Etude du mécanisme de recombinaison des intégrons, et leur utilisation comme générateur de combinaisons génétiques à des fins biotechnologiques

David Bikard

ECOLE DOCTORALE – FRONTIERES DU VIVANT

DOCTORAT

Discipline – Sciences de la vie

Spécialité – Génétique

David Bikard

Etude du mécanisme de recombinaison des intégrons, et leur utilisation comme générateur de combinaisons génétiques à des fins biotechnologiques

Thèse dirigée par le Dr. Didier Mazel

Soutenue le 29/09/2010

JURY

Pr. Philippe Reigner	Président
Dr. Didier Mazel	Directeur de thèse
Pr. Fernando De la Cruz	Rapporteur
Dr. Bénédicte Michel	Rapporteur
Dr. Hervé Isambert	Examinateur

Study of integron recombination mechanism,
and integron use as a genetic shuffling device
for biotechnological purpose

RESUME

Les integrons sont des systèmes de recombinaison génétique bactériens qui jouent un rôle majeur dans la dissémination des gènes de résistances aux antibiotiques. Ces plateformes génétiques sont capables de capturer des cassettes de gène et de les réorganiser afin de trouver des solutions adaptatives à un environnement changeant. Le mécanisme de recombinaison des intégrons est original. Contrairement aux autres systèmes de recombinaison spécifique de site de la même famille (catalysés par les recombinases à tyrosine), le site de recombinaison associé aux cassettes est reconnu sous forme d'ADN simple brin replié. Une large part de mon travail de thèse a été dédiée à la compréhension des mécanismes qui permettent à ces sites de recombinaison de passer de la forme stable qu'est la double hélice d'ADN à la conformation leur permettant d'être reconnu par la recombinase des intégrons. L'autre partie de ma thèse a consisté au développement d'un outil génétique utilisant les remarquables propriétés de recombinaison des intégrons. Ce nouvel outil, nommé « intégron synthétique » permet de générer un grand nombre de combinaisons de séquences hétérologues *in vivo*. Il pourrait être d'une grande utilité aux ingénieurs tentant d'assembler des réseaux et voies génétiques d'intérêt, par évolution dirigée.

4

ABSTRACT

Integrons are bacterial recombination systems that play a major in the spread of antibiotic resistance genes. These genetic platforms are able to capture gene cassettes and reorganize them to find adaptive solution to changing environments. The mechanism of integron recombination is unusual. In contrast to the other site-specific recombination systems of the same family (catalysed by a tyrosine recombinase), the recombination site of gene cassettes is recognized as folded single-stranded DNA. Half of my thesis work was dedicated to understanding how and when these recombination sites are able to go from the stable double-helix form of DNA to a conformation allowing recombination. The other half consisted in using the remarkable recombination capacities of integrons to develop a genetic shuffling device that can be used for the directed evolution of synthetic gene networks and metabolic pathways. Such device can indeed be of great use to genetic engineers trying to assemble new genetic pathways implementing functions of interest.

REMERCIEMENTS

De nombreuses personnes m'ont aidé et soutenu au cours de ces trois années de thèse et il est bien difficile de savoir par où commencer ! Mes premiers hommages reviennent sans doute de droit au grand Dr Didier Mazel pour m'avoir accueilli dans son laboratoire et avoir rendu cette thèse possible. Il a su me faire confiance malgré les nombreux errements scientifiques et projets farfelus dont cette thèse ne représente que la minorité achevée. Je dois beaucoup à mes collaborateurs directs : Marie Bouvier pour m'avoir encadré et formé à la biologie moléculaire et aux intégrons lors de mon premier stage dans le laboratoire en 2005 ; Guillaume Cambray pour m'avoir donné l'idée de l'Integron Synthétique et m'avoir orienté lors de mes débuts sur ce projet ; Céline Loot que je remercie tout particulièrement pour la fructueuse collaboration qui nous a amenés à publier le premier "papier" présenté dans cette thèse. Je me demande encore comment elle a pu supporter mes changements d'avis permanents et contradictoires, mais toujours aussi assurés, au cours de ce travail !

Je remercie aussi l'ensemble des membres du laboratoire PGB présents et passés qui ont su rendre l'ambiance de travail si chaleureuse, en particulier mes voisines de bureau Marie-Ève Val et Zeynep Baharoglu avec qui j'ai également l'honneur de cosigner des articles. Je n'oublie pas non plus Ana Babic et Lydia Robert qui m'ont initié à la microscopie et aidé pour un travail encore inachevé, plus pour longtemps j'espère.

Last but not least, je suis très reconnaissant envers ma famille et mes amis de m'avoir soutenu et de s'être intéressés à mon travail. Je pense particulièrement à ma mère, qui même si elle n'en comprenait pas forcement leurs titres, a imprimé, encadré et diffusé mes articles auprès de toute ma famille. Enfin je n'oublie bien entendu pas Anna, car il n'est pas toujours évident de vivre avec un thésard qui vous prête souvent moins d'attention qu'à ses bactéries.

Du fond du cœur, merci à tous !

Depuis maintenant 30 ans l'homme apprend à écrire sur le support qu'est l'ADN. Nous ne savions au début que faire du copier/coller alors que se profilent maintenant les technologies qui permettent d'écrire n'importe quel texte. Reste à savoir quoi écrire? Car si les génomes d'une multitude d'organismes ont maintenant été séquencés, nous sommes encore loin d'avoir complètement déchiffré la manière dont ces textes sont lus et exécutés pour former les organismes vivants. Cependant, nos maigres connaissances nous permettent déjà de modifier rationnellement des séquences pour forger à façon des organismes réalisant des taches qui nous sont utiles (production de médicaments, de carburants…). L'ingénieur qui poursuit de tels buts est toutefois souvent limité par notre faible connaissance du vivant. Il doit alors produire à l'aveugle un grand nombre de séquences différentes dans l'espoir d'y trouver celle qui accomplira la tâche souhaitée. Ce procédé est appelé évolution dirigée car, à la manière de l'évolution, il procède par mutation aléatoire et sélection. Les mutations introduites ne sont cependant pas toujours complètement aléatoire. Nos connaissances peuvent bien souvent nous guider pour cibler la région dont nous savons qu'elle influencera le comportement que nous cherchons. Une multitude de techniques ont ainsi été développées pour muter et recombiner à souhait les séquences d'intérêts. Ces méthodes ne permettent pas uniquement d'obtenir des organismes remplissant mieux une fonction voulue, mais par l'analyse des mutations sélectionnées, elles permettent aussi de mieux comprendre le fonctionnement de l'organisme vivant. Part l'analyse et la synthèse, un cercle vertueux s'établit. Une meilleure connaissance du vivant permet de mieux le modifier, et l'analyse du comportement des organismes modifiés permet de mieux comprendre le vivant.

Ma thèse s'inscrit directement dans ce cadre. La moitié de mon travail a été dédié à l'étude d'un mécanisme que certaines bactéries possèdent et qui leur permet d'évoluer en réponse à un stress environnemental. Il s'agit des intégrons, qui sont notamment connus pour jouer un rôle prépondérant dans la propagation des gènes de résistance aux antibiotiques. Ces machines génétiques permettent naturellement aux bactéries de capturer des gènes et de les réorganiser afin de s'adapter rapidement à un environnement changeant. La compréhension du mécanisme de recombinaison des

intégrons par mes pairs, et ce que j'ai pu y apporter, m'ont permis de développer un nouvel et performant outil de mutagénèse. L'intégron synthétique permet de générer de manière aléatoire un grand nombre de combinaisons de séquences pour faire de l'évolution dirigée à l'échelle du réseau de gènes. Ce travail est une nouvelle corde à l'arc des ingénieurs et contribuera, je l'espère, à une meilleure compréhension du fonctionnement des organismes vivants.

<div style="border:1px solid black; text-align:center;">

Table of contents

</div>

Introduction

I. Synthetic Biology

I.1 What is Synthetic Biology?

Synthetic biology can be defined as the engineering of living organisms at the molecular level. It is the design and construction of new biological systems that are not found in nature (Endy 2005), and its first purpose is the creation of novel useful functions. Another motivation behind synthetic biology is the raising awareness that descriptive biology has reached several limits, and that one way to overcome these is to study biological systems through the synthesis of artificial systems that reproduce their behavior. Richard Feynman's famous quote is broadly used in the Synthetic Biology community to convey this idea:

"What I cannot create I do not understand"

The idea of relying on synthesis as opposed to analysis to decipher biological phenomenon is not new. Already in 1912, the French chemist Stephan Leduc was writing in his "*La Biologie Synthétique*":

« Jusqu'à présent la biologie n'a eu recours qu'à l'observation et à l'analyse. L'unique utilisation de l'observation et de l'analyse, l'exclusion de la méthode synthétique, est une des causes qui retardent le progrès de la biologie. . . [La méthode synthétique] devoir être la plus féconde, la plus apte à nous révéler les mécanismes physiques des phénomènes de la vie dont l'étude n'est même pas ébauchée. Lorsqu'un phénomène, chez un être vivant, a été observé, et que l'on croit en connaître le mécanisme physique, on doit pouvoir reproduire ce phénomène isolément, en dehors de l'organisme vivant. »

The origins of Synthetic Biology were remarkably discussed by M. Schmitt & al. in their book "*Synthetic Biology*" (Schmidt 2009). They show that the idea of a synthetic, engineering-based approach to life is a prominent and recurring theme in the history of biology of the twentieth century. It is however interesting to note that Synthetic Biology only gained a significant influence in the past decade, mostly due to the presence of charismatic scientists in the field such as Drew Endy, Jay Keasling, Michael Elowitz, Ron Weiss, Timothy Gardner, James J. Collins, Craig Venter or

George Church and their ground-breaking works. It is also worth noting that the iGEM (international Genetically Engineered Machines Competition) student contest has played a major role in Synthetic Biology popularization and in the recent hype around Synthetic Biology.

There has been much arguing about the novelty of Synthetic Biology as a discipline(de Lorenzo and Danchin 2008). After all, most of the tools presently used by genetic engineers were invented back in the 80's. Those are mostly the technologies of PCR and recombinant DNA. Furthermore whole fields, such as metabolic engineering, which started well before the name "Synthetic Biology" was coined, now claim themselves as part of it. One can thus arguably say that Synthetic Biology is just a new brand name for genetic engineering. So, why all the hype around this? The shift, as a matter of fact, is mostly a conceptual one, and was probably, more or less unconsciously, already taking place in the genetic engineering community, explaining the perplexity of some of its members. Nevertheless, all the recent fuss certainly helped raise the awareness about the challenges and opportunities of genetic engineering, and the necessity to build a true engineering framework.

I.2 Tinkering vs Engineering

The goal of Engineering as a discipline is to build a working environment in which you can specify a design that, once executed, will produce the desired outcome with the best possible chance. In other words, as an engineer, you want predictability and reliability. The principles that could allow creating such an environment for biology are best described by Drew Endy in his review "Foundations for engineering biology" (Endy 2005). Three main principles are highlighted: standardization, decoupling and abstraction.

Behind the word **standardization** is the idea that engineers need to speak the same language and must use compatible tools if they want to build upon each other's work. Currently, a promoter described as strong in a given lab might be weak in another lab where the *E.coli* strain and the culture medium are slightly different. Genetic elements would need characterization following standardized protocols. Similarly, it often takes several weeks to assemble genetic elements that were conceived by different scientists because they used different construction methods. A

recent effort in this direction is the registry of standard biological parts (partsregistry.org) developed at MIT. This repository attempts to gather both the physical DNA of genetic elements and their precise characterization according to established standards. Nevertheless, although thousands of genetic elements have been catalogued, the quality of the characterization remains, until now, too weak to trigger a broad usage among the molecular biology community.

Decoupling is the idea that a complex problem can be decomposed into smaller problems that can be worked on independently. Big engineering projects such as the construction of the A380 airbus required thousands of people, but no single individual had the all-encompassing knowledge of everything going on in the design and construction process. However, it was possible to bring together the work of this entire team into something that finally worked at the first trial. It is not trivial that such decoupling can take place in the engineering of biological systems. The different complexity levels in biological systems could be so intertwined that decoupling might be hard to achieve.

Abstraction is the concept of hiding the details that are not relevant to a given level of complexity. Our mind is able to manage only a limited amount of information at the same time. When designing a complex behavior at the cell level, it is impossible to handle all the sequence details and possible interaction between every DNA parts. One needs to rely on high level genetic devices descriptions that take a given input and produce a known output. Such devices are for instance oscillators, sensing modules, switches etc.

The successful development of foundational technologies based on the ideas of standardization, decoupling and abstraction would help the engineering of synthetic biological systems that behave as expected become routine.

I.3 Genetic Circuit Design

Synthetic Biology is a much diversified discipline and has been the object of many reviews(Serrano 2007; Carr and Church 2009; Khalil and Collins 2010), to cite just a few. In this part, I will only focus on the aspects of synthetic biology that have motivated my thesis.

I.3.a Regulatory networks engineering

The first synthetic gene circuits that were successfully designed and implemented consisted in a genetic toggle switch (Gardner, Cantor et al. 2000) and an oscillator made of three transcription factors repressing each other in circle, coined the "repressilator" (Elowitz and Leibler 2000)(Figure 1). Those two systems consisted in small networks of transcription factors that could be predictably assembled into working systems and evaluated against predictions of models. This opened the way to a multitude of other small biological circuits often inspired by electronic circuits, such as genetic switches and oscillators, logic gates, filters and communication modules. Nevertheless designs were rarely behaving as intended the first time, and fine tuning was often required as exemplified by Yokobayashi & al. who used a combination of rational design and directed evolution to make a non-functional circuit functional (Yokobayashi, Weiss et al. 2002).

In a recent approach, Ellis and colleagues show how one can use the guidance of mathematical modeling to choose components from a library of well characterized parts, and construct a system that behaves as predicted (Ellis, Wang et al. 2009). They constructed libraries of tetR- and lacI-regulated promoters, using synthesis-with-degenerated-sequence method, and cloned them in front of a fluorescent reporter to assess by flow cytometry their response to the corresponding transcription factor. They first showed that a mathematical model can accurately predict the behavior of feed-forward loops including the tetR-regulated promoters. They then attempted to show that this methodology can be extended to the construction of genetic timers that are plugged to a gene controlling yeast sedimentation. Such device may be of interest for fermentation processes such as bioethanol production where the precise control of yeast sedimentation is a critical factor. The timers consisted in tetR and lacI genes repressing each other.

Box 1 | Early synthetic biology designs: switches and oscillators

Switches and oscillators that occur in electronic systems are also seen in biology and have been engineered into synthetic biological systems.

Switches

In electronics, one of the most basic elements for storing memory is the reset–set (RS) latch based on logical NOR gates. This device is bistable in that it possesses two stable states that can be toggled with the delivery of specified inputs. Upon removal of the input, the circuit retains memory of its current state indefinitely. These forms of memory and state switching have important functions in biology, such as in the differentiation of cells from an initially undifferentiated state. One means by which cellular systems can achieve bistability is through genetic mutual repression. The natural P_R–P_{RM} genetic switch from bacteriophage λ, which uses this network architecture to govern the lysis–lysogeny decision, consists of two promoters that are each repressed by the gene product of the other (that is, by the Cro and CI repressor proteins). The genetic toggle switch[8] constructed by our research group is a synthetically engineered version of this co-repressed gene regulation scheme. In one version of the genetic toggle, the P_L promoter from λ phage was used to drive transcription of *lacI*, the product of which represses a second promoter, Ptrc2 (a *lac* promoter variant). Conversely, Ptrc2 drives expression of a gene (*cI-ts*) encoding the temperature-sensitive (ts) λ CI repressor protein, which inhibits the P_L promoter. The activity of the circuit is monitored through the expression of a GFP promoter.

The system can be toggled in one direction with the exogenous addition of the chemical inducer isopropyl-β-D-thiogalactopyranoside (IPTG) or in the other direction with a transient increase in temperature. Importantly, upon removal of these exogenous signals, the system retains its current state, creating a cellular form of memory.

Oscillators

Timing mechanisms, much like memory, are fundamental to many electronic and biological systems. Electronic timekeeping can be achieved with basic oscillator circuits — such as the LC circuit (inductor L and capacitor C) — which act as resonators for producing periodic electronic signals. Biological timekeeping, which is widespread among living organisms[120], is achieved with circadian clocks and similar oscillator circuits, such as the one responsible for synchronizing the crucial processes of photosynthesis and nitrogen fixation in cyanobacteria. The circadian clock of cyanobacteria is based on, among other regulatory mechanisms, intertwined positive and negative feedback loops on the clock genes *kaiA*, *kaiB* and *kaiC*. Elowitz and Leibler constructed a synthetic genetic oscillator based not on clock genes but on standard transcriptional repressors (the repressilator)[9]. Here, a cyclic negative feedback loop composed of three promoter–gene pairs, in which the 'first' promoter in the cascade drives expression of the 'second' promoter's repressor, and so on, was used to drive oscillatory output in gene expression.

Figure 1 The first synthetic regulatory networks: the toggle switch and the repressilator (from Khalil and Collins, 2010)

This is in fact a toogle-switch motif (see Figure 1) for which it was shown that changes in opposing repressor levels can disrupt bistability (Gardner, Cantor et al. 2000). This ultimately happens at a rate which depends on promoter characteristics, and memory of induction is then lost as the systems resets to its original default state. In this case, the initial mathematical model of the system did not accurately predict the behavior of the constructed timers. The authors had to first fit their model to the behavior of one timer, and it could only then predict the behavior of the other timers. Also successful prediction based on a mathematical model could not be achieved the first time; this method suggest that it is sufficient to construct a single exemplary networks to fit a model, and then be able to predictably construct variants of this network with different characteristics.

Such approaches are encouraging in our hope to predictably construct more complex genetic systems. Nevertheless, it is important to note that none of the synthetic regulatory networks assembled until now included more than 3 or 4 genes. The prediction capacities demonstrated for simple networks will undoubtedly be challenged by more complex systems.

I.3.b. Metabolic engineering

The implementation of novel metabolic pathways in microorganism for the production of valuable compounds for chemical and pharmaceutical use is the subject of much interest, and one of the cornerstones of Synthetic Biology (Yadav and Stephanopoulos; Keasling 2008). Metabolic engineering generally consist in the introduction of enzymes of interest able to catalyze the reactions leading to the production of a desired compound. The chosen compounds are generally hard to extract from the organisms that produce them naturally. Furthermore, the chemical synthesis of complex organic molecules is often tedious, costly and not environmentally friendly. On the contrary, microorganisms can process very cheap raw materials like glucose into complex molecules at low costs. Alternatively, the production of foreign compounds can be achieved by the introduction of new genes in bacterial or yeast hosts. Transgene expression has been extensively used by pharmaceutical companies since it was first validated for the production of insulin (Goeddel, Kleid et al. 1979; The 1989). Metabolic engineering essentially consists in

building a model of the host cell metabolism and using this model to predict candidate genetic targets that could be deleted or over-expressed to direct the metabolic flux in a desired pathway (Park, Lee et al. 2007; Wang, Isaacs et al. 2009).

There is nonetheless an important difference between the production of pharmaceutical proteins and the production of compounds requiring a whole metabolic pathway. The production of large amounts of proteins can most of the times easily be achieved by controlling the gene of interest with a strong promoter and on a high copy-number plasmid. Nevertheless, when it comes to the engineering of a more complex pathway, the enzymes do not necessarily need to be highly expressed. A strong expression of all the genes in a pathway would most likely deprive the cell of metabolites that might otherwise be useful to produce the molecule of interest (Pfleger, Pitera et al. 2006). The enzymes rather need to be produced in catalytic amounts only sufficient not to limit the metabolic flow. Another constrain is to balance the expression of the different enzymes so as not to accumulate metabolic intermediates to toxic levels(Martin, Pitera et al. 2003; Pitera, Paddon et al. 2007), which is especially true for exogenous compounds.

Recent successful examples of metabolic engineering include the efficient production of L-valine, L-threonine, lycopene, antimalarial drug precursor, and benzylisoquinoline alkaloids (for a comprehensive review see (Lee, Kim et al. 2009)).

We can arguably say that while recombinant protein expression was largely molecular tinkering, the metabolic engineering of complex pathways requires an engineering framework, and thus falls in the realm of synthetic biology. (For further reading see (Keasling 2008), in this review Keasling attempts to lay the bases of biological parts standards that would help make this field a true engineering discipline.)

