UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA VEGETAL

MINING ICE IN GENOMES

COMPARATIVE GENOMICS OF INTEGRATIVE ELEMENTS IN PROKARYOTIC GENOMES

Leonor Maria dos Santos Silva Tomé Quintais

MESTRADO EM BIOLOGIA MOLECULAR E GENÉTICA

2010

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA VEGETAL

MINING ICE IN GENOMES

COMPARATIVE GENOMICS OF INTEGRATIVE ELEMENTS IN PROKARYOTIC GENOMES

Leonor Maria dos Santos Silva Tomé Quintais

DISSERTAÇÃO

Projecto orientado pelo Prof. Doutor Eduardo Rocha, Microbial Evolutionary Genomics group - Insituto Pasteur
e co-orientado Prof. Doutor Pedro Silva, Departamento de Biologia Vegetal – Faculdade de Ciências da Universidade de Lisboa

MESTRADO EM BIOLOGIA MOLECULAR E GENÉTICA

2010

Outline

RESUMO DA DISERTAÇÃO

Elementos Integrativos e Conjugativos

Sabe-se hoje que a transferência horizontal de genes (HGT), processo de transferência de material genético entre organismos não relacionados, desempenha um papel fundamental na evolução dos procariotas. Existem três formas distintas pelas quais HGT pode ocorrer: transformação, transdução e conjugação, sendo este último processo o que se pensa desempenhar o papel mais preponderante. Para que ocorra conjugação as células devem estar em contacto directo, contacto este conseguido através de um complexo multi-proteico produzido pela célula dadora e que se denomina “Mating pair formation” (Mpf). O DNA é geralmente transferido em cadeia simples, sendo posteriormente convertido em cadeia dupla pela maquinaria de replicação da célula receptora. Existem essencialmente dois tipos de elementos que contribuem para este papel determinante da conjugação para a transferência horizontal de genes, plasmídios e ICEs (elementos integrativos e conjugativos). O estudo destes elementos reveste-se de uma importância vital, visto serem os principais vectores de transmissão de, por exemplo, resistência a antibióticos, factores de virulência e produção de produtos antimicrobianos (Burrus and Waldor 2004).

Este projecto incide sobre ICEs, um grupo que inclui todos os elementos que se transferem por conjugação e capazes de se integrar no genoma, independentemente dos mecanismos pelos quais estes dois processos ocorrem (Burrus and Waldor 2004) . Uma vez integrados, os ICEs replicam com o genoma do hospedeiro; quando a sua excisão é induzida, estes elementos circularizam e ocorre um passo de replicação, seguido pela transferência de uma das cópias para a célula receptora por conjugação. Esta cópia integra-se no genoma da célula receptora e a cópia que permanece na célula dadora pode também voltar a integrar-se no seu genoma. Estes elementos apresentam portanto características de transposões, fagos e plasmídios: como transposões, integram-se e sofrem excisão do cromossoma, mas estes elementos não são transferidos de uma célula para outra. Como fagos temperados, integram-se no cromossoma do hospedeiro e replicam com este, mas os fagos são transmitidos por transdução e não por conjugação. Como plasmídios, os ICEs são transmitidos por conjugação, mas os plasmídios não dependem do cromossoma do hospedeiro para replicar e são mantidos como estruturas circulares extracromossómicas.

Os ICEs estão presentes em todas as principais divisões de bactérias e incluem, por exemplo, elementos classificados como transposões conjugativos e ilhas de patogenicidade (Burrus, Pavlovic et al. 2002; Burrus and Waldor 2004). Ao contrário dos plasmídios, descobertos com o surgimento da biologia molecular e estudados desde então, o estudo dos ICEs é recente e não se sabe quantos sistemas existem nos genomas, qual o seu tamanho ou conteúdo génico.

A estrutura central dos ICEs é composta por três módulos: manutenção, transmissão e regulação (Toussaint and Merlin 2002). Para além destas funções essenciais, os ICEs contêm geralmente um grande número de outros genes que conferem potencial adaptativo ao hospedeiro, como acima mencionado. O módulo de manutenção codifica uma recombinase, a proteína responsável pela integração dos ICEs no genoma do hospedeiro. As famílias de recombinases mais amplamente descritas em ICEs são as recombinases de serina e treonina (Wang, Roberts et al. 2000). No entanto estudos recentes revelaram que transposases do tipo DDE podem também desempenhar este papel (Brochet, Da Cunha et al. 2009).

O módulo de transmissão codifica o sistema de conjugação, normalmente um sistema de secreção do tipo IV (T4SS) (Cascales and Christie 2003). Existem quatro tipos principais de T4SS, três deles com base no grupo de incompatibilidade de plasmídios conjugativos: Inc-F (plasmídio F), Inc-P (plasmídio RP4) e Inc-I (plasmídio R64) (Lawley, Klimke et al. 2003). O quarto tipo de T4SS, ICEHin1056, foi recentemente identificado em ilhas genómicas [8].

O módulo de regulação contém genes que regulam a transferência do elemento. Embora pouco se saiba acerca do seu funcionamento, estudos recentes mostram que a presença de tetraciclina ou a activação da resposta SOS induz a conjugação (Beaber, Hochhut et al. 2004).

Apesar de todos os ICEs terem uma estrutura comum, o facto de os módulos e as proteínas por eles codificadas poderem ser muito diferentes confere-lhes plasticidade. Estes elementos são também responsáveis pela plasticidade do genoma do hospedeiro: uma vez que há locais de integração partilhados por diferentes ICEs e devido a estes poderem apresentar uma ampla gama de hospedeiros, tais locais são uma fonte de variabilidade intra-espécie e inter-géneros. Por outro lado, ICEs contêm muitas vezes genes e sequências (tais como a recombinase e sequências de inserção acima mencionadas) que facilitam o recrutamento de outros genes para a estrutura do elemento. Se esta integração ocorrer num locus específico, um conjunto de genes transferidos horizontalmente pode ser então conservado e transmitido entre bactérias (Burrus and Waldor 2004).

Este projecto constitui a primeira iniciativa de identificação e quantificação de ICES em larga escala. Nesta análise foram utilizados todos os 1055 genomas procarioticos sequenciados até à data.

Para identificar a presença de ICEs, a pesquisa centrar-se-á no mecanismo de conjugação, T4SS, isto é, será através da identificação de homólogos de proteínas deste sistema. Esta escolha é justificada porque esta é a característica que distingue inequivocamente os ICEs de todos os outros elementos móveis integrados no genoma. Uma vez identificados os elementos, as recombinases podem ser procuradas na sua vizinhança.

A subfamília de T4SS responsável pela transferência de DNA durante a conjugação bacteriana é conhecida por Mpf/Cp ("Mating pair formation/coupling protein"). O Mpf típico (plasmídio Ti) é composto por onze proteínas conservadas, VirB1-VirB11, que formam o pilus que estabelece contacto entre as duas células. A "Coupling Protein", também designada por VirD4, tem como função promover o transporte do ICE para o sistema Mpf, de modo a que ocorra conjugação (Cascales and Christie 2004; Schröder and Lanka 2005).

VirB4, uma das proteínas do sistema de secreção, parece estar presente nos sistemas conjugativos de Gram-positivas e Gram-negativas, apesar destas últimas não formarem um pilus e utilizarem adesinas para estabelecer o contacto celular (Juhas, Crook et al. 2008). Um estudo prévio do nosso laboratório realizado em plasmídios demonstrou que em 98% dos casos a presença de VirB4 corresponde à presença de todo o complexo (Smillie C., Garcillian M. et al.).

Para que a conjugação se processe é necessária uma relaxase. Em plasmídios estas proteínas são responsáveis pela clivagem inicial do DNA, ao qual permanecem ligadas. É este complexo de nucleoproteína que é reconhecido por VirD4 e transportado para a célula receptora através do sistema T4SS. (Llosa, Gomis-Rüth et al. 2002).

Método

Como atrás descrito, a identificação dos ICEs baseou-se nos sistemas de conjugação tipo IV e na relaxase. Visto que a presença de VirB4 se encontra fortemente associada à presença de todo o complexo, a identificação dos sistemas "Mating pair formation/Coupling protein" será feita através da identificação de VirB4 e VirD4. Em *Proteobacteria*, devido à quantidade de dados disponíveis (representam mais de 50% de todos os genomas procarióticos sequenciados), e ao facto de os sistemas T4SS terem sido identificados neste *clade*, foi possível identificar não só estas duas proteínas mas também as proteínas específicas de cada T4SS.

Para completar a análise, relaxases das seis famílias descritas foram também identificadas (MOBc, MOBf, MOBh, MOBphen, MOBq e MOBv). Relaxases específicas de *Firmicutes* e *Bacteroidetes* foram também analisadas (Flannagan, Zitzow et al. 1994; Xu, Bjursell et al. 2003).

