kinase signaling pathway. MK14 is one of the four p38 MAPKs that play important roles in the cascade of cell responses induced by extracellular stimuli, such as pro-inflammatory cytokines or physical stress, leading to direct activation of transcription factors (Lo et al., 2014). Blocking these cascades could enable the bacterium to evade the innate immune response of the host cell. ATM/MKA14 regulatory networks have also been shown to regulate cytoplasmic targets, resulting in extensive cytoskeletal rearrangements (Pines et al., 2011). Acting on these cascades could favor the maturation of *Ehrlichia*-containing vacuoles, as shown for *L. pneumophila* which controls vesicle trafficking to escape host defenses and counteract the endocytic pathway (Michard et al., 2015). Finally, some candidate T4Es could affect metabolic proteins, like SYVN1, which acts as an E3 ubiquitin-protein ligase. Ubiquitination is a post-traductional biochemical modification that mainly leads to the degradation of ubiquitinated proteins by the proteasome. Moreover, it has been shown that ubiquitination of proteins in the endoplasmic reticulum negatively regulates the stress-induced apoptotic signaling pathway (Kaneko et al., 2002). Interestingly, we found another candidate T4E predicted to interact with the SACSIN molecular chaperone, which is highly expressed in the central nervous system, which regulates HSP70 machinery and interacts with the proteasome (Parfitt et al., 2009; Anderson et al., 2011).

The fact that our analysis of host-interacting proteins revealed putative targets involved in cell signaling, transcriptional regulation, and vesicle trafficking is of particular interest in the context of *Ehrlichia* pathogenesis. Indeed, recent studies on the cellular biology of *E. chaffeensis* infection demonstrated that some *E. chaffeensis* type I effectors interact with similar eukaryotic proteins (Wakeel et al., 2009; Luo et al., 2011). This reinforces the interest of our approach to identify novel type IV effectors and to facilitate their functional characterization, but could also highlight a possible redundancy of action between type I and type IV effectors of *E. chaffeensis* for better infection.

In summary, our results are in accordance with the current knowledge of *Ehrlichia* molecular pathogenesis (Moumène and Meyer, 2016), and the T4Es we predicted using the S4TE algorithm for *E. chaffeensis* are good candidates for further

biological analysis. In addition, the human interactome predicted via HPIDB provides useful information on the possible mode of action of these putative T4Es within the host cell. This study is proof-of-concept that comparative effectomics allows the identification of important host pathways targeted by the bacterial pathogen.

In addition to strain-level variations, allelic diversification in type IV effectors should be further investigated along with variations in regulation or protein expression of these genes. Because type IV effector repertoires are suggested to be major determinants of virulence in *Ehrlichia* (Moumène and Meyer, 2016), it is also important to understand the diversity of type IV effectors present in different species that infect common hosts. Likewise, studying the evolution of type IV effector repertoires among different bacterial species with different host ranges or lifestyles could provide key information to identify the determinants of host specificity.

Based on our results, we hypothesize that the evolution of *E. chaffeensis* intra-species pathogenicity occurs via the acquisition of key regulatory genes. Ultimately, the successive acquisition of type IV effectors could lead to the adaptation of new environmental niches—hosts—resulting in a potential host jump followed by the emergence of new strains in a dynamic environment. However, functional evidence is still lacking for many functions that are hypothetically involved in host specificity.

This study, which focused on type IV effector repertoires in several strains of *E. chaffeensis*, is a step forward in the understanding of *E. chaffeensis* pathobiology. We propose an original approach with rational targets to enable the design of alternative therapies for *ehrlichiae* and other intracellular pathogens.

CONCLUSION

Using S4TE software, we predicted *in silico* the putative type IV effectors from available complete genomes among *E. chaffeensis* species. In particular, we searched for proteins with eukaryotic-like domains, signals for addressing organelles, structural features known to be involved in protein-protein interactions or type IV secretion, and homolog to known T4Es in other bacteria. We identified 47 candidate T4Es in *E. chaffeensis* (45 belonging to the core type IV effectome) with several of the above-cited features. Some presented homologies with known type IV effectors in other bacterial systems and others were annotated as hypothetical proteins with no predicted function. We revealed one strain to be a specific candidate effector in the Liberty strain. The majority of predicted T4Es belonged to plastic regions of the genome. Prediction of protein-protein interactions between *E. chaffeensis* T4Es and human proteome revealed host target proteins that could play a critical role in disease development. Experimental characterization of *E. chaffeensis* candidate T4Es and their targets is now required to confirm these predictions. Yet, our study is the first to show the power of comparative effectomics, even in the case of closely related strains at the intra-species level, in deciphering new cellular pathways potentially involved in host-*Anaplasmataceae* interaction.

AUTHOR CONTRIBUTIONS

CN and DFM conceived the paper, analyzed the results, and wrote the paper.

ACKNOWLEDGMENTS

The authors acknowledge the financial support from European project, FP7-REGPOT-2012-2013-1, grant agreement No. 31598, "EPIGENESIS," One Health approach to integrate Guadeloupe research on vector-borne and emerging diseases in the ERA: From the characterization of emergence mechanisms to innovative approaches for prediction and control (financial support for CN). We are grateful to T. Lefrançois and N. Vachiéry at CIRAD for their confidence and initial input into this project. We thank the reviewers whose insightful comments helped improve our manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fcimb.2016.00204/full#supplementary-material

REFERENCES

Anderson, D. M., Feix, J. B., and Frank, D. W. (2015). Cross kingdom activators of five classes of bacterial effectors. *PLoS Pathog.* 11:e1004944. doi: 10.1371/journal.ppat.1004944

Anderson, J. F., Siller, E., and Barral, J. M. M. (2011). The neurodegenerative-disease-related protein sacsin is a molecular chaperone. *J. Mol. Biol.* 411, 870–880. doi: 10.1016/j.jmb.2011.06.016

Cascales, E., and Christie, P. J. (2003). The versatile bacterial type IV secretion systems. *Nat. Rev. Microbiol.* 1, 137–149. doi: 10.1038/nrmicro753

Cesbron, S., Briand, M., Essakhi, S., Gironde, S., Boureau, T., Manceau, C., et al. (2015). Comparative genomics of pathogenic and nonpathogenic strains of *Xanthomonas arboricola* unveil molecular and evolutionary events linked to pathoadaptation. *Front. Plant Sci.* 6:1126. doi: 10.3389/fpls.2015.01126

Cheng, C., Paddock, C. D., and Reddy Ganta, R. (2003). Molecular heterogeneity of *Ehrlichia chaffeensis* isolates determined by sequence analysis of the 28-kilodalton outer membrane protein genes and other regions of the genome. *Infect. Immun.* 71, 187–195. doi: 10.1128/IAI.71.1.187-195.2003

Cooke, D. E., Cano, L. M., Raffaele, S., Bain, R. A., Cooke, L. R., Etherington, G. J. et al. (2012). Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen. *PLoS Pathog.* 8:e1002940. doi: 10.1371/journal.ppat.1002940

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 5:e11147. doi: 10.1371/journal.pone.0011147

Dawson, J. E., Anderson, B. E., Fishbein, D. B., Sanchez, J. L., Goldsmith, C. S., Wilson, K. H., et al. (1991). Isolation and characterization of an *Ehrlichia*

sp. from a patient diagnosed with human ehrlichiosis. *J. Clin. Microbiol.* 29, 2741–2745.

Dhanasekaran, N., and Premkumar Reddy, E. (1998). Signaling by dual specificity kinases. *Oncogene* 17, 1447–1455. doi: 10.1038/sj.onc.1202251

Dorman, C. J., Colgan, A., and Dorman, M. J. (2016). Bacterial pathogen gene regulation: a DNA-structure-centred view of a protein-dominated domain. *Clin. Sci.* 130, 1165–1177. doi: 10.1042/CS20160024

Dumler, J. S., Sutker, W. L., and Walker, D. H. (1993). Persistent infection with *Ehrlichia chaffeensis*. *Clin. Infect. Dis.* 17, 903–905. doi: 10.1093/clinids/17.5.903

Fouts, D. E., Brinkac, L., Beck, E., Inman, J., and Sutton, G. (2012). PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res.* 40:e172. doi: 10.1093/nar/gks757

Garcia-Garcia, J. C., Rennoll-Bankert, K. E., Pelly, S., Milstone, A. M., and Dumler, J. S. (2009). Silencing of host cell CYBB gene expression by the nuclear effector AnkA of the intracellular pathogen *Anaplasma phagocytophilum*. *Infect. Immun.* 77, 2385–2391. doi: 10.1128/IAI.00023-09

Guyon, K., Balagué, C., Roby, D., and Raffaele, S. (2014). Secretome analysis reveals effector candidates associated with broad host range necrotrophy in the fungal plant pathogen *Sclerotinia sclerotiorum*. *BMC Genomics* 15:336. doi: 10.1186/1471-2164-15-336

Hervet, E., Charpentier, X., Vianney, A., Lazzaroni, J.-C., Gilbert, C., Atlan, D., et al. (2011). Protein kinase LegK2 is a type IV secretion system effector involved in endoplasmic reticulum recruitment and intracellular replication of *Legionella pneumophila*. *Infect. Immun.* 79, 1936–1950. doi: 10.1128/IAI.00805-10

IJdo, J., Carlson, A., and Kennedy, E. (2007). *Anaplasma phagocytophilum* AnkA is tyrosine-phosphorylated at EPIYA motifs and recruits SHP-1 during early infection. *Cell. Microbiol.* 9, 1284–1296. doi: 10.1111/j.1462-5822.2006.00871.x

Kaneko, M., Ishiguro, M., Niinuma, Y., Uesugi, M., and Nomura, Y. (2002). Human HRD1 protects against ER stress-induced apoptosis through ER-associated degradation. *FEBS Lett.* 532, 147–152. doi: 10.1016/S0014-5793(02)03660-8

Kumar, R., and Nanduri, B. (2010). HPIDB–a unified resource for host-pathogen interactions. *BMC Bioinform.* 11(Suppl. 6):S16. doi: 10.1186/1471-2105-11-S6-S16

Lifshitz, Z., Burstein, D., Peeri, M., Zusman, T., Schwartz, K., Shuman, H. A., et al. (2013). Computational modeling and experimental validation of the Legionella and Coxiella virulence-related type-IVB secretion signal. *Proc. Natl. Acad. Sci. U.S.A.* 110, E707–E715. doi: 10.1073/pnas.1215278110

Lin, M., den Dulk-Ras, A., Hooykaas, P., and Rikihisa, Y. (2007). *Anaplasma phagocytophilum* AnkA secreted by type IV secretion system is tyrosine phosphorylated by Abl-1 to facilitate infection. *Cell. Microbiol.* 9, 2644–2657. doi: 10.1111/j.1462-5822.2007.00985.x

Liu, H., Bao, W., Lin, M., Niu, H., and Rikihisa, Y. (2012). *Ehrlichia* type IV secretion effector ECH0825 is translocated to mitochondria and curbs ROS and apoptosis by upregulating host MnSOD. *Cell. Microbiol.* 14, 1037–1050. doi: 10.1111/j.1462-5822.2012.01775.x

Lo, U., Selvaraj, V., Plane, J. M., Chechneva, O. V., Otsu, K., and Deng, W. (2014). p38α (MAPK14) critically regulates the immunological response and the production of specific cytokines and chemokines in astrocytes. *Sci. Rep.* 4:7405. doi: 10.1038/srep07405

Luo, T., Kuriakose, J. A., Zhu, B., Wakeel, A., and McBride, J. W. (2011). *Ehrlichia chaffeensis* TRP120 interacts with a diverse array of eukaryotic proteins involved in transcription, signaling, and cytoskeleton organization. *Infect. Immun.* 79, 4382–4391. doi: 10.1128/IAI.05608-11

Luo, Z.-Q. Q. (2012). Legionella secreted effectors and innate immune responses. *Cell. Microbiol.* 14, 19–27. doi: 10.1111/j.1462-5822.2011.01713.x

Lurie-Weinberger, M. N., Gomez-Valero, L., Merault, N., Glöckner, G., Buchrieser, C., and Gophna, U. (2010). The origins of eukaryotic-like proteins in *Legionella pneumophila*. *Int. J. Med. Microbiol.* 300, 470–481. doi: 10.1016/j.ijmm.2010.04.016

Marcelino, I., Ventosa, M., Pires, E., Müller, M., Lisacek, F., Lefrançois, T., et al. (2015). Comparative proteomic profiling of ehrlichia ruminantium pathogenic strain and its high-passaged attenuated strain reveals virulence and attenuation-associated proteins. *PLoS ONE* 10:e0145328. doi: 10.1371/journal.pone.0145328

McAdam, P. R., Vander Broek, C. W., Lindsay, D. S., Ward, M. J., Hanson, M. F., et al. (2014). Gene flow in environmental *Legionella pneumophila* leads to genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. *Genome Biol.* 15:504. doi: 10.1186/s13059-014-0504-1

Meng, L., Sefah, K., O'Donoghue, M. B., Zhu, G., Shangguan, D., Noorali, A., et al. (2010). Silencing of PTK7 in colon cancer cells: caspase-10-dependent apoptosis via mitochondrial pathway. *PLoS ONE* 5:e14018. doi: 10.1371/journal.pone.0014018

Meyer, D. F., Noroy, C., Moumène, A., Raffaele, S., Albina, E., and Vachiéry, N. (2013). Searching algorithm for type IV secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context. *Nucleic Acids Res.* 41, 9218–9229. doi: 10.1093/nar/gkt718

Michard, C., Sperandio, D., Baïlo, N., Pizarro-Cerdá, J., LeClaire, L., Chadeau-Argaud, E., et al. (2015). The Legionella Kinase LegK2 Targets the ARP2/3 complex to inhibit actin nucleation on phagosomes and allow bacterial evasion of the late endocytic pathway. *MBio* 6, e00354–e00315. doi: 10.1128/mBio.00354-15

Miura, K., and Rikihisa, Y. (2007). Virulence potential of *Ehrlichia chaffeensis* strains of distinct genome sequences. *Infect. Immun.* 75, 3604–3613. doi: 10.1128/IAI.02028-06

Moffatt, J. H., Newton, P., and Newton, H. J. (2015). *Coxiella burnetii*: turning hostility into a home. *Cell. Microbiol.* 17, 621–631. doi: 10.1111/cmi.12432

Moumène, A., and Meyer, D. (2016). Ehrlichia's molecular tricks to manipulate their host cells. *Microbes Infect.* 18, 172–179. doi: 10.1016/j.micinf.2015.11.001

Paddock, C. D., and Childs, J. E. (2003). *Ehrlichia Chaffeensis*: a prototypical emerging pathogen. *Clin. Microbiol. Rev.* 16, 37–64. doi: 10.1128/CMR.16.1.37-64.2003

Paddock, C. D., Sumner, J. W., Shore, G. M., Bartley, D. C., Elie, R. C., McQuade, J. G., et al. (1997) Isolation and characterization of *Ehrlichia chaffeensis* strains from patients with fatal ehrlichiosis. *J. Clin. Microbiol.* 35, 2496–2502.

Parfitt, D. A., Michael, G. J., Vermeulen, E. G., Prodromou, N. V., Webb, T. R., Gallo, J.-M. M., et al. (2009). The ataxia protein sacsin is a functional co-chaperone that protects against polyglutamine-expanded ataxin-1. *Hum. Mol. Genet.* 18, 1556–1565. doi: 10.1093/hmg/ddp067

Pines, A., Kelstrup, C. D., Vrouwe, M. G., Puigvert, J. C., Typas, D., Misovic, B., et al. (2011). Global phosphoproteome profiling reveals unanticipated networks responsive to cisplatin treatment of embryonic stem cells. *Mol. Cell. Biol.* 31, 4964–4977. doi: 10.1128/MCB.05258-11

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/molbev/msp077

Raffaele, S., Win, J., Cano, L. M., and Kamoun, S. (2010). Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of *Phytophthora infestans*. *BMC Genomics* 11:637. doi: 10.1186/1471-2164-11-637

Rikihisa, Y. (2010). *Anaplasma phagocytophilum* and *Ehrlichia chaffeensis*: subversive manipulators of host cells. *Nat. Rev. Microbiol.* 8, 328–339. doi: 10.1038/nrmicro2318

Schwartz, A. R., Potnis, N., Timilsina, S., Wilson, M., Patané, J., Martins, J., et al. (2015). Phylogenomics of *Xanthomonas* field strains infecting pepper and tomato reveals diversity in effector repertoires and identifies determinants of host specificity. *Front. Microbiol.* 6:535. doi: 10.3389/fmicb.2015.00535

Sumner, J. W., Childs, J. E., and Paddock, C. D. (1999). Molecular cloning and characterization of the *Ehrlichia chaffeensis* variable-length PCR target: an antigen-expressing gene that exhibits interstrain variation. *J. Clin. Microbiol.* 37, 1447–1453.

Voth, D. E., Broederdorf, L. J., and Graham, J. G. (2012). Bacterial type IV secretion systems: versatile virulence machines. *Future Microbiol.* 7, 241–257. doi: 10.2217/fmb.11.150

Wakeel, A., Kuriakose, J. A., and McBride, J. W. (2009). An *Ehrlichia chaffeensis* tandem repeat protein interacts with multiple host targets involved in cell signaling, transcriptional regulation, and vesicle trafficking. *Infect. Immun.* 77, 1734–1745. doi: 10.1128/IAI.00027-09

Yu, C.-S. S., Cheng, C.-W. W., Su, W.-C. C., Chang, K.-C. C., Huang, S.-W. W., Hwang, J.-K. K., et al. (2014). CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation. *PLoS ONE* 9:e99368. doi: 10.1371/journal.pone.0099368

Yu, X. J., McBride, J. W., and Walker, D. H. (1999). Genetic diversity of the 28-kilodalton outer membrane protein gene in human isolates of *Ehrlichia chaffeensis*. *J. Clin. Microbiol.* 37, 1137–1143.

Zhang, J. Z., Popov, V. L., Gao, S., Walker, D. H., and Yu, X. J. (2007). The developmental cycle of *Ehrlichia chaffeensis* in vertebrate cells. *Cell. Microbiol.* 9, 610–618. doi: 10.1111/j.1462-5822.2006.00812.x

Zimmermann, P., Tomatis, D., Rosas, M., Grootjans, J., Leenaerts, I., Degeest, G., et al. (2001). Characterization of syntenin, a syndecan-binding PDZ protein, as a component of cell adhesion sites and microfilaments. *Mol. Biol. Cell* 12, 339–350. doi: 10.1091/mbc.12.2.339

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

Partie 3

Etude de la plasticité génomique des effecteurs du système de sécrétion de type IV associée à la famille des *Anaplasmataceae*.

1. Préambule

Dans cette étude, nous avons émis l'hypothèse que les différences de spectre d'hôtes observées entre les espèces du genre *Ehrlichia* pouvaient être liée à l'acquisition de différents répertoires d'effecteurs du SST4. Pour cela nous avons utilisé quatre espèces d'*Ehrlichia* séquencés (*E. chaffeensis, E. canis, E. muris et E. ruminantium*) et quatre autres espèces de la famille des *Anaplasmataceae* comme contrôle. En utilisant le logiciel S4TE 2.0, nous avons identifié 579 ET4 candidats. Parmi ceux-ci, nous avons identifié le super répertoire d'ET4 prédits pour le genre *Ehrlichia* et nous avons montré que 92 effecteurs appartenaient à l'effectome cœur. De plus, nous avons montré que 30% des ET4 étaient spécifiques et présents dans des zones peu denses en gènes avec un pourcentage en base GC différent de la moyenne du génome. Enfin, une étude d'association des domaines protéiques nous a permis de suggérer de nouvelles fonctions pour ces effecteurs de type IV.

2. Manuscrit préliminaire: The super repertoire of type IV effectors in the pangenome of *Ehrlichia* spp. provides insights into host-specificity and pathogenesis.

The super repertoire of type IV effectors in the pangenome of *Ehrlichia* spp. provides insights into host-specificity and pathogenesis

Christophe Noroy[1,2,3] and Damien F. Meyer[1,2]

[1] CIRAD, UMR ASTRE, F- 97170 Petit-Bourg, Guadeloupe, France
[2] ASTRE, CIRAD, INRA, Univ Montpellier, Montpellier, France.
[3] Université des Antilles, Fouillole, BP-250 97157 Pointe-à-Pitre, Guadeloupe, France

ABSTRACT

The identification of bacterial effectors is essential to understand how obligatory intracellular bacteria such as *Ehrlichia* spp. manipulate the host cell for survival and replication. Infection of mammals – including humans – by the intracellular pathogenic bacteria *Ehrlichia* spp. depends largely on the injection of virulence proteins which hijack host cell processes. Although several hypothetical virulence proteins have been identified in *Ehrlichia* spp., only one has been experimentally shown to translocate into host cells via the type IV secretion system so far. Thus, there is a crucial need to identify more type IV effectors (T4Es) to fully understand their role in *Ehrlichia* spp. virulence and host adaptation. Here, we predicted the T4E repertoires of four sequenced *Ehrlichia* spp. and four other *Anaplasmataceae* as controls (pathogenic *Anaplasma* spp. and *Wolbachia* endosymbiont) using previously developed S4TE 2.0 software. This analysis identified 579 pT4Es (228 for *Ehrlichia* spp. only). The effector repertoires of *Ehrlichia* spp. largely overlapped,

thereby defining a conserved core effectome of 92 effectors shared by all strains. In addition, 69 species-specific T4Es were predicted with non-canonical GC% mostly in gene sparse regions of the genomes. We also identified new protein domain combinations, suggesting novel effector functions. This work presenting the predicted effector collection of *Ehrlichia* spp. can serve as a guide for future functional characterisation of effectors and for the design of alternative strategies against these bacteria.