I.4. Directed evolution and Synthetic Biology

The constructivist approach brought by Synthetic Biology proposes to understand biological systems as we build them, and is in this sense in clear opposition to reductionism. Reductionism is often embodied in molecular biology by reverse genetic approaches, which brought much of our present knowledge of living systems. It mainly consists in the dissection of molecular functions through the analysis of the phenotypes of mutants. The knowledge gathered by this method

throughout the past decades constitutes the basis of the models used by synthetic biologists to predict and construct new functions. Unfortunately, all of our current knowledge appears extremely scarce when it comes to predict the phenotype of an organism with a new sequence. There are only very few examples where the knowledge of every parts of a system was enough to predict its outcome (Ellis, Wang et al. 2009), so that it worked at first trial. Hence building a large set of system variants increase the chance that one of them will work. It is then possible to select for the functional designs and repeat the procedure until it converges towards a satisfying solution. This methodology, known as directed evolution, has successfully been used in protein engineering and has only recently been applied to metabolic pathways and gene networks (see part I.4.c). Directed evolution consists in three steps that can be iterated: (i) generation of diversity through mutagenesis, (ii) selection of better performing mutants, (iii) amplification of selected mutants.

Proteins have been engineered in this way to modify their substrate specificity, affinity and reaction rate (Zhao 2007), without the need of an a priori knowledge of the role of selected mutations and there impact on the protein folding and activity. Here lies all the power of this technique; it allows us to engineer better systems without understanding how exactly we did it. But in retrospect, it also allows a better understanding of the engineered system through the analysis of the selected mutants phenotypes in a classical reverse genetics approach.

Nevertheless, when one has a poor idea of potential mutagenesis targets, the number of possible mutations and combination of mutations can very easily exceed the size of the population. In these cases, only a limited landscape of evolution can be assessed. This is especially problematic when several mutations are required to obtain the desired behavior. There, the combinatorial explosion of possibilities is rapidly prohibitive. Rational decisions on mutagenesis targets must then be taken (Yokobayashi, Weiss et al. 2002). The combination of rational and evolutionary strategies is in many cases probably the best way to produce a working system. Furthermore this methodology works in a virtuous circle: the analysis of selected mutations allows a better understanding of the system, which in return allows a better targeting of more random mutations. I will review here the strategies employed in protein directed evolution (for which most of the mutagenesis techniques were developed), RNA directed evolution and finally gene network directed evolution.

I.4.a Protein directed evolution

Creation of novel proteins by rational design is an intricately difficult task. Although progresses are made in this direction, we are still not able to predict the folding of a new amino acid sequence, let alone predict its activity. However, directed evolution of proteins with improved stability and activity has now been a current practice for more than 20 years (Brannigan and Wilkinson 2002).

i. A brief history of random mutagenesis techniques

The mutagenesis of specific genes became possible with the advent of cloning methods and challenged the imagination of scientists. Genes of interest could be isolated and mutagenized with the same agents that were already in use for in vivo random mutagenesis, such as hydroxylamine or nitrous acid (Benzer and Champe 1961). Targeted sequences would be isolated with endonucleases, treated with chemicals and cloned back into a vector. The main drawback of these methods was the strong biais in the spectrum of introduced mutations. Another technique consisted in a brief passage of purified DNA through bacterial strains carrying various mutator activities (Cox 1976; Ruvkun and Ausubel 1981). Alternatively, endonucleases and exonucleases were used to produce a single stranded gap over a region of interest (Shortle 1983). The gap would then be field by a polymerase in the presence of non-natural mutagenic nucleotides or with a bias in the nucleotide pool that would push the polymerase to make errors. The mutations spectra were less biased than with chemicals. Although these techniques were already improvements, the sequence span that could be mutagenized depended on the presence of appropriate restriction sites, and mutagenesis wasn't very efficient.

In the 80s, the use of synthetic oligonucleotides allowed better controlled mutagenesis. Targeted substitutions, insertions, or deletions could readily be achieved. Synthetic oligos could be used as primers to synthesize the complementary strand of a single stranded circular plasmid (isolated with techniques involving ssDNA phages, and called "phagemids"), thereby introducing a mutation on one strand that would segregate upon transformation of the vector into a cell. Alternatively in a method called "cassette mutagenesis", mutations would be introduced at specific positions in pairs of oligonucleotides that would be annealed and cloned in a digested vector (Wells, Vasser et al. 1985). With the earthquake triggered by the PCR development

(Saiki, Gelfand et al. 1988), it became even easier to introduce mutations with synthetic oligos at any position. Furthermore, the controlled synthesis of numerous random oligonucleotide sequences in the same reaction tube (Frank, Heikens et al. 1983) made the introduction of random mutations through PCR or cassette mutagenesis still easier and better controlled. All these mutagenesis techniques were first developed in order to study the relationship between genotype and phenotype (Dalbadie-McFarland, Cohen et al. 1982), but it soon became clear that the same methods could be used to improve protein functions for given purposes.

One of the first successes was the engineering of an alkaline protease (subtilisin) with increased resistance to oxidation, which was later used in washing powders (Wells, Vasser et al. 1985). This mutant was generated in a library where a methionine codon was randomized by saturation mutagenesis using degenerate oligonucleotides that incorporate NNC/G codons. All amino acids can indeed be encoded by NNC/G which allows limiting the redundancy in the library in comparison to a NNN codon (Arkin and Youvan 1992).

This method allowed exploring all possible amino-acids at specific positions, but when there was no rational evidence about positions to target, "brute force" strategies were required. The goal was then to mutagenize the whole sequence and to explore the larger possible amino-acid sequence-space with the smallest possible population. A first step in this direction came from the isolation of error-prone Taq polymerases which allowed much better and less biased mutagenesis than previous methods (Cadwell and Joyce 1992). Using this technique, mutants were isolated in a variety of genes, including recA and crp, which provided insights into the relation between protein structure and function (Lerner and Inouye 1994).

ii. Exploring mutation combinations

However, these strategies were not always successful in improving protein activity or stability. It soon became clear that single amino-acid alterations would not be sufficient to achieve more radical changes and that the combinatorial mutagenesis of several positions would be required. In the 1990s, PCR-based methods were developed to mimic natural recombination to generate a high number of mutation combinations. In particular Stemmer introduced a method consisting in rounds of gene fragmentation, reassembly and selection (Stemmer 1994). As a proof of principle

this method was shown to improve the activity of a β-lactamase 32,000-fold where a classical error-prone PCR only yielded 16-fold improvements. This method enabled to combine beneficial mutations that arouse independently in the population, hence the denomination of "sexual PCR". Mutations in a same protein can have synergistic effects, explaining the incredible success of the method.

The use of this method to shuffle related natural sequences was shown to be particularly efficient to improve proteins. Crameri et al. demonstrated that the shuffling of four related cephalosporinase genes gave a 540-fold improvement in moxalactamase activity in only one single round, whereas single gene shuffling of the four genes independently yielded only 8-fold improvement (Crameri, Raillard et al. 1998).

iii. Broadening the evolutionary landscape

The explored evolutionary landscape remained very limited. Because of the nature of the genetic code, a single mutation in a given codon can only change the encoded amino acid with a limited set of the 20 other possibilities. For example, mutation from AUG (methionine) to UAG (tryptophan) is unlikely to occur. The average number of amino-acids that a codon can access with a single mutation is only 5,8 (Cambray 2009). Engineers thus focused on strategies allowing an unbiased exploration of amino-acids at random positions. For instance, an ingenious method called RID (random insertion/deletion) enables deletion of an arbitrary number of consecutive bases (up to 16 bases) at random positions and, at the same time, insertion of a specific sequence or random sequences of an arbitrary number of bases at the same position (Murakami, Hohsaka et al. 2002).

In another approach, Cambray & Mazel proposed to take advantage of the fact that different codons encoding the same amino-acid often have different evolutionary landscapes. An algorithm was developed to compute synonymous sequences with varying evolutionary landscapes. These sequences can then be synthesized and used in a classical error-prone PCR mutagenesis process to optimize the amino-acid landscape that can be explored by point mutations. They showed that an optimized synonymous *aac6'Ib* gene can evolve a new aminoglycoside resistance phenotype never observed from the WT sequence (Cambray and Mazel 2008).

iv. Neutral drift libraries

These efficient strategies allow increasing the explored sequence space while reducing the redundancy in the genetic library. However, the number of possible genetic combinations remains much larger than the possible library size as soon as the combination of several mutations is explored. Considering this, one may wonder how natural evolution so rapidly succeeds in producing proteins with new functions (new antibiotic resistances, the ability to degrade new artificial chemicals…), when natural population sizes often are much smaller than what engineers can achieve in the lab (Peisajovich and Tawfik 2007).

Selection can only improve functions already present in the pool, but it appears that proteins are far from being rigid structures with a single function. It has been proposed that rather than adopting a single structure with a defined function, proteins can be viewed as an ensemble of conformations in equilibrium. Alternate conformations can mediate alternate functions (Tokuriki and Tawfik 2009). These secondary functions, also named "promiscuous" can provide starting points for the evolution to act upon. It was also discovered that functionally neutral mutations play a central role in protein evolution. Such mutations are frequent in nature and a wealth of them is currently obtained during directed evolution procedures. Two main pathways were identified by which functionally neutral mutations open new adaptive pathways. First, a neutral mutation can enhance a protein's stability, thereby increasing its tolerance for subsequent functionally beneficial but destabilizing mutations. Second, neutral mutations can lead to changes in promiscuous functions that can become starting points for adaptative evolution of new functions (Bloom and Arnold 2009).

Gupta & Tawfik nicely demonstrated how neutral drift libraries can be used to change the substrate specificity of the serum paraoxonase PON1 enzyme to the fluorogenic organophosphate CMP-MeCyC. In a neutral drift library of only 500 sequences, they could find 21 variants with improved activity. Further DNA shuffling of the improved variants was realized and 360 randomly selected clones were individually screened for CMP-MeCyC hydrolysis. This screen yielded sequences with an up to 77-fold improvement in catalytic efficiency, while traditional methods required libraries of 10^4-10^7 variants to achieve the same efficiency.

v. Hybrid proteins and block shuffling

All the methods described above enable to explore substitutions, deletion or insertions all along a gene, however they cannot reorganize and combine blocks of non-homologous sequences. Proteins can often be described in terms of domains representing structural and functional motifs (Panchenko, Luthey-Schulten et al. 1996). It is now admitted that many proteins acquired their functional diversity by combining existing building blocks (Kolkman and Stemmer 2001). This is especially true in eukaryotes where the exon/intron organization of the coding sequences readily allows recombination events to build, so-called "mosaic proteins", out of existing domains (Gilbert 1978). In order to build libraries of random mosaic proteins in the lab, methods are required that enable combining non-homologous sequences. The ITCHY method enables to realize libraries of hybrids from 2 parental proteins with a random cross-over point (Ostermeier, Shim et al. 1999). Each parental sequence is randomly truncated, and protein segments are ligated together. Derived methods include SCRATCHY and SHIPREC. Another method named NRR (Non-homologuous Random Recombination) proposes to fragment desired sequences with DNaseI, to realize blunt-end ligation/extension, and capping using two asymmetrical DNA hairpins to stop extension (Figure 2)(Bittker, Le et al. 2004).
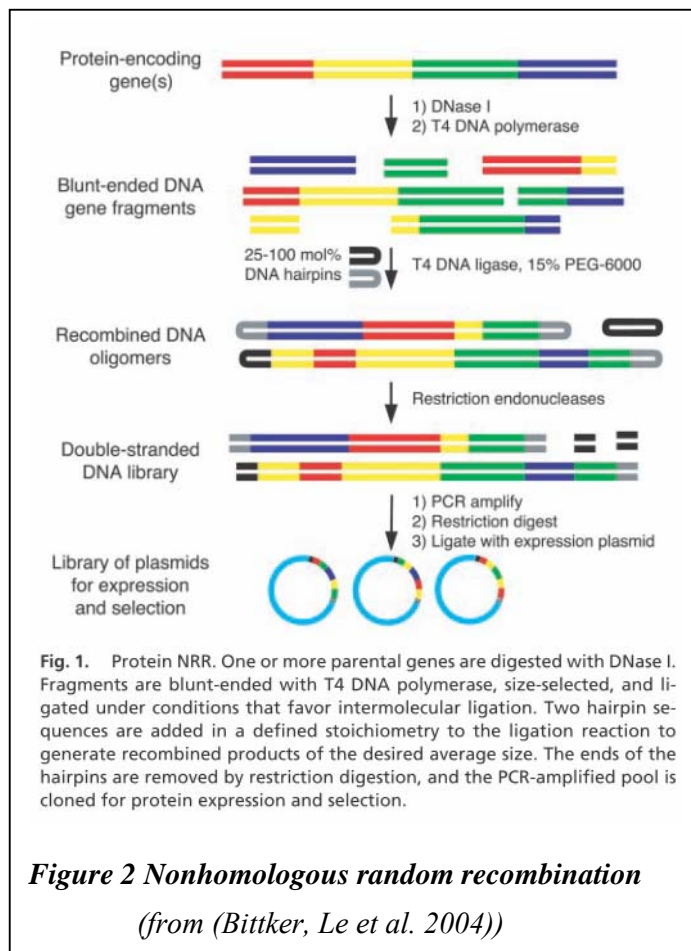
Fig. 1. Protein NRR. One or more parental genes are digested with DNase I. Fragments are blunt-ended with T4 DNA polymerase, size-selected, and ligated under conditions that favor intermolecular ligation. Two hairpin sequences are added in a defined stoichiometry to the ligation reaction to generate recombined products of the desired average size. The ends of the hairpins are removed by restriction digestion, and the PCR-amplified pool is cloned for protein expression and selection.

Figure 2 Nonhomologous random recombination
(from (Bittker, Le et al. 2004))

The main problem of these methods is that frame shifts happen frequently and produce large numbers of non-

Figure 1. A schematic of sequence-independent site-directed chimerageneis (SISDC), illustrating recombination between two parents at three sites (four elements). (a) The nucleotide sequences of the parent genes (A and B) are aligned, and consensus sequences at each targeted site (I, II and III) are determined. (b) Marker tags (I–III) are inserted into targeted sites (I–III). The marker tags contain consensus 5 bp sequence (S_5), a type IIb endonuclease (*Bae* I) recognition sequence (blue), two variable regions (X_{10} and Y_4: pink for site I, cyan for site II and green for site III), a constant region (yellow) and a *Sma* I recognition site (underline). (c) Inserted tag regions are removed by treatment with endonuclease and specific sticky ends are produced. (d) Mixed fragments can identify their neighbors, and the fragments from different parents ligate with each other in the right order. After this step, digestion by *Sma* I can eliminate any untreated tag remaining in the final, chimeric library.

Figure 3 Sequence-Independent Site Directed Chimerageneis (SISDC) (from (Hiraga and Arnold 2003))

functional sequences. Their ability to create diverse libraries of functional proteins has not been demonstrated convincingly. In order to maximize the chances that hybrid proteins will be functional, protein domains integrity should be conserved which is often not the case when recombination points are random. Hence, protein engineering with defined cross-over points is often a preferred route. When domains cannot be clearly defined, computational algorithms can be used to predict crossover points less likely to disrupt the protein folding (Meyer, Silberg et al. 2003).

Methods that allow controlled recombination are based on the assembly of defined sequence blocks either through PCR or ligation. Soon after the development of the PCR, a method called "overlap extension" or "fusion-PCR" enabled to precisely assemble two unrelated sequences in two rounds of PCR (Horton, Hunt et al. 1989). In a first round blocks are separately amplified with oligonucleotides introducing overlaps, and in a second round, the overlapping products are mixed and amplified with external primers. This method was later adapted for the random assembly of multiple blocks (Tsuji, Onimaru et al. 2001). RM-PCR (random multi-recombinant PCR) consists in the separate assembly of every possible block-dimers (for the random assembly of 5 blocks, 25 block-dimers are generated), which are then all mixed together and randomly assembled through PCR.

However, this method is probably unreliable as it was never used beyond the proof of principle. Alternatively, short overlaps of 6nt between synthetic oligonucleotides were used to randomly assemble combinatorial proteins through a PCR-like protocol called "microgene polymerization reaction" (Saito, Minamisawa et al. 2007).

Other methods relying on random ligation have also been developed such as the "Y-Ligation-Based block shuffling" (Kitamura, Kinoshita et al. 2002). It was successful in shuffling a region of a gene coding for GFP divided in 4 or 8 blocks, but introduced strong biases in the assembly and produced only one working sequence among 10^6 screened variants. In another method: sequence-independent site directed chimeragenesis (SISDC); restriction enzymes that cut outside their recognition site are used to produce blocks with homologous overhangs that are shuffled through random ligation (Figure 3)(Hiraga and Arnold 2003). This method was successfully used to generate hundreds of novel beta-lactamases from the shuffling with seven cross-over points of three distant genes (34-42% homology) (Meyer, Hochrein et al. 2006). The choice of the cross-over points was assisted by the SCHEMA algorithm (Voigt, Martinez et al. 2002). Interestingly, Meyer and colleagues could show that many non-functional chimeras could be rescued by random mutagenesis. This suggests a new route for the engineering of novel protein, in which rationally assisted domain shuffling followed by random mutagenesis could produce many new and functional protein variants.

I.4.b RNA directed evolution for synthetic biology

Directed evolution approaches yielded many RNA devices that are now commonly used by synthetic biologists in a variety of functions. Indeed, RNA plays a central role in many cell processes beyond the one of simple messenger between DNA and protein. It can have catalytic activities (Stark, Kole et al. 1978), perform self-cleavage (Bass and Cech 1984) and regulate gene expression (Mironov, Gusarov et al. 2002). Furthermore RNA is an attractive molecule to engineer since its conformations and functions are much easier to model and modify than the one of proteins.

32

i. Riboswitches and RNA Aptamers

Several riboregulators have been rationally developed to control gene expression at the RNA level. Riboregulators commonly work as follow: the 5'UTR of an mRNA is engineered to fold into a hairpin that sequesters the RBS and prevents ribosomes from binding. A trans-activating RNA can then be expressed from another promoter that will bind to the 5'UTR, unfold the hairpin and free the access to the RBS, allowing translation (Isaacs, Dwyer et al. 2006).

Other structures, termed riboswitches, present in the 5'UTR of some RNAs are able to regulate the RNA translation through changes in conformation mediated by the binding of specific molecules (from small molecules like caffeine to whole proteins). RNA can indeed fold into 3D structures termed aptamers able to bind specific ligands. A natural example is the one of thiamine which can bind the RNA encoding the production of the enzymes involved in its biosynthesis (Winkler, Nahvi et al. 2002). Upon binding of thiamine the riboswicth forms an alternative structure that down-regulates translation by sequestering the RBS. Some riboswitches are also able to act as ribozymes and silence expression through mRNA cleavage, while other act as transcriptional terminators (for a comprehensive review see, (Barrick and Breaker 2007))

Aptamers can be used to control gene expression in response to a variety of ligands, and are thus of much interest to biological engineers. A method to develop new aptamers was introduced in 1990, and named SELEX for Systematic Evolution of Ligands by EXponential enrichment (Tuerk and Gold 1990). This method has later been improved and applied to different kinds of nucleic-acids, including non-natural ones (Stoltenburg, Reinemann et al. 2007). Typically the procedures starts from a library of 10^{13} to 10^{15} chemically synthesized oligo-nucleotides. In the SELEX procedure for RNA aptamers, a library of random DNA sequences is transcribed in vitro and the RNAs are passed through a column in which the target ligand is covalently bound. Unbound RNAs are washed away, and the selected RNAs are eluted by competitive binding with the free ligand. The pool of binders is then amplified through reverse-transcription and PCR. The process can then be repeated to find aptamers that bind the ligand with increased affinity (Figure 4). To select for highly specific aptamers, the washing step can be realized with a solution containing molecules structurally similar to the targeted ligand.

Figure 3. Aptamer selection scheme. RNA aptamers may be isolated from large pools of randomized sequences using *in vitro* selection. The pool is enriched for sequences that bind an immobilized ligand and are eluted with free ligand. To obtain high-affinity binders, the selected sequences are reverse transcribed, amplified, and subjected to further rounds of selection.

Figure 4 SELEX method for RNA aptamers *(from (Topp and Gallivan 2010))*

INPUT	A	0	0	1	1	Meaning
	B	0	1	0	1	
OUTPUT	AND	0	0	0	1	Output is true if and only if (iff) both A and B are true.
	XOR	0	1	1	0	True iff *A* is not equal to *B*.
	OR	0	1	1	1	True iff *A* is true, or *B* is true, or both.
	NOR	1	0	0	0	True iff neither *A* nor *B*.
	XNOR	1	0	0	1	True iff *A* is equal to *B*.
	NAND	1	1	1	0	*A* and *B* are not both true.