Todas as MOBs, T4CP, VirB4 e restantes proteínas específicas dos sistemas T4SS que utilizámos neste estudo como base para a pesquisa nos genomas foram identificadas em plasmídios, no trabalho efectuado no nosso laboratório acima mencionado. A pesquisa foi efectuada em 1055 genomas, disponíveis na base de dados do NCBI.

A pesquisa de homólogos das diferentes proteínas nos genomas foi efectuada por HMMER (Durbin, Eddy et al. 1999), um programa que analisa sequencias utilizando "profile hidden Markov models", criando matrizes com informação especifica para cada posição – HMM profiles. Após localização de todas as proteínas nos genomas, procedeu-se a identificação dos sistemas T4SS. Todos os genes específicos que se encontravam a uma distância máxima de 25 posições no genoma foram classificados como pertencentes ao mesmo elemento. Avaliação manual foi efectuada em todos os elementos encontrados, juntando elementos adjacentes que se complementavam. Seguiu-se a pesquisa por VirB4, T4CP e MOB nos 25 genes a montante e a ajuzante dos elementos. Novamente, avaliação manual foi efectuada. Esta análise só pode ser efectuada em *Proteobacteria*, pelo que nos restantes organismos apenas a presença de VirB4, T4CP e MOB foi avaliada. Após concluída a identificação dos elementos, dados obtidos por BLASTP com bom e-value foram também avaliados. Se um cluster não estivesse completo e a proteína que o completaria, encontrada por BLASTP (Altschul, Gish et al. 1990), estivesse próxima, seria incluída no elemento. Estes dados foram utilizados apenas para complementar a análise principal uma vez que resultavam num grande número de falsos positivos.

Uma vez todos os elementos identificados, procedeu-se á sua classificação. Foi identificado o número de genes mínimo que cada sistema T4SS teria de possuir para se poder considerar completo. Um elemento com o sistema T4SS completo, MOB,VirB4 e T4CP foi designado de ICE. Elementos sem MOB mas com um sistema T4SS completos foram designados T4SS, isto é, são sistemas exclusivamente para secreção de proteínas, pois sem MOB não podem ser mobilizados. Elementos incompletos mas possivelmente mobilizáveis, ou seja, que apresentavam MOB, foram classificados como pseudo-ICE (pICE), pois parecem corresponder a ICE em processo de pseudogenização. Elementos incompletos que não apresentavam MOB foram classificados como pseudo-T4SS – não são mobilizáveis e não são capazes de proceder a secreção de proteínas.

Resultados - sumário

Com este trabalho chegámos a duas conclusões principais: Em *Proteobacteria* verificámos uma distribuição de ICES e T4SS dependente do tamanho do genoma. Genomas médios (3-5Mb) e grandes (>5Mb) apresentam sobretudo ICES, e genomas pequenos (<3Mb) apresentam sobretudo T4SS. Este facto pode ser explicado se tivermos em consideração que este grupo de organismos inclui bactérias endossimbiontes, que utilizam os sistemas de secreção de proteínas para mediar a interacção com os seus hospedeiros. Contudo verificamos que em organismos com genomas pequenos também se encontra, contrariamente ao que intuitivamente se esperaria, um elevado número de ICEs em processo de pseudogenização.

Uma das teorias mais aceites é que os ICES seriam os principais responsáveis pela transmissão horizontal de genes em *Firmicutes*, e que em *Proteobacteria* esse papel seria desempenhado por plasmídios conjugativos. Com este trabalho verificamos que de facto em *Firmicutes* o numero de ICES é marcadamente mais elevado que o número de plasmídios conjugativos. Contrariamente ao esperado, também em *Proteobacteria*, apesar de a diferença não ser tão acentuada, foi identificado um maior numero de ICES que de plasmídios conjugativos. Deste modo, podemos agora afirmar que, em *Proteobacteria*, os ICE parecem desempenhar um papel pelo menos tão importante como os plasmídios conjugativos em termos de transferência horizontal de genes, ao contrário do que se supunha até aqui.

RESUMO

Elementos integrativos e conjugativos (ICEs) são um grupo muito diverso de elementos genéticos móveis, que se caracterizam por partilharem características de fagos e plasmídios. Como fagos temperados, os ICES integram-se no genoma do hospedeiro, estando dependentes deste para a sua replicação; como plasmídios, são transmitidos para outras células através de conjugação. Estes elementos são portanto responsáveis por transferência horizontal de genes (HGT) em procariotas.

A sua estrutura é composta por três módulos: manutenção, transmissão e regulação. O módulo de manutenção codifica a recombinase, a proteína responsável pela integração dos ICES no genoma do hospedeiro. O módulo de disseminação inclui o sistema conjugativo, tipicamente um sistema de secreção do tipo IV (T4SS). O módulo de regulação é composto por genes que regulam a transferência dos elementos. No entanto, apesar do estudo dos ICE se revestir de enorme importância clínica, uma vez que transmitem características como resistência a antibióticos, produção de factores de virulência ou mesmo produção de biofilmes, pouco se sabe ainda acerca do seu conteúdo génico, tamanho, e que mecanismos de integração e conjugação utilizam.

Este é o primeiro estudo que identifica ICEs em todos os genomas procarióticos sequenciados. Nos 1055 genomas disponíveis, identificámos e caracterizámos a distribuição de 315 ICEs.

Uma das teorias mais aceites acerca do papel dos ICEs na THG especula que estes terão um papel dominante em *Firmicutes*, mas que em *Proteobactérias* são plasmídios conjugativos os verdadeiros responsáveis pela transferência horizontal de genes. Utilizando dados de um estudo prévio do nosso laboratório que caracterizou a mobilidade em plasmídios, verificámos que esta relação não parece ser verdadeira.

Uma vez que este se trata de um estudo pioneiro, os resultados por nós obtidos podem abrir novas portas na investigação de ICEs.

Palavras-Chave: Transferência Horizontal de Genes (HGT); Elementos Integrativos e Conjugativos (ICE); Procariotas; Genómica Comparativa; Sistemas de Secreção Tipo IV (T4SS).

ABSTRACT

Integrative conjugative elements (ICEs) are a diverse group of mobile genetic elements characterized by their dual phage and plasmid behaviour. Like temperate phages, ICEs can integrate into the host chromosome and replicate with it, and like plasmids they are transferred by conjugation. These elements contribute to horizontal gene transfer (HGT) in prokaryotes, and are responsible for the transmission of traits such as antibiotic resistance, virulence factors and biofilm formation. Its core structure can be divided in three modules: maintenance, dissemination and regulation. The maintenance module encodes a recombinase, which is responsible for ICEs integration into host replicons. The dissemination module includes the conjugating system, typically IV secretion system (T4SS). The regulation module comprises the genes that regulate ICEs transfer. The studies on ICEs are very recent and therefore the knowledge about their cargo content, their size and how they conjugate and integrate into the host genome is still reduced. Therefore, studying these elements is of vital importance.

This is the first large-scale study that identifies integrative conjugative elements in all the sequenced prokaryotic genomes. In the 1055 available genomes, we identified and characterized the distribution of 315 ICEs. We were also able to identify T4SS systems not involved in conjugation, and their distribution was compared to those of ICEs. We used data from a previous work of our laboratory, which characterised plasmid mobility, in order to compare the T4SS systems involved in the conjugation of ICEs and conjugative plasmids. We were able to contradict the mainstream idea of ICE being the major contributors to HGT in *Firmicutes*, whereas that role was played by conjugative plasmids in *Proteobacteria*. Because this is a pioneer study, the obtained results may open new avenues of reasearch in this field.

Key-words: Horizontal Gene Transfer (HGT); Integrative Conjugative Elements (ICE); Prokaryots; Comparative Genomics; Type-IV secretion system (T4SS);

INTRODUCTION

It is now widely accepted that horizontal gene transfer (HGT) has deeply shaped the evolution of prokaryotes. The mechanisms of HGT are transformation, phage transduction and conjugation. The latter is thought to play the major role in HGT, mostly due to both plasmids and integrative conjugative elements (ICEs). Study of these elements is of vital importance since their hosts are known to become resistant to antibiotics and heavy metals (Waldor, Tschape et al. 1996; Rice 1998; Boltner, MacMahon et al. 2002; Whittle, Shoemaker et al. 2002; Davies, Shera et al. 2009), to synthesize antimicrobial products (Burrus and Waldor 2004) or to degrade aromatic compounds (Ravatn, Studer et al. 1998). More complex characteristics were also reported, e.g. the colonization of new hosts (Sullivan J.T. and Ronson C.W. 1998), virulence and biofilm formation or nitrogen fixation (Drenkard and Ausubel 2002; He, Baldini et al. 2004; Davies, Shera et al. 2009). Especially because they are responsible for the antibiotic resistance propagation, investigation of ICE is of great clinical importance (Hochhut, Lotfi et al. 2001; Mohd-Zain, Turner et al. 2004). ICE capacity of antibiotic resistance propagation, makes them an important target for clinical investigation.