INTRODUCTION

Gram-negative intracellular bacteria *Ehrlichia* spp. are pathogens of eukaryotic cells. They have evolved to survive and replicate in a wide range of mammalian and tick hosts and can also infect humans. Following the first publication of the *Ehrlichia chaffeensis* genome sequence in 2006 (Dunning Hotopp et al., 2006), four other species of *Ehrlichia* pathogenic with a versatile host range have been sequenced (https://gold.jgi.doe.gov/). Indeed, *E. chaffeensis,* which is the agent of human monocytic ehrlichiosis, can also cause disease in several other vertebrates, including dogs and deer, and has a broad host range (Paddock and Childs, 2003) whereas *E. canis*, *E. muris* and *E. ruminantium* have narrow host ranges (Braga et al., 2014; Feng and Walker, 2004; Peter et al., 2002).

For infection, *Ehrlichia* spp. depends on a dedicated protein complex, the type IV secretion system (T4SS), which acts as a molecular syringe to translocate bacterial proteins into the host cells (Moumène and Meyer, 2016). The study of these proteins, referred to as type IV effectors (T4Es), provides valuable insights into the mechanisms by which an intracellular pathogen can manipulate eukaryotic cellular processes to survive and replicate in host cells (Rikihisa, 2017; O'Connor et al., 2012; Martinez et al., 2016). In Gram-negative intracellular bacteria, a large number of effectors harbour eukaryotic-like domains (Ninio and Roy, 2007). These proteins interfere in different steps of the infection by mimicking the functions of eukaryotic proteins (Cazalet et al., 2004; de Felipe et al., 2008). Being able to predict these eukaryotic domains, as well as other protein-protein interaction motif or subcellular signalling sequences, is an important step towards understanding the effectors' mode of action. The identification of T4Es, as well as the elucidation of their function inside the host

cell, will help understand the bacterial pathogenesis. A possible approach for the prediction of T4Es is the analysis of T4E protein sequences using machine learning (Wang et al., 2017). For this purpose, our laboratory developed the S4TE2.0 algorithm, which searches for a large number of motifs related to the function (eukaryotic-like domains, protein-protein interaction, *etc*.) and the subcellular location of predicted T4Es (Meyer et al., 2013; Noroy et al., 2018).

In different pathogenic bacteria, some effectors appear to be acquired by pathogenic bacteria from eukaryotic cells by horizontal gene transfer (de Felipe et al., 2005; Lurie-Weinberger et al., 2010; Ruh et al., 2017). Previous comparative effectomics of closely related *Ehrlichia chaffeensis* strains at the intra-species level showed that pT4Es repertoires are strongly conserved in these genomes. Despite this strong conservation and the strong selective pressure due to their obligate intracellular way of life, at least one strain-specific pT4E (ECHLIB_RS02720) has been identified in *E. chaffeensis* str. Liberty. This effector appears to be involved in the differential inter-strain virulence observed between Arkansas and Liberty strains in SCID mice (Noroy and Meyer, 2016). Moreover, some intense recombination events between E. ruminantium strains have been discovered, which may facilitate bacterial adaptation for survival and adaptation under various environmental conditions in both vector and host species (Cangi et al., 2016). These findings suggest that genomic plasticity plays an important role in the evolution of these bacteria by horizontal gene transfer and recombination events.

The genomes of many animal and plant pathogenic bacteria have been completely sequenced in recent years. Comparative genomics studies demonstrated that repertoires of virulence-associated genes comprise a conserved and variable set of genes among bacterial

species but that these genes may have different evolutionary histories and play distinct roles in pathogenicity (Hajri et al., 2009; Burstein et al., 2016). In comparative genomics, it is important to know all bacterial strains and pathovar to fully understand the identification of the molecular determinants of host specificity (Hajri et al., 2009). To overcome the problem of lack of data on new (as yet unpublished/sequenced) species of *Ehrlichia*, in our study, we crossed results obtained for the genera *Ehrlichia* and *Anaplasma*, which share several hosts.

Even though a large number of T4SS-translocated proteins have been identified in numerous bacteria, only two T4Es have been functionally characterised in the family *Anaplasmataceae*, and their role in invasion and pathogenesis is crucial. AnkA, was identified in *Anaplasma phagocytophilum*, based on sequence homology with repeated ankyrin motifs (Caturegli et al., 2000). Once secreted by T4SS, AnkA is tyrosine-phosphorylated and then directed into the nucleus of the host cell to silence CYBB gene expression (Park et al., 2004; Lin et al., 2007; Garcia-Garcia et al., 2009). The other known *Anaplasmataceae* effector, Ats-1, was identified in *A. phagocytophilum* and has an orthologue in *E. chaffeensis* (Etf-1) (Niu et al., 2010). Ats-1 is injected by T4SS into the cytoplasm of the host cell to recruit host autophagosomes to the bacterial inclusion. Another portion of Ats-1 targets mitochondria, where it has an antiapoptotic activity (Niu et al., 2012; Liu et al., 2012). In contrast, *Legionella pneumophila* relies on a set of approximately 300 T4Es (10% of the genome) for efficient pathogenesis (Gomez-Valero et al., 2014). This molecular arsenal targets many cell signalling and biochemical pathways in the host cell and allows the bacterium to hijack host immunity to ensure its survival and development. We thus hypothesised that many effectors remain to be identified in the *Ehrlichia* genus.

In this study, we predicted the pangenome super-repertoire of T4Es and investigated they are related to genome plasticity and host specificity. We showed that T4E gene repertoires of *Ehrlichia* spp. comprise core and variable gene suites which probably have distinct roles in pathogenicity and different evolutionary histories. By analysing the protein architecture of these effectors, we explored new functions of interest potentially involved in the pathogenesis of this important genus of zoonotic bacteria. Our work thus provides resources for functional and evolutionary studies aiming at understanding the host specificity of *Anaplasmataceae*, functional redundancy between T4Es and the driving forces shaping T4E repertoires.

METHODS

Retrieval of genome sequences and prediction of type IV effectors

The complete genome sequences of the eight *Anaplasmataceae* studied were obtained from the National Center for Biotechnology Information (NCBI) database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Of the eight bacteria, four are *Ehrlichia* species: *E. chaffeensis* str. Arkansas (NC_007799.1), *E. canis* str. St Jake (NC_007354.1), *E. muris* AS145 (NC_023063.1) and *E. ruminantium* str. Gardel (NC_006831.1). Three are *Anaplasma* species: *A. phagocytophilum* str. HZ (NC_007797.1), *A. marginale str. Florida (NC_012026.1) and A. centrale str. Israel (NC_013532.1).* One is an *Wolbachia* species: *W. endosymbiont* of *D. melanogaster* (NC_002978.6). The repertoires of predicted type IV effectors (pT4Es) were determined using the S4TE 2.0 algorithm with default parameters (Noroy et al., 2018). S4TE 2.0 predicts and ranks candidate T4Es by using a combination of 11 independent modules to explore 14 characteristics of type IV effectors. One

module searches for consensus motifs in promoter regions; three modules search for five canonical features of the type IV secretion signal (C-terminal basicity, C-terminal charges, C-terminal hydrophobicity, overall hydrophilicity, and E-blocks); six modules search for several protein domains in known T4Es (eukaryotic–like domains, DUF domains, EPIYA motifs, nuclear localisation signals, mitochondrial localisation signals, prenylation domains, coiled-coil domains); and one module searches for global homology with known T4Es (Noroy et al., 2018).

Phylogenetic reconstruction and plasticity related to putative effectors

An initial evolutionary tree was reconstructed based on alignment of the concatenated core genome of *Anaplasmataceae*. The core genomes of the eight studied bacteria were defined using PanOCT software (Fouts et al., 2012) and resulted in 554 orthologous genes (Fig. 1). To evaluate whether the core predicted type IV effectomes resulted in strong evolutionary events, the tree was reconstructed based on alignment of five concatenated core effectors. Core effectomes were determined using PanOCT software with the following parameters: E-value 10^{-5}, percent identity ≥ 30, and length of match ≥ 65. Only orthologous genes present in the eight bacteria were used for this phylogenetic reconstruction (Fig. S1). All tree reconstructions were done using MAFFT multiple alignments with default parameters (Katoh et al., 2017) and RAxML under the GAMMA BLOSUM62 model with 100 bootstrap resamplings (Stamatakis, 2014). *W. endosymbiont* of *D. melanogaster* was used as an out group to root the tree. The Circos algorithm (Krzywinski et al., 2009) was used to represent effector rearrangements between the four *Ehrlichia* species (Fig. 3). Homologies between effectors were defined using the S4TE-CG algorithm (S4TE 2.0 comparative

genomic tool) (Noroy et al., 2018), and PanOCT software (Fouts et al., 2012), and homologous pT4ES were also plotted on a Venn diagram (Fig. S2).

Comparison of predicted type IV effector repertoires

The similarity of pT4E repertoires was calculated as the mean of (1) the fraction of orthologous pT4Es shared by species A and B, out of all effectors represented in species A, and (2) the fraction of orthologous pT4Es shared by species A and B, out of all effectors represented in species B. Results of side-by-side comparisons were plotted on colour-coded heatmaps and sorted according to the order defined by the phylogenetic tree.

Analysis of *Ehrlichia* spp. genomic architectures and distribution of predicted effectomes

The distribution of pT4Es was analysed in two different ways. First, S4TE 2.0 enables analysis of the predicted effectome distribution according to local gene density (Fig. 5). Second, the distribution predicted effectomes was determined according to their ΔGC content and genome architecture (Fig. 6).

To visualise the distance between each gene and its closest neighbours on the five prime and three prime borders in a single representation, S4TE 2.0 sorted genes into two-dimensional bins defined by the length of their 5’ and 3’ intergenic regions (5’FIR and 3’FIR) (Meyer et al., 2013; Noroy et al., 2018). The colour-coded heat map depicts the gene density distribution. S4TE2.0 used the median length of FIRs to distinguish between gene-dense regions (GDRs) and gene-sparse regions (GSRs) and in-between regions (IBRs). Effectors were then drawn on this heat map according to their flanking intergenic region 5 'and 3'. This method makes it possible to visualise the position of pT4Es according to genome density (Fig. 5,

Fig S3). An additional Circos graph was used to visualise the distribution of effectors along the genome according to local gene density.

To visualise the local GC content according to the genome architecture in a single representation, we calculated the G+C content of each gene and then subtracted the mean of G+C content of all the genes in the genome, giving the ΔGC content of each gene. The genes were sorted into two-dimensional bins defined by the length of their 5' and 3' FIRs. The mean ΔGC content of genes in the same bin was calculated and is represented by a colour-coded heat map. GDRs, GSRs and IBRs were defined as described above. Similarly, pT4ES were plotted on the heatmap according to their 5'FIR and 3'FIR. This method makes it possible to visualise the position of pT4Es according to genome architecture and in relation with their GC content (Fig. 6, Fig S4). An additional density graph was used to quantify the ΔGC content of effectors compared to that of other genes. This density graph was constructed using R graphics. The red line represents the density of effectors according to ΔGC content and the black line represents density of all other genes.

Identification of protein domains in *Ehrlichia* spp.

Protein domains of *Ehrlichia* pT4Es were identified using S4TE2.0 and the Pfam database. S4TE2.0 proposes six different modules to find several domains (eukaryotic-like domains, the DUF domain, EPIYA motifs, the nuclear localisation signal (NLS), the mitochondrial localisation signal (MLS), the prenylation domain, coiled-coil domains). In this study, only EPIYA, NLS, MLS, prenylation and coiled-coil domains were used. We also used PfamScan to search for protein domains in the Pfam database.

We analysed the protein architecture of pT4Es using hive plots, which make it possible to obtain a linear layout of the network of the

various domain combinations among these proteins (Krzywinski et al., 2012). The network of protein architectures connected by shared domains was families with Jhive plot software (Krzywinski et al., 2012)

. Domains were plotted on following rules: On the a1 axis, the axes (a1, a2, a3) according to the number of links between

different domains is strictly less than 15; on axis a2, the number of links is between 15 and 40; and on axis a3, the number of links is strictly more than 40 (Fig. 7). To get clearer view of less frequently represented domains, a second hive plot was created by omitting the four most frequently represented domains (NLS, Coiled-coils, EPIYA, Eblock) with the following parameters. On axis a1, the number of links between domains is strictly less than 3; on the a2 axis, the number of links is between 3 and 5; and on the a3 axis, the number of links is strictly more than 5 (Fig. 8). On both hive plots, the position of the domains on the axes (nodes) is sorted according to the increasing number of links from the centre to the outside.

The analysis of these two graphs prompted us to choose two families of putative effectors for a detailed study of their architectural diversity. The ankyrin-containing putative effector family was chosen to represent one of the two known effectors to be secreted in a type IV dependent manner in *Anaplasmataceae*, i.e. AnkA, ECH_0684 (Fig. 9A). The HATPase_c-containing putative effector family was chosen to represent an effector family with a low number of links and containing a species-specific pT4E (Fig.10). In order to determine the evolution of Ank-containing predicted T4Es, a time tree (Fig. 9B) was constructed using the Reltime-ML method (Tamura et al., 2012) and the Tamura-Nei model. The analysis involved 11 nucleotide sequences. All positions containing gaps and missing data were eliminated. A total of 1 836 positions comprised the final dataset. Evolutionary analyses were conducted in MEGA7

(Kumar et al., 2016). To check the evolution of homologous gene sequences to AnkA, dot plots were built using dotmatcher software in the EMBOSS package (Rice et al., 2000) with a window size of 50 and a threshold of 50. All positions from the first input sequence were compared with all positions from the second input sequence using a specified substitution matrix. Only the dot-plot comparing ECH-0684 (x-axis) and ERGA_CDS_03830 (y-axis) sequences is presented here.

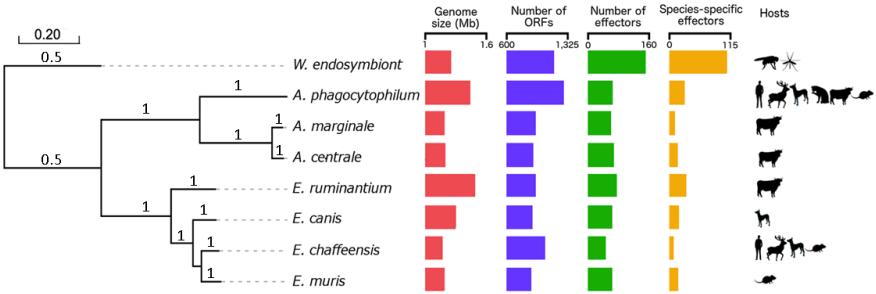


Figure 1. Phylogenetic tree of *Ehrlichia* and *Anaplasma* genus shows three different clades. A maximum likelihood tree of four *Ehrlichia* species (*E. chaffeensis* str. Arkansas, *E. canis* str. Jake, *E. muris* AS145, *E. ruminantium* str. Gardel), three *Anaplasma* species (*A. phagocytophilum* str. HZ, *A. marginale* str. Florida, *A. centrale* str. Israel) and *W. endosymbiont* of *D. melanogaster* (out group) was reconstructed on the basis of concatenated nucleic acid alignment of proteins shared by all species (core genomes) with 100 bootstrap resamplings. The following are represented for each bacterium: genome size (red), number of ORFs (blue), number of predicted T4 effectors (green), number of unique predicted effectors (yellow) and known major hosts (black symbols).

RESULTS

The host spectrum matches the number of ORFs in *Anaplasmataceae*. The *Anaplasmataceae* phylogenetic tree suggests a clear divergence between two distinct clades (Fig. 1): a clade containing four *Ehrlichia* species including the most widely studied species *E. chaffeensis* and a clade containing three *Anaplasma* species including *A. phagocytophilum*, *A. marginale* and *A. centrale*. *Wolbachia endosymbiont* of *Drosophila melanogaster* was used as out group. The length of the genomes ranged from 1.50 Mbp in *A. phagocytophilum* to 1.17 Mbp in *E. chaffeensis*, *E. muris* and *A. marginale* (Fig. 1). The GC content differed considerably between

Ehrlichia species and *Anaplasma* species. The GC content of *Ehrlichia* species ranged from 27.5% in *E. ruminantium* to 30.1% in *E. canis*. The GC content of *Anaplasma* species was also highly variable, ranging from 41.7% for *A. phagocytophilum* to 50% for *A. central (data not shown)*. The number of predicted Type IV Effectors (pT4Es) was almost constant (8% of the genome) whatever the genome considered and ranged from 44 for *E. chaffeensis* to 70 for *E. ruminantium* (Fig. 1). S4TE 2.0 identified a total set of 579 pT4Es in the *Ehrlichia* and *Anaplasma* genera. Despite this significant number, only five pT4Es are shared by all species (core effectome). The phylogenetic tree of core effectome is in agreement with the core genome phylogenetic tree (Fig. S1) and with the 16S tree (not shown). The pathogenic bacteria belonging to the *Anaplasmataceae* family have a wide range of hosts. We showed that this range of hosts is in accordance with the number of ORFs in the genomes (Fig. 1). The more ORFs present is a given species, the broader the host range, as for example, *E. chaffeensis*, which contains the most ORFs, has the broadest host range. Inversely, *E. ruminantium* with a large genome but fewer ORFs, has a limited host range. However, we point out that the number of species-specific effectors appears to be linked to the size of the genome (Fig. 1).

Analysis of groups of putative effectors and their corresponding *Anaplasmataceae* hosts suggests several host-specific effectors. To identify a possible link between pT4E repertoires and host specificity, we performed pairwise comparison of effector gene repertoires by calculating the fraction of shared effectors, and then clustered species based on the similarities in their effector pools (Fig. 2). The resulting clusters strongly agree with the phylogenetic clades. Although pT4Es repertoires are versatile, they show a certain level of conservation within a genus and between phylogenetically close species.

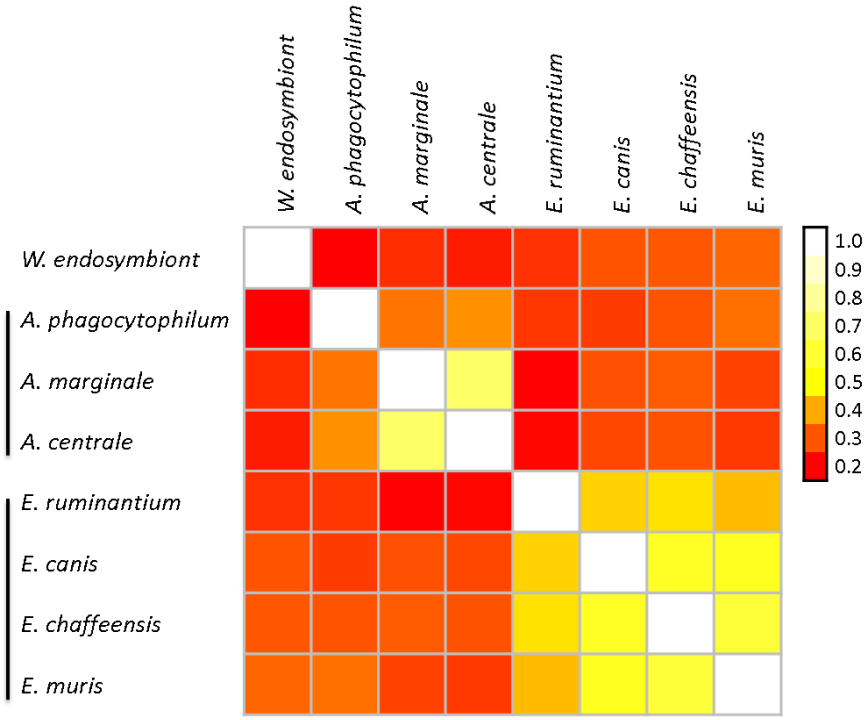


Figure 2. Comparison of the pools of predicted Type IV effectors among *Anaplasmataceae* species revealed strong conservation in *Ehrlichia* spp. The colour gradient represents the similarity between sets of effectors (pale colours mean high similarity). The different species are ordered according to the phylogenetic tree (Fig. 1). Clusters defined on the basis of similar effectors repertoires are marked on the left by black lines.

This is especially true for *Ehrlichia* genus as highlighted by the pale yellow cells in the matrix diagram (Fig. 2). In addition, it is interesting to note that *E. muris* has the same similarity value (0.32) as *A. marginale* with *A. phagocytophilum*. The similarity between *E. muris* and *A. phagocytophilum* could be related to their common rodent host. On the other hand, although *E. ruminantium* and *A. marginale* and *A. centrale* have the same hosts (ruminants), there is less similarity between their corresponding repertoires (0.18).

Within *Ehrlichia* species, *E. chaffeensis* has the broadest host range. Among these preferential hosts, *E. chaffeensis* infects canids and rodents, which are also the specific hosts of *E. canis* and *E. muris* respectively (Fig. 1). Indeed, some preferential hosts of *Ehrlichia* species are shared by *Anaplasma* species. Ruminants are infected by all *Anaplasma* species and by *E. ruminantium*, rodents are infected by *A. phagocytophilum*, *E. muris* and *E. chaffeensis* and canids are infected by *A. phagocytophilum*, *E. canis* and *E. chaffeensis* (Fig. 1). The different preferential hosts are plotted on the a2 axes of a hive plot (Fig. 4). The number of species-specific pT4Es of *Ehrlichia* and *Anaplasma* is represented by the size of each node on the a1 and a3 axes, respectively. Links between the a1 or a3 and a2 axes represent effectors possibly involved in host specificity. Links between the a1 and a3 axes represent the homology between *Ehrlichia* species-specific effectors and *Anaplasma* species-specific effectors. For example, *E. ruminantium* (green) shows 33 species-specific pT4Es, which could be involved in host-specificity (ruminants). Among these effectors, some share homology with *Anaplasma* species-specific pT4Es.