Table 1 Truth table of some logic functions

Perhaps the most widely used aptamers in synthetic biology is an aptamers isolated through SELEX in 1994 that specifically binds theophylline (Jenison, Gill et al. 1994). Since then, aptamers have been incorporated within the 5'UTR of genes to generate synthetic riboswitches that respond to theophylline and a variety of new ligands (for review see (Isaacs, Dwyer et al. 2006)). An increasing number of studies are developing such new riboswitches through a combination of rational design, random mutagenesis and selection. Known riboswitches are randomized at carefully chosen positions, cloned in the 5'UTR of a fluorescent reporter gene, and the best ones are isolated with screens based on fluorescence-activated cell sorting (FACS). Lynch & Gallivan have been able to screen in this way libraries of riboswitches with up to 12 randomized bases ($4^{12} \sim 10^7$ sequences) (Lynch and Gallivan 2009), and could isolate riboswitches with very low basal expression and activation factors up to 96x upon binding with theophylline. The choice of the randomized bases is founded on our continuously improving knowledge of the relationships between aptamers sequence, structure and binding affinity (Carothers, Goler et al. 2010). Even more interestingly, riboswitches carrying several aptamers have been shown to behave cooperatively, i.e to have allosteric properties enabling the encoding of logic functions in riboswitches (Sudarsan, Hammond et al. 2006). Synthetic riboswitches have been constructed that realize AND, NOR, NAND, or OR gates between two input molecules (Win and Smolke 2008) (Table 1). Here again *in vivo* directed evolution for allosteric riboswitches featuring improved behaviors, proved to be very useful (Wieland and Hartig 2008).

ii. Engineering of orthogonal ribosome-RBS pairs.

In order to have a better predictability over an artificial system, engineers seek to limit the interactions with the host cell (often termed "chassis" by synthetic biologists). Orthogonal systems are designed to have the fewer possible interaction with the rest of the cell. In this spirit, Rackham et Chin engineered, in a directed evolution approach, a set of orthogonal ribosome-mRNA pairs (Rackham and Chin 2005), that proved very useful from the design of new genetic circuits (Rackham and Chin 2006) to the incorporation of unnatural amino-acids in proteins (Neumann, Wang et al. 2010). The idea is that the orthogonal ribosome binding site should only be recognized by its cognate artificial 16S rRNA and not by natural 16rRNA. On the other hand, the artificial 16s rRNA should not interfere with natural RBS. To achieve

this goal, they randomized 6 bases of RBS sequences and 8 bases of the 16s rRNA, generating a library of 10^9 possible mRNA x RBS pairs. The selection was then realized in two steps. First, a negative selection ensured that in the absence of a cognate artificial 16S rRNA, the artificial mRNA is not translated. This was realized through the expression of the uracil phosphoribosyltransferase gene (UPRT), which is toxic when 5-fluorouracil (5-FU) is introduced into the medium. Second, a positive selection through the expression of the chloramphenicol resistance gene which was fused to UPRT allowed finding ribosomes that effectively translate a cognate RBS. Three orthogonal pairs of ribosome x mRNA were identified and although this was not selected for, they are also orthogonal with one another.

I.4.c Gene network directed evolution

The first example of directed evolution applied to a system comprising several genes came from the team of Stemmer (Crameri, Dawes et al. 1997). They showed that the sexual-PCR method they had developed for gene shuffling could also be applied to operons. As a proof of principle, they improved in *E. coli* an arsenate resistance operon isolated from *Staphylococcus aureus*. Point mutations leading to substitutions were selected in 2 of the 3 genes of the operon and increased the resistance up to 40 fold. Along these lines, F. Arnold's team applied directed evolution to the production of new carotenoids to show that assembling genes into a metabolic pathway and evolving key enzymes is an efficient strategy for the synthesis of new compounds in *E. coli* (Schmidt-Dannert, Umeno et al. 2000).

The main problem for the directed evolution of systems comprising multiple genes is that the space of possibilities is enormous. In every directed evolution experiment, the mutagenesis technique must maximize the number of phenotype explored within a limited population. This entails limiting redundancy as much as possible. In other words, we need to make sure that we do not waste individuals by exploring mutations of which we know the outcome, or by exploring several mutations that will have the same outcome. Therefore, the better our knowledge of a system is, the better we will be able to target mutations. In protein directed evolution, the evolutionary landscape one usually wants to explore is all the possible amino-acids changes at a given number of specific and/or random positions. One thus wants

to avoid stop codons, of which we know the outcome, and synonymous mutations, which explore identical points in the space of possibilities.

However, in gene networks, not only is it possible to vary the individual proteins activity and specificity, but the outcome of the circuit will also depend on the relative level of genes expression, on the possible interactions between each protein, between each protein and their respective genes, and even possible interactions with and between RNAs. Engineers have thus devised several strategies to match the desired space of possibilities to what can actually be explored with reasonable population sizes.

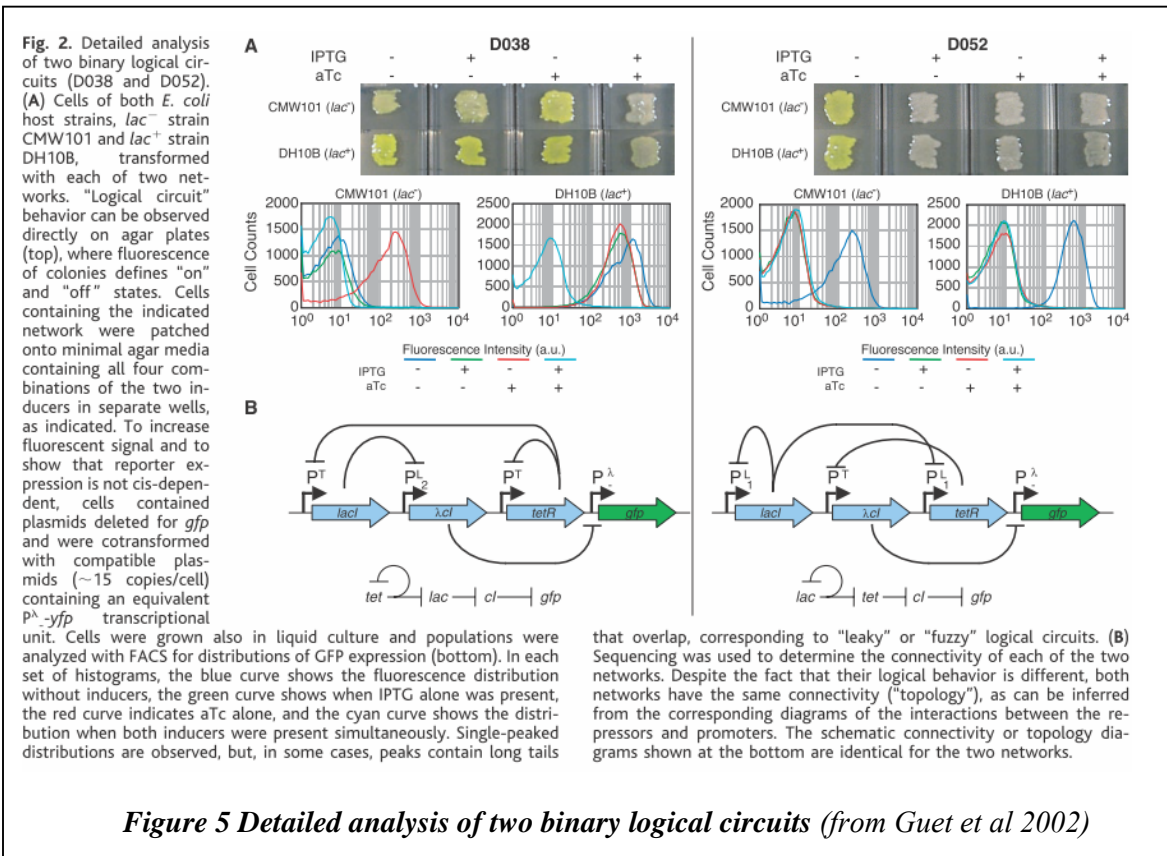### i. Fixing a regulatory networks by directed evolution

In 2000 the papers of the first synthetic genetic circuits (the repressilator of Elowitz (Elowitz and Leibler 2000) and the switch of Gardner (Gardner, Cantor et al. 2000)) attracted much attention toward artificial networks of transcription factors (Figure 1). In order to obtain a specified behavior, the traditional engineering approach consists in constructing a circuit from different functional parts in a plug and play fashion. Unfortunately, this almost never works at first trial, and the few published successes are only the tip of an iceberg of unpublished failures. In regulatory circuits one of the major problems is that the input range of one part will often not match the output of another part. It has therefore been proposed to use directed evolution in a second step to fix non-functional circuits. This was achieved by Yokobayashi and colleagues on a simple circuit consisting in a constitutively expressed *lacI* gene repressing the expression of CI, itself repressing the expression of a reporter GFP (Yokobayashi, Weiss et al. 2002). When this circuit was first constructed, no GFP was expressed at any IPTG input concentration. The expression of CI from the leaky P_{lac} was sufficient to repress GFP expression even in the absence of inducer. Random mutagenesis was realized on the CI gene and functional circuits were selected by assessing their fluorescence with or without IPTG.

By limiting mutations to a specific region, one can rapidly test how that part of the circuit contributes to overall circuit function, and find functional variants. This was further demonstrated by the Beijing iGEM team who published a genetic circuit behaving as a push-on push-off switch controlled by UV light (Lou, Liu et al.). The design consisted in a memory module based on the toggle switch (CI and CI434

mutually repressing each other) plugged to a NOR logic gate (a promoter repressed by LexA or LacI, and controlling the expression of CI_{ind}-). The toggle switch and the NOR gate were constructed by rational design, but the interconnection of the two modules didn't work at first and required a directed evolution procedure to ensure that the output of each module was compatible with the input of the other module. The RBS of the interconnecting parts *lacI* and CI_{ind}- were mutagenized and a screen was designed to select circuits behaving as wanted, i.e switching back and forth between the two stable states (each expressing a different fluorescent reporter) upon induction with UV light.

ii. Combinatorial synthesis of regulatory networks

All these approaches are based on the introduction of point mutations in the elements that are suspected to cause the initial failure of the system. The space of possibilities explored only consisted in changing relative gene expression and transcription factor binding affinity, but the network architecture remained unchanged. If one wants to explore new relations betweens the genes of a network, then point mutations won't do.

Fig. 2. Detailed analysis of two binary logical circuits (D038 and D052). (A) Cells of both *E. coli* host strains, *lac⁻* strain CMW101 and *lac⁺* strain DH10B, transformed with each of two networks. "Logical circuit" behavior can be observed directly on agar plates (top), where fluorescence of colonies defines "on" and "off" states. Cells containing the indicated network were patched onto minimal agar media containing all four combinations of the two inducers in separate wells, as indicated. To increase fluorescent signal and to show that reporter expression is not cis-dependent, cells contained plasmids deleted for *gfp* and were cotransformed with compatible plasmids (~15 copies/cell) containing an equivalent Pλ-*yfp* transcriptional unit. Cells were grown also in liquid culture and populations were analyzed with FACS for distributions of GFP expression (bottom). In each set of histograms, the blue curve shows the fluorescence distribution without inducers, the green curve shows when IPTG alone was present, the red curve indicates aTc alone, and the cyan curve shows the distribution when both inducers were present simultaneously. Single-peaked distributions are observed, but, in some cases, peaks contain long tails that overlap, corresponding to "leaky" or "fuzzy" logical circuits. (B) Sequencing was used to determine the connectivity of each of the two networks. Despite the fact that their logical behavior is different, both networks have the same connectivity ("topology"), as can be inferred from the corresponding diagrams of the interactions between the repressors and promoters. The schematic connectivity or topology diagrams shown at the bottom are identical for the two networks.

Figure 5 Detailed analysis of two binary logical circuits *(from Guet et al 2002)*

Combinatorial approached are needed. Guet and colleagues synthesized with a randomized ligation assembly method all the 125 possible networks with varying connectivity in a circuit consisting of 3 transcription factors (LacI, TetR, and lambda CI) each controlled by one of 5 possible promoters (P_{lac1}, P_{lac2}, P_{Tet}, $P_{\lambda+}$, $P_{\lambda-}$) (Guet, Elowitz et al. 2002). The authors were mainly interested in evaluating how the different networks would respond to information inputs. They assessed the behaviour of each circuit in response to the two inductors, IPTG and anhydrotetracyclin (aTc), and measured the output with a GFP reporter gene under the control of $P_{\lambda-}$. They looked for circuits where the output was a binary logical function of both inducers. They could isolate "logical circuits" behaving as NAND, NOR, or NOT IF gates (Table 1). Interestingly, circuits with a same architecture but different components could behave very differently and the other way round: several circuits with different architectures behaved similarly (Figure 5).

This work showed that circuit's behaviours cannot simply be inferred from boolean models based solely on their architecture, as was previously attempted (Thomas 1973). Furthermore, this approach proved that combinatorial approaches can yield libraries of functional circuits from which we may chose the one corresponding to our specifications. However, only very few different logic gates were observed and most of the circuits behaved similarly.

Another approach to create new connectivity and new signal integration in a network is to engineer new promoters controlled by several transcription factors. Such promoters are commonly used in nature to integrate several signals, a classical example being the one of the Lac operon which is regulated by LacI and CRP. The response of a promoter to several transcription factors will depend on the relative position of the operator sequences in a manner that is hard to predict. Cox and colleagues proposed a synthetic library-based approach for construction and analysis of modular combinatorial promoters (Cox, Surette et al. 2007). They divided promoters into 3 regions: the 45-bp region upstream of the -35 box (distal), the 25-bp region between the -35 and -10 boxes (core), and the 30-bp region downstream of the -10 box (proximal). For each position (distal, core, and proximal), they designed 16 units containing or not operator sequences of the AraC, LuxR, LacI, or TetR transcription factors, and assembled the 16*16*16=4096 possible promoters following a randomized assembly ligation method.

Figure 5 Combinatorial promoter architecture reveals rules for programming gene expression. The architecture and function of several dual-input promoters. The architecture of each promoter (colored according to Figure 1) is shown with its functional operators and −10 and −35 boxes. (**A**) RR promoters respond to both LacI and TetR. The fourth induction column (+ IPTG, + aTc) corresponds to the unregulated state. (**B**) AR promoters respond to AraC and one of the two repressors, as indicated. Here, the third column (+ IPTG/aTc, −Lara) corresponds to the unregulated state.
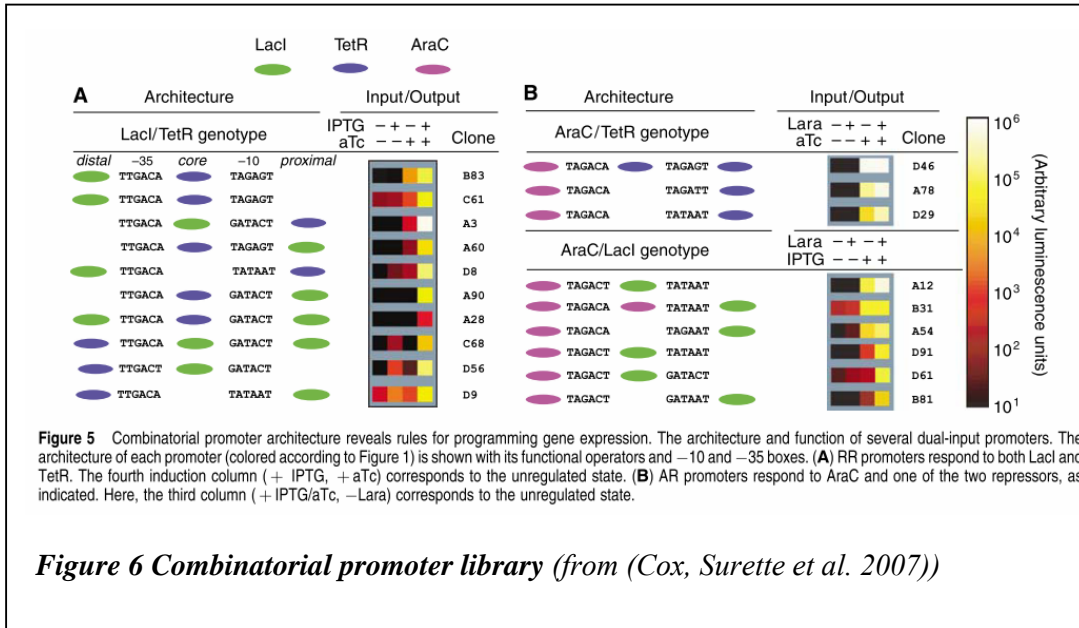
Figure 6 Combinatorial promoter library *(from (Cox, Surette et al. 2007))*

They analysed the behaviour of 288 randomly chosen networks in response to each of the 16 combinations of the four chemical inducers (Arabinose, aTc, IPTG and oxo-C6-homoserine lactone (VAI)). This study allowed to confirm previous studies about promoters functioning and enabled to gain new insights as well. For instance, activation worked only when operator sequences were present at the distal position, whereas repression worked best in the core region. Repression always dominated activations, and further design rules could be inferred from this analysis. A wide variety of promoter logic types was observed among the 50 promoters that displayed dual-input logic (Figure 6).

Such approaches provide new components for synthetic gene circuits and unveil at the same time design rules of promoters that can then be used in models to achieve better computational prediction of promoter regulation and activity. This was recently realized by Gertz and colleagues, who synthesized and analyzed 2087 different promoters using 18 different building blocks in yeast (Gertz, Siggia et al. 2009), and show that a thermodynamic model based on protein–DNA and protein–protein interactions can generally accurately predict the expression driven by combinations of binding sites. This is again a perfect example of how combinatorial approaches can work hand in hand with rational methods to increase our knowledge of biological systems and create new building blocks for synthetic circuit engineering.

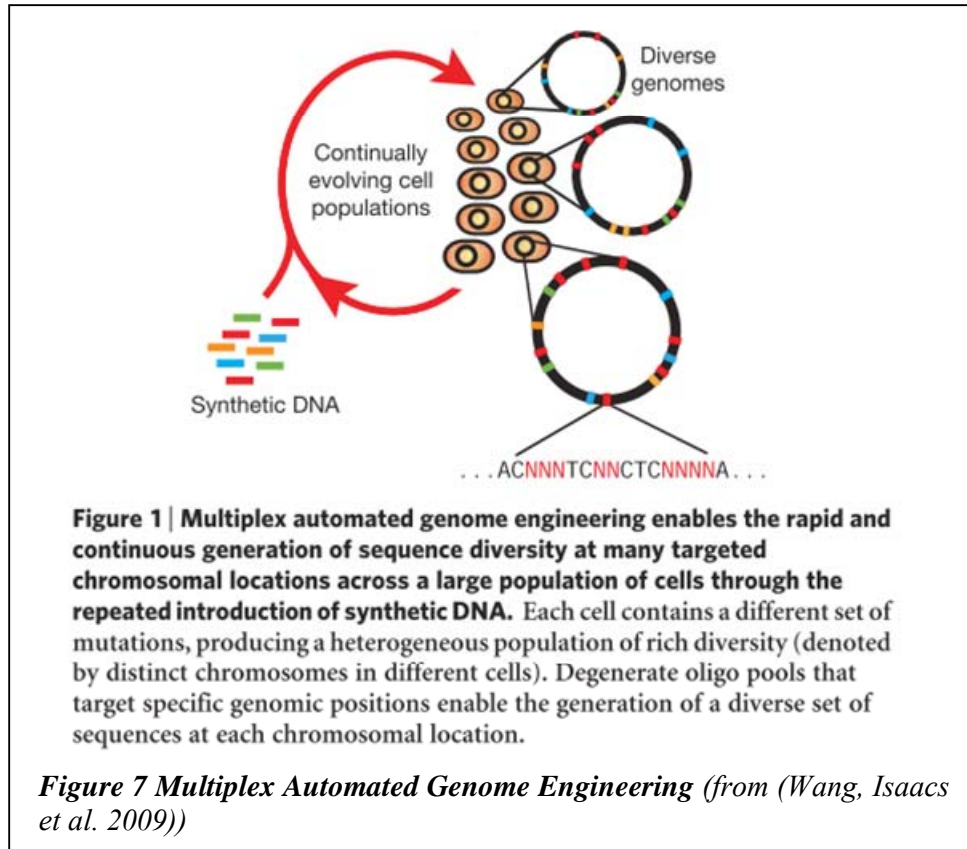### iii. Combinatorial methods in metabolic engineering

Metabolic engineering is one of the cornerstones of synthetic biology. As mentioned above, directed evolution methods were used early on to improve the activity of single enzymes in pathways of interest, and for the synthesis of new compounds through the evolution of new enzyme specificities. Nevertheless, targeting one enzyme at a time in a particular pathway can be extremely time consuming and often requires good knowledge of the limiting steps in the synthesis process. One of the flagship projects of synthetic biology is the engineering of plant isoprenoids in microbial strains.