The focus of the current project is on integrative conjugative elements, a diverse group including all integrative and conjugative self-transmissible elements, independently of the mechanisms by which integration and conjugation occurs (Burrus and Waldor 2004). They encode not only the machinery for excision and conjugation, but also complex regulatory systems to control these processes. ICE integrates into the host chromosome and replicates with it, and when excision is induced they circularize, replicate and are transmitted by conjugation to a recipient cell. The result of this process is the insertion of one copy of the element into the new host chromosome, while the other copy which remains in the donor cell can again be reintegrated. ICEs are characterized by their transposon, phage and plasmid like features. Similar to transposons, they integrate into the chromosome and excise from it, differently transposons are not transferred from one cell to another. Like temperate phages, they integrate into the host chromosome and replicate with it, but phages are not transmitted by conjugation. In common with plasmids they are transferred by conjugation, although ICE are dependent on the chromosome to replicate and are not kept in the circular form.

ICEs are present in all major divisions of bacteria and include, for example, elements classified as conjugative transposons (normally require minimal sequence specificity), such as Tn*916* (Lu and Churchward 1995), and mobile pathogenicity islands, such as ICE*clc*B13

(Ravatn, Studer et al. 1998; Burrus, Pavlovic et al. 2002; Burrus and Waldor 2004). Contrary to plasmids, which were discovered in the early days of molecular genetics and studied ever since, the study of ICE is comparatively recent, and consequently there is a large gap in knowledge regarding them. Some fundamental questions are still to be answered, such as: the number of systems existent in genomes, their size, their gene content and their processes of conjugation and integration. Despite their importance, there are very few studies on comparative and evolutionary genomics of ICE.

The core structure of ICE consists of three modules for maintenance, dissemination and regulation (Toussaint and Merlin 2002). Apart from these essential functions, ICEs often contain a large number of unrelated genes conferring adaptive changes in bacterial genome repertoires, as mentioned above.

The maintenance module encodes the proteins responsible for integration and excision of the ICE into host replicons, such as chromosomes or plasmids. The integration is mediated by an integrase, which is necessary and sufficient for this process to occur. This protein is also responsible for the excision of the element from the chromosome, but in most cases requires the presence of other factors. Tyrosine recombinase family is the most widely described recombinase family of ICE, and its prototypical recombinase is the λ phage integrase. This protein recognizes identical or highly similar sequences both in the host chromosome (the *attB* sites) and the phage (the *attP* sites), promoting site-specific recombination without deletions or sequence duplications (Kikuchi and Nash 1979). Several integrases from ICE, such as proteins of the SXT-R391 family, use a mechanism similar to λ phage (Beaber, Hochhut et al. 2002), and promote the integration into the 3' end of transfer RNAs (tRNAs). In most of the described cases, integration occurs only in one particular locus, even though the bacteria possess multiple alleles of the same tRNA. However, exceptions are known such as the integrase of ICE*clc*B13 (Gaillard, Vallaeys et al. 2006), which does not depend completly on the typical *attB* sequence (Burrus and Waldor 2003; Lee, Auchtung et al. 2007). The tyrosine recombinase family also includes proteins with a different origin from the λ phage integrase (Rajeev, Malanowska et al. 2009), such as the integrase of the Tn*916*. This integrase presents less sequence specificity, integrating for example in AT-rich or bent sequences (Lu and Churchward 1995). There are however proteins responsible for integration of ICE which do not belong to the tyrosine recombinases family. This is the case for the proteins encoded by Tn*GBS2*, a DDE-type transposase (Brochet, Da Cunha et al. 2009), and by Tn*5397*, a serine recombinase (Wang and Mullany 2000).

As mentioned above, the integrase is necessary but not always sufficient for the excision process, which is required to create the circular extrachromosomal form of the ICE that is transferred to another host. For excision to occur, the presence of recombination directionality factors (RDF) if often required. RDFs are small DNA-binding proteins which bias the action of integrase towards excision rather than integration by influencing the formation of specific protein-DNA architectures (Lewis and Hatfull 2001). The ICE excision may also be influenced by environmental factors, as shown for ICE*clc*B13, whose excision increases in stationary phase (Ravatn, Studer et al. 1998).

If the host cell undergoes replication after ICE excision, the element can be lost. Therefore, some ICEs also encode factors that prevent their own loss from the chromosome. One such example is a homolog of Soj, a protein implicated in plasmid maintenance, present in the ICE PAPI-1. Wild type PAPI-1 is lost in 0,16% of the cells, whereas all host cells lose the element in the absence of the Soj homolog (Klockgether, Reva et al. 2004; Qiu, Gurkar et al. 2006). Although the mechanism is not yet fully understood, since this protein is only expressed after excision it has been proposed that its role is to stabilize the extrachromosomal form of the ICE.

The dissemination module encodes the proteins responsible for the DNA processing after excision and for the transference of the element copy. Most models of DNA processing in ICE are derived from those of plasmids, in which the conjugative DNA processing starts with relaxase, a protein responsible for the cleavage of the DNA at the origin of transfer, initiating of the rolling circle replication. The relaxase remains attached to the single-stranded DNA (ssDNA), and the resulting nucleoprotein complex is transported to the recipient cell via the mating pore (Llosa, Gomis-Rüth et al. 2002). Since the copy number of the element in the donor cell does not increase, ICE are regarded as not truly replicative. Even though ICE are thought to transfer as ssDNA there are exceptions: as for plasmids, some ICEs from *Actinobacteria* are transferred as double-stranded DNA (dsDNA) by a different mobilization mechanism active in the mycelia (Grohmann, Muth et al. 2003).

The conjugation system of ICE is typically a type IV secretion system (T4SS) (Cascales and Christie 2003). The subfamily of T4SS responsible for DNA transfer during bacterial conjugation is known by Mpf/CP (Mating pair formation/Coupling Protein or VirD4). The prototypical Mpf (from the Ti plasmid) consists of eleven conserved proteins, VirB1-VirB11, which form the membrane-spanning complex and the surface pilus that establish contact with the recipient bacteria. (Schröder and Lanka 2005). VirD4 is a NTP-binding protein that probably plays two roles in the conjugation: initially, it is the first component of the secretion machinery that comes into contact with the nucleoprotein complex (Cascales and Christie

2004), and secondly it couples it with the secretion pore formed by the Mfp system (Schröder and Lanka 2005), where it is thought to help to energize the secretion machinery (Schroder, Krause et al. 2002). The only protein of this complex that was found to be ubiquitous in conjugative systems of both Gram-negative and Gram-positive bacteria is VirB4 (Juhas, Crook et al. 2008), even though the surface proteins produced by the latter work as non-specific adhesins instead of forming a pilus. VirB4 is an inner membrane protein that energizes the secretion machinery (Dang, Zhou et al. 1999; Schröder and Lanka 2005).
There are four major types of T4SS, three of them were identified based on the incompatibility group of the representative conjugative plasmids: IncF (plasmid F), IncP (plasmid RP4) and IncI (plasmid R64) (Lawley, Klimke et al. 2003). The fourth type of T4SS, ICEHin1056, was recently identified in genomic islands (Juhas, Crook et al. 2007). These systems will be referred to as T4SS-F, T4SS-T, T4SS-I and T4SS-G, respectively.

The regulation modules comprise the genes that regulate ICE transfer. Although little is known about their activity, studies have revealed induction of conjugation in the presence of tetracycline or the activation of the SOS response by DNA damaging agents (Stevens, Shoemaker et al. 1990; Beaber, Hochhut et al. 2004).

Although all ICEs have a common backbone, their structure is plastic, as the modules and the proteins they encode may be very different. They are also responsible for genome plasticity, because the same integration sites are shared by related ICEs. Since they may have a broad host range, such sites increase the variability within both bacterial species and genera, increasing intra-species and inter-genus locus variability. On the other hand, they often contain genes and sequences, such as the above-mentioned recombinases and insertion sequences, which facilitate the recruitment of other genes to the ICE backbone. If this integration occurs in a specific locus, a cluster of horizontally transferred genes may be conserved and transmitted between bacteria (Burrus and Waldor 2004).

The major goal of this project is to quantify and characterize the distribution of ICE in the 1055 prokaryotic genomes available. In order to identify the presence of ICEs, we will search for the key elements of the conjugation machinery, the T4SS system and the relaxosome. Centering our attention on conjugation is reasonable because within all integrated mobile elements in genomes, such as prophages, and prophage-like elements, the presence of a conjugative apparatus is the very defining feature of ICE.