The *Ehrlichia* pan-genome effector super-repertoire includes core and variable effectors with 30% of species-specific effectors. Figure 3 shows the position of the pT4Es on the genomes of the four *Ehrlichia* species and the homologies between the effectors clearly indicate species-specific effectors (i.e. no orthologues) and effectors belonging to the core effectome (i.e. present in all the strains) (Fig. 3). Interestingly, the colour code highlights an inversion of part of the genomes compared to *E. chaffeensis* as reference (with the biggest number of ORFs). The super repertoire of pT4Es in the pangenome of *Ehrlichia* spp. comprises 52 groups of orthologues, hereafter referred to as effector orthologue groups (EOGs). In other words, 70% of *Ehrlichia* predicted effectors belong to EOGs. Moreover, we found that most predicted effectors were shared by a small subset of species (29 EOGs) and 23 effectors were "core effectors", i.e. had orthologues in every *Ehrlichia* species (Fig. 3, Fig. S2). Although the evolutionary tree of *Ehrlichia* core effectors (Fig. S1A) is somewhat congruent with the *Ehrlichia* phylogenetic tree (Fig. 1), it is interesting to note some discrepancies in the percentage identity between the 23 EOGs of the core effectome (Fig. S1B). Indeed, 14 EOGs showed high sequence identity between *E. canis*, *E. muris* and *E. ruminantium*. Four EOGs showed high pairwise identity between *E. canis* and *E. muris* and *E. chaffeensis* and *E. ruminantium*. A single EOG showed strong identity between *E. canis* and *E. muris*. Surprisingly, only four EOGs showed strong identity between all pT4Es, particularly between *E. chaffeensis* and *E. muris*. For example, the EOG corresponding to AnkA (ECH_0684, black dot) showed high sequence identity between *E. canis*, *E. muris* and *E. ruminantium* rather than between *E. chaffeensis* and *E. muris* and contrary to what could be inferred from the phylogenetic tree of core effectors (Fig. S1A and B).

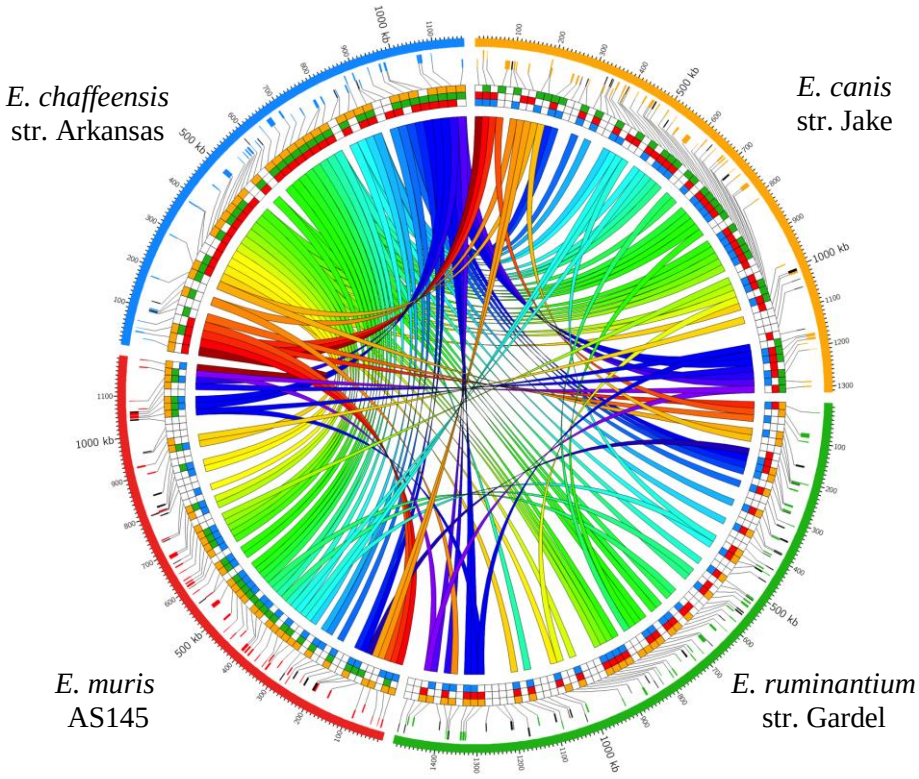


Figure 3. Mapping of *Ehrlichia spp.* predicted Type IV effectors (pT4Es) and their homologies highlights the genomic plasticity of this genus. Genomes of *E. chaffeensis* str. Arkansas (blue), *E. canis* str. Jake (orange), *E. ruminantium* str. Gardel (green) and *E. muris* AS145 (red) are represented in the outer circle of this Circos graph. The second and third circles represent the genes encoding the pT4Es (sense and antisense genes, respectively). The genes are colour coded depending on the genome in which they originated and species-specific genes are in black. Links show homologies between pT4Es of the four genomes. The homologies between the pT4Es of the four genomes are also represented by squares of the corresponding genome colour.

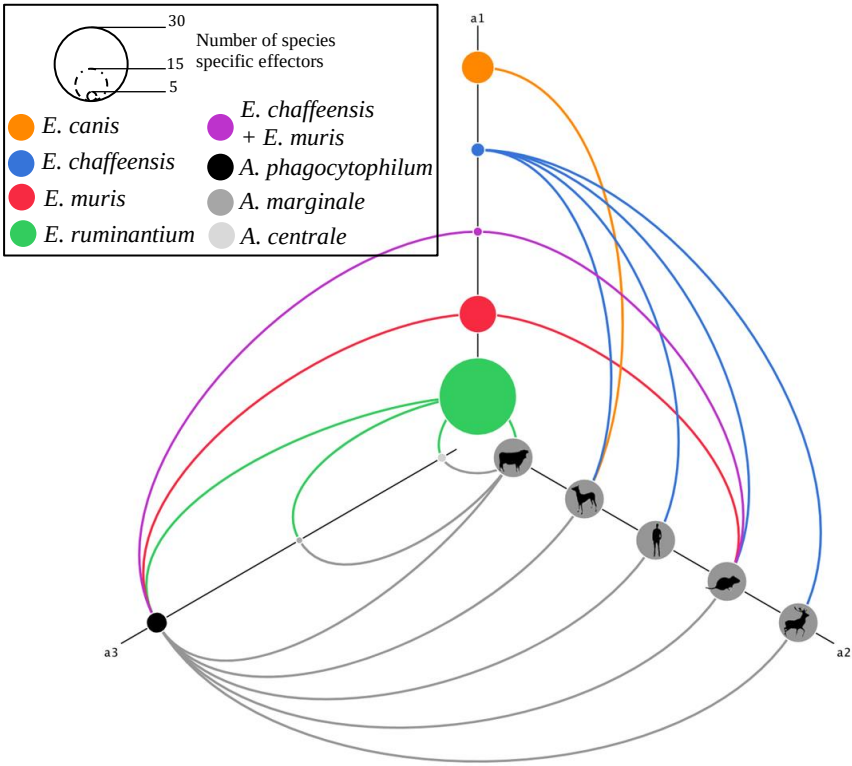


Figure 4. Network analysis of *Anaplasmataceae* species-specific pT4Es and host range suggests the existence of host-specific pT4Es. This network of species-specific pT4Es was drawn using the hive plot algorithm, which is a rational visualization method for drawing networks based on their structural properties. Nodes are mapped to and positioned on radially distributed linear axes and edges are drawn as curved links. *Ehrlichia* species-specific pT4Es are represented by nodes on the a1 axis of the hive plot. The size of each node is linked to the number of species-specific pT4Es for a given *Ehrlichia* species which is colour coded as follow: *E. chaffeensis* str. Arkansas (blue), *E. canis* str. Jake (orange), *E. ruminantium* str. Gardel (green) and *E. muris* AS145 (red). Purple node is the subset of pT4Es specific to *E. chaffeensis* str. Arkansas and *E. muris* AS145. *Anaplasma* species-specific pT4E are represented by nodes on the a3 axis whose size and nuance of grey depend on the *Anaplasma* species. The different hosts of these 7 *Anaplasmataceae* bacteria are represented by grey nodes on a2 axis. Curved links a1-a2 and a3-a2 show the putative host specificity of each bacterium. Links between a1-a3 represent host-specific homologies. The colour of each link is related to the node from which it emerges.

The analysis of effectors revealed that 30% of them (69) were only observed in one of the *Ehrlichia* species analysed. The species with the highest number of unique pT4Es is *E. ruminantium*, with 33 species-specific effectors. Notably, each *Ehrlichia* genome contains at least six species-specific effectors (Fig. 3, Fig. S2). In addition, species-specific effectors appear to be randomly located in the genome relative to other effectors. This is notable in the genome of *E. ruminantium*, which has the most species-specific effectors (Fig. 3, Fig. S2

Predicted type IV effectors of *Ehrlichia* species are overrepresented in gene sparse regions and in high GC content regions of the genome. In order to understand how genomic plasticity influences the distribution of pT4Es, we first analysed the genome architecture of *Ehrlichia* species by looking at local gene density (Fig. 5, Fig. S3). The gene architecture of *E. canis* shows 29.2% of genes in gene dense regions (GDRs) and in gene sparse regions (GSRs). So, 41.6% of genes are in 'in between' regions (IBRs) (Fig. 2). This genome architecture is also representative of other *Ehrlichia* species (Fig. S3). Although 29.2% of *E. canis* genes belong to GSRs, 42% of pT4Es and 42.9% of species-specific pT4Es are in GSRs (Fig. 2). Thus, compared to the whole genome, the GSR showed a 1.44-fold enrichment in candidate type IV effector genes. The proportion of candidate T4Es in IBRs is not significantly different, with 42.8% of *E. canis* genes belonging to IBRs and 43.5% of pT4Es - 42.9% of species-specific pT4Es - in IBRs (Fig. 2). Consequently, the proportion of candidate T4Es in GDRs is lower than the proportion of genes of the whole genome. These results suggest that plastic regions with low gene density harbour pathogenicity genes and could play a role in host-bacteria interactions.

Contemporary methods used to infer horizontal gene transfer events are based on analyses of genomic sequence data. One

interesting method consists in searching for a section of a genome that significantly differs from the genomic average, such as GC content (Lawrence, 2002). To understand how genomic plasticity and horizontal gene transfers could influence the distribution of predicted T4Es, we analysed the genome architecture of *Ehrlichia* species according to the local GC content (Fig. 6, Fig. S4). These figures show that there is an enrichment of pT4Es in regions with high GC content. Indeed, 67.7%, 69.4% and 73.3% of the pT4Es of *E. canis*, *E. muris* and *E. chaffeensis*, respectively, were found to have a higher ΔGC than the average of the other genes (50% of genes less than or equal to the average ΔGC). It is noteworthy that species-specific pT4Es are also over-represented in high GC content regions with 53.3% and 75% for *E. canis* and *E. muris*, respectively. For *E. ruminantium*, the difference in GC content between effectors and other genes is less pronounced because this genome has a mosaic structure with zones of high density GC content (in red in Fig. S4 A). In addition, unlike the curve of the other genes in the genome (normal curve), several density peaks are highlighted by the ΔGC-density curve of putative effectors (Fig. 6 Fig. S4). These results suggest that pT4Es (and species-specific pT4Es) are mostly located in high GC content regions and could have been linked to the acquisition of new genes during the evolution of the host-pathogen relationship.

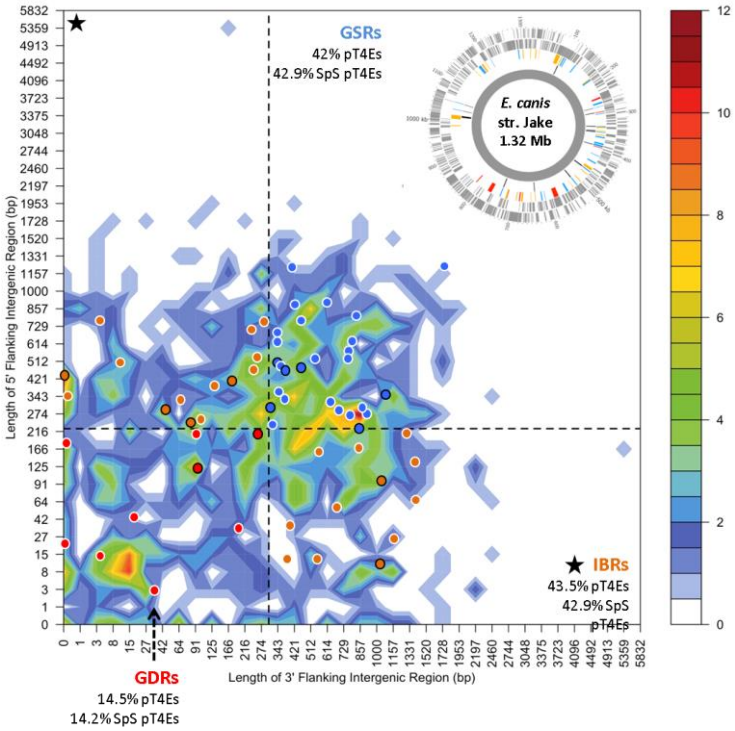


Figure 5. The distribution of *Ehrlichia* type IV effectomes according to local gene density shows an enrichment of pT4Es in gene sparse regions. Distribution of *E. canis str*. Jake genes according to the length of their flanking intergenic regions (FIRs). All *E. canis* genes were sorted in two-dimensional bins according to the length of their 5′ (*y*-axis) and 3′ (*x*-axis) FIRs. The number of genes in the bins is represented by a colour-coded density graph. Genes whose FIRs were both longer than the median length of FIRs were considered as gene-sparse region (GSR) genes. Genes whose FIRs were both below the median value were considered as gene-dense region (GDR) genes. In between region (IBR) genes are genes with a long 5′ FIR and short 3′ FIR, and inversely. For *E. canis*, this median value is 225 bp for 5′ FIRs and 304 bp for 3′ FIRs. The dashed line showing the median length of FIR delimits the genes in GSR, GDR and IBR. Candidate effectors predicted using the S4TE 2.0 algorithm were *s* plotted on this distribution according to their own 3′ and 5′ FIRs. A colour was assigned to each of the three following groups: red to GDRs, orange to IBRs, and blue to GSRs. Specific pT4Es are represented by a dot circled in black. In the top right corner, a Circos graph shows the distribution of *E. canis* str. Jake putative effectors along the genome. The outermost and second circles (in grey) represent *E. canis* antisense and sense genes, respectively. The third and innermost circles represent pT4Es. The black, red, orange and blue colour of each putative T4 effector corresponds to species-specific effectors located in GDRs, IBRs and GSRs, respectively.

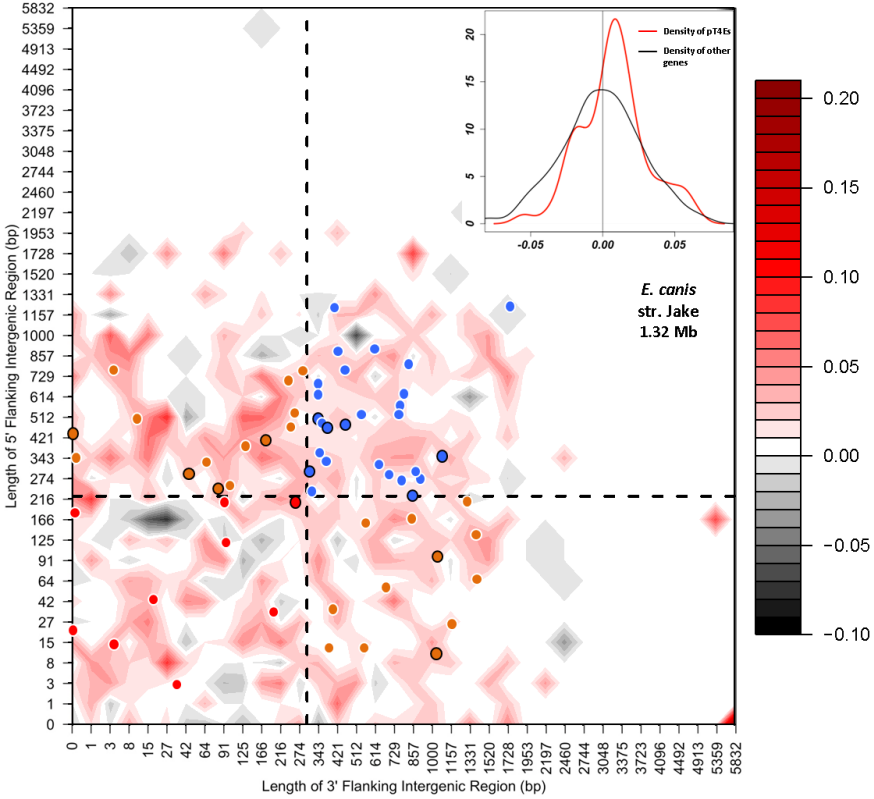


Figure 6. The distribution of *Ehrlichia* type IV effectomes according to local gene density and ΔGC content shows an enrichment of pT4Es in high GC content regions. Distribution of *E. canis* str. Jake genes according to the length of their flanking intergenic regions (FIRs). All *E. canis* genes were sorted in two-dimensional bins according to the length of their 5′ (*y*-axis) and 3′ (*x*-axis) FIR lengths. For each gene, the ΔGC content was calculated by subtracting the GC content of a gene by the average of GC content of all the genes. The mean of ΔGC of genes in the bins is represented by a colour-coded density graph. GSR, GDR and IBR were defined as described for the analysis of local gene density. A colour was assigned to each of the three following groups: red to GDRs, orange to IBRs, and blue to GSRs. Specific pT4Es are represented with a dot circled in black. In the top right corner, a density graph indicates the density of pT4Es according to ΔGC content (red line) and the density of other genes (black line).

Domain shuffling seems to play a major role in effector evolution. Previous studies of T4Es revealed that they harbour numerous eukaryotic domains as well as effector-specific domains (Burstein et al., 2016; Gomez-Valero et al., 2011). The high number of effectors predicted by S4TE 2.0 made it possible to identify and analyse conserved effector domains across the *Ehrlichia* genus. We identified the domains present in *Ehrlichia* pT4Es using a similarity search of Pfam databases, and outputs of S4TE 2.0 software. Conserved domains were detected in 97% of the pT4Es. A total of 116 distinct domains were identified by the two methods combined. When analysing the protein architectures (different domain combinations), we noticed that the same domains were often shared among different architectures. We visualised this phenomenon as a network of protein architectures connected by shared domains (Fig. 7). The network of protein architecture of pT4Es was drawn using the hive plot algorithm. The network clearly demonstrates that several domains are present in numerous effectors (indicated by the occurrence and the width of the links), as well as in numerous different architectures (indicated by the number of links). The most common domains in *Ehrlichia* T4Es are domains for protein subcellular location like NLS or MLS, domains using protein-protein interactions like coiled-coils or Ank domains, and domains for post-translational modification like EPIYA or HATPase_c. NLS, which are known to address proteins to the nucleus of eukaryotes, were found in 67% (162) of the pT4Es. The coiled-coils domains, known to mediate protein-protein interactions, were found in 36% (87) of pT4Es. The EPIYA domains, known to be phosphorylation sites, were found in 31% (76) of pT4Es. Finally, the Eblock domain, known as a secretion signal in *Legionella pneumophila*, was found in 29% (70) of pT4Es. In addition to these four most abundant and best-known domains, the network represents a wealth of domains with

both known and unknown functions. To facilitate the reading of the least represented domains, a second network was built by removing the four most frequently represented domains described above (NLS, coiled-coils, EPIYA and Eblock) (Fig. 8). After removing these domains, interestingly, some domains like Ank repeats disappeared from the network. In other words, Ank repeat domains were always associated with at least one of the four most widely represented domains.

In *Ehrlichia* species, Ankyrin repeats (Ank) appear to be associated with two different domains (NLS and EPIYA). Overall, Anks were found in 11 pT4Es in the four species (Fig. 9A). Although there is strong homology between different Ank-containing T4Es, differences in the position of the different protein domains were observed. For example, the first of Ank-containing pT4Es, Ecaj_0365, EMUR_01925 and ERGA_CDS_03830 are homologous with the known effector AnkA ECH_0684 of *E. chaffeensis*, but the position of the Ank, EPIYA and NLS domains differs, some domains are even absent in certain proteins (Fig. 9A). Further, the analysis of ECH_0684 and ERGA_CDS_03830 revealed seven duplications of about 200 amino acids of ECH_0684 (from 200 to 400) in ERGA_CDS_03830 (Fig. 9C). Despite these differences, we were able to highlight three groups of proteins with the time tree (Fig. 9B). The second group on Ank-containing pT4Es comprises Ecaj_0221, ECH_0877 and ERGA_CDS_02160 that show conserved architectures, while the last group of Ank-containing pT4Es show versatile architectures (Fig. 9A). The evolutionary tree of the Ank-containing pT4Es is congruent with the phylogenetic tree of *Anaplasmataceae* species (Fig. 1) for the first (AnkA orthologues) and second groups. Moreover, these two groups appeared to derive from a common ancestral gene and their difference may be linked to different coevolution in the bacterium and its host.

Another family of special interest is Ehrlichia pT4Es whose architecture harbours a HATPase_c domain. The HATPase_c was found in only six pT4Es but is present in three different architectures. Two domain architectures are conserved among the different species of Ehrlichia, while the third domain architecture is only represented by E. ruminantium ERGA_CDS_03390, a species-specific pT4E (Fig. 10). Some domains, including HisKA, HSP90, Toprim, DNA_J, Reponse_ reg, were found adjacent to the HATPase_c domain. HATPase_c is present in several ATP-binding proteins including histidine kinase, DNA gyrase B or the heat shock protein HSP 90. The variety of protein domains carried by these pT4Es indicates that this family may have a different mode of action. It is interesting to note that effectors with domains related to DNA binding (DNA_j, Toprim) also have one or more nuclear localisation sequences (NLS) (Fig. 10A).

When we analysed the genomic environment of the 'orphan' pT4E ERGA_CDS_03390 in more depth, we found that this gene is part of a cluster of six genes, three of which are sense and three (including ERGA_CDS_03390) are antisense (Fig. 10B). For 5' to 3' the cluster is composed as follows (i) a sense gene (ERGA_CDS_03370) encoding an enzyme involved in the methylerythritol phosphate pathway (isopentenyl-pyrophosphate biosynthetic metabolic pathway), (ii) an antisense gene (ERGA_CDS_03380) encoding CutA, an ion transporter, (iii)

ERGA_CDS_03390 encoding a protein homologous to VirA, which is a protein sensor secreted by Agrobacterium tumefaciens, (iv) the last antisense gene (ERGA_CDS_03400) coding for a putative O-methyltransferase, (v) a sense gene encoding a hypothetical protein, and (vi) a sense gene encoding a NAD(P)H-hydratase dehydratase, which acts on both hydrated NADH and hydrated NADPH (Fig. 10B). Although the average GC content of

this part of genome is 25.75%, the antisense cluster showed several peaks of GC content (mean GC content of 26.4% for the three antisense genes).