Isoprenoids constitute a very diverse group of natural products that have various agronomical, industrial and medical applications. Amongst the most famous isoprenoids, we find carotenoids (with many applications in food, animal feed and cosmetics), paclitaxel also known as Taxol® (an effective anti-cancer chemotherapy drug) and arteminsine (the most frequently used drug against malaria) (Kirby and Keasling 2009). Pathway fluxes are improved in microbial strains by two main means, the knockout of genes that reroute metabolites to other ends, and the improvement of relative gene expression in the pathway to suppress bottlenecks and the potential accumulation of toxic intermediates.

However, there can be numerous combinations of potential gene knockouts and their outcome can be hard to predict. Moreover balancing the expression of several genes is a very complex task with an almost infinite space of possibilities. Combinatorial approaches have recently been extensively used in the engineering of the production of isoprenoids in *E. coli* in order to improve pathway fluxes. This was realized by two means. In their study, Alper and colleagues used modeling to identify potential knockout targets, predicted to increase cofactor or precursor supply for the lycopene production pathway (Alper, Miyaoku et al. 2005). Seven genes were combinatorially knocked out using transposons, and 64 combinations were assessed for improved production. The best combination yielded a 8,5-fold product increase over the *E.coli* K12 wild type strain.

Another study proposed a new method called MAGE (multiplex automated genome engineering) for the combinatorial introduction of random substitutions, insertions or deletions of a few base pairs at several target sites simultaneously

(Wang, Isaacs et al. 2009). The method relies on rounds of transformation of designed oligonucleotides into an engineered strain. The introduced oligonucleotides were homologous to target sites excepted for the introduced mutations. In this procedure, each oligo can mutagenize its target site with a probability of $\sim 10^{-1}$ per round, or lower (depending on the introduced mutation). This method was used to improve lycopene production by a factor of 5 in only 3 days during which 35 automated transformation cycles were accomplished (Figure 7).

Figure 1 | Multiplex automated genome engineering enables the rapid and continuous generation of sequence diversity at many targeted chromosomal locations across a large population of cells through the repeated introduction of synthetic DNA. Each cell contains a different set of mutations, producing a heterogeneous population of rich diversity (denoted by distinct chromosomes in different cells). Degenerate oligo pools that target specific genomic positions enable the generation of a diverse set of sequences at each chromosomal location.

Figure 7 Multiplex Automated Genome Engineering (from (Wang, Isaacs et al. 2009))

24 sites were simultaneously targeted: 20 of them were RBS of genes involved in the lycopene production pathway that were targeted with random oligos of the form (5'homolgy-DDRRRRRDDDD-3'homology, D=(A, G, T); R=(A, G)), and 4 of them were genes targeted for knockout by oligos introducing non-sense mutations. An estimated 15 billion genetic variants were generated and screening of variants was done by isolating colonies that produced intense red pigmentation. The MAGE method presents an important step forward in our ability to introduce mutations at several loci simultaneously and rapidly, and we can expect similar methods to be broadly used in the future for the accelerated directed evolution of microorganisms.

Balancing gene expression in a pathway can be achieved at different levels. We have seen above how RBS could be targeted to play on translation efficiency. Alternatively, transcription efficiency can be modified, either through promoter or transcription machinery engineering. For instance, it was shown that the σ70 sigma factor can be engineered to allow for global perturbations of the transcriptome in a directed manner. Ethanol tolerance, metabolite overproduction, and multiple phenotypes were improved by this method (Alper and Stephanopoulos 2007). In this study, an additional copy of the *rpoD* gene encoding σ70 was cloned on a plasmid and mutagenized through error-prone PCR. A single round of mutagenesis and selection for improved lycopene production enabled to select for improved mutants that were outperforming existing mutants formerly obtained after multiple rounds of knockouts and overexpression modifications. Interestingly, it was also shown that this method can be used to engineer multiple phenotypes. A library of *rpoD* was selected for improved ethanol resistance and another for improved SDS resistance. Surprisingly the combined expression of both selected *rpoD* genes in the same strain displayed a markedly improved resistance to both ethanol and SDS, and performed better than single *rpoD* mutants selected for both ethanol and SDS resistance simultaneously. Global transcription machinery engineering (gTME) appears to be a very promising route to modify the expression of multiple genes simultaneously and to select for improved phenotypes.

Improved balance in gene expression can also be achieved at the mRNA level. Multiple genes of the same pathways are often found in operons, i.e. their expression is driven by the same promoter. Several factors will impact the relative expression of genes present on the same transcription unit. First, the different genes may have different translation efficiencies that may not be directly correlated to individual RBS strength. Indeed in some cases the start codon of one gene can overlap the stop codon of the previous gene (for instance **TA**<u>A</u>TG or <u>**ATGA**</u>**,** start codon is underlined and stop codon is in bold font). Such configurations have been shown to enable translation restart: the ribosome that translated the previous gene can directly initiate translation of the next gene regardless of the strength of the RBS. Furthermore, when genes do not overlap, the intergenic region can influence the mRNA stability through the presence of secondary structures or RNAse cleavage sites. Based on this, a combinatorial approach was developed to improve expression balance in synthetic

operons (Pfleger, Pitera et al. 2006). Libraries of tunable intergenic regions (TIGRs) combining various secondary structures and RNAse cleavage sites were cloned in an operon of three genes encoding a heterologous mevalonate biosynthetic pathway. This method allowed selecting sevenfold increase in mevalonate production.

Finally gene order in operon will also affect the relative gene expression, since the further away a gene is from the promoter the less chance it has to be transcribed. An assembly method called OGAB (ordered gene assembly in *B.subtilis*) was used to assemble the five genes of the zeaxanthin pathway (a carotenoid) in five different orders. Gene order was shown to affect the mRNA expression levels of the different genes, and zeaxanthin level consequently varied fourfold between the five operons.
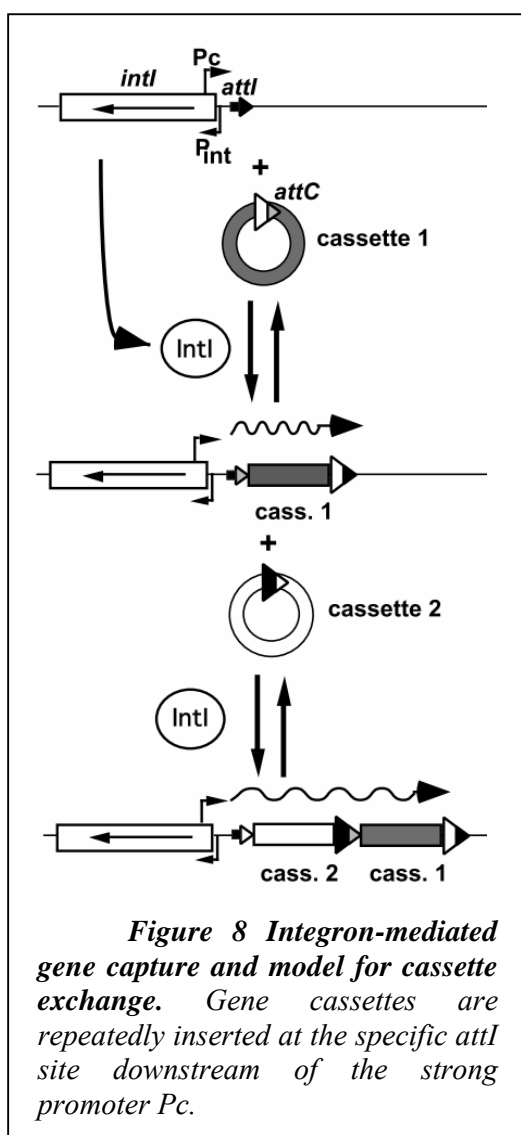
iv. Combinatorial DNA assembly

Exploring different possible gene orders requires time consuming cloning work, and only a very limited number of combinations can be assessed. New methods of DNA assembly have recently been developed that must make this kind of approach more readily exploitable. The SLIC (sequence and ligation–independent cloning) method harnesses homologous recombination *in vitro* for the simultaneous assembly of up to 10 fragments (Li and Elledge 2007). The DNA assembler method achieves similar assemblies taking advantage of *Saccharomyces cerevisiae* homologous recombination machinery to assemble whole metabolic pathways in a single step (Shao, Zhao et al. 2009). Engineers have also started to make use of high-throughput robotic platforms to assemble large DNA construct through classical but automated cloning methods such as the BioBrick[TM] assembly standard (Densmore, Hsiau et al. 2010). However, even though these methods make the cloning of large constructs much quicker, they lack the power to harness large numbers of combinations. Robots may allow the rapid construction of hundreds of combinations simultaneously, but this will hardly be enough as soon as more than five/six genes are involved. For instance, there are 6!=720 possible arrangements of 6 genes and 7!=5040 possible arrangements of 7 genes. In order to explore thousands of variants in the design space, new methods are thus required. Part of my thesis work was dedicated to the construction of an *in vivo* recombination device based on integrons to address this particular point, and to propose a new solution to this challenging problem.

II. Integrons

Integrons are genetic platforms able to capture, stockpile and rearrange gene cassettes by a site-specific recombination mechanism. They were first identified in the late 80s, when they were found to be responsible for the gathering of antibiotic resistance genes in mobile elements (Martinez and de la Cruz 1988; Stokes and Hall 1989).

II.1 Structure of integrons

Figure 8 Integron-mediated gene capture and model for cassette exchange. Gene cassettes are repeatedly inserted at the specific attI site downstream of the strong promoter Pc.

Integrons are composed of a gene coding for the tyrosine recombinase IntI and a primary recombination site *attI*. The integrase is able to excise and capture discrete genetic elements, known as gene cassettes, that are characterized by the presence of an *attC* recombination site (formerly called 59-base element) (Stokes and Hall 1989). Successive integration of cassettes at the *attI* site results in the formation of cassette arrays that constitute the variable part of the integron (Collis, Grammaticopoulos et al. 1993; Recchia, Stokes et al. 1994)(Figure 8). Cassette integration occurs through intermolecular recombination between an *attC* site and the *attI*, while excision occurs either through intramolecular recombination between the *attI* and an *attC* site, or between two *attC* sites. The excision of cassettes by the IntI integrase leads to non-replicative covalently closed circular intermediates

(Collis and Hall 1992), that can eventually be recaptured by the integron. Integration of circular intermediates has been shown to preferentially occur at the *attI* site compared to an arbitrary *attC* site within the array (Collis, Grammaticopoulos et al. 1993; Collis, Recchia et al. 2001). Such excision and reintegration events result in a reordering of the cassettes in the array.

Cassettes are generally constituted by a single ORF immediately followed by an *attC* site. ORFs are found generally promoterless. Expression of the first cassettes of the array is driven by a promoter located upstream of the *attI* site, called the Pc (Levesque, Brassard et al. 1994; Collis and Hall 1995; Jove, Da Re et al. 2010).

II.2 The different integrons types

The first integrons discovered are termed "mobile" integrons, because of their association with transposons, and constitute the major vectors of antibiotic multi-resistance in gram-negative and to a lesser extent in gram-positive bacteria (Partridge, Tsafnat et al. 2009). In the late 90s, a new type of integrons that are not involved in resistance phenotype was identified: sedentary chromosomal integrons. They are found in the genome of a significant fraction of environmental bacteria (Vaisvila, Morgan et al. 1999; Rowe-Magnus, Guerout et al. 2001; Boucher, Labbate et al. 2007); the first discovered being the integron of *Vibrio cholerae* (Mazel, Dychinco et al. 1998). It is now clear that this chromosomal integrons have been maintained in the genome of gram-negative bacteria for a long while, to help facing changing environments and that these chromosomal elements are the source of the mobile integrons' backbones and of their antibiotic resistance gene cassettes (Mazel 2006).

II.2.a Multiresistant integrons

Five different classes of mobile integrons have been defined to date, based on the sequence of the encoded integrases (40–58% identity). They all have in common to be associated with transposons and to carry essentially antibiotic resistance genes. A pool of >130 different cassettes harboring various antibiotics resistance genes have been identified in mobile integrons (based on a 98% nucleotide identity threshold) (Fluit and Schmitz 2004; Partridge, Tsafnat et al. 2009). Together, these cassettes provide resistance to most classes of antibiotics including β-lactams, all

46

aminoglycosides, chloramphenicol, trimethoprim, streptothricin, rifampin, erythromycin, fosfomycin, lincomycin, quinolones and antiseptics of the quaternary-ammonium-compound family (Fluit and Schmitz 2004; Mazel 2006; Partridge, Tsafnat et al. 2009).

However, several mobile integrons that are not associated with resistance genes have been recovered in environmental bacteria (Stokes, Nesbo et al. 2006; Xu, Davies et al. 2007; Gillings, Boucher et al. 2008). This suggests that mobile integrons are not specifically dedicated to antibiotic resistance, but are likely to be broadly involved in mediating bacterial adaptation. The prevalence of resistance functions probably results from biased sampling focused on clinically relevant environment and reflects the evolutionary success of integrons in these settings.

Class 1 integrons are associated with functional and non-functional transposons derived from Tn402 which can be further embedded in larger transposons, such as Tn21. They represent the most widespread and clinically important class of integrons, as they are detected in 22 to 59% of Gram-negative clinical isolates (Labbate, Case et al. 2009)}, and they have also been occasionally identified in Gram-positive bacteria (Martin, Timm et al. 1990; Nesvera, Hochmannova et al. 1998; Nandi, Maurer et al. 2004; Shi, Zheng et al. 2006).

Class 2 integrons are almost exclusively associated with Tn7 derivatives and show a dozen of different cassette arrays (Biskri and Mazel 2003; Ramirez, Pineiro et al. 2010). The integrase gene of class 2 integrons, *intI2*, generally contains a nonsense mutation in codon 179 that yields a non-functional protein, which can be rescued by a single mutation (Hansson, Sundstrom et al. 2002). However, it is not known whether the cassette recombination in the different Tn7 derivatives is mostly due to occasional natural suppression of the ochre179 codon, leading to an active integrase or by the trans acting recombination activity of another IntI, such as IntI1, which has been shown to recognize and recombine the *attI2* site of class2 integron (Collis, Kim et al. 2002; Hansson, Sundstrom et al. 2002). Class 2 integrons encoding a functional integrase have been isolated on two occasions (Barlow and Gobius 2006; Marquez, Labbate et al. 2008).

Class 3 integrons (Arakawa, Murakami et al. 1995) are also thought to be located in a transposon (Collis, Kim et al. 2002) and are less prevalent than class 2.

The other two classes of mobile integrons have been identified through their role in the development of trimethoprim resistance in *Vibrio* species.

The class 4 integron is embedded in a subset of the integrative and conjugative element SXT found in *Vibrio cholerae* (Hochhut, Lotfi et al. 2001). The class 5 is located in a compound transposon carried on the pRSV1 plasmid of *Alivibrio salmonicida* (GenBank AJ277063 (Sorum, Roberts et al. 1992)).

II.2.b Chromosomal Integrons

Chromosomal integrons (CI) have been found in several bacterial phyla. 17% of the 1189 complete bacterial genomes available at the NCBI in March 2010 carry an integron integrase. A phylogenic analysis revealed that the branching pattern of integrases is in good agreement with the organismal phylogeny. This clearly shows that integrons are ancient and generally stable genomic structures (Rowe-Magnus, Guerout et al. 2001; Rowe-Magnus, Guerout et al. 2003; Mazel 2006; Nemergut, Robeson et al. 2008). Three major groups were identified: i) the *soil-freshwater proteobacteria* group of integrons, mostly composed of proteobacteria from freshwater and soil environments; ii) the *marine γ-proteobacteria* group; and iii) the *inverted integrase* group, characterized by the co-linear orientation of the integrase with respect to the cassette array, so that the *attI* site is found in the 3' end of the integrase.

The number of cassettes present in chromosomal integrons is very variable. Some integron also termed superintegrons can contain up to 217 cassettes (*V. vulnificus*) while other contain only few or even no cassettes. The vast majority of the ORFs contained in cassettes are unknown genes, and the few that have known homologues are involved in a variety of functions (iron scavenging, DNA modification, isochorismatases, acetyltransferases, methylases…) (Rowe-Magnus, Guerout et al. 2003). A significant part of the encoded proteins carry a signal peptide region or/and a transmembrane domains (Rowe-Magnus, Guerout et al. 2003; Koenig, Boucher et al. 2008). Altogether, these data indicate that chromosomal integrons carry important functions to mediate interactions with external environments. Finally, a special category of cassette was identified. Toxin-antitoxin modules are often found in large integrons. They consist in two genes, one encoding a stable toxin and the other

48

an unstable antitoxin. If the cassette is lost, or its expression disturbed, the antitoxin is quickly degraded and the toxin acts as a bacterio-static. These module are also referred to as post-segregation killing systems has they were identified to stabilize plasmids carrying them. Toxin-antitoxin systems have been shown to stabilize integron with large cassette arrays. For instance 13 of them are found in *V.cholerae* superintegron (Heidelberg, Eisen et al. 2000; Pandey and Gerdes 2005).

Superintegrons are also characterized by the high homology existing between the cassettes *attC* sites (>80%), which allows to define species specific *attC* sites. This is the case of most CIs found in *Vibrio* species genomes, in *Pseudomonas alcaligenes* and related, in several *Xanthomonas* and in *Treponema denticola* (Rowe-Magnus, Guerout et al. 2001; Vaisvila, Morgan et al. 2001; Rowe-Magnus, Guerout et al. 2003; Coleman, Tetu et al. 2004; Gillings, Holley et al. 2005). In contrast, mobile integrons present very diverse *attC* sites that can sometime be identified as coming from various superintegrons. This suggests that mobile integrons are able to capture cassettes as they move from one organism to the other. It was for instance experimentally shown that a class 1 mobile integron could capture a chloramphenicol resistance cassette present in the *V.cholerae* superintegron (*catB9* (Rowe-Magnus, Guerout et al. 2002)).