Since we have data from previous studies of our laboratory regarding conjugative plasmids (Smillie C., Garcillian M. et al.), and because the chosen approach allowed us to distinguish between complete and non complete elements, we are able to ask relevant questions such as: are there differences in the secretion systems used by ICE and conjugative plasmids? Is there really, as hypothesized, a predominant role of ICE for horizontal gene transfer in *Firmicutes*, whereas in *Proteobacteria* this function is essentially performed by conjugative plasmids?

This study is also a first step towards understanding the secretion systems present in symbionts – is it possible that they derive from ICE?

The number of sequenced genomes available is exponentially increasing, mainly due to the next-generation sequencing techniques. But along with the creation of data, new methods for its analysis must also be developed. The informatics tools available nowadays to treat biological data may be the key to its efficient integration, and allow the formulation of new questions. In this project we performed comparative genomics analysis, i.e., we used well characterized proteins, which we knew that could allow us to discover the ICEs, as templates to search for their homologs across entire genomes.

This is the first large-scale study that identifies integrative conjugative elements in all the available prokaryotic genomes, and the obtained results may therefore open new avenues of reasearch in this field.

DATA AND METHODS

Main objective

As described in the Introduction, we identified the presence of ICEs by searching the T4SS system and the relaxosome. According to previous literature mentioned about, we consider that conjugation involves the Mpf system and the transfer of ssDNA, which is brought to the complex by the coupling protein or VirD4. Some systems found in *Actinobacteria* transfer dsDNA in micelia by a different mechanism using an FtsK-like system and will not be considered here.

A previous study from the laboratory (Smillie C., Garcillian M. et al.), in plasmids of *Proteobacteria*, showed that in 98% of the cases when VirB4 in found, it corresponds to the presence of the entire complex. Therefore, to identify the Mpf/CP we will focus on both VirB4 and VirD4.

In *Proteobacteria*, because of the amount of information available about the proteins that constitute the T4SS systems, it was possible to search not only for VirB4 but also for the other proteins that are specific of the different T4SS systems.

To complete the analyses we searched for the proteins responsible for the initialization of the conjugative process: relaxases or MOBs (from mobilization). They are responsible for the initial cleavage of the DNA and then remain attached to it. These form the nucleoprotein complex that is transported to the recipient cell via T4SS system. The relaxases are classified in six families: MOBc, MOBf, MOBh, MOBphen, MOBq and MOBv.

In addition, in *Firmicutes* and in *Bacteroidetes*, two specific MOBs of this clades were also searched: ORF20 (YP_133675.1), from *Enterococcus faecalis*, and mobilization protein B (NP_818960.1), from *Bacteroides thetaiotaomicron VPI-5482* (Xu, Bjursell et al. 2003),(Flannagan, Zitzow et al. 1994). The protein designations that start with an YC or NC prefixes is in fact a RefSeq accession number, i.e., an unique identifier that classifies a molecule (in this case a protein) in the NCBI database. Both prefixes indicate that these molecules are proteins, and that they result from both automated processing and expert curation. For the proteins with the prefix YC a corresponding transcript record was provided.

Data

Due to the previous study of our laboratory that determined plasmid mobility, we obtained a data set of plasmidic VirB4, T4CP, MOBs and, for *Proteobacteria*, specific T4SS system genes, that we could use to effectuate the search for ICE in the genomes.

There are four prototypical systems previously described (Cascales and Christie 2003): Plasmid F (NC_002483) for T4SS type F (T4SS-F), Plasmid Ti (NC_002377) for T4SS type T (T4SS-T), Plasmid R64 (NC_005014) for T4SS type I (T4SS-I) and ICEHin1056

(NC_008739) for T4SS type G (T4SS-G). These protein identifiers are also RefSeq accession numbers, and the NC prefix stands for complete genomic molecules that result from both automated processing and expert curation.

The prototypical systems above mentioned were in fact described and classified based on the incompatibility type of the plasmids. It is important to refer that Mating pair formation and incompatibility type are two different and independent concepts. The incompatibility type is a propriety attributed to plasmids (both conjugative and non-conjugative), often related to each other, that are unable to stably coexist in the same cell. Even though the Mpf complexes were identified in these plasmids and its classification derives from there (as F, T and I type), the Mpf is a type IV secretion system implicated exclusively in conjugation.

The specific genes from the different T4SS systems that we searched for are, respectively: TraN, TraU, traE, traH, traK, traL, traV, traW and trbC for T4SS-F system; virB3, virB6, virB8 and virB9 for T4SS-T system; traI, traK, traL, traM, traN, traP, traQ, traR, traW and traY for T4SS-I system; for T4SS-G system, ICEHIN1056_000310, ICEHIN1056_000410, ICEHIN1056_000440, ICEHIN1056_000510 and ICEHIN1056_000520.

It is important to refer, however, that in the T4SS-I system the role of VirB4 is played by another ATPase, TraU, with very low sequence similarity with VirB4.

The sequences with more than 95% of similarity were removed to facilitate the analysis, using a end-gap free global alignment BALI.

We search for these proteins in the 1055 genomes available at the NCBI database when we started the work. Figure 1 shows the clade distribution of these genomes.

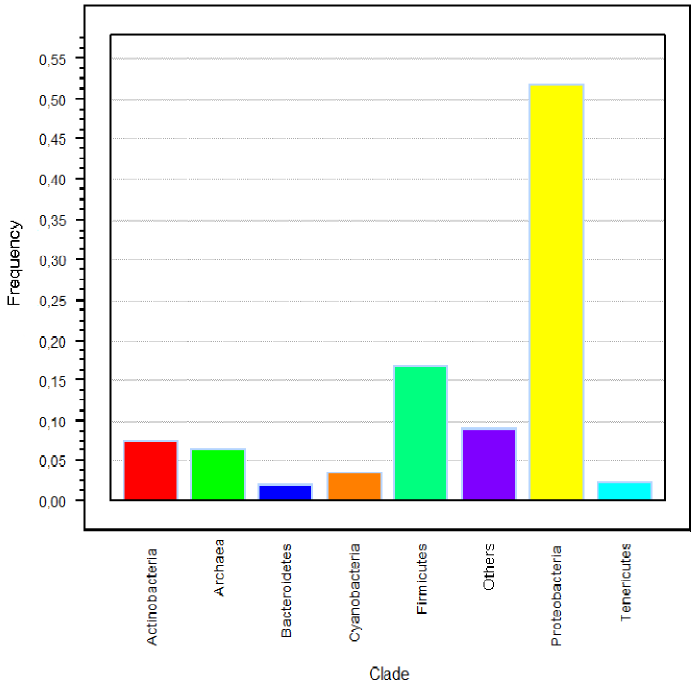


Figure 1. Frequency of organisms from each clade in our 1055 genomes database.

Methodological essays

Based on the previous study of our laboratory, which inferred mobilization and conjugation of plasmids, we initially tried to find ICE using PSI-BLAST (Position specific iterative BLAST) (Altschul, Madden et al. 1997) to search for MOBs, and BLASTP (Basic Local Alignment Search Tool for protein) (Altschul, Gish et al. 1990) to search for both VirB4 and T4CP. Since PSI-BLAST, due to its matrix based proprieties, searches for distant relatives, this was the most powerful method in that it is able to identify very weak similarities. However, it is also the more error prone, for it can create many false positives. If in the end of an iteration several of the distant relatives of the protein of interest are found and included in the matrix, the next round of iteration will obtain even more evolutionarily distant proteins, and so forth. Therefore, the ideal situation is that the results converge, i.e., that in one given round of the iteration only proteins that had already been found are retrieved, terminating the search.

As an initial approach, we effectuated PSI-BLAST for each MOB family in all the genomes available. Several thousands of proteins were found by this methodology, but only the MOBc family converged. It was not possible to use these results since there was not a clear way to decide, for each family of MOBs, how many iterations to accept. As a result this methodology was abandoned, and new methods were tried.

We therefore developed another approach to retrieve fewer false positives. The first attempt was to use BLASTP, since this was the methodology proposed to search the other proteins. The use of a single well-known method would simplify the whole subsequent analysis. Because this method was not used in the previous study on plasmid data, we had to first validate it. Plasmids were used as a control study, and a BLASTP that allowed a maximum e-value of 0.1 was performed using all the protein of the different families of MOBs with less than 95% of similarity. The objective was, since we knew already which proteins to find, to define for each family a way to distinguish the true from the false positives. Since we had several proteins from the same family, the idea was to count the number of times a protein in the family was found in all the BLASTPs with the plasmid data. Taking MOBf as an example, we performed BLAST only with those proteins with less than 95% of polipeptide sequence similarity. By doing this we exclude proteins that are highly similar and that would likely simplify the analysis. From the 155 MOBf proteins identified in plasmids only 76 have less than 95% of similarity, and therefore 76 BLASTP were performed. If, say, all the true positives were found in at least 60 of these BLASTs and the false positives always found less times, then we could use 60 as a cut-off and, when using this methodology in the genomes, admit that all the proteins found at least 60 times were true MOBf, and the ones found in less BLASTs were false positives. Figure 2 shows an example of the search results obtained in plasmids, where we have enough data to distinguish between true and false positives. Even though all 155 real MOBf were found, we also retrieved other 161 proteins, some of them MOBs from different families – more than 50% are false positives. We then tried to define a threshold to separate these two types of proteins.