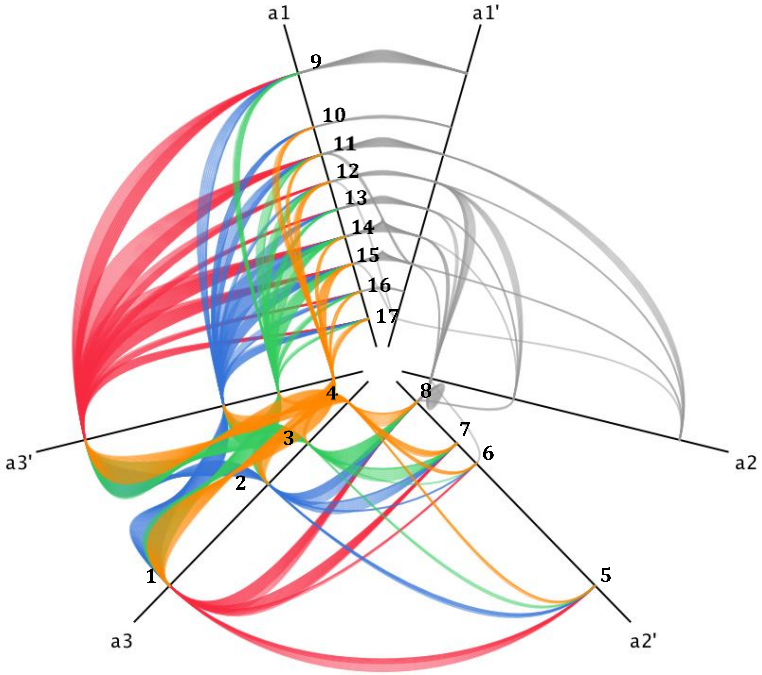


Figure 7. Protein architecture network of *Ehrlichia* pT4Es shows a large number of interactions between protein domains. This network of protein architecture of pT4Es is drawn using the hive plot algorithm, which is a rational visualization method for drawing networks based on their structural properties. Nodes are mapped to and positioned on radially distributed linear axes, and edges are drawn as curved links. Each node represents a specific domain or a list of specific domains (see table S1) found in *Ehrlichia* T4Es predicted by S4TE 2.0. Links between domains represent the association of these domains in the architecture of *Ehrlichia* pT4Es. Links between NLS, Coiled-coils, EPIYA and Eblock domains (the most abundant protein domains in *Ehrlichia*) and other nodes are red, blue, green and orange, respectively. Other links are pale grey. The table of domains identifies the protein domains for each node (numbered) and their occurrences in *Ehrlichia* spp predicted T4 effectomes. All the domains are ranked on the three axes (a1, a2, a3) according to the number of their links. Let X be the number of links between one domain and the others, $X < 15$ was represented on a1, $15 \geq X \leq 40$ was represented on a2, and $X > 40$ was represented on a3, thus defining the most abundant domains.

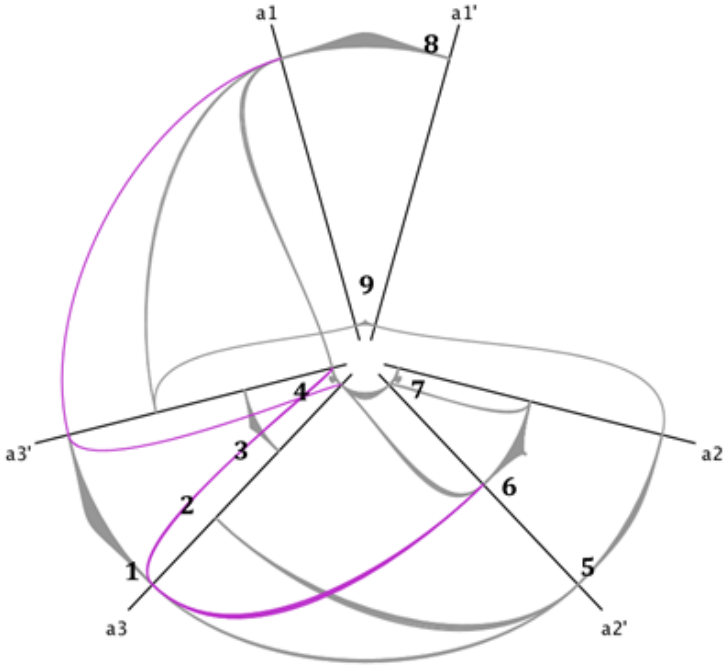


Figure 8. Protein architecture network of putative effectors for rarely occurring domains. This network of pT4E protein architectures was drawn using the hive plot algorithm to produce a rational visualization method based only on the network structural properties. Each node represents a specific domain or a list of specific domains (see table S2) found in *Ehrlichia* T4Es predicted by S4TE 2.0. Links between domains represents the association between these domains in the architecture of *Ehrlichia* pT4Es. Nodes representing the most abundant domains presented in fig.7 (NLS, Coiled-coils, EPIYA and Eblock domains) and their corresponding links to other nodes are not included in this graph to highlight the less abundant protein domains in *Ehrlichia* spp predicted T4 effectomes. The table of domains identifies the protein domains for each node (numbered) and their occurrences. All the domains are ranked on the three axes (a1, a2, a3) according to the number of their links. Let X be the number of links between one domain and the others, $X < 3$ was represented on axis a1, $3 \geq X \leq 5$ was represented on axis a2, and $X > 5$ was represented on axis a3.

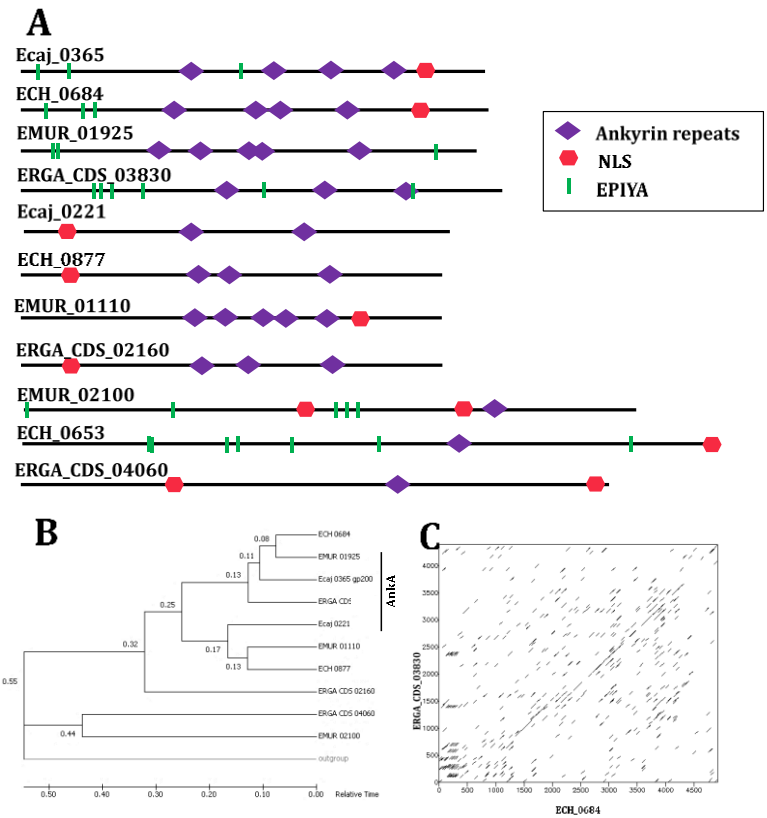


Figure 9. Ankyrin-containing predicted type IV effectors show diverse architectures and inter- and intragenic rearrangements in *Ehrlichia* spp.

A. Each protein in *Ehrlichia* spp. pT4E whose architecture includes an ankyrin domain is represented. **B**. Relative time phylogenetic tree build from 11 nucleotide sequences of Ankyrin-containing predicted type IV effectors (pT4Es). ECH_0653 was used as outgroup. The numbers in front of each node represent the relative time from the putative common ancestor of two branches. Evolutionary analyses were conducted in MEGA7 (Tamura et al., 2012). **C**. Dot plot of regions of similarities between ECH-0684 (x-axis) and ERGA_CDS_03830 (y-axis). This graph was constructed with the dotmatcher software included in the EMBOSS package, where all positions from the first input sequence are compared with all positions from the second input sequence using a specified substitution matrix and using a window size of 50 and a threshold of 50.

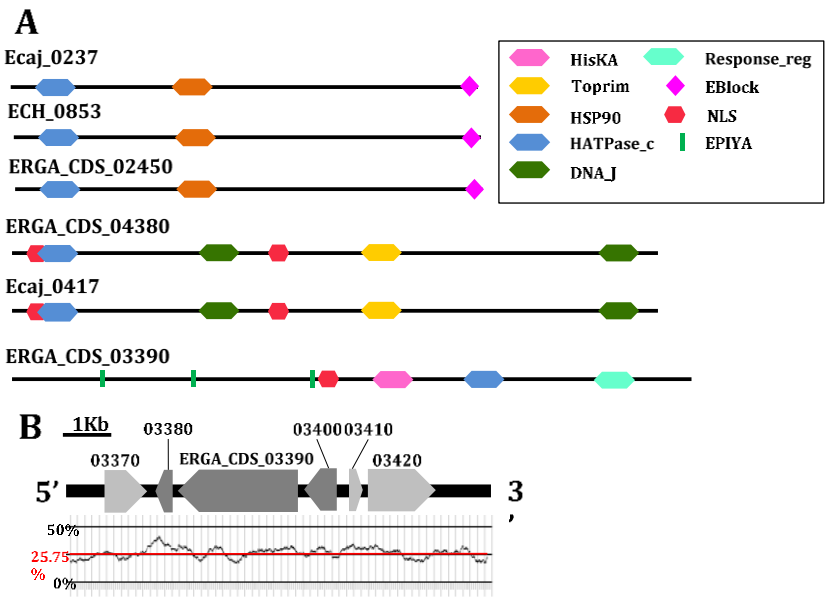


Figure 10. The strong domain diversity of HATPase_c-containing putative type IV effectors defines three conserved families of effectors in *Ehrlichia* spp.

A. Each protein in *Ehrlichia* spp. type IV effectome whose architecture includes an HATPase_c domain is represented. **B**. Representation of the genomic context surrounding pT4E ERGA_CDS_03390 between position 557733 and 563669 of *E. ruminantium* str. Gardel genome. Pale grey arrows represent sense genes and dark grey arrows represent anti-sense genes. The GC content of this region of the genome was calculated using 200 bp windows and is represented by the black curve. The average GC content of this gene cluster (25.75 % of GC) is indicated by the horizontal red line.

DISCUSSION

The obligatory intracellular pathogens of the *Ehrlichia* genus cause potentially fatal emerging infectious diseases. Although they are strictly restricted to their mammalian host or tick vector cells, *Ehrlichia* spp. exhibit marked differences in host range (Moumene and Meyer 2016). Moreover, despite the reductive evolution of their genome, they have evolved a versatile type IV secretion system which translocates effector proteins (T4Es) into the cytoplasm of host cells, hijacks innate immunity, and manipulates numerous cellular pathways to its own advantage (Rikihisa, 2017). Thus, it is tempting to speculate that T4E repertoires shape the *Ehrlichia* spp. host range. In order to tentatively identify candidate genes involved in host specificity, we predicted *Ehrlichia* spp. T4Es and analysed these T4E gene repertoires.

Here, we report that the number of hosts appears to be related to the number of ORFs in the genome. Bacteria can acquire and maintain a diverse repertoire of accessory (variable) genes as a key feature to better adapt to changing environments and to colonise a wider range of ecological niches. This is a well-known phenomenon among opportunistic or ubiquitous pathogens which have bigger genomes than bacteria living in specific more constrained specific

environments (Bobay and Ochman, 2017). However, for the broad host spectrum *E. chaffeensis*, we only found six variable pT4Es that are species-specific effectors. The host range thus does not appear to be only related to effector repertoires but to a wider set of genes, and may also be governed by other pathogenicity factors such as host cell adhesion or other type of effectors which ensure bacterial survival (Luo et al., 2011; Farris et al., 2017). Similar observations have been made when *Pseudomonas syringae* pathovars T3E repertoires were compared (Baltrus et al., 2012), thereby reinforcing the hypothesis that a complex genetic basis underlies host range

evolution in bacterial pathogens. Some of the variable effectors of the *Ehrlichia* pangenome effector super-repertoire, species-specific effectors, could harbour the strongest virulence phenotypes. Although we identified species-specific genes but not strong host-specificity candidates, we highlighted a strong similarity between the effector repertoires of *E. muris* and *A. phagocytophilum*, two bacteria which share the same rodent host. Species-specific effectors common to both strains necessarily carry this similarity, and could therefore be specific to this host. Similarly, we found 12 *Ehrlichia*-specific effectors in *Anaplasma* which share the same hosts, thus supporting the hypothesis of a model for coevolving molecular dialogues between effector repertoires and host immunity. Regarding the question of effector adaptation to different hosts, most *Ehrlichia* effectors may function as 'generalists' in a broad range of hosts, and immune response to effectors may be the primary driver of effector repertoire diversification (Tago and Meyer, 2016).

On the other hand, the function of the core effectors remains largely unknown but their high conservation within the evolution of the *Ehrlichia* family shows that the core effectome may fulfil a critical function during infection. We further reveal that T4E gene repertoires of these pathogens comprise core and variable gene suites which probably have distinct roles in pathogenicity and different evolutionary histories. A frequent hallmark of genes with an extrinsic origin is the difference in GC content of these genes compared with the mean content of the host genome (Dufraigne et al., 2005; Kado, 2009). Around 70% of T4E genes exhibit a high mean GC%, whereas the genomic mean content in *Ehrlichia* spp. is around 28%. The GC content of *Ehrlichia* predicted T4Es is consistently higher than the genomic GC content, suggesting these genes were recently acquired from an exogenous source by horizontal gene transfer (HGT), possibly from natural hosts or the natural tick vector of *Ehrlichia*,

which are typically characterised by high GC content. In addition, in the case of the 69 *Ehrlichia* species-specific effectors, the fact that none showed significant sequence similarity to another *Ehrlichia*-encoded protein underlines the magnitude of the functional novelty of the putative effectors we found. The high GC content of species-specific effectors combined with the fact that most of them contain a secretion signal (Eblock), suggest that recently acquired genes can adapt to function as effectors in a relatively short evolutionary time and could be linked to the change in host. Our results also suggest that a majority of species-specific effectors may be part of the flexible genome in *Ehrlichia* spp.

Despite their obligate intracellular nature, *Ehrlichia* ssp. show an high level of genomic plasticity. As an example, the large numbers of species-specific pT4Es of *E. ruminantium* appear to be randomly inserted into the genome and may therefore be the result of multiple acquisition events. Similarly, we observed a large chromosomal inversion in the genome of *E. chaffeensis*. It has been shown in numerous bacteria that effector-coding genes in close genomic proximity can function together in the host cell (Ingmundson et al., 2007; Kubori et al., 2010; Siamer and Dehio, 2015). Our search for pairs of effectors in *Ehrlichia* spp. led to the identification of 18 pairs of which only seven were present in at least one other *Ehrlichia* genome (data not shown). This suggests a certain degree of co-evolution of these pairs during the evolution of *Ehrlichia* genus, indicating they could have related functions important for *Ehrlichia* spp. pathogenesis. Interestingly, one of these pairs was found in three *Ehrlichia* genomes, and one of the gene had orthologues in the four *Ehrlichia* species and encoded a lipoprotein which plays a role in *Ehrlichia* pathogenesis (Huang et al., 2008).

While analysing the protein architectures (different domain combinations), we noticed that the same domains were often shared

by different architectures. We visualised this phenomenon using hive plots to reveal the network of protein architectures based on their structural properties. This network clearly showed that several domains are present in many effectors and form different architectures. Among the most connected architectures, we found well-known effector domains including nuclear localization signals (NLS), coiled-coil domains, EPIYA domains and E-block domains. Similarly, among the less abundant architectures, we found domains including HATPase_c domains, mitochondrial localization signals (MLS), and BRCT domains which are closely connected. These networks represent a plethora of domains of known and unknown function which could provide information on possible effector functions.

Ankyrin-repeat (Ank) proteins mediate protein-protein interactions involved in a multitude of host processes. Ank domain-containing effectors are crucial for the pathogenesis of several obligate intracellular bacteria (Rikihisa and Lin, 2010). In addition, when associated with NLS domains, such as *Anaplasma phagocytophilum* AnkA, some Ank-containing effectors target the host cell nucleus to directly interfere with the host defences at the gene and chromatin level (Bierne and Cossart, 2012). In *Ehrlichia* spp., we showed that the Ank-containing pT4Es are associated with few other domains (NLS and EPIYA) but they show a myriad of different protein architectures with some sequence rearrangements (duplications). Hence, the sequence and position of the different domains on the Ank-containing effectors point to the many protein interactions and cellular functions they may have in the host cell. Thus, the modular architecture of Ank-containing effectors, as well as the intragenic recombination we showed, may have played a critical role in the evolution of *Ehrlichia* virulence related to host specificity, as shown for SidJ in *Legionella pneumophila* (Costa et

al., 2014). Such high modularity may also be linked to some loss of function, as we showed for *E. ruminantium* ERGA_CDS_03830, which is not secreted in a type IV-dependent manner in the *Legionella pneumophila* heterologous system (data not shown). The great variety and polymorphism of Ank-containing pT4Es may be crucial for increasing the *Ehrlichia* spp. genetic pool and may contribute to the resilience of the bacteria. Further experiments should be conducted to clarify the importance of this polymorphism.

The HATPase_c domain is found in several ATP-binding proteins including histidine kinase and DNA gyrase. This domain is of particular interest because it could define a novel effector function. Even though numerous type III effector kinases have been characterised (Dean, 2011), no HATPase_c-containing T4E has been described to date. These effectors could be highly novel proteins with previously unseen biochemical properties. In contrast to Ank-containing effectors, this family of pT4E shows conserved protein architectures with a wide variety of domain associations. ERGA_CDS_03390 is a species-specific pT4Es of the *E. ruminantium* containing the HATPase_c domain. Even if orthologues of this gene are present in *E. chaffeensis* and *E. canis*, they were not predicted as T4Es by SATE 2.0 software. Indeed, ERGA_CDS_03390 had a score of 101 while ECH_0755 and Ecaj_0319 (homologous proteins of ERGA_CDS_03390) scored below 72, the threshold of S4TE 2.0. The differential score of ERGA_CDS_03390 is due to the presence of a PmrA promoter domain upstream from the gene. The response regulator PmrA is a major regulator of the *icm/dotA* type IV secretion system in *Legionella pneumophila* and *Coxiella burnetii* (Zusman et al., 2007). PmrA is important for the regulation of the effectors and could enable positive regulation of the ERGA_CDS_03390 gene. Its absence in the two other species and the disappearance of the gene in *E. muris*

suggests that this gene no longer has any function in these species. In this sense, ERGA_CDS_03390 could encode a putative effector related to host specificity. ERGA_CDS_03390 is homologous to VirA. VirA is a sensory component of a two-component signal transduction system. This protein is composed of a HisKA domain, a HATPase_c domain and a response_reg domain. The response_reg domain receives the signal from the sensor partner in bacterial two-component systems. It is usually found in the N-terminal of a DNA binding effector domain (Pao and Saier, 1995). Upstream and downstream from this gene are two antisense genes: (I) a gene encoding the periplasmic divalent cation tolerance protein CutA (ERGA_CDS_03380) and (ii) an O-methyltransferase (ERGA_CDS_03400). These two genes may be related to the function of VirA. Indeed, in *A. tumefaciens* crosstalk has been demonstrated between chemotaxis and virulence induction signalling thanks to the interaction modulated by CheR methyltransferase between VirA and transmembrane chemoreceptors MCPs (Guo et al., 2017). In *E. ruminantium*, the three-gene cluster containing ERGA_CDS_03390 pT4E could therefore be related to the perception of the external environment and the subsequent induction of virulence. Moreover, this cluster is antisense, surrounded by ERGA_CDS_03370 and ERGA_CDS_03420, two enzymes involved in the synthesis of methylerithriol phosphate. Finally, the three genes of the cluster (ERGA_CDS_03380, ERGA_CDS_03390, and ERGA_CDS_03400) have high ΔGC content and are located in a gene-sparse region of the genome. Altogether, this suggests that the ERGA_CDS_03390 cluster could be a functional cassette involved in *E. ruminantium* virulence and may have been acquired by horizontal gene transfer (HGT).

The two other conserved families of HATPase_c-containing effectors are exemplified by ERGA_CDS_02450, which encodes the

heat shock protein 90 (HSP90), and ERGA_CDS_04380 *gyrB*, which encodes the beta subunit of DNA gyrase.

The HSP90 is an important cofactor in the response to oxidative stress and has been shown to have a crucial function in innate immune response (Mayor et al., 2007). This family of pT4Es may therefore be important for *Ehrlichia* to manipulate host immunity, possibly by avoiding or delaying pathogen recognition.

GyrB is a topoisomerase that is necessary for relaxation of DNA during replication. GyrB family of *Ehrlichia* pT4Es are expected to act in the nucleus and indeed it is noteworthy that they harbour NLS directly upstream from the HATPase_c domain. Controlling the state of DNA topology could be an efficient way for the bacterium to regulate specific host promoters. These proteins could thus mimic host chromatin-regulatory factors, as described for other effectors (Bierne and Cossart, 2012). The property of subtle mimicry of the activities of cellular proteins is also one of the most common features of type III secreted effectors (Galán, 2009). Moreover, a remarkable feature to note is that *Bartonella* toxin VbhT is secreted into target cells in a type IV-dependent manner and acts as a gyrase inhibitor. This VbhT effector represents the missing link in the evolution of *Bartonella* effectors from inter-bacterial conjugative toxins to inter-kingdom host-targeted effectors (Harms et al., 2017) .