Mobile integrons arose several times independently from CIs. Indeed, their integrase do not branch together, but are scattered in the phylogenic tree (Rowe-Magnus, Guerout et al. 2001). Several scenarios have been proposed that explain how CIs could have been mobilized by mobile elements inserting nearby (Labbate, Case et al. 2009).
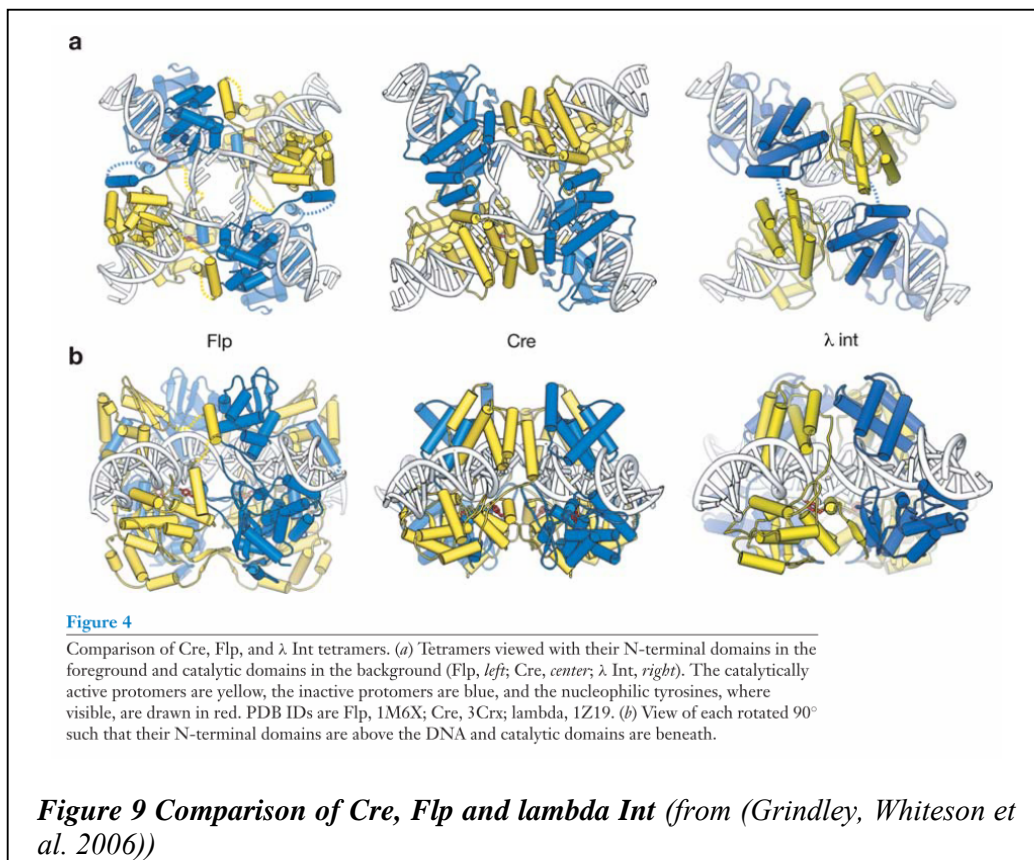
II.3 Integron recombination mechanism

II.3.a Tyrosine recombinases

The integron integrase is a member of the tyrosine recombinase family. Other members of this family include the phage lambda integrase, the Cre recombinase of the P1 phage, the Flp recombinase of the *Saccharomyces cerevisiae* 2-µm plasmid and the XerCD resolvases responsible for solving chromosomes concatemers that may arise through homologous recombination during replication. Their typical core
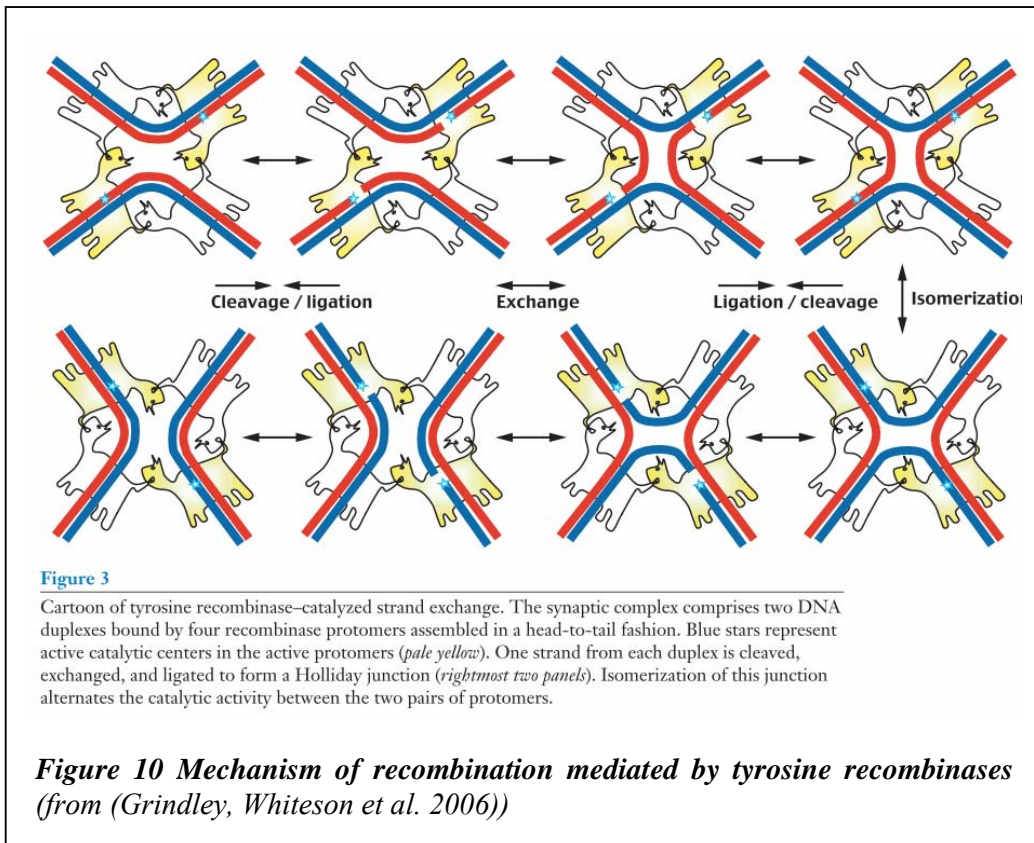
recombination sites consist of a pair of highly conserved 9-13 bp inverted binding sites separated by a 6-8 bp central region where the strands exchange occurs. Additional accessory sequences can be found and are implicated in host factors binding contributing the proper conformation of the recombination site.

The tyrosine recombinases present C-terminal catalytic domain with recognizable sequence motifs. In particular a nucleophilic tyrosine is responsible for attacking the phosphodiester bound of DNA and initiate strand exchanges. The less conserved N-terminal domain plays a role in DNA-protein interaction (recognition of the binding sites) and protein-protein interaction (formation of dimers and tetramers). Tyrosine recombinases indeed act as tetramers, tow monomers binding each recombination site. Although the amino-acid sequence of the various members of the tyrosine recombinase family is poorly conserved, their 3D structure shows comparable organization (Figure 9). They all form C-shaped clamps around the DNA substrate.

Figure 4

Comparison of Cre, Flp, and λ Int tetramers. (*a*) Tetramers viewed with their N-terminal domains in the foreground and catalytic domains in the background (Flp, *left*; Cre, *center*; λ Int, *right*). The catalytically active protomers are yellow, the inactive protomers are blue, and the nucleophilic tyrosines, where visible, are drawn in red. PDB IDs are Flp, 1M6X; Cre, 3Crx; lambda, 1Z19. (*b*) View of each rotated 90° such that their N-terminal domains are above the DNA and catalytic domains are beneath.

Figure 9 Comparison of Cre, Flp and lambda Int *(from (Grindley, Whiteson et al. 2006))*

Recombination goes as follow (Figure 10): A synaptic complex consisting of tow recombination substrates and 4 recombinases is formed. Only tow opposing

monomers are active in a first step and recombination is initiated when one strand of each substrate is cleaved by the nucleophilic tyrosine. Covalent phosphotyrosine bounds are formed between the attacking monomers and the 3' ends of DNA, while the 5'OH ends remains free. The next steps consist in the 5'ends attacking the opposing 3'phosphodiester bounds resulting in a first strand exchange forming a Holliday junction. The complex can then isomerize and inactive monomers become active and vice versa. A second strand exchange can then proceed following the exact same mechanism. This last step resolves the junction, frees the proteins and the recombination reaction is complete.

Figure 3

Cartoon of tyrosine recombinase–catalyzed strand exchange. The synaptic complex comprises two DNA duplexes bound by four recombinase protomers assembled in a head-to-tail fashion. Blue stars represent active catalytic centers in the active protomers (*pale yellow*). One strand from each duplex is cleaved, exchanged, and ligated to form a Holliday junction (*rightmost two panels*). Isomerization of this junction alternates the catalytic activity between the two pairs of protomers.

Figure 10 Mechanism of recombination mediated by tyrosine recombinases (from (Grindley, Whiteson et al. 2006))

II.3.b The integron recombination sites

i. The attI primary recombination site

The *attI* site is present in only one copy adjacent to the integrase gene. No sequence consensus can be established for the *attI* site which differs greatly between integron classes (Rowe-Magnus, Guerout et al. 2001). It is minimally composed of two integrase binding sites termed L and R, the L box being always degenerated with respect to R. The recombination point is located in a conserved 5'-GTT-3' triplet between G and TT in the R box (Figure 11) (Martinez and de la Cruz 1990; Hall, Brookes et al. 1991). The *attI* site of the class 1 integron was well characterized by in vitro experiments. In addition to the L and R box, the integrase was shown to bind two direct repeats DR1 and DR2. These direct repeats are essential for the *attI* x *attC* recombination reaction. However *attI* x *attI* recombination, which happens 100x less frequently than *attC* x *attI*, do not require the DR1 and DR2 boxes. These repeats are not conserved in other *attI* sites, and their precise role is not clear.
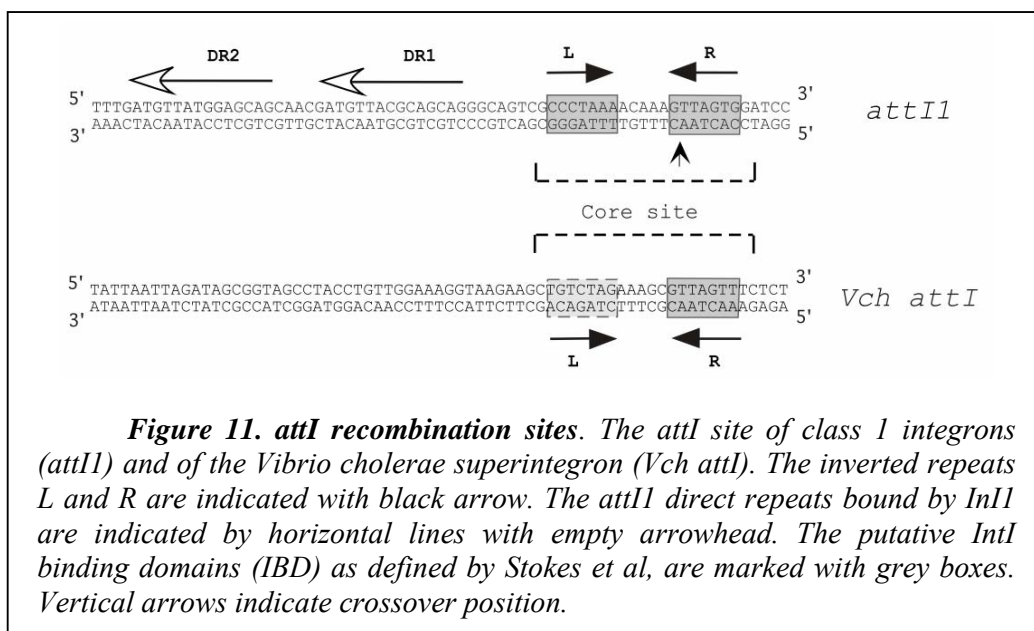
Figure 11. attI recombination sites. The attI site of class 1 integrons (attI1) and of the Vibrio cholerae superintegron (Vch attI). The inverted repeats L and R are indicated with black arrow. The attI1 direct repeats bound by InI1 are indicated by horizontal lines with empty arrowhead. The putative IntI binding domains (IBD) as defined by Stokes et al, are marked with grey boxes. Vertical arrows indicate crossover position.

ii. The attC recombination site of cassettes

attC sites define the integron cassettes. They are almost always present downstream of the coding sequence of the cassette and are in the same orientation. This organization ensures that upon integration of a cassette at the *attI* site, the ORF is in the proper orientation to be transcribed by the Pc promoter.

attC sites are very variable both in sequence and in size (57-147 bp). They consist in an inverted repeat with a variable central spacer sequence (20-104 bp)(Mazel 2006). In each repeat a degenerate core site is found, named R"-L" and L'-R' respectively. The only conserved sequence among all *attC* sites is a 5'-AAC-3' in the R" box and a 5'-GTT-3' in the R' box. Consistently with the *attI* site, the recombination point is located between the G en TT of the R' box (Figure 12). Gel-shift assays showed that the integrase binds *attC* sites only when they are single stranded, and that the bottom strand only is recognized (Francia, Zabala et al. 1999; Johansson, Kamali-Moghaddam et al. 2004). Experiments where either the bottom or the top strand of the *attC* site was delivered through conjugation to a recipient cell showed that the bottom strand is recombined 1000x time more efficiently than the top strand (Bouvier, Demarre et al. 2005). Furthermore, mutations disrupting the inverse repeat so that the *attC* site cannot fold a hairpin, strongly hindered recombination. From these experiments it was clear that *attC* sites recombine as a folded single strand that is recognized by the integrase.

In contrast to canonical core recombination sites, the genetic information required for proper recombination is not contained in the primary sequence of *attC* sites, but mostly in their secondary structures. All known folded single-stranded *attC* sites present an almost canonical core site consisting of R and L boxes (formed by the pairing between R"/R' and L"/L') separated by an unpaired central segment (UCS), two or three extrahelical bases (EHB) and a variable terminal structure (VTS) (Bouvier, Ducos-Galand et al. 2009) (Figure 12). The recognition of structural determinants of the *attC* hairpin by the integrase was further confirmed when the structure of the integrase bond to *attC* x *attC* synaptic complex was obtained (MacDonald, Demarre et al. 2006). These structural data have shown the importance of extra helical bases (EHB) inside the stem-loop structure formed from the bottom strand (MacDonald, Demarre et al. 2006), which are specifically interacting with hydrophobic pockets formed by IntIA residues, conserved among all integron integrases.
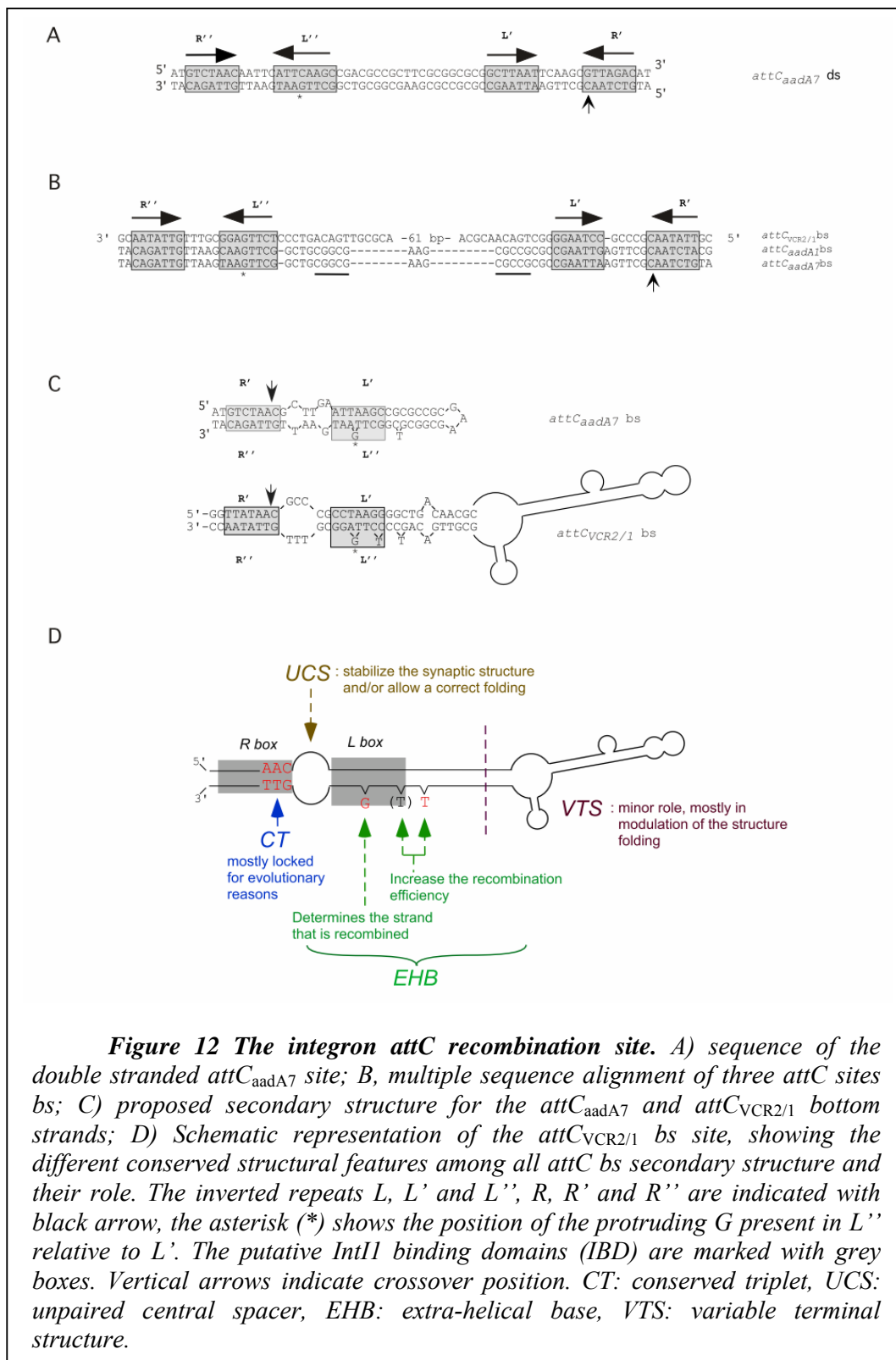
Figure 12 The integron attC recombination site. *A) sequence of the double stranded attC*$_{aadA7}$ *site; B, multiple sequence alignment of three attC sites bs; C) proposed secondary structure for the attC*$_{aadA7}$ *and attC*$_{VCR2/1}$ *bottom strands; D) Schematic representation of the attC*$_{VCR2/1}$ *bs site, showing the different conserved structural features among all attC bs secondary structure and their role. The inverted repeats L, L' and L'', R, R' and R'' are indicated with black arrow, the asterisk (*) shows the position of the protruding G present in L'' relative to L'. The putative IntI1 binding domains (IBD) are marked with grey boxes. Vertical arrows indicate crossover position. CT: conserved triplet, UCS: unpaired central spacer, EHB: extra-helical base, VTS: variable terminal structure.*

Bouvier and colleagues determined the precise contribution of the three structural elements (EHBs, UCS and VTS) for the strand choice and its recombination (Bouvier, Ducos-Galand et al. 2009) (Figure 12). This study showed that strand choice is primarily directed by the first EHB, while the presence of the two other EHBs also serves to increase strand selection but to a lesser extent. In particular, when an *attC* sequence is modified in such a way that the EHBs are appropriately oriented on the top strand in place of the bottom one, the integrase strand specificity is switched. This work also established that the structure of the UCS is essential to achieve high level of recombination of the bottom strand, suggesting a dual role for this structure in active site exclusion and for hindering the reverse reaction after the first strand exchange. Although the VTS can greatly influence the recombination frequency; it has apparently no role in strand selectivity. One of the objectives of my PhD work was to decipher the mechanisms by which the VTS influences recombination, and more precisely the *attC* site folding (see the first article in the results section).

Finally a recent study showed that even the conserved "GTT/ACC" core site is can undergo many mutations without altering the recombination efficiency (Frumerie, Ducos-Galand et al. 2010). The only requirement is that the first nucleotide of the triplet must be the same in both recombination partners. This sequence is thus likely conserved for evolutionary reasons rather than mechanistic ones.

iii. Secondary recombination sites

Recombination events can occur at secondary sites at low frequencies (Francia, de la Cruz et al. 1993; Recchia, Stokes et al. 1994; Francia and Garcia Lobo 1996; Francia, Avila et al. 1997; Hansson, Skold et al. 1997), and a small number of cassettes has been identified outside of integron platforms (Recchia and Hall 1995; Segal and Elisha 1997). Events have been detected at GNT sequences, which is similar to the consensus "GTT" of the R box (Recchia and Hall 1995; Francia, Avila et al. 1997). After an integration event at a secondary recombination site, the absence of other recombination sites nearby likely makes the cassette stable.