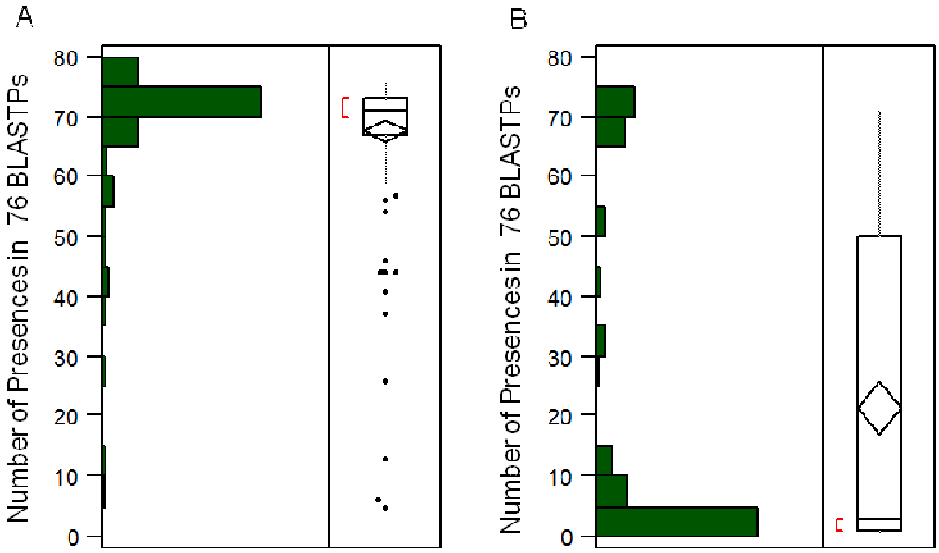


Figure 2. BlastP results of MOBf – number of times a protein is found in 76 performed BlastPs. Distribution of the 155 true positives (A) and 161 false positives (B) obtained, according to the number of times they were retrieved by the 76 independent BlastP effectuated using plasmidic MOBf as queries.

As we can see in figure 2A, the true positives are mainly found in at least 65 of the 76 BLASTPs effectuated; on the other hand, the false positives (Figure 2B) are mainly found in less than 15 BLASTPs. If we define a cut-off of 65, we apparently obtain a good separation. However, from the 155 MOBf we know that exist in plasmids, 23 are above this number, meaning that almost 15% of the true MOBf would be missed in our analyses. As for the false positives, 28 of these proteins are found by at least 65 BLASTPs, meaning that 17% of all the proteins wrongly retrieved would be included in our results.

Even though this was not a completely satisfactory result we decided to perform the BLASTP analysis in the genomes not only for the MOB families but also for VirB4 and T4CP, always using the proteins with less than 95% of similarity, and this information was kept as a backup. We did this because in our attempts to minimize the number of false positives, some potentially true positives may as well be lost. When it is possible incomplete elements are retrieved by other methodologies, and when in its periphery there is a protein identified by BLASTP with a good e-value, then we could consider this protein as a true positive and include it in the element.

Adopted Methodology

The program we tested next was HMMER (Durbin, Eddy et al. 1999), that analyzes sequences using profile hidden Markov models – profile HMMs.

Profile HMMs are statistical models of multiple sequence alignments. This profiles capture position-specific information about how conserved each column of the alignment is, and which residues are likely to occur in a certain position. The multiple alignments of the known proteins from plasmids necessary to create the HMM profiles were obtained with MUSCLE (Edgar 2004), and for each type of protein that we searched for – T4SS specific proteins, MOBs, VirB4 and T4CP – all the plasmidic proteins with less than 95% identity were used to create the HMM profile. The search of these profiles in the database was then performed as a "glocal" alignment, i.e., global with respect to the profile, so that we know that all the protein must align, but local with respect to the sequence.

In the control tests with plasmids, allowing e-values up to 0.1, this method resulted in really few or even none false positives. When used in the complete genomes, the proteins obtained were in the expected order of magnitude, and hence we decided to adopt this methodology.

The utilization of HMMER had not been considered before because this is a slow methodology. The new version of HMMER is faster (that was not available at the time of the analyses), but does not perform "glocal" alignments.

Identification and Characterization of Elements

When all the proteins of interest were localized in the genomes, we started the identification of the possible elements. The first step was to verify if the T4SS specific genes were near each other, forming possibly functional T4SS systems. In order to do this, an awk scrip clustered the proteins localized in the genome up to 25 genes apart. The first step was to list all the specific genes according to the position they occupy in the genome. For each identified gene, if its distance to the previous gene in the list was 25 or less positions, these two genes were clustered together. Since in organisms such as the ones from genus *Rickettsia* the conjugation systems are known to be scattered throughout the genomes (Weinert, Welch et al. 2009), a manual curation was required to join the different clusters formed automatically in order to create the complete one. On the other hand, some clusters that were not complete and presented the complementary genes within more than 25 positions were also merged together.

Once concluded the identification of the T4SS systems, we searched within 25 genes upstream and downstream of the clusters for VirB4, T4CP and MOBs. Also these results were manually curated in order to include proteins that, if not within 25 genes in the genome, were close enough to be considered part of the element. It is important to remember that the T4SS systems were described in detail only in *Proteobacteria*, and therefore the specific genes were searched only in this clade. In the other organisms the elements were classified according to the presence or absence of only VirB4, T4CP and MOB.

We decided to complement this analysis with the results from the less restrictive methodology, BLASTP. For this, we searched for proteins located in the genome up to 25 positions apart and manual curation was performed as described above. This allowed us to create more complete elements, because we know that some proteins may not be retrieved with the previous methodology, which is more conservative.

A global view of our results, however, made us realise that some of the elements with apparently functional T4SS systems were incomplete, lacking for example a T4CP or a MOB. This could indicate that a pseudogenization process was occurring, and therefore the element was no longer functional. In order to understand if this assumption was true, we tried to identify possible pseudogenes in the vicinity using TBLASTN (Altschul, Gish et al. 1990). This program uses BLASTP to compare a protein sequence against a database of nucleotide sequences translated in all six reading frames. The database used to search for pseudogenes of the different proteins was created using the nucleotidic sequences that covered 50 Kb upstream and downstream of the elements lacking that given protein. The fact that we found pseudogenes with this method does not influence the classification of the elements, since are likely to code for non-functional proteins; it only helps to consolidate the idea that element was indeed an ICE. There is also the possibility that these pseudogenes are in fact the product of sequencing errors, but it was beyond the scope of this work to re-sequence such loci.

In first classed the T4SS systems of the *Proteobacteria*. For each of the four systems we had several elements with all or near all the proteins we searched for, and elements with very few of those proteins. This clear bimodal distribution of our hits suggested that there was a minimum number of proteins required to the system to be functional, so that if some genes were lost the system would no longer be functional and the other genes would be rapidly lost as well, leading to us finding only near complete or really degraded systems. Also, isolated genes might simply be false positives. We identified five genes that seem to be present in the vast majority of known T4SS-F and absent in the other systems, four for the T4SS-G, also four for the T4SS-I and three for the T4SS-T.

Table 1. Element classification in *Proteobacteria*.

MOB	Complete T4SS	VirB4/TraU	T4CP	Classification
+	Yes	+	+	ICE
+	No	+/-	+/-	pICE
-	Yes	+	+/-	T4SS
-	No	+/-	+/-	pT4SS

After the classification of the T4SS we evaluated the presence or absence of a MOB. If an element has a MOB it can be mobilized, and can be or have been an ICE. In other words, if an element presents a complete T4SS, a MOB, a T4CP and a VirB4 is considered an ICE; if it has a MOB and some but not all of the other components, we consider it a pseudo-ICE (pICE), since it presents the relaxase and, even if not complete, a conjugation system. If there is no MOB in an element, then it can only function as a protein secretion system, a T4SS or, if not complete, a pseudo-T4SS (pT4SS). This classification is summarized, for the *Proteobacteria* clade, in Table 1.

For organisms other than *Proteobacteria*, where we have less information, the classification is based only in the presence or absence of VirB4, T4CP and MOB. If all the three were present, the element is considered an ICE; if one of these proteins is absent, the element is a pICE. In this analysis, *Archaea* are the exception, since no MOB is known in these organisms. Therefore, if in an element we did not find a MOB this fact is not enough for us to state that there is none, and we consider it to be an ICE-A (ICE from *Archaea*).