Similarly, the family of *Ehrlichia* HATPase_c-containing pT4Es could be a further example of the wonderful functional plasticity of effector proteins. It is worth mentioning that ERGA_CDS_04380 has been shown to be overexpressed in the virulent Gardel strain of *E. ruminantium* at early stages of infection (Emboulé , 2010).This is reminiscent of the early expression of *Ehrlichia* T4SS (Cheng et al., 2008) and compatible with T4SS-dependent secretion into the host cell. Thus, unravelling the role of

HATPase_c-containing effectors in *Ehrlichia* pathogenesis will be a major challenge but could also highlight the importance of chromatin remodelling in *Ehrlichia* infections.

In this study, we showed that genomic plasticity *sensu largo*, i.e. from gene presence/absence polymorphisms to protein domain shuffling, with hallmarks such as local gene density, G+C content, intergenic recombination, is a major tool for *Ehrlichia* spp. to acquire and evolve new potential virulence functions. Our study revealed that *Ehrlichia* T4Es repertoires comprise core and variable gene suites which likely have distinct roles in pathogenicity as well as different evolutionary histories. We also identified several DNA rearrangements in T4E genes, some of which could be correlated with the host specificity of *Ehrlichia* species. Despite the fact that more genomic sequences of *Ehrlichia* strains and functional validations are needed to provide evidence for robust associations between host range and T4E repertoires, our observations already suggest that the host range is controlled by multiple or differential combinations of T4E determinants, or determinants other than T4E, or that differences in the T4E protein sequence (and even expression) may also be involved. We also identified new protein domain associations among *Ehrlichia* pT4Es, probably rearranged over the course of evolution, which could contribute to the numerous potential functions exerted by *Ehrlichia* effectors. A noticeable feature of *Ehrlichia* effectors is their modular architecture, comprising domains or motifs that could confer an array of subversive functions within eukaryotic cells. These domains/motifs therefore represent a fascinating repertoire of molecular determinants with important roles during infection.

Our study illustrates the power of comparative genomics to understand the host specificity of this important family of pathogens. The properties in the cytoplasmic effector repertoires of *Ehrlichia* may predict the basis for pathogen host specificity. The study of these

T4Es will provide insights not only into fundamental aspects of host-pathogen interactions, but also into the basic biology of eukaryotic cells.

Although the underlying basis for host specialization remains largely unresolved, adaption to a host clearly requires coevolutionary maintenance of a compatible effector repertoire. Further experimental determination of minimal *bona fide* T4Es repertoires required for *Ehrlichia* to cause to disease in a given host is needed. A major challenge in the future will be using systems-level knowledge in pathogen genomics, for different host-bacteria interactions, to predict the potential threat of emerging pathogens or to imagine science-based rapid response plans.

ACKNOWLEDGEMENTS

This study was partly conducted in the framework of the project MALIN “Surveillance, diagnosis, control and impact of infectious diseases of humans, animals and plants in tropical islands” supported by the European Union in the framework of the European Regional Development Fund (ERDF) and the Regional Council of Guadeloupe. We thank E. Albina and D. Goodfellow for reading of this manuscript.

SUPPLEMENTARY MATERIAL

Table S1 : Node ids related to domain names and their occurrences in figure 7.

Id	Domains	Occurrence
1	NLS	334
2	Coiled coils	226
3	EPIYA	182
4	Eblock	143
5	MLS	29
6	HATPase_c	20
7	Ank, TrbL	19
8	GTP_EFTU, IF-2, OxoGdeHyase_C, Toprim	15
9	E1_dh, Response_reg, Rotamase_2, SurA_N_3, Transket_pyr, BRCT, DNA_ligase, HHH_2	14-12
10	PCRF, RF-1	11
11	RNA_pol_Rpb, DUF3514, Ribosomal_L12, SEC-C, SecA, tRNA_SAD, tRNA-synt_2b	10-9
12	HGTP_anticodon, T4SS-DNA_transf, Topoisom_bac, UvrD_C, UvrD-helicase, zf-C4_Topoisom, CaaX	8
13	DNA_gyraseB, HSP90, MutS_I, MutS_II	7
14	Anticodon_1, CarD_CdnL_TRCF, DEAD, Helicase_C, MutS_III, MutS_IV, MutS_V, Sigma70, TRCF, tRNA-synt_1, Bac_GDH, DNA_pol3_alpha, DNA_processg_A, GGDEF, GTA_TIM, HHH_6, HisKA, Lon_C, LON_substr_bdg, Peptidase, Phage-tail_3, PHP	6-5
15	Aconitase, Aconitase_C, Acyltransferase, Aminotran_5, Band_7, CPSase_L_D2, CPSase_L_D3, DUF3023, Hexapep, HTH_8, MGS, PEP-utilizers, PPDK_N, Radical_SAM, Rrf2, S1, Sigma54_activat, SRP_SPB, SRP54, SRP54_N, TGS, TRAM, UPF0004	4
16	HemY_N, HSP70, iPGM_N, Metalloenzyme, Ribosomal_S2, TPR_8, Transcrip_reg, YidC_periplas	3
17	ABC_tran, CagE_TrbE_VirB, DUF2312, GIDA, NAD_synthase, Pyr_redox_2, Pyr_redox_dim, Ribosomal_L28, SBP_bac_8, tRNA-synt_1c, 2-oxogl_dehyd_N, 60KD_IMP, AAA, Abhydrolase_2, AIRS, DnaJ, DUF2497, FtsK, Methyltrans_RNA, Pentapeptide, Phage_capsid, RmuC, RNase_E_G	2-1

Table S2 : Node ids related to domain names and thier occurrences in figure 8.

Id	Domains	Occurrence
1	2-oxogl_dehyd, E1_dh, HATPase_c, OxoGdeHyase_C, Transket_pyr	10
2	MLS	9
3	BRCT, DNA_ligase, HHH_2, Toprim	8
4	HGTP_anticodon, Response_reg, tRNA_SAD, tRNA-synt_2b	7
5	HTH_8, CaaX, SEC-C, SecA, UvrD_C, UvrD-helicase	5
6	CPSase_L_D3, DNA_gyraseB, GTP_EFTU, IF-2, MutS_I, MutS_II, MutS_III, MutS_IV, MutS_V, PCRF, RF-1, Rotamase_2, SurA_N_3, Topoisom_bac, zf-C4_Topoisom	4
7	Anticodon_1, CarD_CdnL_TRCF, DEAD, GGDEF, Helicase_C, TGS, TRCF, tRNA-synt_1	3
8	AAA, Aconitase_C, Aminotran_5, DNA_pol3_alpha, GTA_TIM, HHH_6, HisKA, Lon_C, LON_substr_bdg, MGS, Phage-tail_3, PHP, Radical_SAM, Rrf2, Sigma54_activat, SRP_SPB, SRP54, SRP54_N, TRAM, UPF0004	2
9	60KD_IMP, HemY_N, iPGM_N, Metalloenzyme, Methyltrans_RNA, PEP-utilizers, PPDK_N, Pyr_redox_dim, RNase_E_G, S1, TPR_8, YidC_periplas	1

SUPPLEMENTAL FIGURE LEGENDS

Figure S1. Phylogenetic tree of *Ehrlichia* and *Anaplasma* T4Es shows three different clades. A. A maximum-likelihood tree of 4 *Ehrlichia* species, 3 *Anaplasma* species and *W. endosymbiont* of *D. melanogaster* (out group) was reconstructed on the basis of concatenated nucleic acid alignments of pT4Es shared by all species (core effectome) with 100 bootstrap resamplings. **B.** The identity percentage was calculated for each effector ortholog group (EOG) of the *Ehrlichia* core effectome, and is represented by a heat map. The colour gradient represents the identity between effectors (pale colours mean high similarity).

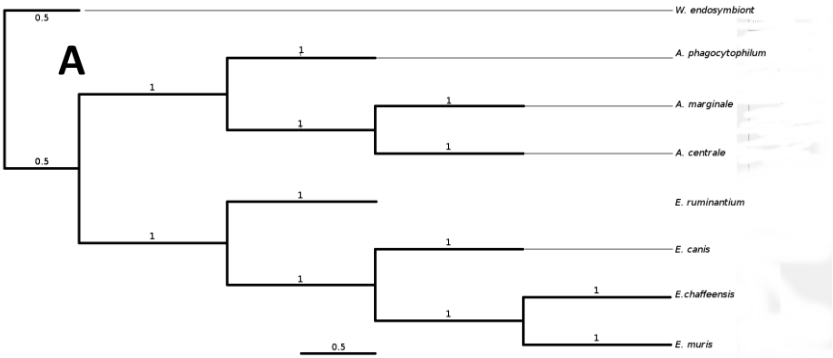
Figure S2. Venn diagram of *Ehrlichia* spp. pT4Es shows a large number of specific effectors. Predicted T4 effectomes of four *Ehrlichia* species compared with S4TE-CG and PanOCT to find homologous proteins in each species. Results are plotted on a Venn diagram and a number indicates the occurrence of predicted effectors is each intersection.

Figure S3. Distribution of predicted type IV effectomes according to local gene density based on the length of flanking intergenic regions (FIRs). Distribution of *E. ruminantium str.* Gardel, *E. chaffeensis* str. Arkansans and *E. muris* AS145 genes according to the length of their flanking intergenic regions (FIRs). All the genes of each species were sorted into two-dimensional bins according to the length of their 5′ (*y*-axis) and 3′ (*x*-axis) FIR lengths. The number of genes in the bins is represented by a colour-coded density graph. Genes whose FIRs were both longer than the median length of FIRs were considered as gene-sparse region (GSR) genes. Genes whose FIRs were both below the median value were considered as gene-dense region (GDR) genes. In between (IBR) genes are genes with a long 5′ FIR and short 3′ FIR, and inversely. For *E. ruminantium*, *E. chaffeensis* and *E. muris*, median values are 246 bp, 156 bp and 207 bp for 5′ FIRs respectively and 405 bp, 138 bp and 219 bp for 3′ FIRs respectively. The dashed line stands for the median length of FIR and delimits the genes in GSR, GDR and IBR. Candidate effectors predicted using the S4TE 2.0 algorithm were *s* plotted on this distribution according to their own 3′ and 5′ FIRs. A colour was assigned to each of the three following groups: red to GDRs, orange to IBRs, and blue to GSRs. Specific pT4Es are represented with a dot outlined in black. On the right, a Circos graph shows the distribution of *E. ruminantium* str. Gardel, *E. chaffeensis*

str. Arkansans and *E. muris* AS145 pT4Es along the genome. The colour (red, orange or blue) of each gene corresponds to their location in GDR, IBR or GSR regions respectively.

Figure S4. Distribution of the effectome according to the length of their flanking intergenic region (FIR) and ΔGC content. Distribution of *E. ruminantium* str. Gardel, *E. chaffeensis* str. Arkansans and *E. muris* AS145 genes according to the length of their flanking intergenic regions (FIRs). All the genes of each species were sorted into two-dimensional bins according to the length of their 5′ (*y*-axis) and 3′ (*x*-axis) FIRs. For each gene, the ΔGC content was calculated by subtracting the GC content of a gene by the average of GC content of all the genes. The mean of ΔGC of genes in the bins is represented by a colour-coded density graph. Genes whose FIRs were both longer than the median length of FIRs were considered as gene-sparse region (GSR) genes. Genes whose FIRs were both below the median value were considered as gene-dense region (GDR) genes. In between (IBR) genes are genes with a long 5′ FIR and short 3′ FIR, and inversely. For *E. ruminantium*, *E. chaffeensis* and *E. muris*, median values are 246 bp, 156 bp and 207 bp for 5′ FIRs, respectively, and 405 bp, 138 bp and 219 bp for 3′ FIRs, respectively. The dashed line showing the median length of FIR delimits the genes in GSR, GDR and IBR. A colour was assigned to each of the three following groups: red to GDRs, orange to IBRs, and blue to GSRs. Specific pT4Es are represented with a dot outlined in black. A density graph is plotted in the top right corner. The red line represents the density of pT4Es according to ΔGC content and the black line represents the density of the other genes.

SUPPLEMENTAL FIGURE



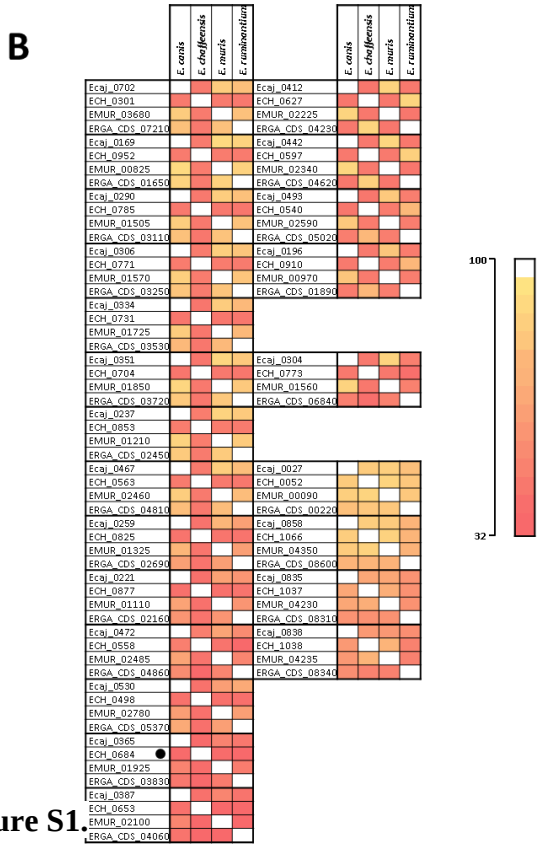


Figure S1.

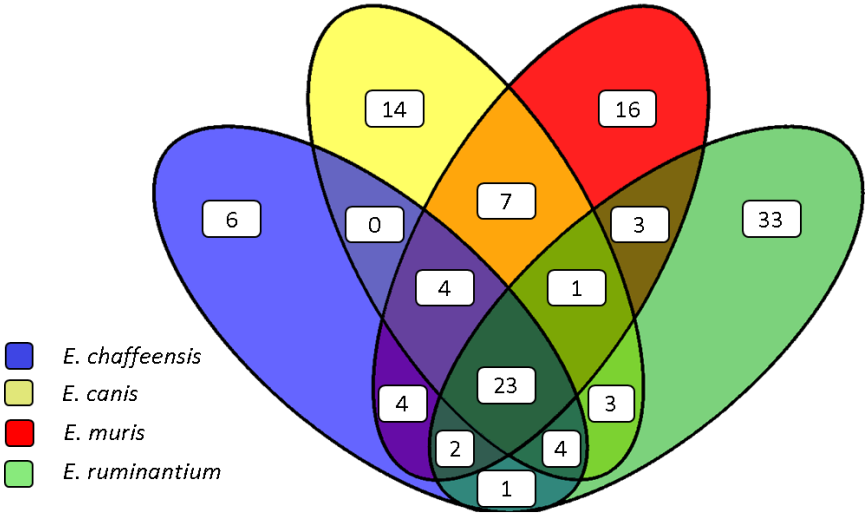

14
16
6
0
7
3
33
4
1
23
4
3
2
4
1
E. chaffeensis
E. canis
E. muris
E. ruminantium

Figure S2.

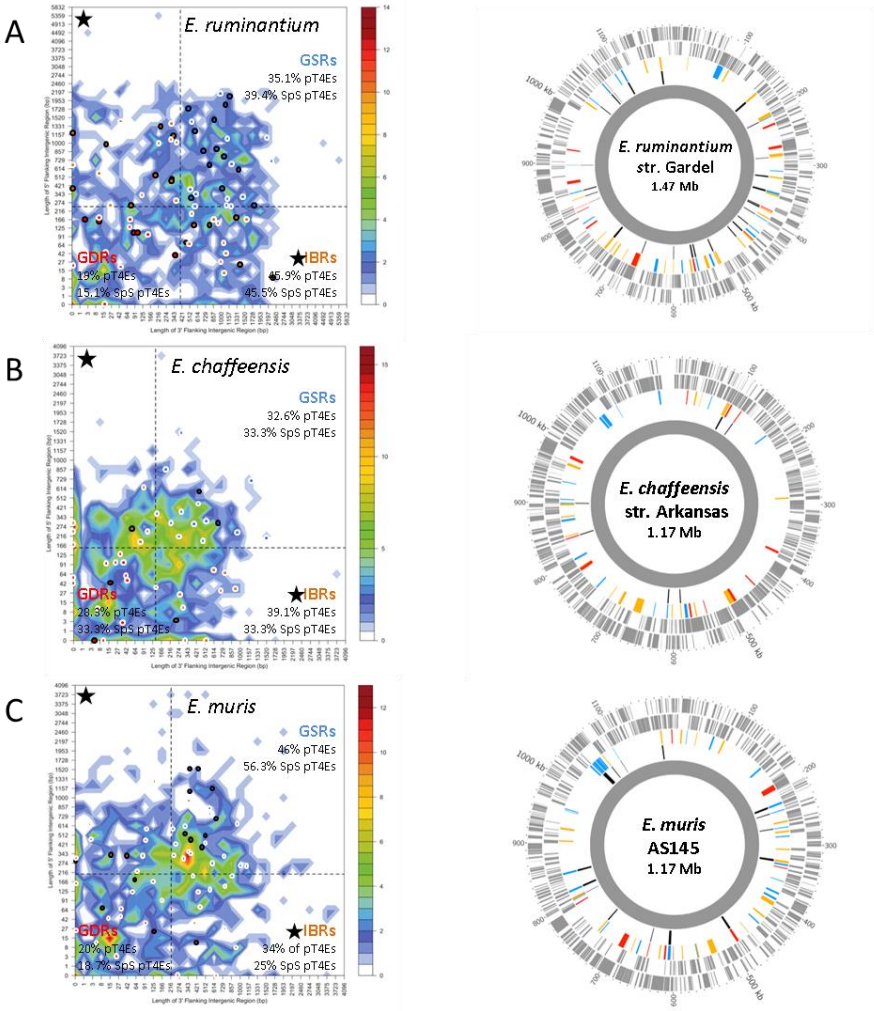

A
E. ruminantium
GSRs
35.1% pT4Es
39.4% SpS pT4Es
GDRs
19% pT4Es
15.1% SpS pT4Es
IBRs
45.9% pT4Es
45.5% SpS pT4Es
E. ruminantium
str. Gardel
1.47 Mb
B
E. chaffeensis
GSRs
32.6% pT4Es
33.3% SpS pT4Es
GDRs
28.3% pT4Es
33.3% SpS pT4Es
IBRs
39.1% pT4Es
33.3% SpS pT4Es
E. chaffeensis
str. Arkansas
1.17 Mb
C
E. muris
GSRs
46% pT4Es
56.3% SpS pT4Es
GDRs
20% pT4Es
18.7% SpS pT4Es
IBRs
34% of pT4Es
25% SpS pT4Es
E. muris
AS145
1.17 Mb

Figure S3.

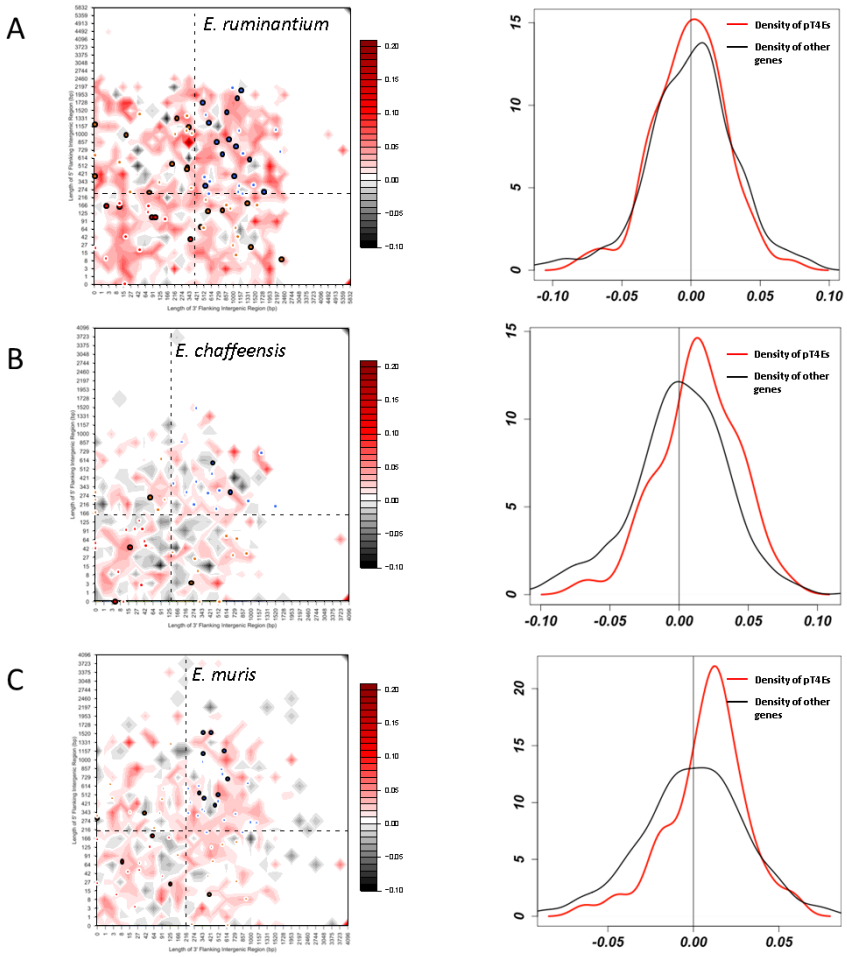

A
E. ruminantium
Density of pT4Es
Density of other genes
B
E. chaffeensis
Density of pT4Es
Density of other genes
C
E. muris
Density of pT4Es
Density of other genes

Figure S4.

REFERENCES

Baltrus, D. A., Nishimura, M. T., Dougherty, K. M., Biswas, S., Mukhtar, M. S., Vicente, J., Holub, E. B., and Dangl, J. L. (2012). The molecular basis of host specialization in bean pathovars of Pseudomonas syringae. *Mol. Plant Microbe Interact.* 25, 877–88.