II.3.c The integron integrases structural properties

As mentioned above the integron integrases belong to the family of tyrosine recombinases (Nunes-Duby, Kwon et al. 1998). However, they all contain an additional domain in their C-terminal part (Boyd, Almagro-Moreno et al. 2009), which is essential for their activity (Messier and Roy 2001). This segment has a specific role in the folding of the hydrophobic pockets that stabilize the two EHBs involved in the specific recognition of the *attC* bottom strand (MacDonald, Demarre et al. 2006). This segment also carries an alpha helix, called I2 (MacDonald, Demarre et al. 2006), absent in all the other tyrosine recombinase. The analysis of the recombination activity of integrase mutants, together with its structure, enabled to establish the role of key catalytic residues (Figure 13)(Gravel, Messier et al. 1998; Collis, Recchia et al. 2001; Messier and Roy 2001; MacDonald, Demarre et al. 2006; Johansson, Boukharta et al. 2009). Furthermore, a directed evolution experiment targeted at improving the *attC* x *attI* recombination efficiency, enabled to demonstrate the critical role of the conserved aspartic acid in position 161 which is important for the multimer assembly (Demarre, Frumerie et al. 2007). Mutations of this residue could increase *attC* x *attI* recombination efficiency but at the same time decreased the efficiency of *attC* x *attC* events. These results indicate that protein/protein interactions in the recombination synapse are critical to the specificity of the reaction, and reveal a tradeoff between the different reactions catalyzed by the integrase.
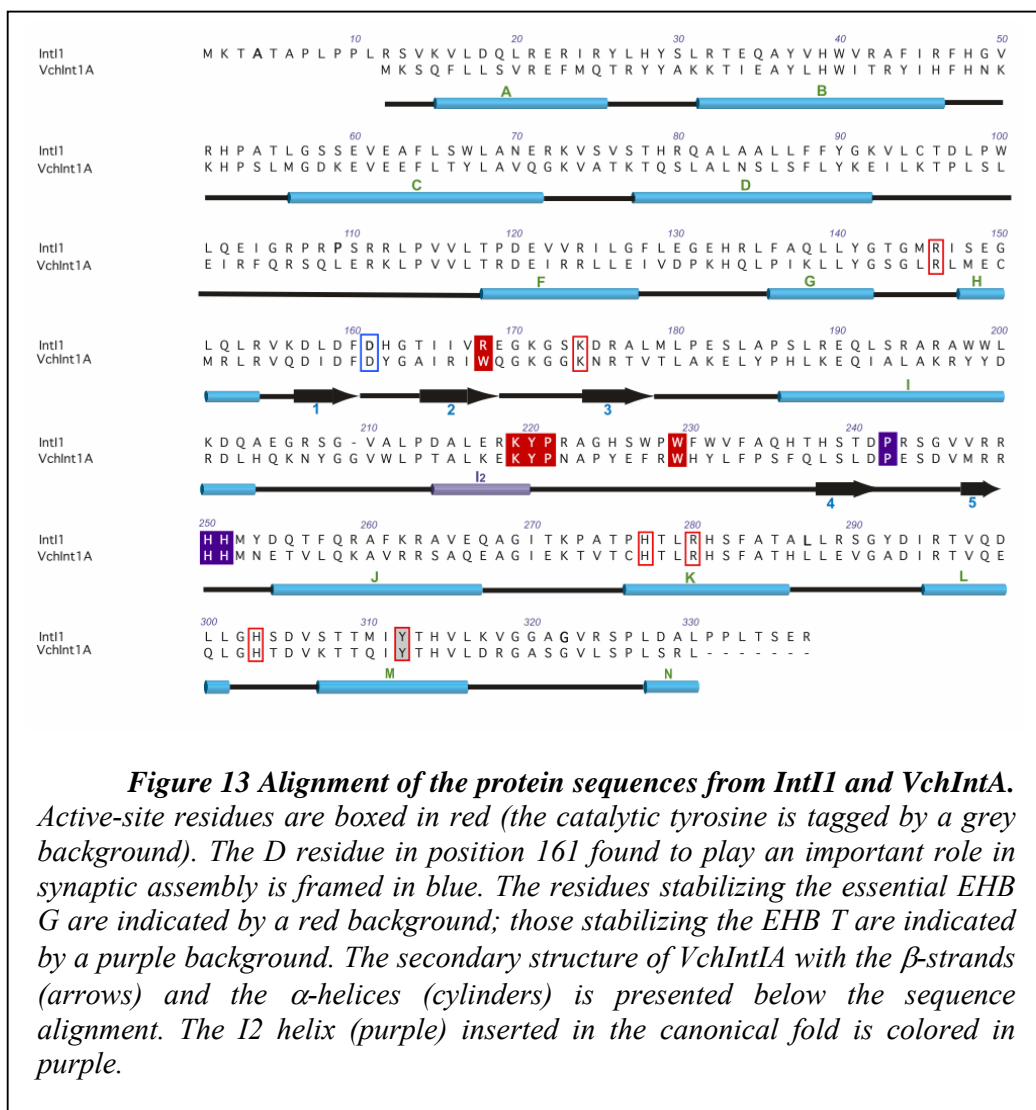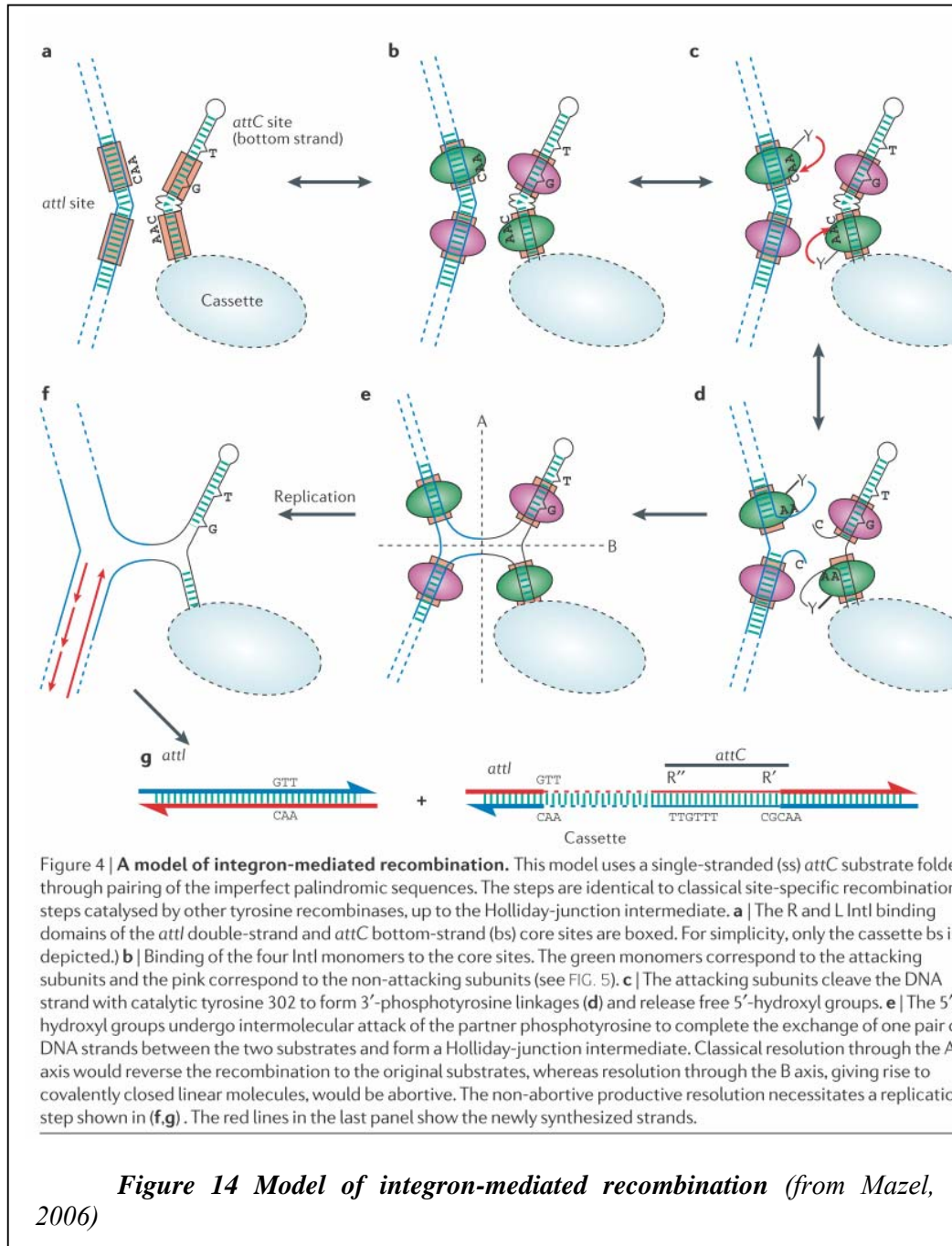
Figure 13 Alignment of the protein sequences from IntI1 and VchIntA.
Active-site residues are boxed in red (the catalytic tyrosine is tagged by a grey background). The D residue in position 161 found to play an important role in synaptic assembly is framed in blue. The residues stabilizing the essential EHB G are indicated by a red background; those stabilizing the EHB T are indicated by a purple background. The secondary structure of VchIntIA with the β-strands (arrows) and the α-helices (cylinders) is presented below the sequence alignment. The I2 helix (purple) inserted in the canonical fold is colored in purple.

II.3.c Single stranded recombination of *attC* sites

The *attI* sites are recognized under classical double stranded form (Francia, Zabala et al. 1999). A canonical site-specific recombination reaction between a single stranded *attC* and a double stranded *attI* would lead to abortive products, as the second strand exchange would generate linearized products with covalently closed ends. Bouvier and colleagues have thus proposed a model, where recombination stops after the first strand exchange and replication solves the junction (Bouvier, Demarre et al. 2005)(Figure 14). The tridimensional structure of the VchIntIA/*attC* synaptic complex gave some clue to understand how the second strand exchange is prevented. Indeed, two of the four subunits show a disorganization of their catalytic domain,

which pulls the catalytic tyrosine away from the phosphate links (MacDonald, Demarre et al. 2006). However, this models still requires experimental validation.

Figure 4 | **A model of integron-mediated recombination.** This model uses a single-stranded (ss) *attC* substrate folded through pairing of the imperfect palindromic sequences. The steps are identical to classical site-specific recombination steps catalysed by other tyrosine recombinases, up to the Holliday-junction intermediate. **a** | The R and L IntI binding domains of the *attI* double-strand and *attC* bottom-strand (bs) core sites are boxed. For simplicity, only the cassette bs is depicted.) **b** | Binding of the four IntI monomers to the core sites. The green monomers correspond to the attacking subunits and the pink correspond to the non-attacking subunits (see FIG. 5). **c** | The attacking subunits cleave the DNA strand with catalytic tyrosine 302 to form 3′-phosphotyrosine linkages (**d**) and release free 5′-hydroxyl groups. **e** | The 5′-hydroxyl groups undergo intermolecular attack of the partner phosphotyrosine to complete the exchange of one pair of DNA strands between the two substrates and form a Holliday-junction intermediate. Classical resolution through the A axis would reverse the recombination to the original substrates, whereas resolution through the B axis, giving rise to covalently closed linear molecules, would be abortive. The non-abortive productive resolution necessitates a replication step shown in (**f,g**). The red lines in the last panel show the newly synthesized strands.

Figure 14 Model of integron-mediated recombination *(from Mazel, 2006)*

58

II.4 Regulation of the integrase expression

Expression of the integrase (*intI*) has recently been shown to be controlled by the SOS regulon (Guerin, Cambray et al. 2009). The main regulator of SOS is the LexA transcription factor which represses the genes of the regulon (Kelley 2006). The main trigger of the SOS response is the presence of ssDNA in the cell. The formation of a RecA nucleofilament on ssDNA stimulates self-cleavage of LexA, leading to its inactivation. Promoters from the SOS regulon, controlling mostly DNA repair, recombination and mutagenic polymerases, are than de-repressed (Figure 23).

The fact that integron *attC* site recombines as ssDNA hairpins and that the integrase expression is triggered in the presence of ssDNA is certainly not a coincidence. Integrons appear to be integrated systems, which have evolved to sense environmental stresses and to recombine in response (this point is further discussed part III.4 and in the discussion). Guerin, Cambray and colleagues showed that some antibiotics known to induce the SOS response in Gram-negative and Gram-positive bacteria (Kelley 2006), such as quinolones, trimethoprim and beta-lactams, induce the integrase expression. This probably explains how rapidly multi-resistant integrons where able to recruit resistant cassettes to the very antibiotic that trigger the SOS response. Indeed, resistance cassettes to trimethoprim, quinolones and β-lactams, are prevalent in multi-resistant integrons (Fluit and Schmitz 2004; Partridge, Tsafnat et al. 2009).

III. Folded DNA in action: hairpin formation and biological functions in prokaryotes

This part was written as a review for Microbiology and Molecular Biology Reviews. It will hopefully be published under the same title by the time I defend my thesis.

The B-helix form of DNA proposed by Watson and Crick accounts for most of DNA's behavior in the cell. Nevertheless, it is now obvious that DNA isn't always present in this canonical structure, but can also form alternative structures such as Z-DNA, cruciforms, triple-helix H-DNA, quadruplex G4-DNA and slipped-strand DNA (Zhao, Bacolla et al. 2010). This review focuses on DNA hairpins, i.e. DNA with intrastrand base pairing, their functions and properties, in light of the specific behavior of DNA in horizontal gene transfer between bacterial cells.

Hairpin structures can be formed by sequences with inverted repeats (IRs), also termed palindromes, following two main mechanisms. Firstly, in several cellular processes, DNA is single-stranded (ssDNA); for instance, on the template for the synthesis of the lagging strand during replication, during DNA repair or, more importantly, during rolling circle replication, bacterial conjugation, natural transformation and virus infection. ssDNA is not simply a transient inert state of DNA, but can fold into secondary structures recognized by proteins, notably involved in site-specific recombination, transcription and replication. A second mechanism is the formation of hairpins from double-stranded DNA (dsDNA) as a cruciform, i.e. two opposite hairpins extruding through intrastrand base pairing from a palindromic sequence. The existence of cruciforms was already hypothesized soon after Watson and Crick's discovery (Platt 1955): negative supercoiling of double-stranded DNA (dsDNA) could provide free energy to stabilize cruciforms that would otherwise be unstable. Cruciforms then attracted much attention in the 1980's when their existence was experimentally assessed in vitro under natural superhelical densities (Panayotatos and Wells 1981). But most studies at that time rejected their possible implication in cellular processes because of the slow kinetics of cruciform formation, which made

60

them theoretically very unlikely to occur in vivo (Courey and Wang 1983; Sinden, Broyles et al. 1983). Nonetheless, this point of view was revised when techniques revealing cruciforms in vivo were developed and biological functions involving DNA secondary structures were discovered.
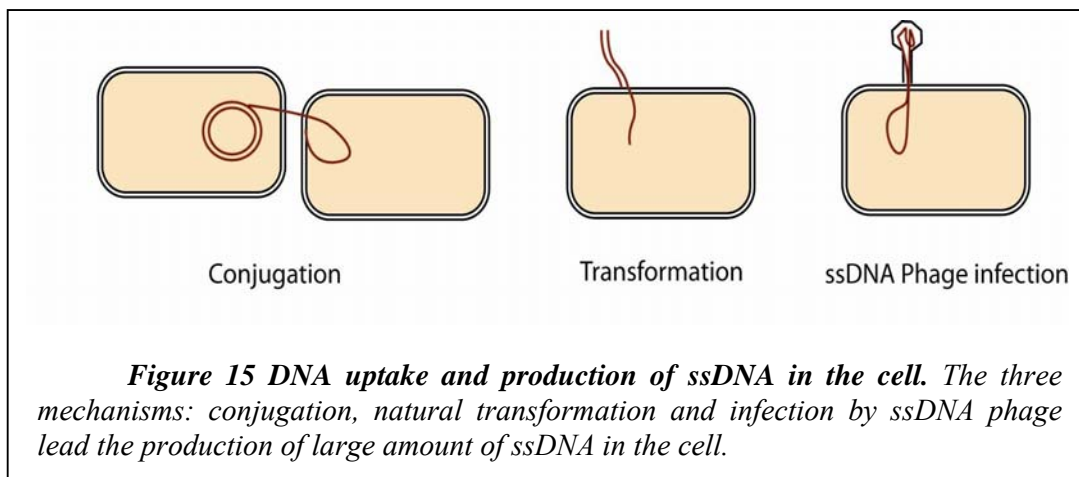
There are three ways in which DNA hairpins can interact with proteins and impact cell physiology. (i) Cruciform formation modifies the coiling state of DNA (White and Bauer 1987), which is known to affect the binding of regulatory proteins for transcription, recombination and replication (Cozzarelli and Wang 1990; Hatfield and Benham 2002); (ii) the DNA-protein interaction can be inhibited if a hairpin overlaps a protein recognition site (Horwitz and Loeb 1988). (iii) Proteins can directly recognize and bind DNA hairpins (Masai and Arai 1997; Val, Bouvier et al. 2005; MacDonald, Demarre et al. 2006; Gonzalez-Perez, Lucas et al. 2007; Barabas, Ronning et al. 2008).

We describe here the cellular processes leading to DNA hairpin formation, biological functions involving hairpins, and the mechanisms of protein-hairpin recognition. Finally, we try to shed light on the evolution of folded DNA with biological functions and their cognate proteins.

III.1 DNA hairpin formation

III.1.a Hairpin formation from ssDNA

The production of a large amount of single-stranded DNA (ssDNA) in the cell occurs mainly during the entry of exogenous DNA, macromolecular synthesis and repair. The three mechanisms of DNA uptake, namely, natural transformation, conjugation and, occasionally, bacteriophage infection, involve the production of ssDNA (Figure 15). The processes of replication and transcription also involve the unwinding of duplex DNA; finally, DNA repair can lead to the production of large quantities of ssDNA. The amount of single strand available, its lifetime and the bound proteins are different properties of these processes that may affect the possibility of hairpins to fold.

Figure 15 DNA uptake and production of ssDNA in the cell. *The three mechanisms: conjugation, natural transformation and infection by ssDNA phage lead the production of large amount of ssDNA in the cell.*

Conjugation. Conjugation is the process by which one bacterium can actively transfer DNA to a neighboring cell (Figure 15). The mechanism of conjugation is conserved across all described systems. A protein called relaxase binds and nicks a cognate origin-of-transfer site (*oriT*). This reaction results in a complex between the relaxed plasmid and the relaxase (together with accessory factors), called the relaxosome. Only the strand that is covalently bound by the relaxase is transferred to the recipient cell as ssDNA. The transferred strand (T-strand) is excreted from the donor cell through the type IV secretion system and the relaxase then directs recircularization of the T-strand in the recipient cell (for a comprehensive review, see (Alvarez-Martinez and Christie 2009)). Two main families of conjugative elements have been described: self-transmissible plasmids and "integrative and conjugative elements" (ICEs). ICEs cannot autonomously replicate and are thus carried by chromosomes or other replicons. These elements are able to excise themselves as circular intermediates through the action of a recombinase/excisionase and are then transferred following the same mechanism. In the recipient cell, they can be integrated through homologous recombination or through the action of a site-specific recombinase (Burrus and Waldor 2004; Juhas, Crook et al. 2008). The length of the DNA molecule that is transferred is usually the size of the whole conjugative element (usually <200kb).

Occasionally, chromosomal DNA can be transferred. This happens when conjugative plasmids are integrated into the chromosome, a famous example being the plasmid F/Hfr system (Tatum and Lederberg 1947; Low 1972). Alternatively, the conjugation functions carried by ICEs can also promote transfer of chromosomal or plasmid DNA, as demonstrated for the SXT element in *Vibrio cholerae* (Hochhut, Marrero et al. 2000). In this case, the length of the transferred strand is limited by the conjugation bridge strength and the contact time between the bacteria. Since the time of early genetic mapping trough Hfr conjugation of the *Escherichia coli* chromosome by Nelson, we have learned that it takes about 100 min to transfer the whole *E.coli* chromosome (4.6 Mb) (Nelson 1951). Although very long DNA fragments can be transferred, the average length of ssDNA region is unknown. Indeed, the ssDNA length and its lifetime depend on the speed of complementary strand synthesis in the recipient strain. The only direct data available comes from microscopy experiments enabling visualization of complementary strand synthesis and showing that synthesis already starts within 5 min after the donor and recipient cells are mixed (Babic, Lindner et al. 2008). Nevertheless, the number of ssDNA replication origins is unknown in most cases. Single-stranded origins of replication have been studied in the case of rolling-circle replication, which is discussed later (part III.2.1). The fact that specific origins of replication have evolved for initiation of complementary strand synthesis suggests that this process does not happen easily at random sequences. It is therefore unlikely that complementary strand synthesis is initiated at numerous loci. Conjugation thus massively produces ssDNA and conjugative plasmids are probably a place of choice for the evolution of functions where hairpins are involved. Indeed, the very process of conjugation, for instance, implies DNA secondary structures (Gonzalez-Perez, Lucas et al. 2007) (see "Hairpin and conjugation").

Transformation. Bacterial competence for natural transformation is a physiological state that permits uptake and incorporation of naked exogenous DNA (Figure 15). Many Gram-negative bacteria (species of *Haemophilus, Neisseria, Helicobacter, Vibrio* and *Acinetobacter*) as well as Gram-positive bacteria (species of *Bacillus, Mycobacterium* and *Streptomyces*) are capable of natural competence. In all cases, one strand of the transformed DNA is degraded, providing the energy for transport of the complementary strand across the cytoplasmic membrane (Chen, Christie et al. 2005). Some bacteria have been shown to fragment exogenous DNA so

that they take only small bits, while others can take up long DNA molecules (Dubnau 1999). Monitoring of ssDNA fate during transformation in *Streptoccocus pneumoniae* revealed that ssDNA does not subsist in the cell more than 15 min (Mejean and Claverys 1984). Globally, the length of the incoming DNA and the lifetime of ssDNA in the recipient cell are probably shorter than for conjugation. The entering single strand is protected from the action of nucleases essentially by the binding of SSB (Claverys, Martin et al. 2009), whereas, during conjugation, the relaxase is covalently bound to the T-strand, effectively protecting it from exonucleases. However, in some bacteria including *B.subtilis* and *S.pneumoniae*, a protein named DprA has been found to bind the incoming ssDNA, protecting it from both endo- and exonucleases and facilitating further homologous recombination (Mortier-Barriere, Velten et al. 2007). All in all, during transformation, ssDNA is not long-lived in the cell; it is either quickly integrated into the chromosome through homologous recombination or it is degraded.