All the programming was performed in UNIX, and the statistical analyses with JMP.

RESULTS AND DISCUSSION

Complete and Non Complete Elements

Using the classification defined in *Data and Methods* we found 652 elements: 315 ICE, from which 198 in *Proteobacteria*, 243 pICE, from which 114 in *Proteobacteria*, and 67 T4SS and 27 pT4SS, only in *Proteobateria* since these organisms are the only ones were we can search for the T4SS specific genes. The clade *Proteobacteria* includes more than half of the elements we found but it is important to keep in mind that, as shown in Figure 1, this clade represents more than 50% of the available genomes. These numbers are, therefore, not enough *per se* to take any conclusions regarding higher frequency of ICE in *Proteobacteria*.

Comparison of ICE and T4SS in *Proteobacteria*

Biased distribution of ICE and T4SS

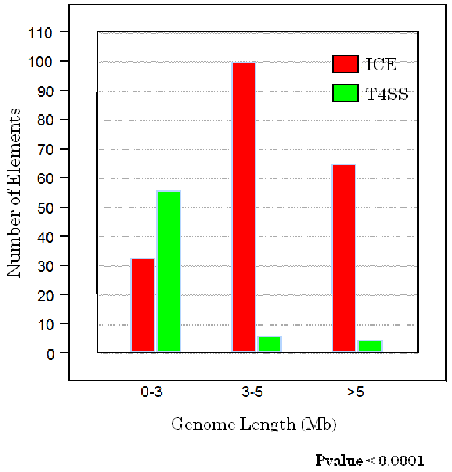


Figure 3. Number of ICE and T4SS present in each genome length category. There are, in *Proteobacteria*, 220 genomes with less than 3 Mb (small genomes), 229 genomes with 3 to 5 Mb (medium genomes), and 98 genomes with more than 5 Mb (large genomes). The distribution of the number of elements is dependent from the length of the genome (pvalue of qui-square test is <0.0001).

The first question we tried to answer was if there were differences in the distribution of ICE and T4SS in *Proteobaceria*. The approach was to divide the genomes into three categories - small (less than 3 Mb), medium (between 3 and 5 Mb) and large (more than 5 Mb) - and compare the number of ICE and T4SS between them. As shown in Figure 3 we observed a biased distribution, with clear predominance of ICE in medium and large genomes, and the inverse in small genomes, where not only there are significantly less ICE than in the other categories, but also more T4SS. As a result, in the small genomes there is a predominance of T4SS, contrary to the remaining genomes. A qui-square test confirms that the differences are statistically significant, and that the distribution of the elements is indeed dependent from the size of the genomes.

This result was somewhat expected. Organisms with smaller genomes tend to require symbiotic relationships and endure little horizontal transfer. Such organisms use T4SS systems in their interactions with eukaryotic hosts, by exporting effectors to the eukaryotic cytoplasm. On the other hand, reduced horizontal transfer leads to fewer ICE.

Are T4SS derived from ICE?

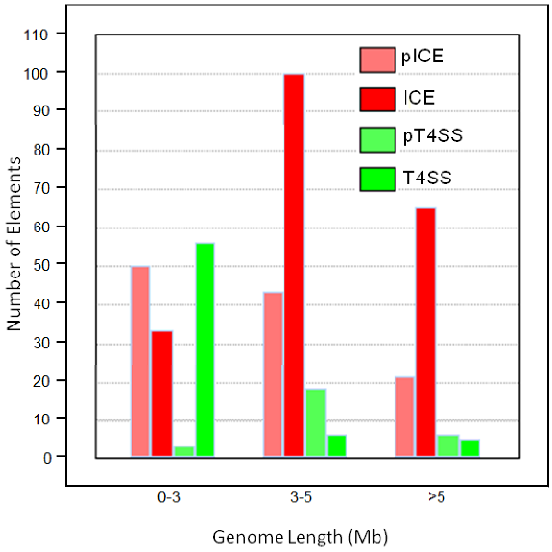


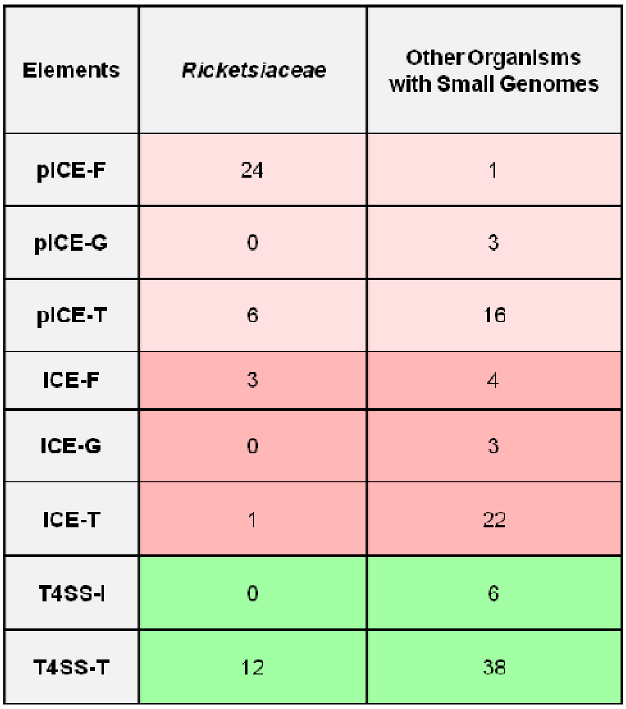
Figure 4. Number of complete and non complete elements in each genome length category.

We have shown a biased distribution of the T4SS systems and ICE. In larger genomes, the constraints of carrying non functional elements are lower, so it is not unexpected to see pICE and pT4SS in such organisms. In smaller genomes, however, the constraints are much higher, and as we can see in Figure 4, the number of pT4SS is really reduced. The number of pICE, however, is the highest of the three length categories. Unlike medium and large genomes, in the smaller genomes a T4SS>pICE>ICE relation is observed. This observation leaded us to the question: is it possible that T4SS are retained in these organisms, after the remaining ICE is degraded?

We therefore focused our attention in the organisms that seems to unbalance this equilibrium – the symbionts.

Since we do not have habitat information for all the genomes, we performed this first analysis with a rough selection of animal endosymbionts, the family *Ricketsiaceae.* Table 2 presents a more detailed list of elements found in these organisms (15 genomes) and in the other organisms with small genomes (205 organisms).

Table 2. Elements Present in Organisms with Small Genomes



Elements	*Ricketsiaceae*	Other Organisms with Small Genomes
pICE-F	24	1
pICE-G	0	3
pICE-T	6	16
ICE-F	3	4
ICE-G	0	3
ICE-T	1	22
T4SS-I	0	6
T4SS-T	12	38

As we can see in Table 2, 24 of the 25 pICE-F present in organisms with small genomes are in fact in the 15 genomes of the family *Ricketsiaceae*. At a closer look to these elements, we realized that these pICE-F were highly similar between them, with the same orientation and the same genes missing. This result may point to functionality, being either an ICE or a complete T4SS. If ICE, this would be the first F-type ICE described in these organisms, and the reason why this particular type is retained could be investigated. This option, however, contradicts the intuitive thought of organisms with small genomes having preferentially less ICE, since they have really restrict interaction with their hosts and not with other bacteria. If T4SS, these would be the firsts T4SS-F systems ever described that do not play a role in conjugation. With the data available, however, we cannot yet decide which of the hypotheses is correct, and we keep the classification as pICE.

We can observe that, excluding the 25 pICE-F that appear to a biological role rather than being pseudogenized ICE, we observe the relation T4SS > pICE > ICE, with respectively 12, 6 and 4 elements, whereas in the remaining organisms with small genomes the ICEs are the second most abundant elements.

In *Ricketsiales*, even though the horizontal gene transfer is really reduced, the genes of the T4SS system are proven not to be result of vertical transference (Weinert, Welch et al. 2009). Given this discovery and the relation T4SS>pICE>ICE that we observed in *Ricketsiaceae*, our theory of transition from ICE to T4SS seems plausible.

This study may however be improved by using other endosymbionts. Therefore, more data is needed regarding the habitat and the bacteria-host interactions.

Comparison of ICE and Conjugative Plasmids

Predominant Role: ICE in *Firmicutes*, Conjugative Plasmids in *Proteobacteria*?

One idea often conveyed in the literature, but not yet statistically tested, is that horizontal transfer is most frequently caused by conjugative plasmids in *Proteobacteria*, and by ICE in *Firmicutes.* Since we have data from both conjugative plasmids (previous work of our laboratory) and ICE in both clades, we can test this hypothesis.