Bierne, H., and Cossart, P. (2012). When bacteria target the nucleus: the emerging family of nucleomodulins. *Cell. Microbiol.* 14, 622–33.

Bobay, L.-M. M., and Ochman, H. (2017). The Evolution of Bacterial Genome Architecture. *Front Genet* 8, 72.

Braga, Í. A. A., dos Santos, L. G., de Souza Ramos, D. G., Melo, A. L. L., da Cruz Mestre, G. L., and de Aguiar, D. M. (2014). Detection of Ehrlichia canis in domestic cats in the central-western region of Brazil. *Braz. J. Microbiol.* 45, 641–5.

Burstein, D., Amaro, F., Zusman, T., Lifshitz, Z., Cohen, O., Gilbert, J. A., Pupko, T., Shuman, H. A., and Segal, G. (2016). Genomic analysis of 38 Legionella species identifies large and diverse effector repertoires. *Nat. Genet.* 48, 167–75.

Cangi, N., Gordon, J. L., Bournez, L., Pinarello, V., Aprelon, R., Huber, K., Lefrançois, T., Neves, L., Meyer, D. F., and Vachiéry, N. (2016). Recombination Is a Major Driving Force of Genetic Diversity in the Anaplasmataceae Ehrlichia ruminantium. *Front Cell Infect Microbiol* 6, 111.

Caturegli, P., Asanovich, K. M., Walls, J. J., Bakken, J. S., Madigan, J. E., Popov, V. L., and Dumler, J. S. (2000). ankA: an Ehrlichia phagocytophila group gene encoding a cytoplasmic protein antigen with ankyrin repeats. *Infect. Immun.* 68, 5277–83.

Cazalet, C., Rusniok, C., Brüggemann, H., Zidane, N., Magnier, A., Ma, L., Tichit, M., Jarraud, S., Bouchier, C., Vandenesch, F., et al. (2004). Evidence in the Legionella pneumophila genome for exploitation of host cell functions and high genome plasticity. *Nat. Genet.* 36, 1165–73.

Cheng, Z., Wang, X., and Rikihisa, Y. (2008). Regulation of type IV secretion apparatus genes during Ehrlichia chaffeensis intracellular development by a previously unidentified protein. *J. Bacteriol.* 190, 2096–105.
Costa, J., Teixeira, P. G., d' Avó, A. F., Júnior, C. S. S., and Veríssimo, A. (2014). Intragenic recombination has a critical role on the evolution of Legionella pneumophila virulence-related effector sidJ. *PLoS ONE* 9, e109840.
Dean, P. (2011). Functional domains and motifs of bacterial type III effector proteins and their roles in infection. *FEMS Microbiol. Rev.* 35, 1100–25.
Dufraigne, C., Fertil, B., Lespinats, S., Giron, A., and Deschavanne, P. (2005). Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res.* 33, e6.
Emboulé L., Ph.D. dissertation, Université des Antilles et de la Guyane, 2010. Retrieved from https://agritrop.cirad.fr/562394/1/document_562394.pdf.
Farris, T. R., Zhu, B., Wang, J. Y., and McBride, J. W. (2017). Ehrlichia chaffeensis TRP32 Nucleomodulin Function and Localization Is Regulated by NEDD4L-Mediated Ubiquitination. *Front Cell Infect Microbiol* 7, 534.
De Felipe, K. S., Glover, R. T., Charpentier, X., Anderson, O. R., Reyes, M., Pericone, C. D., and Shuman, H. A. (2008). Legionella eukaryotic-like type IV substrates interfere with organelle trafficking. *PLoS Pathog.* 4, e1000117.
Feng, H.-M. M., and Walker, D. H. (2004). Mechanisms of immunity to Ehrlichia muris: a model of monocytotropic ehrlichiosis. *Infect. Immun.* 72, 966–71.
Fouts, D. E., Brinkac, L., Beck, E., Inman, J., and Sutton, G. (2012). PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and

closely related species. *Nucleic Acids Res.* 40, e172.
Galán, J. E. (2009). Common themes in the design and function of bacterial effectors. *Cell Host Microbe* 5, 571–9.
Garcia-Garcia, J. C., Rennoll-Bankert, K. E., Pelly, S., Milstone, A. M., and Dumler, J. S. (2009). Silencing of host cell CYBB gene expression by the nuclear effector AnkA of the intracellular pathogen Anaplasma phagocytophilum. *Infect. Immun.* 77, 2385–91.
Gomez-Valero, L., Rusniok, C., Cazalet, C., and Buchrieser, C. (2011). Comparative and functional genomics of legionella identified eukaryotic like proteins as key players in host-pathogen interactions. *Front Microbiol* 2, 208.
Gomez-Valero, L., Rusniok, C., Rolando, M., Neou, M., Dervins-Ravault, D., Demirtas, J., Rouy, Z., Moore, R. J., Chen, H., Petty, N. K., et al. (2014). Comparative analyses of Legionella species identifies genetic features of strains causing Legionnaires' disease. *Genome Biol.* 15, 505.
Guo, M., Huang, Z., and Yang, J. (2017). Is there any crosstalk between the chemotaxis and virulence induction signaling in Agrobacterium tumefaciens? *Biotechnol. Adv.* 35, 505–511.
Hajri, A., Brin, C., Hunault, G., Lardeux, F., Lemaire, C., Manceau, C., Boureau, T., and Poussier, S. (2009). A "repertoire for repertoire" hypothesis: repertoires of type three effectors are candidate determinants of host specificity in Xanthomonas. *PLoS ONE* 4, e6632.
Harms, A., Liesch, M., Körner, J., Québatte, M., Engel, P., and Dehio, C. (2017). A bacterial toxin-antitoxin module is the origin of inter-bacterial and inter-kingdom effectors of Bartonella. *PLoS Genet.* 13, e1007077.
Huang, H., Lin, M., Wang, X., Kikuchi, T., Mottaz, H., Norbeck, A., and Rikihisa, Y. (2008). Proteomic analysis of and immune

responses to Ehrlichia chaffeensis lipoproteins. *Infect. Immun.* 76, 3405–14.
Ingmundson, A., Delprato, A., Lambright, D. G., and Roy, C. R. (2007). Legionella pneumophila proteins that regulate Rab1 membrane cycling. *Nature* 450, 365–9.
Kado, C. I. (2009). Horizontal gene transfer: sustaining pathogenicity and optimizing host-pathogen interactions. *Mol. Plant Pathol.* 10, 143–50.
Katoh, K., Rozewicki, J., and Yamada, K. D. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinformatics*.
Krzywinski, M., Birol, I., Jones, S. J., and Marra, M. A. (2012). Hive plots--rational approach to visualizing networks. *Brief. Bioinformatics* 13, 627–44.
Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–45.
Kubori, T., Shinzawa, N., Kanuka, H., and Nagai, H. (2010). Legionella metaeffector exploits host proteasome to temporally regulate cognate effector. *PLoS Pathog.* 6, e1001216.
Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 1870–4.
Lawrence, J. G. (2002). Gene transfer in bacteria: speciation without species? *Theor Popul Biol* 61, 449–60.
Lin, M., den Dulk-Ras, A., Hooykaas, P., and Rikihisa, Y. (2007). Anaplasma phagocytophilum AnkA secreted by type IV secretion system is tyrosine phosphorylated by Abl-1 to facilitate infection. *Cellular Microbiology* 9, 26442657.
Liu, H., Bao, W., Lin, M., Niu, H., and Rikihisa, Y. (2012).

Ehrlichia type IV secretion effector ECH0825 is translocated to mitochondria and curbs ROS and apoptosis by upregulating host MnSOD. *Cell. Microbiol.* 14, 1037–50.
Luo, T., Kuriakose, J. A., Zhu, B., Wakeel, A., and McBride, J. W. (2011). Ehrlichia chaffeensis TRP120 interacts with a diverse array of eukaryotic proteins involved in transcription, signaling, and cytoskeleton organization. *Infect. Immun.* 79, 4382–91.
Martinez, E., Allombert, J., Cantet, F., Lakhani, A., Yandrapalli, N., Neyret, A., Norville, I. H., Favard, C., Muriaux, D., and Bonazzi, M. (2016). Coxiella burnetii effector CvpB modulates phosphoinositide metabolism for optimal vacuole development. *Proc. Natl. Acad. Sci. U.S.A.* 113, E3260–9.
Mayor, A., Martinon, F., De Smedt, T., Pétrilli, V., and Tschopp, J. (2007). A crucial function of SGT1 and HSP90 in inflammasome activity links mammalian and plant innate immune responses. *Nat. Immunol.* 8, 497–503.
Meyer, D. F., Noroy, C., Moumène, A., Raffaele, S., Albina, E., and Vachiéry, N. (2013). Searching algorithm for type IV secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context. *Nucleic Acids Res.* 41, 9218–29.
Moumène, A., and Meyer, D. F. (2016). Ehrlichia's molecular tricks to manipulate their host cells. *Microbes Infect.* 18, 172–9.
Ninio, S., and Roy, C. R. (2007). Effector proteins translocated by Legionella pneumophila: strength in numbers. *Trends Microbiol.* 15, 372–80.
Niu, H., Kozjak-Pavlovic, V., Rudel, T., and Rikihisa, Y. (2010). Anaplasma phagocytophilum Ats-1 is imported into host cell mitochondria and interferes with apoptosis induction. *PLoS Pathog.* 6, e1000774.
Niu, H., Xiong, Q., Yamamoto, A., Hayashi-Nishino, M., and Rikihisa, Y. (2012). Autophagosomes induced by a bacterial Beclin

1 binding protein facilitate obligatory intracellular infection. *Proc. Natl. Acad. Sci. U.S.A.* 109, 20800–7.
Noroy, C, and Meyer, DF (2016). Comparative genomics of the zoonotic pathogen Ehrlichia chaffeensis reveals candidate type IV effectors and putative host cell targets. *Frontiers in Cellular and Infection …*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5263134/.
Noroy, C., Lefrançois, T., and Meyer, D. F. (2018). Searching Algorithm for Type IV Effector proteins (S4TE) 2.0: improved tools for type IV effector prediction, analysis and comparison. *bioRxiv*.
O'Connor, T. J., Boyd, D., Dorer, M. S., and Isberg, R. R. (2012). Aggravating genetic interactions allow a solution to redundancy in a bacterial pathogen. *Science* 338, 1440–4.
Paddock, C. D., and Childs, J. E. (2003). Ehrlichia chaffeensis: a prototypical emerging pathogen. *Clin. Microbiol. Rev.* 16, 37–64.
Pao, G. M., and Saier, M. H. (1995). Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution. *J. Mol. Evol.* 40, 136–54.
Park, J., Kim, K. J., Choi, K. S., Grab, D. J., and Dumler, J. S. (2004). Anaplasma phagocytophilum AnkA binds to granulocyte DNA and nuclear proteins. *Cell. Microbiol.* 6, 743–51.
Peter, T. F., Burridge, M. J., and Mahan, S. M. (2002). Ehrlichia ruminantium infection (heartwater) in wild animals. *Trends Parasitol.* 18, 214–8.
Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16, 276–7.
Rikihisa, Y. (2017). Role and Function of the Type IV Secretion System in Anaplasma and Ehrlichia Species. *Curr. Top. Microbiol. Immunol.* 413, 297–321.
Rikihisa, Y., and Lin, M. (2010). Anaplasma phagocytophilum and

Ehrlichia chaffeensis type IV secretion and Ank proteins. *Curr. Opin. Microbiol.* 13, 59–66.
Siamer, S., and Dehio, C. (2015). New insights into the role of Bartonella effector proteins in pathogenesis. *Curr. Opin. Microbiol.* 23, 80–5.
Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–3.
Tago, D., and Meyer, D. F. (2016). Economic Game Theory to Model the Attenuation of Virulence of an Obligate Intracellular Bacterium. *Front Cell Infect Microbiol* 6, 86.
Tamura, K., Battistuzzi, F. U., Billing-Ross, P., Murillo, O., Filipski, A., and Kumar, S. (2012). Estimating divergence times in large molecular phylogenies. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19333–8.
Wang, J., Yang, B., An, Y., Marquez-Lago, T., Leier, A., Wilksch, J., Hong, Q., Zhang, Y., Hayashida, M., Akutsu, T., et al. (2017). Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. *Brief. Bioinformatics*.
Zusman, T., Aloni, G., Halperin, E., Kotzer, H., Degtyar, E., Feldman, M., and Segal, G. (2007). The response regulator PmrA is a major regulator of the icm/dot type IV secretion system in Legionella pneumophila and Coxiella burnetii. *Mol. Microbiol.* 63, 1508–23.

DISCUSSION GENERALE

Discussion générale et perspectives

Les bactéries pathogènes ont acquis lors de leur évolution des facteurs de virulence essentiels à leur mode de vie. En particulier, les effecteurs sécrétés par ces bactéries possèdent de nombreuses fonctions afin de manipuler leur hôte pour échapper à la réponse du système immunitaire inné. L'identification de ces protéines d'intérêt constitue un pas majeur pour une meilleure compréhension de la pathogenèse de ces bactéries.

Au sein d'une famille de bactéries pathogènes, la plasticité génomique peut être étudiée sous différents angles, généralement au niveau des gènes et des évènements de recombinaison de séquence. De façon plus originale, on peut étudier la plasticité au niveau de l'architecture du génome ou encore au niveau des répertoires d'effecteurs. Dans cette partie de discussion, les résultats décrits précédemment seront replacés dans un contexte plus global en utilisant les facteurs de virulence comme marqueurs de la plasticité des génomes.

1. Mutations, recombinaisons, réarrangements : vers une analyse classique de la plasticité génomique

L'architecture des gènes bactériens peut être modifiée par de multiples évènements de recombinaison d'insertion/délétion, de duplication, d'inversion ou de translocation (Darmon and Leach, 2014). Ces réarrangements peuvent entrainer l'apparition ou la perte de gènes (ou de fonctions) au sein du génome bactérien (Ma et al., 2006). Pour pallier à cette perte de fonction, il existe une redondance fonctionnelle des effecteurs clés de la pathogénèse chez les bactéries pathogènes comme *Legionella pneumophila*, *Pseudomonas syringae*

(O'Connor et al., 2011; Cunnac et al., 2011) . Chez *Legionella pneumophila,* une étude portant sur le criblage génétique d'une banque de mutants a permis d'identifier les gènes essentiels à la multiplication de la bactérie. Cette étude a mis en évidence que la perte de 31% des ET4 connus chez *Legionella* n'entrainait pas de chute de croissance perceptible entre la souche mutée et la souche sauvage dans la cellule hôte (O'Connor et al., 2011). De plus, chez *Pseudomonas syringae* pv. *tomato* DC3000, une bactérie pathogène de plante dépendante du système de sécrétion de type III, les 28 effecteurs connus de ce pathovar ont été supprimés. En réassociant de façon aléatoire les effecteurs avec le génome mutant, Cunnac *et al.* (2011) ont montré que seulement huit effecteurs parmi les 28 de départ étaient suffisants pour induire la maladie (Cunnac et al., 2011). Chez les bactéries, cinq types de redondances fonctionnelles ont été mis en évidence (Ghosh and O'Connor, 2017). Tout d'abord, les gènes portant des redondances fonctionnelles peuvent être des gènes dupliqués ayant une fonction équivalente sur une même protéine cible. Mais en grand majorité, les fonctions redondantes sont portées par des gènes codant pour des protéines différentes mais ciblant diverses protéines dans une même voie métabolique. Des études montrent que les gènes redondants sont maintenus dans le génome et ne subissent pas la dérive génétique observable chez les protéines dupliquées (Clark, 1994; Lynch et Force, 2000; Bergthorsson et al., 2007). En effet, deux protéines redondantes peuvent avoir une fonction commune et chacune un autre rôle distinct. Cela a notamment pu être mis en évidence chez *Legionella* entre les protéines SidI et Lgt1 qui inhibent la synthèse protéique en ciblant eEF1A mais qui ont aussi d'autres rôles comme l'interaction avec eEF1Bγ pour SidI (Shen et al., 2009) et la surveillance des mRNA pendant la traduction des protéines pour Lgt1 (Belyi et al., 2009). Ces mécanismes de redondance fonctionnelle ne semblent pas être

courant chez les bactéries intracellulaires obligatoires, dont celles de la famille des *Anaplasmataceae.* En effet, ces bactéries ont un génome beaucoup plus restreint (<1,5Mb) que d'autre bactéries pathogènes (>3MB pour *Legionella pneumophila*). De plus, la réduction extensive de la taille du génome au cours de l'évolution des bactéries intracellulaires obligatoires maintient de grandes familles de gènes ainsi qu'une diversité importante de domaines protéiques avec une perte simultanée de la redondance génétique (Mendonça et al., 2011). Ainsi, l'ensemble des facteurs de virulence, en particulier les ET4, pourrait avoir un rôle unique essentiel dans la pathogénèse des bactéries intracellulaires obligatoires. En ce sens, on observe une forte conservation des effecteurs prédits entre les souches d'*E. chaffeensis.* En effet, 96% des effecteurs prédits sont partagés entre toutes les souches. D'autre part, chez *E. ruminantium,* nous avons montré qu'un effecteur prédit (ERGA_CDS_03830), homologue à l'effecteur connu AnkA, ne semble pas être sécrété en système hétérologue chez *Legionella*. L'analyse de la conservation de la séquence ERGA_CDS_03830 avec son orthologue d'*E. chaffeensis* montre un grand nombre de répétitions dans la partie N-terminale de cette séquence protéique (Noroy and Meyer, 2016). De plus, l'analyse des séquences des ET4 via une approche d'apprentissage automatique a mis en évidence la présence de séquences consensus en N-terminal avec une surreprésentation des acides aminés lysine et asparagine et une absence de glycine et d'alanine (Wang et al., 2017). La présence de répétitions N-Terminal de la protéine ERGA_CDS_03830 pourrait donc expliquer l'absence de sécrétion de cette protéine en système hétérologue. Bien que nous ayons vu que les effecteurs des bactéries intracellulaires obligatoires semblent avoir des fonctions clés dans la pathogénèse en raison la forte diminution des redondances fonctionnelles, il est possible qu'un autre effecteur ait le rôle de la protéine AnkA chez *E. ruminantium*. Cela

est conforté par le fait que le génome d'*E. ruminantium* est 25% plus grand que celui d'*E. chaffeensis* et que l'on retrouve environ 50% d'effecteurs prédits en plus. D'autre part, un grand nombre d'effecteurs prédits chez *E. ruminantium* possèdent des Signaux de Localisation Nucléaire (NLS) et peuvent donc potentiellement agir en tant que nucléomoduline et avoir une fonction équivalente à AnkA.

Une étude portant sur l'analyse de la diversité génétique chez *E. ruminantium* a montré que la recombinaison génétique était un moteur important de la diversité dans la famille des *Anaplasmataceae* (Cangi et al., 2016). De plus, ces mécanismes de recombinaison semblent avoir un rôle important la pathogénèse. En effet, chez la souche Senegal d'*E. ruminantium,* il a été montré que le gène *ntrX* (régulateur à deux composants, exprimé lors de l'infection de la cellule hôte) subissait une délétion de quatre paires de bases non aléatoire lorsque la bactérie était cultivée *in vitro* de façon répétitive (processus d'atténuation). Cette délétion serait induite par la conversion d'une duplication inversée partielle du gène *ntrX* contenant déjà la délétion. Cette mutation entrainerait ainsi une modification de la régulation de la bactérie et permettrait une meilleure croissance (meilleur fitness) de la bactérie dans cet environnement sans pression de sélection liée à l'hôte (Gordon et al., *en preparation*).

Au delà de la plasticité au niveau des séquences, nous avons observé une plasticité au niveau de la position des domaines protéiques. Au sein de la famille protéique ankyrine dans le genre *Ehrlichia,* la comparaison des effecteurs montre une grande variabilité de ces domaines (article Partie 3). Les domaines ankyrines sont des domaines d'interaction protéine-protéine les plus communs dans l'ensemble du vivant (Mosavi et al., 2004). Ils interviennent dans un grand nombre de protéines aux fonctions diverses,

principalement dans les cellules eucaryotes (Bork, 1993). La position de ces motifs semble être essentielle au bon fonctionnement de la protéine. En effet, une étude de cristallographie a montré l'importance de leurs positions dans la structure tertiaire des protéines ankyrines (Mosavi et al., 2004). La variabilité dans la position des domaines ankyrines observée dans le genre *Ehrlichia* (Partie 3, Rikihisa and Lin, 2010) mais aussi pour les légionnelles (Burstein et al., 2016) est sans doute liée à la variabilité des cellules hôtes ciblées et/ou des protéines cibles elles-mêmes. Ainsi, cette variabilité pourrait être une conséquence de la pression évolutive exercée sur ces gènes pour correspondre au mieux à leurs cibles.

En plus de la variabilité associée à la position des domaines sur les effecteurs, nous avons observé une variabilité au niveau de l'association des différents domaines (Partie 3, Burstein et al., 2016). La variabilité de l'architecture de domaines des ET4 dans le genre *Ehrlichia* est grande. En effet, dans la séquence protéique des 228 ET4 prédits chez *Ehrlichia spp.*, on dénombre plus de 400 paires de domaines associés. Les domaines les plus représentés sont les NLS, séquence protéique d'adressage des protéines dans le noyaux de la cellule hôte (Nguyen Ba et al., 2009), les coiled-coils, domaines agissant dans la structure des protéines et dans l'interaction protéine-protéine (Boysen et al., 2002), les domaines EPIYA, domaines de phosphorylation des tyrosine impliqué dans la maturation et la stabilisation des protéines (Papadakos et al., 2013), et le domaines Eblock , reconnu comme étant un domaine associé à la sécrétion des effecteurs chez les légionnelles (Huang et al., 2011). L'ensemble de tous les domaines prédits dans le genre *Ehrlichia* présente des associations inédites comme nous l'avons montré chez *E. ruminantium* (Partie 3). Par exemple, un effecteur candidat, ERGA_CDS_03530, appartenant à la famille des prolyl isomérase, présente une combinaison de domaines insolites chez les effecteurs

connus. Les prolyl isomérases sont des enzymes capables d'inter-convertir les isomères cis- et trans- des liens peptidiques de l'acide aminée proline afin de modifier le repliement des protéines (Schmid, 2001). Cet effecteur candidat est composé du domaine actif de la prolyl-isomérase (Rotamase_2), d'un domaine d'interaction protéine-protéine (Coiled-Coils), d'un motif de protéine chaperonne permettant de faciliter les repliements des protéines (Bitto and McKay, 2002) et d'un domaine de localisation subcellulaire (NLS). Ce type d'effecteurs (ainsi que son architecture de domaines) n'a encore jamais été décrite dans la famille des *Anaplasmataceae* et pourrait agir dans le noyau de la cellule hôte en se fixant et en changeant la configuration tertiaire d'une protéine cible.