Phage infection. Single-stranded phages encapsidate their genome and deliver it to newly infected cells in this form (Figure 15). The maximum amount of DNA that can be transferred is equivalent to the size of the phage genome (generally <10 kb), but here again, little is known about the timing of complementary strand synthesis. Nevertheless, hairpins have been found to play important roles at all steps of ssDNA phage life cycles, from synthesis of the complementary strand (Wickner and Hurwitz 1975; Lambert, Waring et al. 1986) to phage DNA encapsidation (Russel, Linderoth et al. 1997) (see the part "DNA hairpin biological functions").

(ii) Macromolecule synthesis and repair.

Transcription. RNA synthesis requires the opening of the DNA duplex. The size of the transcription bubble ranges between 12 and 25 bp, covered by the transcription complex (Gamper and Hearst 1982). This small opening leaves very little room for secondary structure formation, and transcription is thus unlikely to foster hairpin formation. On the contrary, the transcription bubble needs to unfold hairpins that it may encounter so as to enable production of the correct transcripts by the RNA polymerase (RNAP).

Replication. In contrast to transcription, DNA synthesis produces large amounts of ssDNA. Firstly, the replication initiation step often requires melting of a

large DNA region around the origin of replication. Multiple hairpins have been found to play important roles at replication origins (Masai and Arai 1996; Carr and Kaguni 2002) (see

Figure 16 Hairpin formation during replication. *Hairpins can fold on the ssDNA formed by the discontinuous replication of the lagging strand, or on ssDNA gaps remaining after lesion bypass.*
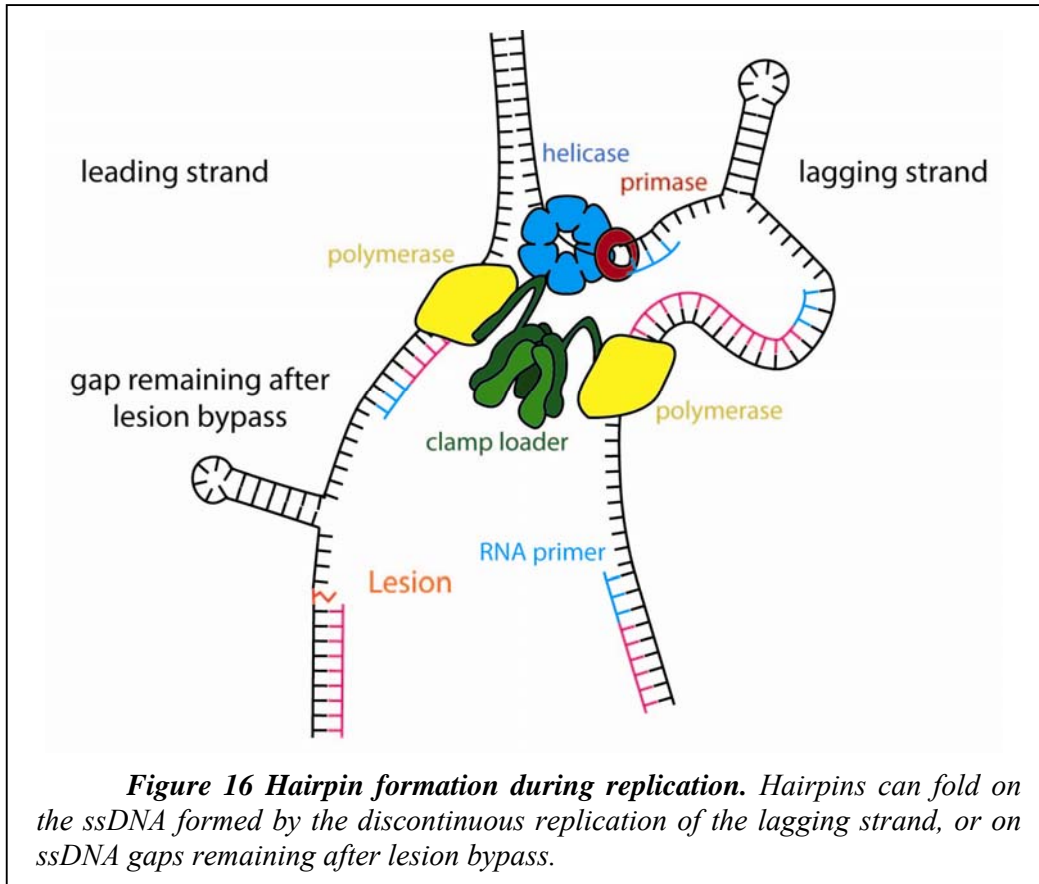
the part "Hairpins and replication origins"). Secondly, lagging strand replication is not continuous and an ssDNA loop is formed to place the DNA in the correct orientation for DNA polymerase. Half of the replication loop consists of nascent Okazaki fragment and the other half of ssDNA extruded by the helicase (Figure 16). In *E.coli*, Okazaki fragments are 1 kb to 2 kb nucleotides long, and the replication fork speed is about 1 kb.s-1 in optimal conditions (Kornberg and Baker 1992). The lifetime of ssDNA should thus be on the order of a second. Evidence that inverted repeats (IRs) can fold into stable hairpins in vivo during replication came from the observation that large and perfect IRs are genetically unstable on plasmids in *E.coli*. Indeed, they are the cause of mismatched alignment or slippage during replication (Sinden, Zheng et al. 1991; Leach 1994). In particular, deletions of IRs occur preferentially on the lagging strand (Trinh and Sinden 1991).

Finally, a special mode of replication, called rolling circle replication (RCR), involves unwinding of the full lagging strand template into ssDNA (Khan 2005). Multiple hairpins have been found to play important roles in RCR (Noirot, Bargonetti et al. 1990; Kramer, Khan et al. 1997; Kramer, Espinosa et al. 1998; Kramer, Espinosa et al. 1999).

DNA repair. A major source of ssDNA in the cell is through DNA repair. Double-strand breaks are processed by the RecBCD enzyme which produces ssDNA tails through its exonuclease activity. These ssDNA tails can then be bound by RecA and may be involved in homologous strand invasion and replication-dependent repair (Kowalczykowski 1994; Kowalczykowski, Dixon et al. 1994; Kreuzer 2005). Double-strand breaks can be caused by many agents, including ionizing radiation, UV light and oxygen radicals, but in normally growing cells as well, double-strand breaks are formed in almost every cell cycle as a consequence of replication through imperfect DNA templates (for a comprehensive review, see (Dillingham and Kowalczykowski 2008)).

It has also been shown that when replication forks encounter a lesion, the replication of the lagging and leading strands can be uncoupled in order to bypass the lesion, leaving ssDNA gaps on the damaged strand (Pages and Fuchs 2003; Heller and Marians 2006; Langston and O'Donnell 2006). These gaps are around 1 kb in length and can be processed by RecA-mediated recombinational repair (Figure 16).

(iii) Single-strand binding proteins.

In all these processes, ssDNA in the cell is not left naked. Several proteins bind ssDNA without sequence specificity. The most important ones are the RecA and single-strand binding (SSB) protein. SSB coats any ssDNA present in the cell and prevents intrastrand pairing, i.e. hairpin formation. The RecA protein also binds ssDNA forming a straight nucleoproteic filament. RecA can then promote strand invasion of homologous dsDNA and catalyze recombination (Kowalczykowski 1994). Furthermore, SSB directs RecA binding to ssDNA (Kowalczykowski and Krupp 1987; Reddy, Vaze et al. 2000). Recent single molecule studies have shown how tetrameric SSB can spontaneously migrate along ssDNA, melting unstable hairpins while stimulating RecA filament elongation (Roy, Kozlov et al. 2009).

Although ssDNA is present on many occasions in the cell, hairpin formation is strongly constrained by SSB and RecA binding. Proteins that ensure their function through hairpin binding are thus in competition with SSB and RecA for substrate availability. Hairpins that are formed need to be stable enough to resist SSB melting and coating. For instance, it was demonstrated that SSB can inhibit the activity of the plasmid pT181 RepC protein at secondary cleavage sites on ssDNA, but not at its primary binding site (Koepsel and Khan 1987) (see the part "Rolling Circle Replication" and Figure 19).

III.1.b Cruciform extrusion

(i) Mechanism of cruciform extrusion

The formation of DNA hairpins in the cell does not necessarily require the production of ssDNA. Extrusion of cruciforms occurs through the opening of the DNA double helix to allow intrastrand base pairing. Base opening in relaxed DNA is both infrequent and transient. However, negatively supercoiled DNA molecules are much more active, because their topology facilitates both large- and small-scale opening of the double helix (Furlong, Sullivan et al. 1989). Two main mechanisms for cruciform extrusion have been proposed (Figure 17)(Lilley 1985). The first (type S) implies small-scale melting of the double helix at the dyad of the IR (~10bp). This small opening allows a few bases to pair with their cognate base in the repeat. The stem can then be elongated through branch migration, which is also facilitated by negative supercoiling. The other mechanism (type C) involves the melting of a large region, which is favored by nearby AT-rich sequences. This large melting would allow hairpins to fold on both strands leading to cruciform formation (Figure 17). The S-type mechanism is highly dependent on the IR sequence (it is favored by the AT-rich sequence at the dyad), and works in physiological ionic conditions (Sullivan and Lilley 1987). On the other hand, C-type extrusion takes place in low-salt solutions and is highly dependent on the presence of AT-rich neighbor sequences, but should theoretically be suppressed at physiological ion concentrations (Murchie and Lilley 1987). Nevertheless, this mechanism could possibly take place in DNA regions with propensities to undergo substantial denaturation, such as replication origins.

***Figure 17 Mechanisms of cruciform extrusion.** In the C-type pathway, a substantial region of dsDNA is denatured allowing the folding of the whole hairpins on both strands in one step. In the S-type pathway, a small region is denatured (~10bp) allowing the folding of a small hairpin that can then be elongated through branch migration.*
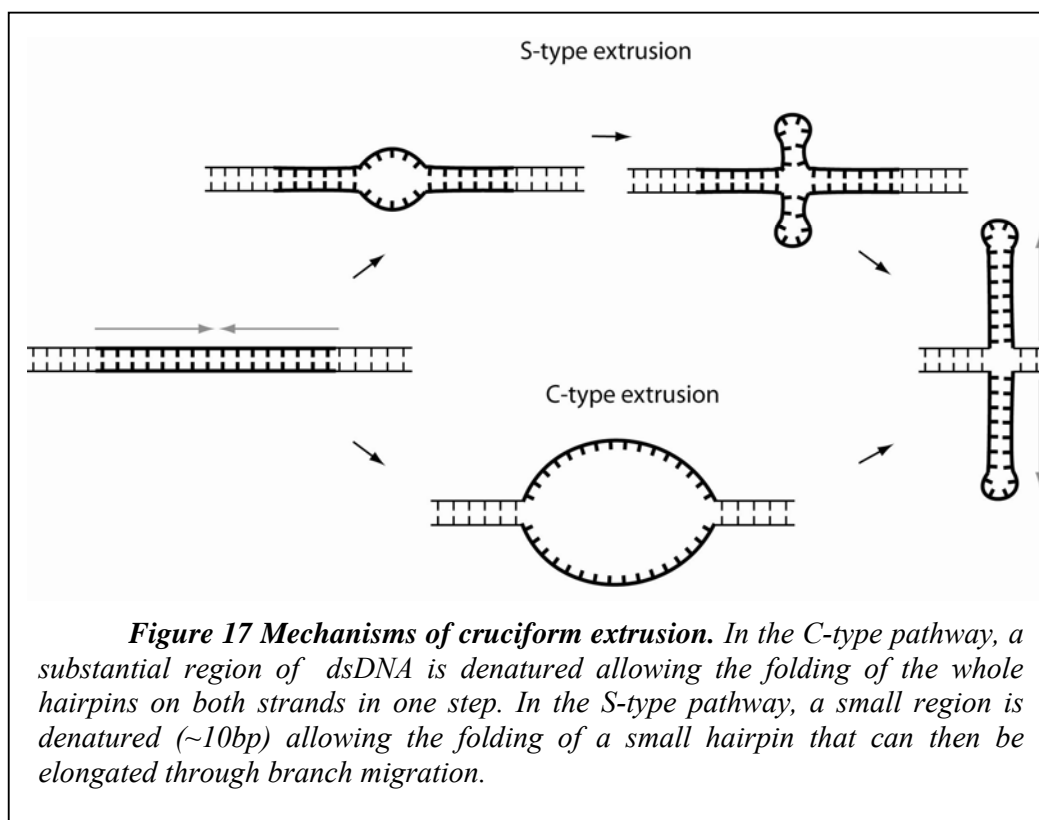
(ii) Regulation of cruciform extrusion.

Cruciforms were extensively studied in the 1980's when techniques enabling their observation *in vitro* were developed, such as S1 sensitivity and 2D electrophoresis. Although cruciform extrusion can be energetically favorable under moderate superhelical densities, the slow kinetics of cruciform extrusion raises questions as to their relevance *in vivo* (Courey and Wang 1983). However, several techniques later developed led to the demonstration of cruciform formation *in vivo* under natural superhelical densities (Horwitz and Loeb 1988; Noirot, Bargonetti et al. 1990; Dayn, Malkhosyan et al. 1991; Dayn, Malkhosyan et al. 1992). In particular, cruciforms that were tuned to fold stably at different superhelical densities have even been used to measure the natural superhelix densities of plasmids. *In vivo* cross-linking with psoralen demonstrated that the propensity of an IR to fold into a cruciform strongly depends on its sequence and context, and that some IRs can exist as cruciforms at levels as high as 50% in plasmids in living *E.coli* cells (Zheng, Ussery et al. 1991; Zheng, Kochel et al. 1991).

Nevertheless, most reported cruciform detection involved artificial conditions favoring hairpin extrusion: small loops, IR in AT-rich regions, perfect palindromes with AT-rich centers and GC-rich stems, *topA* background or salt shock to increase supercoiling (Zheng and Sinden 1988; Sinden, Zheng et al. 1991; Zheng, Kochel et al. 1991). Random IRs do not seem to fold cruciforms at significant rates under average *in vivo* supercoiling. However, many factors may transiently increase local superhelical density to a critical level sufficient for cruciform extrusion (see review (Pearson, Zorbas et al. 1996)). Biological processes such as transcription and replication may generate local and temporal domains of supercoiling on circular DNA (Liu and Wang 1987; Dayn, Malkhosyan et al. 1992; Schvartzman and Stasiak 2004). Indeed, during replication and transcription, enzymes alter the structure of DNA, such that additional twists are added (positive supercoiling) or subtracted (negative supercoiling). Negative supercoiling favors the unwinding of the DNA double helix, which is required for initiation of transcription and replication processes (Pruss and Drlica 1989; Hirose and Matsumoto 2005). As transcription proceeds, DNA in front of the transcription machinery becomes positively supercoiled, and DNA behind becomes negatively supercoiled. Similarly, during replication, strand separation by the helicase leads to positive supercoiling of the duplex ahead of the fork (see review (Schvartzman and Stasiak 2004)).

Changes in supercoiling in response to external and/or internal stimuli could also play a significant role in the formation and stability of cruciforms. In *E.coli,* superhelicity has been shown to vary considerably during cell growth and to change under different growth conditions (Balke and Gralla 1987; Jaworski, Higgins et al. 1991). Moreover, topology analysis of reporter plasmids isolated from strains where the SOS stress response regulon is constitutively expressed revealed higher levels of negative supercoiling (Majchrzak, Bowater et al. 2006). Finally, the level of superhelicity is known to be variable between bacterial strains. For instance, the average supercoiling density of a pBR322 reporter plasmid extracted from mid-log cultures of WT *Salmonella* is 13% lower (σ=-0.060) than that from *E.coli* (σ=-0.069) (Champion and Higgins 2007).

(iii) Effect of cruciform extrusion on DNA topology dynamics.

The positioning of IRs within topological domains appears to be another parameter that influences cruciform extrusion. Studies involving visualization of the cruciform on supercoiled plasmids through atomic force microscopy have shown that extrusion is favored when IRs are positioned at the apex of a plectonemic supercoil (Oussatcheva, Pavlicek et al. 2004). Furthermore, cruciforms can exist in two distinct conformations, an X-type conformation and a planar conformation. In the X-type conformation, the cruciform arms form an acute angle and the main DNA strand is sharply bent, whereas in the planar conformation, the arms are present at an angle of 180°C (Shlyakhtenko, Hsieh et al. 2000). It has been shown that the rest of the DNA molecule is deeply affected by the conformation adopted by the cruciform. X-type cruciforms tend to localize at the apex of the plectonemic supercoil and restrict slithering of the molecule, i.e. they reduce the possibility of distant sites coming into contact. Environmental conditions, such as salt concentration and protein binding, are factors influencing the conformation choice. For instance, the RuvA protein tetramer which binds to the Holliday junction at the base of cruciforms forces them into a planar conformation in which the constraints upon DNA movements are relieved (Shlyakhtenko, Hsieh et al. 2000). It has thus been proposed that cruciform extrusion may act as a molecular switch that can control DNA transactions between distant sites. Such long-range contacts are known to be essential in many cellular processes, including site-specific recombination, transposition or control of gene expression through DNA-loop formation (Gellert and Nash 1987; Adhya 1989; Schleif 2000; Liu, Bondarenko et al. 2001).

III.1.c Genetic instability of inverted repeats

It was quickly noticed that long palindromes are impossible to maintain in vivo (for a review, see (Leach 1994)), either because they are not genetically stable and will be partially mutated or deleted, or because they are not viable, i.e. the molecule carrying them cannot be replicated (Collins, Volckaert et al. 1982). It is assumed that instability and inviability are caused by the inability of the replication fork to process secondary structures that are too stable, and by the presence of proteins destroying these structures. In particular, the SbcCD enzyme can cleave hairpins forming on ssDNA, leading to double-strand breaks that are then repaired by

recombination (Chalker, Leach et al. 1988; Cromie, Millar et al. 2000). This leads to constraints on the size and perfection of the inverted repeats that can be maintained in vivo. Typically, a size of 150-200 bp is a limit for IRs, although the presence of mismatches and spacers between the repeats strongly improves their maintenance. However, a mutation mechanism was identified, which tends to restore perfection to quasi-palindromes during chromosomal replication (Dutra and Lovett 2006). The model proposes that during replication, the nascent DNA strand dissociates from its template strand, forming a partial hairpin loop structure. The nascent strand is then extended by DNA synthesis from the hairpin template, forming a more fully paired hairpin. IRs are thus balanced between a mechanism than tends to perfection and the fact that perfect IRs are not genetically stable.

III.2 DNA hairpin biological function

III.2.a Hairpins and replication origins

Hairpins play an essential and common role in replication initiation. Indeed, they have been found to be indispensable for initiation of complementary strand synthesis on single-stranded phages as well as for replication of dsDNA replicons, in particular, during rolling circle replication.

(i) Priming on single strand.

The first evidence for the role of DNA hairpins in a biological function came from the early studies of the primosome. The inability of DNA polymerases to initiate *de novo* replication makes the independent generation of a primer necessary (Kornberg and Baker 1992). The primosome is a complex of proteins which carries out this priming through *de novo* synthesis of a small RNA whose 3' end can be used by the DNA polymerase as a starting point. The role of RNA in priming DNA replication was discovered primarily through studies of single-stranded phages, notably G4 and ɸX174 (Wickner and Hurwitz 1975; Lambert, Waring et al. 1986). Single-stranded phages are delivered to the infected cells and have evolved diverse mechanism for priming synthesis of the complementary strand, but all the strategies described to date involve DNA hairpins.

G4 type priming. Phage G4 carries, in the region of replication initiation, three hairpins with stems of 5 to 19 bp and loops of 4 to 8 bases. Early models invoked these structures as recognition sites for the primase, DnaG (Lambert, Waring et al. 1986). However, it was later shown that none of these hairpins are required for DnaG to initiate primer synthesis in the absence of SSB in *E.coli* (Swart and Griep 1993). The hairpins seem, in fact, to direct the binding of SSB so that primase recognition site 5'-CTG-3' is exposed (Sun and Godson 1998). This mechanism is likely to be at stake for a large number of G4-like phages, including a3, St-1 and ɸK. This is an illustration of how hairpins can direct protein binding and structure an ssDNA region (Figure 18).