In order to make this analysis we need to be sure that the data-sets are comparable, because if in one of the clades the frequency of plasmids is lower, this fact could alone explain the presence of less conjugative plasmids in that clade.

Several plasmids were sequenced by its intrinsic biological interest and not with its correspondent genome, and if we used all the sequenced plasmids to make the comparisons with ICE the analysis would be plasmid biased. As an example, there are 957 plasmids of *Proteobacteria* available but only 406 were sequenced with the genomes - these are the plasmids that we are going to include in our analysis. Therefore, we first selected only the plasmids that were sequenced with the genomes.

In *Proteobacteria* we have 547 sequenced genomes and 406 plasmids, and in *Firmicutes* we have 178 genomes and 119 plasmids. Even though the amount of data is significantly different, the frequency of genomes and plasmids in both clades is comparable (pvalue of qui-square test is 0,4396). Therefore, plasmids are equally represented in the two clades and we can proceed with the analysis.

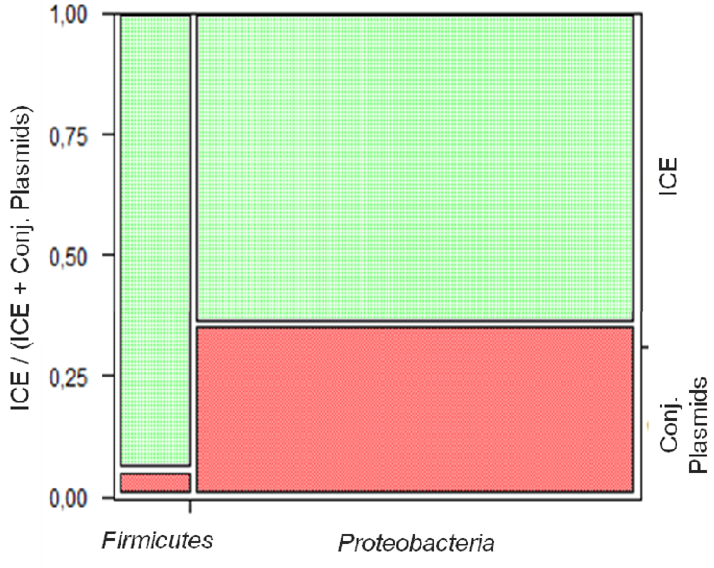


Figure 5. Proportion of ICE and Conjugative Plasmids in *Proteobacteria* and *Firmicutes*. In the y axis is shown the proportion of ICE and conjugative Plasmids according to the rule ICE / (ICE + Conjugative Plasmids): 1 means only ICE, 0 means only Conjugative Plasmids. The x axis represents the amount of data, larger for *Proteobacteria* as expected.

The second step is to verify if in fact there are more ICE than conjugative plasmids in *Firmicutes*. With our study we found 50 ICE in this clade, and only 3 of the 119 plasmids were classed as conjugative. We observe, therefore, 16.6 times more ICE than ICE in *Firmicutes* (proportion shown in Figure 5), in agreement with the hypothesis that we wish to test.

The third step is to understand if *Proteobacteria* have indeed a significantly larger number of conjugative plasmids than ICE. As we can see by Figure 5, the answer to this question is no. In fact, we found 198 ICE in these organisms, and only 110 conjugative plasmids. Therefore our data suggests that ICE are more frequent than conjugative plasmids in both clades, albeit the difference is much more important in *Firmicutes*.

It is important to note, however, that in culture the segregation of plasmids is thought to be higher than that of ICE. So it is possible that, even with the precautions taken, the data is biased because at the moment of sequencing the plasmid has already been lost. In any case, the data suggests that at best ICE and conjugative plasmids have comparable frequency in P*roteobacteria*.

Are there differences in the T4SS systems of Conjugative Plasmids and ICE?

As a final question, since in *Proteobacteria* we were able to distinguish between the different T4SS systems present in both conjugative plasmids and ICE, we can try to understand if there are differences in their distribution.

Indeed, we show that T4SS-T is not only the predominant type in conjugative plasmids and ICE, but is also equally distributed in both elements (Figure 6, pvalue of Fisher's exact test is 0.8071). There is, however, an important difference in the frequencies of T4SS-F and T4SS-G (pvalue of Fisher's exact test is <0.0001 in both cases). In conjugative plasmids T4SS-F is the second more frequent system, a position occupied by T4SS-G in ICE. Since T4SS-G is not really well known from a molecular point of view, is not yet possible to give a biological interpretation to these results.

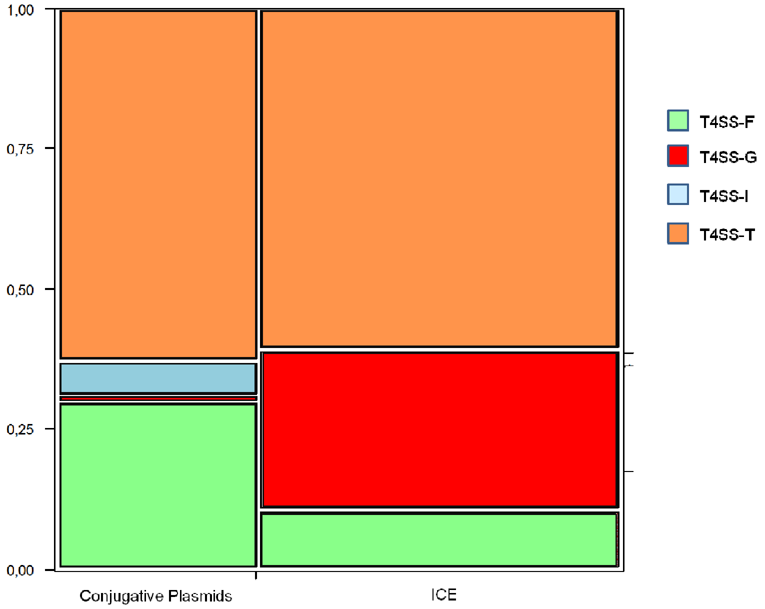


Figure 6. Proportion of the different T4SS types in both conjugative Plasmids and ICE from Proteobacteria. From the 110 Conjugative Plasmids, 69 present a T4SS-T, 33 a T4SS-F, 7 a T4SS-I and 1 a T4SS-G. From the 198 ICE, 120 present a T4SS-T, 57 a T4SS-G, 21 a T4SS-F and there is no Ice with T4SS-I. The graphic shows the proportions, according to the rule ICE / (ICE + Conjugative Plasmids). A Fisher's exact test was performed to compare each T4SS type in both samples. All pvalues were significative (<0.0001) except the one from T4SS-T, which means that this is the only system equally distributed in both ICE and Conjugative Plasmids.

It would be particularly interesting to investigate the cargo content of both ICE and conjugative plasmids with T4SS-T in order to understand if their predominant presence in both kinds of elements is due to mechanisms that prevent loss, or if there is a more specific and yet unknown reason for the predominance of this system.

FUTURE PERSPECTIVES

There are three main studies that can be performed with the ICEs identified with this project.

The first is the delimitation of the ICEs, i.e., exactly where in the genome do the self-transmissible elements begin and end. A program based in syntenic blocks (group of genes found in the same order in different species) is currently being developed in our laboratory. Such an approach could allow not only to define the borders of the ICEs, if the genes surrounding the elements constitute a syntenic block, but also the definition of the ICEs themselves, if homologous genes occupy the same position within the element, constituting one or several syntenic blocks. One possible example would be the identification of conserved modules across different elements. Using well characterized ICEs as a training set, the program can be optimized and used to delimit the elements described in this work. Such an analysis will allow the systematic study of the functions coded in the cargo regions of ICE, which has never been achieved before in a large-scale study. This is possibly the most clinically relevant study to be made with the obtained data.

The second possible study, already being performed in our laboratory, is a phylogenetic analysis using the identified recombinases and T4SS systems. This will allow to frame the evolutionary history of ICE and to test their relative relatedness with phages, plasmids and transposons.

The third study is related with our hypothesis of the T4SS systems in organisms with small genomes, particularly the ones of endosymbionts, being derived from ICEs. It would imply a phylogenetic analysis of the T4SS machinery used to secrete proteins in these organisms and the T4SS systems of ICEs.

REFERENCES

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." Journal of Molecular Biology **215**(3): 403-410.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Beaber, J. W., B. Hochhut, et al. (2002). "Genomic and Functional Analyses of SXT, an Integrating Antibiotic Resistance Gene Transfer Element Derived from Vibrio cholerae." J. Bacteriol **184**(15): 4259-4269.

Beaber, J. W., B. Hochhut, et al. (2004). "SOS response promotes horizontal dissemination of antibiotic resistance genes." Nature **427**(6969): 72-74.