Cette variabilité d'architectures et ces réarrangements au niveau de la séquence des effecteurs, nous montre que la plasticité au sein même d'un gène est très importante et pourrait être induite par la pression évolutive exercée par l'environnement sur la bactérie. De plus, il est intéressant de noter que les effecteurs, protéines agissant à l'interface hôte-bactérie et ayant un rôle majeur dans la pathogénèse bactérienne, sont en constante évolution due à la course à l'armement induite par la pathogénèse (Bliven and Maurelli, 2016). Les effecteurs sont donc des marqueurs privilégiés de la plasticité chez les bactéries pathogènes.

Pour étudier la plasticité génomique au niveau de l'inter-connectivité des domaines nous avons mis au point une nouvelle représentation graphique intuitive en utilisant le logiciel Hiveplot (Krzywinski et al., 2012). Hiveplot permet de représenter des réseaux interconnectés selon 3 axes distincts. Cette représentation devient ainsi plus facile à lire, plus intuitive et moins aléatoire que la représentation classique des réseaux. Le réseau d'interaction ainsi représenté ne dépend que de la structure même du réseau. Ce type de

représentation n'a encore jamais été utilisé pour représenter l'architecture des domaines au sein des protéines. Cependant, elle permet regarder l'association des domaines entre eux et de voir la variabilité et l'importance de chaque combinaison de domaine protéique identifié. Ce type de représentation devient essentiel lorsque le réseau représenté est trop complexe (trop de connexions entre les différents domaines), ce qui était le cas pour notre étude.

2. Mise en évidence de la plasticité génomique par l'analyse de l'architecture des génomes

Le génome bactérien, est composé d'une mosaïque de régions denses et clairsemées en gènes. Avec une représentation classique du génome, il est difficile de se représenter l'architecture du génome bactérien. Afin de mieux appréhender cette architecture, nous avons utilisé une approche originellement développée pour visualiser l'architecture des génomes des champignons phytopathogènes (Raffaele et al., 2010; Meyer et al., 2013; Noroy et al., 2018). Grâce a cette représentation, nous sommes capables de définir des Régions Dense en gènes (RDG), correspondant aux gènes en cluster de la bactérie, et des Régions Clairsemés en gènes (RCG) correspondant aux gènes dans les zones plastiques du génome. Ainsi, au lieu de regarder l'architecture du génome de façon linéaire, nous pouvons l'observer selon une nouvelle dimension : celle de la densité locale de gènes. Le point délicat de cette représentation réside dans la définition des différents cadrans. En effet, en fonction de la taille des gènes et de l'architecture du génome la valeur de la médiane des longueurs des RIF peut être sous-évaluée ou surévaluée. Par exemple, entre deux génomes différents, le premier ayant un génome de petite taille avec un grand nombre de gènes et le deuxième ayant un génome de grande taille et moins de gènes, on observerait une grande

différence au niveau des médianes de ces deux génomes. Afin de minimiser ce biais, il faudrait pouvoir répondre a cette question : quelle est la longueur maximale que peuvent avoir les RIF dans un cluster ? Ou en d'autres termes, quelle serait la valeur minimale de la limite (5' et 3') pour que 100% des gènes en cluster soient dans les RDG? Cependant, les bactéries étudiées dans le cadre de cette thèse appartiennent toutes à la famille des *Anaplasmataceae*, et possèdent une taille de génome et un nombre de gènes suffisamment proches pour être comparables et ne pas poser de problème lors de l'analyse. Lorsque l'on reporte la position des effecteurs prédits par le programme S4TE sur ce graphique, on observe une proportion plus importante (d'environ 1,5 fois) des effecteurs dans les RCG par rapport au génome. Cet enrichissement a été observé chez d'autres bactéries pathogènes comme *Legionella pneumophila* (Meyer et al., 2013) ou encore chez *Phytophtora infestans*, pathogène de plantes où 82,8% du secrétome est exclu des RGD (Raffaele et al., 2010). Il est donc particulièrement pertinent de se demander si la position d'un gène dans un RCG ne pourrait pas être un critère de prédiction d'effecteur.

Alors que nous avons identifié les régions plastiques du génome et mis en évidence un enrichissement des ET4 prédits dans ces régions, il est pertinent de se demander si les transferts horizontaux de gènes (THG), qui ont lieu préférentiellement dans les RCG, ne pourraient pas être une source importante d'acquisition de nouveaux effecteurs. Afin d'essayer de répondre à cette question, nous avons analysé la composition en Guanine et en Cytosine (GC) des gènes des bactéries du genre *Ehrlichia*. En effet, le taux de GC entre les différentes espèces animales étant très variable, une caractéristique fréquemment observable des gènes d'origine extrinsèque est la différence du taux de GC de ces gènes par rapport à la moyenne du génome (Lawrence and Ochman, 1998). Par

exemple, au sein de la famille des *Anaplasmataceae*, une grande variabilité du taux de GC est observée. Le genre *Ehrlichia* possède un taux de GC n'excédant pas 30,1% alors que le genre *Anaplasma*, pourtant proche phylogénétiquement, possède un taux de GC d'environ 45%. Le taux de GC du genre *Ehrlichia* étant très faible, nous avons recherché des régions génomiques codant pour des gènes avec un haut taux de GC afin d'essayer de retracer les THG. En ce sens, nous avons développé une représentation graphique permettant d'associer (i) le ΔGC c'est à dire la différence entre la composition en GC d'un gène et la composition moyenne de l'ensemble des gènes et (ii) la densité locale des gènes. Une analyse complémentaire de la densité des ΔGC par rapport au GC moyen des gènes a aussi été effectuée afin de quantifier les différences observées. Grâce à cette analyse, nous avons observé qu'environ 70% des gènes codant des ET4 candidats d'*E. canis*, d'*E. muris* et d'*E. chaffeensis* étaient présents dans des zones riches en GC. De plus, sur la courbe de densité des ET4 prédits, nous avons pu mettre en évidence chez toutes les espèces du genre *Ehrlichia*, la présence de pics de densité dans des hauts et bas taux de GC. La composition en GC des ET4 prédits chez *Ehrlichia* est régulièrement supérieure au taux de GC moyen des génomes et la présence de différents pics distincts dans les courbes de densité suggère que ces gènes pourraient être acquis via des THG depuis différentes sources exogènes dans un temps évolutif relativement court.

D'autre part, une partie des domaines identifiés chez les effecteurs des *Ehrlichia* possèdent des domaines archétypiques des eucaryotes, comme par exemple les domaines Ankyrines. Ces domaines sont donc évidemment présents chez les eucaryotes mais aussi dans de rares cas, dans un petit nombre de génomes bactériens (particulièrement chez les bactéries pathogènes ou endosymbiotiques) (Ponting et al., 1999; Lurie-Weinberger et al.,

2010). Chez *Legionella pneumophila*, l'analyse phylogénétique des différents gènes codant pour des protéines Euk-like a permis de démontrer que ces gènes dérivaient à la fois de THG mais aussi de l'évolution de gènes bactériens devenus Euk-like par une adaptation graduelle au fil du temps ou grâce à l'acquisition de fragments de gènes lors de l'évolution (Lurie-Weinberger et al., 2010).

Enfin, nous avons pu observer chez les 69 effecteurs prédits spécifiques du genre *Ehrlichia* – c.a.d les effecteurs putatifs n'ayant pas homologies parmi les ET4 des autres espèces d'*Ehrlichia* - que les gènes semblaient être insérés de façon aléatoire dans le génome. De plus, on observe aussi qu'environ 50% des ces effecteurs prédits sont insérés dans les zones faiblement denses en gènes (RCG) et que par ailleurs, ces gènes sont aussi présents dans des zones à haut taux de GC. Tous ces résultats combinés et le fait qu'une partie de ces protéines possède des domaines Euk-like, suggèrent fortement que ces gènes pourraient avoir été acquis récemment par transfert horizontal.

Au début de la thèse, une des hypothèses principales était de dire que les effecteurs conservés et les effecteurs spécifiques pouvaient être présent dans des zones de plasticité différentes. Plus précisément, la question que nous nous posions était de savoir si les effecteurs spécifiques d'espèce et /ou de souches étaient principalement présents dans les RCG et les effecteurs conservés dans les RDG. Une autre hypothèse importante était de dire que la distribution selon la densité génique locale (en particulier les longueurs des RIFs) pouvait être un critère prédictif pour les effecteurs comme cela a pu être mis en évidence chez un oomycète pathogène de plante *Phytophtora infestans* (Haas et al., 2009). Pour confronter ces hypothèses, nous avons développé pour la première fois pour les génomes bactériens, une représentation graphique

permettant de visualiser la plasticité du génome en fonction de la longueur des RIFs. Ce graphique ne nous a pas permis de montrer que les effecteurs conservés (effectome cœur) étaient présents dans les zones denses en gènes, mais montre bien que les effecteurs spécifiques de souche/ espèces sont présents en majorité dans les zones du génome plastiques peu denses en gènes. Concernant l'hypothèse de prédiction des effecteurs en fonction de la longueur des RIFs, et alors que l'on observe bien un enrichissement des effecteurs dans les zones peu denses en gènes, la longueur des RIFs semble être une bonne variable indicatrice de la famille de gènes « effecteurs », mais ne semble pas être suffisante pour prédire seule les effecteurs.

3. La spécificité d'hôte comme facteur influençant la plasticité génomique

Lors de nos travaux de génomique comparative au sein de l'espèce *Ehrlichia chaffeensis* ainsi que dans la famille des *Anaplasmataceae*, nous cherchions à identifier au niveau d'un genre et d'une famille bactérienne, les différences de répertoire d'effecteurs pouvant impliquer une différence de virulence ou de spécificité d'hôte. Pour cela, nous avons utilisé l'algorithme de comparaison génomique du logiciel S4TE2.0 (Noroy et al., 2018) qui nous a permis de mettre en évidence des effecteurs présents dans toutes les souches/espèces et définissant l'effectome cœur. Ces effecteurs conservés dans l'ensemble des génomes bactériens que nous avons étudiés sont primordiaux et pourraient avoir des fonctions essentielles dans la pathogénèse bactérienne (Cunnac et al., 2011). Les autres ensembles d'effecteurs remarquables sont les effecteurs n'ayant aucune homologie parmi les souches/espèces étudiées. Ces effecteurs uniques et variables, définissant l'effectome accessoire, sont très

intéressants à étudier car ils pourraient être liés aux différences de virulences/ spécificité d'hôtes observées (Hajri et al., 2009).

L'étude des répertoires d'E4Ts putatifs au niveau des souches *d'Ehrlichia chaffeensis* a montré une très forte similarité entre les répertoires d'effecteurs avec seulement un effecteur candidat spécifique de la souche Liberty (Noroy et Meyer, 2016). Cette souche bactérienne montre une différence de virulence chez les souris SCID. Cet effecteur putatif pourrait être impliqué dans le changement de pathogenèse observé. Cependant, l'élément le plus frappant de cette étude est la similarité des génomes étudiés. En effet, chez les différentes souches *E. chaffeensis,* et bien que l'on observe un certain niveau de plasticité au niveau de l'architecture du génome, les répertoires d'effecteurs, notamment pour les effecteurs spécifiques de souche d'*E. chaffeensis,* semblent très conservés, contrairement à ce que l'on peut observer chez *L. pneumophila* (Gomez-Valero et al., 2014). En effet chez cette bactérie, l'ensemble de l'effectome compte pour environ 10% du génome bactérien contrairement à *Ehrlichia chaffeensis* qui ne compte que 5% d'effecteurs prédits. Comme nous avons pu le discuter dans la première partie de la discussion, la redondance fonctionnelle chez les légionnelles est importante et pourrait expliquer cette différence. Ainsi, la diminution de l'effectome observé chez *Ehrlichia* pourrait être en liée à la réduction de la taille de son génome en raison de son mode de vie intracellulaire et seuls les gènes essentiels auraient été sauvegardés lors de l'évolution de cette bactérie.

Bien qu'il soit généralement admis que le transfert horizontal de gènes, et donc l'acquisition de nouveaux gènes, joue un rôle important dans l'évolution des répertoires effecteurs, l'impact de ce processus et son importance par rapport à la patho-adaptation est rarement quantifié. La patho-adaptation est le mécanisme selon

lequel l'acquisition de nouveaux gènes proviendrait d'évènements de mutation de gènes (Sokurenko et al., 1999). Cependant quelques études de phylogénies on montré que les deux mécanismes d'évolution (THG et patho-adaptation) étaient visibles chez différents pathogènes (Prager et al., 2000; Lavie et al., 2004; Rohmer et al., 2004).

Le rôle des répertoires d'effecteurs en lien avec la spécificité d'hôte a été mis en évidence chez certaines bactéries pathogènes comme *Xanthomonas* (Schwartz et al., 2015). En comparant les différents génomes de la famille des *Anaplasmataceae*, nous avons mis en évidence des répertoires d'effecteurs putatifs très différents. Parmi ces répertoires, il est à noter que 30% des effecteurs prédits étaient des effecteurs spécifiques du genre *Ehrlichia*. De plus, nous avons montré une forte similarité entre les répertoires d'effecteurs d'*E. muris* et d'*A. phagocytophilum*. Ces deux bactéries possédant le même hôte commun (rongeurs), on peut se demander si les effecteurs spécifiques d'*Ehrlichia* muris et ceux d'*Anaplasma* phagocytophilum ne seraient pas en lien avec la spécifié d'hôte. Un autre indice qui pourrait mettre en relation les effecteurs spécifique d'E*hrlichia* avec une spécificité d'hôte est que 12 ET4 putatifs spécifiques du genre *Ehrlichia* sont homologues avec des ET4 chez les espèces d'*Anaplasma* qui partagent les mêmes hôtes, soutenant ainsi l'hypothèse d'un modèle de coévolution moléculaire entre les répertoires d'effecteurs et l'immunité de l'hôte.

La plasticité des répertoires d'effecteurs entre souches d'une même espèce chez les bactéries intracellulaires semble très faible alors qu'au niveau des espèces d'un même genre, cette plasticité augmente. En effet, on retrouve des effecteurs spécifiques d'espèces partagées par plusieurs genres bactériens. Ces indices nous laissent entrevoir l'importance de la plasticité des répertoires d'effecteurs dans de possibles changements d'hôtes.

4. Vers une étude à N-dimensions de la plasticité génomique

Nous avons vu que l'on pouvait étudier la plasticité à différentes échelles et sous différents angles. Dans cette partie, nous allons proposer d'utiliser des techniques d'écologies spatiale pour tenter de comprendre au mieux les variables impliquées dans la plasticité des génomes.

En supposant que les gènes puissent partager certaines propriétés écologiques avec des espèces et que la plasticité joue un rôle majeur dans la distribution de celles-ci, nous avons considéré le génome d'une bactérie, *Legionella pneumophila*, comme un territoire et ses familles de gènes comme des espèces évoluant sur ce territoire.

Sur la base de ce postulat, on peut objectivement séparer les gènes de ces bactéries en différentes familles, 17 familles dans ce cas d'étude. Deux familles, liées à leur distribution dans les autres souches de *Legionella pneumophila,* correspondent aux gènes du génome coeur et du génome accessoire. Ensuite, 15 familles sont liées à leur fonction dans cette bactérie telles que les effecteurs, les enzymes, les régulateurs, les transporteurs… Dans cette étude, nous voulions évaluer si ces familles de gènes pouvaient avoir des habitats préférentiels dans le génome, c'est-à-dire voir si les gènes montraient des préférences environnementales. Il a donc fallu décrire des variables (ici huit variables) binaires ou quantitatives pouvant intervenir dans ces préférences d'habitats. Ces huit variables ont été choisies pour tenter de caractériser l'environnement génique de chaque gène et ainsi avoir une idée de la plasticité associée à ce gène. Ainsi, nous avons choisi comme variable, la longueur du gène, la longueur des RIFs 5' et 3', le GC% du gènes, les GC% des RIFs 5' et 3', la présence du gène sur le brin sens, la présence d'un autre gène

(ou d'une partie d'un gène) sur le brin opposé et la présence d'un gène imbriqué, c'est-a-dire un gène présentant un recouvrement par un autre gène. Ces huit variables seront utilisées dans cette étude et considérées comme des variables environnementales. Nous avons ensuite cherché à analyser si, avec ces variables environnementales, les familles de gènes ainsi définies avaient un habitat préférentiel dans le génome.

A titre d'exemple, nous allons nous intéresser plus particulièrement à la famille de gènes 'effecteurs'. Pour analyser la préférence d'habitat de cette famille, nous avons utilisé une Analyse Factorielle des Niches Ecologiques (**ENFA**) (Hirzel et Arlettaz, 2003) (Figure1A). L'ENFA est une méthode d'analyse factorielle très proche de l'analyse en composante principale (ACP). L'originalité de cette méthode par rapport à une ACP classique vient du fait que les axes sont construits afin d'avoir un sens écologique clair.

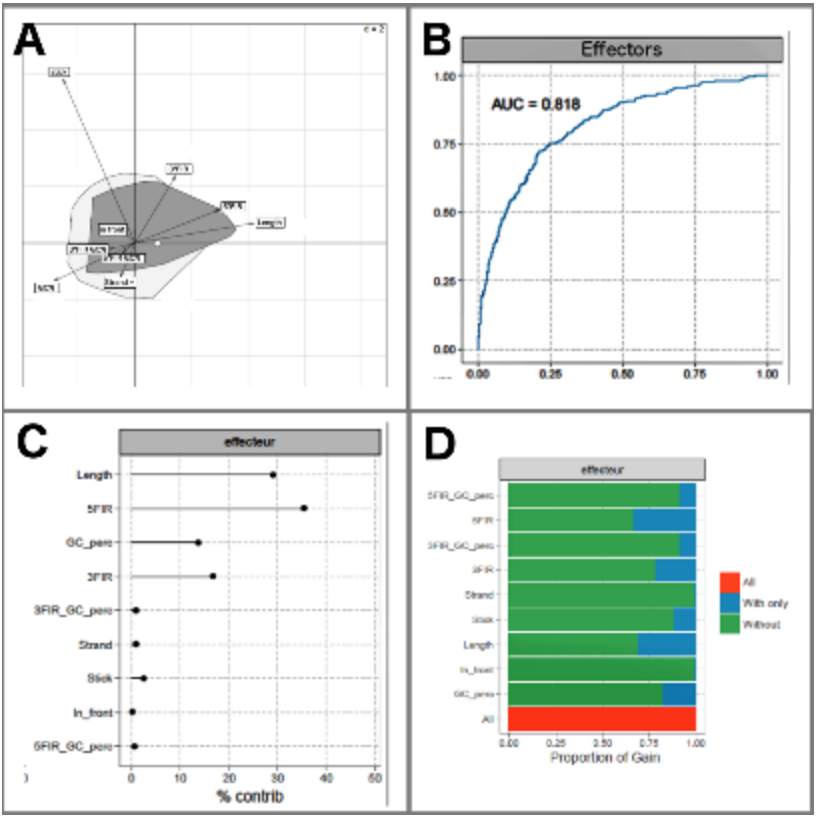


Figure 1 : Analyse du contexte génomique de la famille écologique des effecteurs de *Legionella pneumophila* par les méthodes d'écologie spatiale. A. présentation de la famille écologique des effecteurs via une représentation en ENFA. Les 8 variables écologiques prises en compte dans cette étude sont : la longueur du gène (length), la longueur des régions intergéniques flanquantes 5' (5'FIR), la longueur des régions intergéniques flanquantes 3' (3'FIR), le taux de GC du gènes (GC%), le taux de GC des régions intergéniques flanquantes 5'(5' FIR GC%), le taux de GC des régions intergéniques flanquantes 3'(3' FIR GC%), la présence sur le brin sens du génome (strand +), la présence de gènes ou d'une partie d'un gène sur le brin opposé (in front) et la présence d'un gène imbriqué (stick) **B.** modèle de prédiction de l'habitat de la famille des effecteurs. Un modèle de prédiction est excellent lors que la valeur AUC >0,8. **C.** Contribution des différentes variables environnementales pour la prédiction de l'habitat de la famille des effecteurs **D.** Gain associé à chaque variable environnementale lorsqu'elle est seule lors de la prédiction.

La construction de ces axes est basée sur deux concepts importants : la *marginalité* et la *spécialisation*. La **marginalité** est un indicateur de position qui mesure la déviation de la niche de l'espèce par rapport à l'espace disponible. Plus elle est grande et plus l'espèce préfère des conditions qui s'écartent fortement des conditions moyennes disponibles dans la zone d'étude. La **spécialisation** est une mesure complémentaire de la marginalité, de la même façon qu'une mesure de dispersion l'est par rapport à une mesure de position. Elle donne la forme de la niche en mesurant son degré d'étroitesse. Une spécialisation forte dans une direction (*i.e.* une variable) de l'espace écologique implique que la variance de nuage de l'espace disponible est grand par rapport à celui de l'espace utilisé, donc la niche est *étroite* par rapport à ce qui est disponible à l'espèce pour cette variable.

L'ENFA permet de réaliser une projection optimale de la niche dans un plan. L'interprétation se fait graphiquement en utilisant le biplot, ce qui permet de donner directement l'influence des variables sur le choix de l'habitat par l'espèce. Les variables représentées par des vecteurs de normes élevées sont celles qui jouent un rôle critique dans le choix de l'habitat. L'angle réalisé par ce même vecteur dans le plan est une mesure du degré de marginalité ou/et de spécialisation, plus cette angle est faible ou plat (proche de l'axe des abscisses) et plus cette variable est caractéristique de la marginalité. D' autre part plus l'angle est droit (proche de l'axe des ordonnées) et plus la variable représente un facteur de spécialisation chez l'espèce.