Boltner, D., C. MacMahon, et al. (2002). "R391: a Conjugative Integrating Mosaic Comprised of Phage, Plasmid, and Transposon Elements." J. Bacteriol **184**(18): 5158-5169.

Brochet, M., V. Da Cunha, et al. (2009). "Atypical association of DDE transposition with conjugation specifies a new family of mobile elements." Molecular Microbiology **71**(4): 948-959.

Burrus, V., G. Pavlovic, et al. (2002). "Conjugative transposons: the tip of the iceberg." Molecular Microbiology **46**(3): 601-610.

Burrus, V. and M. K. Waldor (2003). "Control of SXT Integration and Excision." J. Bacteriol **185**(17): 5045-5054.

Burrus, V. and M. K. Waldor (2004). "Shaping bacterial genomes with integrative and conjugative elements." Research in Microbiology **155**(5): 376-386.

Cascales, E. and P. J. Christie (2003). "The versatile bacterial type IV secretion systems." Nat Rev Micro **1**(2): 137-149.

Cascales, E. and P. J. Christie (2004). "Definition of a Bacterial Type IV Secretion Pathway for a DNA Substrate." Science **304**(5674): 1170-1173.

Dang, T. A., X. R. Zhou, et al. (1999). "Dimerization of the Agrobacterium tumefaciens VirB4 ATPase and the effect of ATP-binding cassette mutations on the assembly and function of the T-DNA transporter." Molecular Microbiology **32**(6): 1239-1253.

Davies, M. R., J. Shera, et al. (2009). "A Novel Integrative Conjugative Element Mediates Genetic Transfer from Group G Streptococcus to Other {beta}-Hemolytic Streptococci." J. Bacteriol **191**(7): 2257-2265.

Drenkard, E. and F. M. Ausubel (2002). "Pseudomonas biofilm formation and antibiotic resistance are linked to phenotypic variation." Nature **416**(6882): 740-743.

Durbin, R., S. Eddy, et al. (1999). Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids, {Cambridge University Press}.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Flannagan, S. E., L. A. Zitzow, et al. (1994). "Nucleotide Sequence of the 18-kb Conjugative Transposon Tn916 from Enterococcus faecalis." Plasmid **32**(3): 350-354.

Gaillard, M., T. Vallaeys, et al. (2006). "The clc Element of Pseudomonas sp. Strain B13, a Genomic Island with Various Catabolic Properties." J. Bacteriol **188**(5): 1999-2013.

Grohmann, E., G. Muth, et al. (2003). "Conjugative Plasmid Transfer in Gram-Positive Bacteria." Microbiol. Mol. Biol. **67**(2): 277-301.

He, J., R. L. Baldini, et al. (2004). "The broad host range pathogen Pseudomonas aeruginosa strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes." Proc. Natl. Acad. Sci. USA **101**(8): 2530-2535.

Hochhut, B., Y. Lotfi, et al. (2001). "Molecular Analysis of Antibiotic Resistance Gene Clusters in Vibrio cholerae O139 and O1 SXT Constins." Antimicrob. Agents Chemother. **45**(11): 2991-3000.

Juhas, M., D. W. Crook, et al. (2007). "Novel Type IV Secretion System Involved in Propagation of Genomic Islands." J. Bacteriol **189**(3): 761-771.

Juhas, M., D. W. Crook, et al. (2008). "Type IV secretion systems: tools of bacterial horizontal gene transfer and virulence." Cellular Microbiology **10**(12): 2377-2386.

Kikuchi, Y. and H. A. Nash (1979). "Nicking-closing activity associated with bacteriophage lambda int gene product." Proc. Natl. Acad. Sci. USA **76**(8): 3760-3764.

Klockgether, J., O. Reva, et al. (2004). "Sequence Analysis of the Mobile Genome Island pKLC102 of Pseudomonas aeruginosa C." J. Bacteriol **186**(2): 518-534.

Lawley, T., W. Klimke, et al. (2003). "F factor conjugation is a true type IV secretion system." FEMS Microbiology Letters **224**(1): 1-15.

Lee, C. A., J. M. Auchtung, et al. (2007). "Identification and characterization of int (integrase), xis (excisionase) and chromosomal attachment sites of the integrative and conjugative element ICEBs1 of Bacillus subtilis." Molecular Microbiology **66**(6): 1356-1369.

Lewis, J. A. and G. F. Hatfull (2001). "Control of directionality in integrase-mediated recombination: examination of recombination directionality factors (RDFs) including Xis and Cox proteins." Nucleic Acids Res **29**(11): 2205-2216.

Llosa, M., F. X. Gomis-Rüth, et al. (2002). "Bacterial conjugation: a two-step mechanism for DNA transport." Molecular Microbiology **45**(1): 1-8.

Lu, F. and G. Churchward (1995). "Tn916 target DNA sequences bind the C-terminal domain of integrase protein with different affinities that correlate with transposon insertion frequency." J. Bacteriol **177**(8): 1938-1946.

Mohd-Zain, Z., S. L. Turner, et al. (2004). "Transferable Antibiotic Resistance Elements in Haemophilus influenzae Share a Common Evolutionary Origin with a Diverse Family of Syntenic Genomic Islands." J. Bacteriol **186**(23): 8114-8122.

Qiu, X., A. U. Gurkar, et al. (2006). "Interstrain transfer of the large pathogenicity island (PAPI-1) of Pseudomonas aeruginosa." Proc. Natl. Acad. Sci. USA **103**(52): 19830-19835.

Rajeev, L., K. Malanowska, et al. (2009). "Challenging a Paradigm: the Role of DNA Homology in Tyrosine Recombinase Reactions." Microbiol. Mol. Biol. Rev. **73**(2): 300-309.

Ravatn, R., S. Studer, et al. (1998). "Chromosomal Integration, Tandem Amplification, and Deamplification in Pseudomonas putida F1 of a 105-Kilobase Genetic Element Containing the Chlorocatechol Degradative Genes from Pseudomonas sp. Strain B13." J. Bacteriol **180**(17): 4360-4369.

Ravatn, R., S. Studer, et al. (1998). "Int-B13, an Unusual Site-Specific Recombinase of the Bacteriophage P4 Integrase Family, Is Responsible for Chromosomal Insertion of the 105-Kilobase clc Element of Pseudomonas sp. Strain B13." J. Bacteriol **180**(21): 5505-5514.

Rice, L. B. (1998). "Tn916 Family Conjugative Transposons and Dissemination of Antimicrobial Resistance Determinants." Antimicrob. Agents Chemother. **42**(8): 1871-1877.

Schroder, G., S. Krause, et al. (2002). "TraG-Like Proteins of DNA Transfer Systems and of the Helicobacter pylori Type IV Secretion System: Inner Membrane Gate for Exported Substrates?" J. Bacteriol **184**(10): 2767-2779.

Schröder, G. and E. Lanka (2005). "The mating pair formation system of conjugative plasmids - A versatile secretion machinery for transfer of proteins and DNA." Plasmid **54**(1): 1-25.

Smillie C., Garcillian M., et al. "Unpublished."

Stevens, A. M., N. B. Shoemaker, et al. (1990). "The region of a Bacteroides conjugal chromosomal tetracycline resistance element which is responsible for production of plasmidlike forms from unlinked chromosomal DNA might also be involved in transfer of the element." J. Bacteriol **172**(8): 4271-4279.

Sullivan J.T. and Ronson C.W. (1998). "Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene." Proc. Natl. Acad. Sci. USA **95**: 5145 - 5149.

Toussaint, A. and C. Merlin (2002). "Mobile Elements as a Combination of Functional Modules." Plasmid **47**(1): 26-35.

Waldor, M. K., H. Tschape, et al. (1996). "A new type of conjugative transposon encodes resistance to sulfamethoxazole, trimethoprim, and streptomycin in Vibrio cholerae O139." J. Bacteriol **178**(14): 4157-4165.

Wang, H. and P. Mullany (2000). "The Large Resolvase TndX Is Required and Sufficient for Integration and Excision of Derivatives of the Novel Conjugative Transposon Tn5397." J. Bacteriol **182**(23): 6577-6583.

Weinert, L. A., J. J. Welch, et al. (2009). "Conjugation genes are common throughout the genus Rickettsia and are transmitted horizontally." Proc Biol Sci. **276**(1673): 3619-3627.

Whittle, G., N. B. Shoemaker, et al. (2002). "The role of &lt;SMALL&gt;Bacteroides&lt;/SMALL&gt; conjugative transposons in the dissemination of antibiotic resistance genes." Cellular and Molecular Life Sciences **59**(12): 2044-2054.

Xu, J., M. K. Bjursell, et al. (2003). "A Genomic View of the Human-Bacteroides thetaiotaomicron Symbiosis." Science **299**(5615): 2074-2076.