Ainsi, pour l’espèce « effecteur », l’analyse des ENFAs révèle que la taille de ces gènes et que les longueurs des RIFs 5’ et 3’ sont plus longues que la moyenne des autres gènes. D’autre part, l’ENFA montre que le taux de GC associé à ces gènes est plus faible que la moyenne. Chez *Legionella pneumophila*, le taux de GC moyen étant fort (environ 40%), ces données nous laissent à penser que ces gènes,

codant pour de grandes protéines avec un taux de GC faible dans des zones du génome à faible densité et n'ayant pas de recouvrement avec un autre gène, pourraient avoir été acquis chez cette espèce par THG, comme cela a pu être mis en évidence par des études de génomique comparative (Burstein et al., 2016). Ces gènes pourraient donc avoir un rôle dans la spécificité d'hôte, ou encore dans le changement ou saut d'hôte.

En plus de l'analyse selon la méthode des ENFAs, nous effectué une analyse prédictive (MaxEnt) de la niche des différentes familles de gènes afin de voir si les variables environnementales explicatives choisies permettaient de prédire la position de chaque famille (Fourcade et al., 2014). Ainsi sur la figure 1B, nous observons le pouvoir prédictif du modèle a l'aide d'une métrique appellée AUC. Une AUC de 0,5 signifie que la prédiction du modèle est aléatoire et que les variables ne permettent pas d'expliquer la prédiction de la niche écologique. A partir d'une valeur de 0,7, on considère que le modèle a un bon pouvoir prédictif et au dessus de 0,9, le pouvoir prédictif est considéré comme excellent.

Pour le cas de la famille d'effecteurs, nous observons une AUC de 0,8, ce qui nous permet d'affirmer que le pouvoir prédictif du modèle est très bon et que les variables environnementales choisies au départ permettent d'expliquer cette prédiction. La contribution apportée par chaque variable environnementale à la prédiction a ensuite été calculée (Figure 1C) et nous observons pour la famille des effecteurs, que la contribution de la longueur des RIF 5' pour la prédiction du modèle est très importante (plus de 35%). En plus de cela, nous avons calculé le gain associé (Figure 1D) à chaque variable environnementale lorsqu'elle est seule à prédire la niche écologique. Pour les effecteurs, les meilleures variables prédicatrices sont la longueur des RIF 5', suivie par la longueur des gènes et la longueur des RIFs 3'. Lorsque l'on regarde la niche des ENFA (Figure 1A),

cette famille est effectivement associée a de grandes valeurs de 3’ et 5’ RIFs, indiquant que ces gènes sont statistiquement dans des régions peu denses en gènes et donc plastiques.

Cependant, il faut noter que dans ce modèle, aucune variable ne peut être utilisée seule pour prédire la position de la famille des effecteurs sur le génome. Cela indique qu’une approche multi-variée comme celle présentée ici pourrait être un bon moyen d’étude du génome de prédiction des différentes niches écologiques, sous réserve de trouver et définir les bonnes variables environnementales. Un des atouts de cette approche est dû au fait que le génome est un ‘territoire’ fini et que par définition, il n’existe donc pas de biais d’échantillonnage, à l’inverse de ce qui se passe sur le terrain pour des études d’écologie spatiale classique. Ceci constitue un atout de robustesse non négligeable et permet d’avoir des modèles prédictifs beaucoup plus puissants.

Ce type d’approche, que nous proposons d’appeler Ecologie Spatiale Génomique (*Spatial Ecological Genomics*), pourrait à l’avenir et pour peu que l’on identifie toutes les variables environnementales pertinentes pour la prédiction des niches des différentes familles de gènes, permettre l’analyse sans *a priori* des génomes bactériens et définir un nouveau champ thématique en génomique bactérienne.

5. Conclusion générale

Le pari que nous nous étions lancé, d’aborder l’étude de la plasticité génomique des *Anaplasmataceae* en lien avec des déterminants de virulence que sont les effecteurs de type IV n’était pas aisé, au vu de la présence de données biologiques en quantité limitées sur ces modèles bactériens.

Ce n'est que grâce à l'exploitation de l'effectome exhaustif de *Legionella* spp. que nous avons pu confirmer la pertinence de notre approche bioinformatique. Les nombreuses collaborations nouées à l'issue de la publication de la première version du logiciel S4TE, ainsi que les propres travaux développés par l'équipe au laboratoire, nous ont permis de valider la robustesse de notre algorithme et la validité quelques unes de nos prédictions dans différentes espèces bactériennes (*E. ruminantium*, *Coxiella burnetii*, *Brucella* spp. *Liberibacter*, etc)

Le travail présenté dans ce manuscrit permet donc de répondre à certaines des questions que nous nous posions mais le champ d'investigation reste immense pour avoir des réponses encore plus précises et génériques. Ainsi, même si nous comprenons mieux comment la plasticité génomique influe sur les super-répertoires d'effecteurs et comment elle modèle les génomes pour permettre une meilleure adaptation du pouvoir pathogène à l'hôte, nous percevons maintenant que l'exploration de la planète 'génome' nécessite un saut conceptuel majeur. Un génome bactérien est en effet une structure mouvante, évoluant sans cesse en fonction des conditions environnementales et du temps. Nos résultats permettent donc de proposer une nouvelle définition de la plasticité génomique en prenant en compte tous ces facteurs.

Ainsi, nous pourrions définir la **plasticité génomique** comme la capacité de tout ou partie d'un génome, d'acquérir, de perdre ou de dupliquer, à un instant t, tout ou une partie d'un gène ou d'un groupe de gènes. Les zones plastiques du génome sont distribuées de façon non linéaire dans le génome et sont caractérisées par l'environnement génique de chaque gène.

Cette capacité évolue au cours du temps et en fonction du (ou des) gène(s) considérés.

L'étude de cette plasticité des génomes bactériens nécessitera donc une approche multifactorielle, multi-dimensionnelle et transdisciplinaire, utilisant des outils de visualisation des données complexes. Ainsi, comprendre le rôle précis et sans ambiguïté des effecteurs bactériens dans la pathogenèse ne pourra se faire que par l'expérimentation combinatoire des répertoires d'effecteurs. De plus, nous pensons que l'écologie spatiale génomique est un nouveau champ d'expérimentation scientifique qui permettra d'avoir un regard neuf sur les interactions entre espèces de gènes dans le paysage génomique.

REFERENCES BIBLIOGRAPHIQUES

Belyi, Y., Niggeweg, R., Opitz, B., Vogelsgesang, M., Hippenstiel, S., Wilm, M., and Aktories, K. (2006). Legionella pneumophila glucosyltransferase inhibits host elongation factor 1A. *Proc. Natl. Acad. Sci. U.S.A.* 103, 16953–8.
Belyi, Y., Stahl, M., Sovkova, I., Kaden, P., Luy, B., and Aktories, K. (2009). Region of elongation factor 1A1 involved in substrate recognition by Legionella pneumophila glucosyltransferase Lgt1: identification of Lgt1 as a retaining glucosyltransferase. *J. Biol. Chem.* 284, 20167–74.
Bennett, P. M. (2004). Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement. *Methods Mol. Biol.* 266, 71–113.
Bergthorsson, U., Andersson, D. I., and Roth, J. R. (2007). Ohno's dilemma: evolution of new genes under continuous selection. *Proc. Natl. Acad. Sci. U.S.A.* 104, 17004–9.
Bierne, H., and Cossart, P. (2012). When bacteria target the nucleus: the emerging family of nucleomodulins. *Cell. Microbiol.* 14, 622–33.
Bitto, E., and McKay, D. B. (2002). Crystallographic structure of SurA, a molecular chaperone that facilitates folding of outer membrane porins. *Structure* 10, 1489–98.
Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? *Proteins* 17, 363–74.
Boysen, R. I., Jong, A. J., Wilce, J. A., King, G. F., and Hearn, M. T. (2002). Role of interfacial hydrophobic residues in the stabilization of the leucine zipper structures of the transcription factors c-Fos and c-Jun. *J. Biol. Chem.* 277, 23–31.
Burstein, D., Amaro, F., Zusman, T., Lifshitz, Z., Cohen, O., Gilbert, J. A., Pupko, T., Shuman, H. A., and Segal, G. (2016).

Genomic analysis of 38 Legionella species identifies large and diverse effector repertoires. *Nat. Genet.* 48, 167–75.
Cangi, N., Gordon, J. L., Bournez, L., Pinarello, V., Aprelon, R., Huber, K., Lefrançois, T., Neves, L., Meyer, D. F., and Vachiéry, N. (2016). Recombination Is a Major Driving Force of Genetic Diversity in the Anaplasmataceae Ehrlichia ruminantium. *Front Cell Infect Microbiol* 6, 111.
Clark, A. G. (1994). Invasion and maintenance of a gene duplication. *Proc. Natl. Acad. Sci. U.S.A.* 91, 2950–4.
Cunnac, S., Chakravarthy, S., Kvitko, B. H., Russell, A. B., Martin, G. B., and Collmer, A. (2011). Genetic disassembly and combinatorial reassembly identify a minimal functional repertoire of type III effectors in Pseudomonas syringae. *Proc. Natl. Acad. Sci. U.S.A.* 108, 2975–80.
Darmon, E., and Leach, D. R. (2014). Bacterial genome instability. *Microbiol. Mol. Biol. Rev.* 78, 1–39.
Dorman, C. J. (2014). H-NS-like nucleoid-associated proteins, mobile genetic elements and horizontal gene transfer in bacteria. *Plasmid* 75, 1–11.
Filloux, A. (2010). Secretion signal and protein targeting in bacteria: a biological puzzle. *J. Bacteriol.* 192, 3847–9.
Fourcade, Y., Engler, J. O., Rödder, D., and Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PLoS ONE* 9, e97122.
Garcia-Garcia, J. C., Rennoll-Bankert, K. E., Pelly, S., Milstone, A. M., and Dumler, J. S. (2009). Silencing of host cell CYBB gene expression by the nuclear effector AnkA of the intracellular pathogen Anaplasma phagocytophilum. *Infect. Immun.* 77, 2385–91.

Ghosh, S., and O'Connor, T. J. (2017). Beyond Paralogs: The Multiple Layers of Redundancy in Bacterial Pathogenesis. *Front Cell Infect Microbiol* 7, 467.
Gomez-Valero, L., Rusniok, C., Rolando, M., Neou, M., Dervins-Ravault, D., Demirtas, J., Rouy, Z., Moore, R. J., Chen, H., Petty, N. K., et al. (2014). Comparative analyses of Legionella species identifies genetic features of strains causing Legionnaires' disease. *Genome Biol.* 15, 505.
Gordon, J. L., Chavez, A. O., Martinez, D., Vachiery, N., and Meyer, D. F. In vitro virulence attenuation in Ehrlichia ruminantium caused by gene conversion of ntrX. *In preparation*.
Haas, B., Kamoun, S., Zody, M., Jiang, R., Handsaker, R., Cano, L., Grabherr, M., Kodira, C., Raffaele, S., Torto-Alalibo, T., et al. (2009). Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. *Nature* 461, 393–398.
Hajri, A., Brin, C., Hunault, G., Lardeux, F., Lemaire, C., Manceau, C., Boureau, T., and Poussier, S. (2009). A "repertoire for repertoire" hypothesis: repertoires of type three effectors are candidate determinants of host specificity in Xanthomonas. *PLoS ONE* 4, e6632.
Hirzel, A., and Arlettaz, R. (2003). Modeling Habitat Suitability for Complex Species Distributions by Environmental-Distance Geometric Mean. *Environmental Management* 32, 614–623.
Huang, L., Boyd, D., Amyot, W. M., Hempstead, A. D., Luo, Z.-Q. Q., O'Connor, T. J., Chen, C., Machner, M., Montminy, T., and Isberg, R. R. (2011). The E Block motif is associated with Legionella pneumophila translocated substrates. *Cell. Microbiol.* 13, 227–45.
Jansen, A., Gemayel, R., and Verstrepen, K. J. (2012). Unstable microsatellite repeats facilitate rapid evolution of coding and regulatory sequences. *Genome Dyn* 7, 108–25.

Krzywinski, M., Birol, I., Jones, S. J., and Marra, M. A. (2012). Hive plots--rational approach to visualizing networks. *Brief. Bioinformatics* 13, 627–44.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–45.

Lavie, M., Seunes, B., Prior, P., and Boucher, C. (2004). Distribution and sequence analysis of a family of type ill-dependent effectors correlate with the phylogeny of Ralstonia solanacearum strains. *Mol. Plant Microbe Interact.* 17, 931–40.

Lawrence, J. G., and Ochman, H. (1998). Molecular archaeology of the Escherichia coli genome. *Proc. Natl. Acad. Sci. U.S.A.* 95, 9413–7.

Lifshitz, Z., Burstein, D., Peeri, M., Zusman, T., Schwartz, K., Shuman, H. A., Pupko, T., and Segal, G. (2013). Computational modeling and experimental validation of the Legionella and Coxiella virulence-related type-IVB secretion signal. *Proc. Natl. Acad. Sci. U.S.A.* 110, E707–15.

Low, H. H., Gubellini, F., Rivera-Calzada, A., Braun, N., Connery, S., Dujeancourt, A., Lu, F., Redzej, A., Fronzes, R., Orlova, E. V., et al. (2014). Structure of a type IV secretion system. *Nature* 508, 550–553.

Lurie-Weinberger, M. N., Gomez-Valero, L., Merault, N., Glöckner, G., Buchrieser, C., and Gophna, U. (2010). The origins of eukaryotic-like proteins in Legionella pneumophila. *Int. J. Med. Microbiol.* 300, 470–81.

Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459–73.

Ma, W., Dong, F. F., Stavrinides, J., and Guttman, D. S. (2006). Type III effector diversification via both pathoadaptation and

horizontal transfer in response to a coevolutionary arms race. *PLoS Genet.* 2, e209.
Madsen, J. S., Burmølle, M., Hansen, L. H., and Sørensen, S. J. J. (2012). The interconnection between biofilm formation and horizontal gene transfer. *FEMS Immunol. Med. Microbiol.* 65, 183–95.
McFall-Ngai, M., Hadfield, M. G., Bosch, T. C., Carey, H. V., Domazet-Lošo, T., Douglas, A. E., Dubilier, N., Eberl, G., Fukami, T., Gilbert, S. F., et al. (2013). Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl. Acad. Sci. U.S.A.* 110, 3229–36.
Mendonça, A. G. G., Alves, R. J., and Pereira-Leal, J. B. B. (2011). Loss of genetic redundancy in reductive genome evolution. *PLoS Comput. Biol.* 7, e1001082.
Metzgar, D., Liu, L., Hansen, C., Dybvig, K., and Wills, C. (2002). Domain-level differences in microsatellite distribution and content result from different relative rates of insertion and deletion mutations. *Genome Res.* 12, 408–13.
Meyer, D. F., Noroy, C., Moumène, A., Raffaele, S., Albina, E., and Vachiéry, N. (2013). Searching algorithm for type IV secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context. *Nucleic Acids Res.* 41, 9218–29.
Middleton, R., Sjölander, K., Krishnamurthy, N., Foley, J., and Zambryski, P. (2005). Predicted hexameric structure of the Agrobacterium VirB4 C terminus suggests VirB4 acts as a docking site during type IV secretion. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1685–90.
Mit'kina, L. N. (2003). [Transposition as a way of existence: phage Mu]. *Genetika* 39, 637–56.
Mosavi, L. K., Cammett, T. J., Desrosiers, D. C., and Peng, Z.-Y. Y. (2004). The ankyrin repeat as molecular architecture for protein

recognition. *Protein Sci.* 13, 1435–48.
Moumène, A., and Meyer, D. F. (2016). Ehrlichia's molecular tricks to manipulate their host cells. *Microbes Infect.* 18, 172–9.
Mozhayskiy, V., and Tagkopoulos, I. (2012). Horizontal gene transfer dynamics and distribution of fitness effects during microbial in silico evolution. *BMC Bioinformatics* 13 Suppl 10, S13.
Nguyen Ba, A. N., Pogoutse, A., Provart, N., and Moses, A. M. (2009). NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics* 10, 202.
Niu, H., Kozjak-Pavlovic, V., Rudel, T., and Rikihisa, Y. (2010). Anaplasma phagocytophilum Ats-1 is imported into host cell mitochondria and interferes with apoptosis induction. *PLoS Pathog.* 6, e1000774.
Niu, H., Xiong, Q., Yamamoto, A., Hayashi-Nishino, M., and Rikihisa, Y. (2012). Autophagosomes induced by a bacterial Beclin 1 binding protein facilitate obligatory intracellular infection. *Proc. Natl. Acad. Sci. U.S.A.* 109, 20800–7.
Noroy, C, and Meyer, DF (2016). Comparative genomics of the zoonotic pathogen Ehrlichia chaffeensis reveals candidate type IV effectors and putative host cell targets. *Frontiers in Cellular and Infection …*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5263134/.
Noroy, C., Lefrançois, T., and Meyer, D. F. (2018). Searching Algorithm for Type IV Effector proteins (S4TE) 2.0: improved tools for type IV effector prediction, analysis and comparison. *bioRxiv*.
O'Connor, T. J., Adepoju, Y., Boyd, D., and Isberg, R. R. (2011). Minimization of the Legionella pneumophila genome reveals chromosomal regions involved in host range expansion. *Proc. Natl. Acad. Sci. U.S.A.* 108, 14733–40.
Papadakos, K. S., Sougleri, I. S., Mentis, A. F., Hatziloukas, E., and

Sgouras, D. N. (2013). Presence of terminal EPIYA phosphorylation motifs in Helicobacter pylori CagA contributes to IL-8 secretion, irrespective of the number of repeats. *PLoS ONE* 8, e56291.
Papke, R. T., Corral, P., Ram-Mohan, N., de la Haba, R. R., Sánchez-Porro, C., Makkay, A., and Ventosa, A. (2015). Horizontal gene transfer, dispersal and haloarchaeal speciation. *Life (Basel)* 5, 1405–26.
Periwal, V., and Scaria, V. (2015). Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics* 31, 1–9.
Peterson, P. A. (2013). Historical overview of transposable element research. *Methods Mol. Biol.* 1057, 1–9.
Polz, M. F., Alm, E. J., and Hanage, W. P. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* 29, 170–5.
Ponting, C. P., Aravind, L., Schultz, J., Bork, P., and Koonin, E. V. (1999). Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J. Mol. Biol.* 289, 729–45.
Prager, R., Mirold, S., Tietze, E., Strutz, U., Knüppel, B., Rabsch, W., Hardt, W. D., and Tschäpe, H. (2000). Prevalence and polymorphism of genes encoding translocated effector proteins among clinical isolates of Salmonella enterica. *Int. J. Med. Microbiol.* 290, 605–17.
Raffaele, S., Win, J., Cano, L. M., and Kamoun, S. (2010). Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of Phytophthora infestans. *BMC Genomics* 11, 637.
Rikihisa, Y., and Lin, M. (2010). Anaplasma phagocytophilum and Ehrlichia chaffeensis type IV secretion and Ank proteins. *Curr.*

Opin. Microbiol. 13, 59–66.
Rikihisa, Y., Lin, M., and Niu, H. (2010). Type IV secretion in the obligatory intracellular bacterium Anaplasma phagocytophilum. *Cell. Microbiol.* 12, 1213–21.
Rikihisa, Y., Lin, M., Niu, H., and Cheng, Z. (2009). Type IV secretion system of Anaplasma phagocytophilum and Ehrlichia chaffeensis. *Ann. N. Y. Acad. Sci.* 1166, 106–11.
Rohmer, L., Guttman, D. S., and Dangl, J. L. (2004). Diverse evolutionary mechanisms shape the type III effector virulence factor repertoire in the plant pathogen Pseudomonas syringae. *Genetics* 167, 1341–60.
Schmid, F. X. (2001). Prolyl isomerases. *Adv. Protein Chem.* 59, 243–82.
Schwartz, A. R., Potnis, N., Timilsina, S., Wilson, M., Patané, J., Martins, J., Minsavage, G. V., Dahlbeck, D., Akhunova, A., Almeida, N., et al. (2015). Phylogenomics of Xanthomonas field strains infecting pepper and tomato reveals diversity in effector repertoires and identifies determinants of host specificity. *Front Microbiol* 6, 535.
Shen, X., Banga, S., Liu, Y., Xu, L., Gao, P., Shamovsky, I., Nudler, E., and Luo, Z.-Q. Q. (2009). Targeting eEF1A by a Legionella pneumophila effector leads to inhibition of protein synthesis and induction of host stress response. *Cell. Microbiol.* 11, 911–26.
Sokurenko, E. V., Hasty, D. L., and Dykhuizen, D. E. (1999). Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends Microbiol.* 7, 191–5.
Thomas, R. J., Dumler, J. S., and Carlyon, J. A. (2009). Current management of human granulocytic anaplasmosis, human monocytic ehrlichiosis and Ehrlichia ewingii ehrlichiosis. *Expert Rev Anti Infect Ther* 7, 709–22.

Wang, J., Yang, B., An, Y., Marquez-Lago, T., Leier, A., Wilksch, J., Hong, Q., Zhang, Y., Hayashida, M., Akutsu, T., et al. (2017a). Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. *Brief. Bioinformatics*.
Wang, Y., Guo, Y., Pu, X., and Li, M. (2017b). Effective prediction of bacterial type IV secreted effectors by combined features of both C-termini and N-termini. *J. Comput. Aided Mol. Des.* 31, 1029–1038.
Zhu, B., Nethery, K. A., Kuriakose, J. A., Wakeel, A., Zhang, X., and McBride, J. W. (2009). Nuclear translocated Ehrlichia chaffeensis ankyrin protein interacts with a specific adenine-rich motif of host promoter and intronic Alu elements. *Infect. Immun.* 77, 4243–55.

ANNEXES