Figure 18 Priming of replication on ssDNA hairpins. *G4-type priming: hairpins structure the region, directing the binding of SSB and allowing access to the dnaG primase. Phi174-type priming: a ssDNA hairpin forms a Y-fork recognized by PriA which directs the formation of a primosome. Filamentous phage type priming: an hairpin mimicking a promoter is recognized by the RNA polymerase (RNAP), which synthesizes an RNA primer for replication.*

ɸX174-type priming. Although ɸX174 is a close relative of G4, the priming mechanism leading to the synthesis of the complementary strand cannot be realized by DnaG alone. The PriA protein, which is now known to play a major role in stalled replication fork restart, was first identified as an essential component of the ɸX174 primosome (Wickner and Hurwitz 1975). It catalyzes priming from a specific primosome assembly site (PAS) which can adopt a stable secondary structure (Arai and Kornberg 1981). However, it is now clear that the main PriA substrates are not PAS sites but D-loops and R-loops encountered during replication, DNA repair and recombination events. It has thus been proposed that PAS sequences have evolved to mimic the natural targets of PriA (McGlynn, Al-Deib et al. 1997). A stem-loop formed on a single strand can indeed be viewed as a branched structure between a

double strand and two single-strand components (a Y-fork). PriA was recently shown to bind Y-forks (Tanaka, Mizukoshi et al. 2007). This is an illustration of hairpins that have evolved to be recognized by a host protein, to direct primosome assembly (Figure 18).

Filamentous phage type priming. In the case of the M13 phage and other filamentous phages (f1 and fd), synthesis of the complementary strand is primed neither by DnaG nor PriA, but by the host RNA polymerase (RNAP) holoenzyme containing the sigma70 subunit which synthesizes a 20 nt long RNA primer (Kaguni and Kornberg 1982; Higashitani, Higashitani et al. 1996). The RNAP recognizes a double hairpin structure mimicking a promoter with a -35 and a -10 box (Higashitani, Higashitani et al. 1997) (Figure 18). Here again, hairpins have evolved to be recognized by a host protein. Hairpins recognized by the RNAP have now been associated with several functions (see 2.a).

(ii) Double-strand DNA replication

The first step in dsDNA replication is the melting of a region where the replication priming complex can load. This melting event is favored, with some exceptions, by a complex of proteins (DnaA for the chromosome, or Rep for plasmids), which binds the DNA (usually at direct repeats: DnaA boxes or iterons) and bends it (Katayama, Ozaki et al.; Konieczny 2003; Mott and Berger 2007). This bending promotes DNA melting, but also formation of alternative DNA structures.

A common feature of many origins of replication is the presence of inverted repeats (IRs). The extrusion of IRs as cruciforms is energetically more favorable than the simple DNA melting and is thus very likely to occur, absorbing a part of the strain generated. Furthermore, when DNA melting actually occurs (which is favored by AT-rich regions present in most *oris*) IRs are free to fold into hairpins. There is thus ample opportunity at origins of replication for a DNA structure to arise and interact with proteins.

As a matter of fact, hairpins have also been shown to play essential roles in primosome assembly in dsDNA replication. The generation of a primer occurs in two major ways: opening of the DNA double helix followed by RNA priming (chromosomal, theta and strand displacement replications) or cleavage of one of the DNA strands to generate a 3'-OH end (rolling-circle replication (RCR)) (del Solar,

Giraldo et al. 1998; Khan 2005). In both mechanisms, cases where hairpins play essential roles have been described.

Chromosomal and theta replication. The DnaA protein plays a central role in the replication of the bacterial chromosome and of several plasmids. It is involved in the control of replication initiation, unwinding of the helix and recruitment of the priming complex (for a review see (Mott and Berger 2007)). It has been proposed that in some replication origins, a hairpin structure carrying a DnaA box folds in the region unwound by DnaA itself. This hairpin, named M13-A, is at the core of the ABC priming mechanism first described for the R6K plasmid (Masai, Nomura et al. 1990). M13-A is specifically bound by DnaA, which then recruits DnaB, DnaC and finally initiates RNA priming. This mechanism was later proposed to occur at the *E.coli* origin of replication (Carr and Kaguni 2002), and putative M13-A hairpins are present in a large number of theta-replicating plasmids.

Inverted repeats other than M13-A and called single-stranded initiators (*ssi*) are often present at replication origins and can be involved in RNA priming. In the same way that filamentous phages prime complementary strand synthesis, the F plasmid origin of replication has a hairpin (*ssiD* or *Frpo*) recognized by *E.coli* RNAP which synthesizes an RNA primer (Masai and Arai 1997). Other *ssi* have been isolated from a variety of plasmids and shown to use a ϕX174 type priming involving PriA (for a review, see (Masai and Arai 1996)).

Strand displacement replication. The best described example of strand displacement replication is plasmid RSF1010. The plasmid-encoded RepC protein binds to iterons and unwinds the DNA in a region carrying two single-stranded initiators (*ssiA* & *ssiB*). These sequences are IRs which fold into hairpins. The secondary structures of these hairpins and parts of their sequences have been shown to be essential for replication (Miao, Honda et al. 1993). The current model states that plasmid-encoded RepB primase specifically recognizes *ssiA* and *ssiB* and primes continuous replication from these sequences (Honda, Sakai et al. 1988; Honda, Sakai et al. 1989; Honda, Sakai et al. 1991). However, it is not clear whether *ssiA* and *ssiB* fold when the region is largely single-stranded or whether they extrude as a cruciform, thanks to the action of RepC.

74

Rolling circle replication (RCR). RCR is widely present among plasmids and viruses (including the filamentous phages previously mentioned), with the model being plasmid pT181 (for a review see, (Khan 2005)). The plasmid-encoded Rep protein binds to the double-stranded origin of replication (*dso*) and bends the DNA, producing a strain leading to the extrusion of a hairpin carrying the Rep nicking site. This structure was among the first cruciforms probed in vivo (Noirot, Bargonetti et al. 1990). Rep nicks DNA in the hairpin and becomes covalently attached to the 5' phosphate (Figure 19). The free 3'-OH end serves as the primer for leading strand synthesis. No synthesis occurs on the lagging strand until it is completely unwound by the helicase and released as ssDNA. The synthesis of the complementary strand is then initiated at the single-strand origin (*sso*). Four classes of *sso* have been described (*ssoA*, *ssoW*, *ssoT* and *ssoU*). These classes have little nucleotide sequence homology, but share structural features (Kramer, Espinosa et al. 1999) necessary for their recognition by the host RNA polymerase which primes complementary strand synthesis (Kramer, Khan et al. 1997; Kramer, Khan et al. 1998; Kramer, Espinosa et al. 1999).

Figure 19 Rolling Circle Replication. A)The Rep protein binds an hairpin formed by double-stranded origin (dso) and extruded from dsDNA as a cruciform. Rep nicks DNA and covalently binds the 5'-end, leaving a 3'-end for replication to proceed. The leading strand is replicated while the lagging strand is extruded and remains single-stranded until the single-stranded origin (sso) is reached. The RNA polymerase (RNAP) binds the sso hairpin and synthesizes an RNA primer for replication. B) The pT181 dso in cruciform conformation. C) The pT181 sso as folded by the mFOLD software.

III.2.b Hairpins and transcription

There are essentially three ways in which hairpins and cruciforms can affect transcription. (i) The extrusion of a cruciform dramatically reduces the local supercoiling of DNA. Since superhelical density is known to affect the activity of promoters, cruciform extrusions in promoter regions could reduce their activity (Wang and Lynch 1993). (ii) A cruciform could prevent proteins from binding to their cognate site if it overlaps the extruding sequence. (iii) RNA polymerases or transcription factors could recognize hairpins present on ssDNA or extruded from dsDNA. Since there is as yet no documented case for the first possibility, only the two other mechanisms are discussed here.

(i) Hairpin promoters

We have already discussed how the RNAP can recognize hairpin promoters to prime DNA replication (rolling circle replication / filamentous phage type priming / F plasmid replication). The RNAP primes F plasmid replication through recognition of the F*rpo* hairpin, but under certain conditions, it can produce transcripts longer than the one needed for priming and express the downstream genes (Masai and Arai 1997).

Accordingly, transcription from a structured single-stranded promoter was suggested to occur during conjugative DNA transfer for several *oriT*-associated genes of enterobacterial conjugative plasmids, namely *ssb*, *psiB* and sometimes *ardA*. Considering that conjugation consists of ssDNA entry into the recipient cell, the product of these genes - respectively single-strand binding, anti-SOS and anti-restriction - could be needed for maintaining the plasmid in the recipient. Indeed, the transcriptional orientation of these genes, always on the leading strand, means that the transferred strand is destined to be the transcribed strand (Chilley and Wilkins 1995). Moreover, conjugative induction of these first loci so as to enter the recipient bacterium was shown to be transfer-dependent (Jones, Barth et al. 1992). The burst of activity observed shortly after initiation of conjugation led to the proposal that this early transcription could be mediated by the presence of a secondary structure in the transferred ssDNA (Nomura, Masai et al. 1991; Althorpe, Chilley et al. 1999)that mimics an RNA polymerase promoter recognized by the F*rpo* sigma factor (Masai and Arai 1997).

Figure 20 N4 virion hairpin promoters. *The three promoters of N4 controlling the expression of the early genes as cruciform structures.*
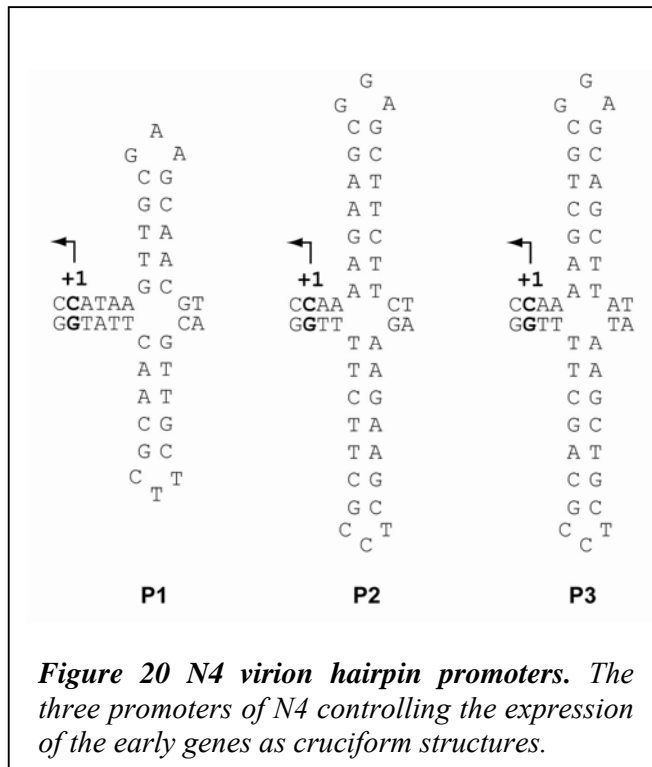
Other hairpin promoters which are not involved in priming have been described. Notably, the N4 virion carries three hairpin promoters specifically recognized by the virion RNA polymerase (vRNAP) and used to direct the transcription of the phage early genes (Figure 20). Upon infection of *E.coli*, the N4 double-stranded DNA injected into the cell is supercoiled by the host DNA gyrase, which leads to the extrusion of hairpin promoters as cruciforms (Dai, Greizerstein et al. 1997; Dai and Rothman-Denes 1998).

(ii) Promoter inhibition through cruciform extrusion

Early studies have shown how an artificial IR overlapping a promoter can regulate transcription by superhelix-induced cruciform formation (Horwitz and Loeb 1988). Although promoters usually have higher activity with increasing superhelix density, such a promoter has a lower expression level at high superhelix density because of the extrusion of the IR as a cruciform preventing RNAP binding. It has also been shown that the N4 hairpin placed between the -10 and -35 boxes of the *rrnB* P1 promoter can repress its activity in a supercoil-dependent manner (Dai, Greizerstein et al. 1997). DNA cruciform extrusion seems likely to be a mechanism for the regulation of genes repressed by supercoiling. However, it is not clear how common this mechanism of regulation is, since no compelling natural example has been reported. The *bgl* operon promoter, which presents a 13bp IR, was first though to be a natural example of such regulation (Singh, Mukerji et al. 1995). However, it was

78

later shown that no cruciform is required to account for its supercoiling-dependent repression (Caramel and Schnetz 1998).

III.2.c Hairpins and conjugation

IRs are present in a majority of origins of transfers (*oriT*) (Francia, Varsaki et al. 2004). The best described is the origin of transfer of R388, where an IR named IR2 located 5' to the nicking site plays an essential role (Guasch, Lucas et al. 2003). Conjugation occurs as follow: DNA is nicked at the *oriT* and bound covalently by the plasmid-encoded relaxase protein TrwC. The T-strand is then unwound through rolling circle replication and transferred to the recipient cell. Although the folding of IR2 into a hairpin is not required for the initial nicking of the *oriT*, the recircularization of the T-strand requires folding of IR2 into a hairpin specifically recognized by the relaxase (Gonzalez-Perez, Lucas et al. 2007). In addition to IR2, other IRs important for transfer efficiency are present in R388 *oriT* (Llosa, Bolland et al. 1991), but their exact role remains to be elucidated. It is not yet known whether their sequence or structure is important. They probably help adapt *oriT* into a potentially active state through cruciform formation.

Two relaxases other than TrwC have been crystallized: the F plasmid relaxase TraI (Datta, Larkin et al. 2003) and the R1162 plasmid relaxase MobA (Monzingo, Ozburn et al. 2007). Although they show poor sequence homology to TrwC, the 3D structure of all this relaxases is very similar. These enzymes are evolutionarily homologous to certain identical mechanisms of action.

III.2.d Hairpins and recombination

To date, there are three compelling examples of recombination systems using DNA hairpins as substrates: the CTX phage recombination site, the IS*200*/IS*605* insertion sequence family, and integron *attC* recombination sites.

(i) The single-stranded CTX phage of Vibrio cholerae

CTX is a single-stranded phage involved in *V. cholerae* virulence. In lysogenic phase, it integrates the *V. cholerae* chromosome I or II at its respective *dif1* and *dif2* sites.

Figure 21 The* V. cholerae *chromosome I dif site and the CTX phage hairpin. *The CTX attP region folds into a forked hairpin mimicking the V. cholerae dif1. This enables the CTX phage to use the host XerC/D recombinase to catalyze its integration into the chromosome.*
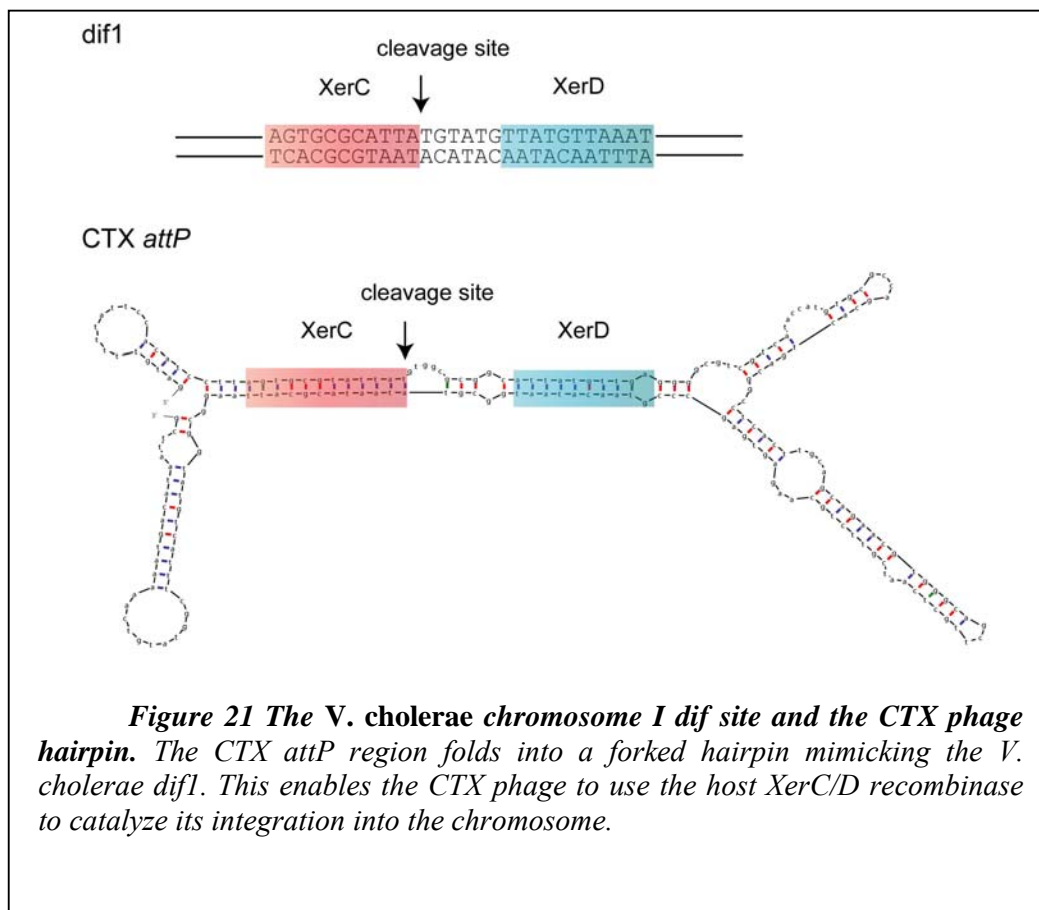
Chromosomal *dif* sites are recombination sites recognized by the XerCD protein complex which solves concatemers and allows proper chromosome segregation. CTX enters the infected cells as ssDNA, and the single-stranded form is directly integrated into one of the chromosomes (Val, Bouvier et al. 2005). The *attP* recombination site of CTX carries a ~150 bp forked hairpin, which is homologous to *dif* sites (Figure 21). The phage uses this hairpin to hijack the host XerCD protein complex which catalyzes a strand exchange between *attP* and the *dif* site (Das, Bischerour et al. 2010).

(ii) The IS200/IS605 insertion sequence family

The mechanism of transposition of the recently discovered IS200/IS605 insertion sequence family greatly differs from systems already described, in particular those using DDE transposase catalysis (Gueguen, Rousseau et al. 2005).

Figure 22 Organization of IS608 and Overall Transposition Pathway (adapted from Guynet, 2009). *(A) Organization. tnpA and tnpB open reading frames (light and dark arrows, respectively). Left end (LE) and right end (RE) (red and blue boxes, respectively). (B) Sequence of LE and RE. Sequence and secondary structures, IPL and IPR, at the LE and RE IS608 are shown. Left and right tetranucleotide cleavage sites are boxed in black (CL) and underlined in blue (CR). Position of cleavage and of formation of the 5' phosphotyrosine TnpA-DNA intermediate (vertical arrows). (C) Transposition pathway. (i) Schematized IS608 with IPL and IPR, left (TTAC; CL) and right (TCAA; CR) cleavage sites. (ii) Formation of a single-strand transposon circle intermediate with abutted left and right ends. The* transposon junction (TCAA) and donor joint (TTAC) are shown. (iii) Pairing with the target (TTAC) and cleavage (vertical arrows). (iv) Inserted transposon with new left and right flanks (dotted black lines).

The best studied representative of this family, IS*608,* was originally identified in *H*. pylori (Kersulyte, Velapatino et al. 2002). It presents at its ends short palindromes recognized as hairpins by the TnpA transposase. "Top strands" of the two IS ends are nicked and joined together by TnpA a few base pairs away from the hairpins (19 nt upstream from the left hairpin and 10 nt downstream from the right hairpin) (Barabas, Ronning et al. 2008; Guynet, Hickman et al. 2008). TnpA then catalyzes the formation of a single-stranded transposon circle intermediate which is then inserted specifically into a single-stranded target. This target site is not recognized directly by TnpA, but by four bases at the foot of the hairpin in the transposition circle (Figure 22 and (Guynet, Achard et al. 2009)) that realize unconventional base pairing with the ssDNA target sequence.

(iii) The IS91 insertion sequence

IS*91* is a member of an insertion sequence family displaying a unique mechanism of transposition. The IS*91* transposase is related to replication proteins of RCR plasmids. IS*91* transposition involves an ssDNA intermediate generated in a rolling circle fashion (Mendiola, Bernales et al. 1994). Short palindromes have been identified in the regions essential for transposition just a few base pairs away from the recombination sites. Their exact functions have not been studied. Nevertheless, striking similarities between these regions, RCR plasmids *dso* and conjugation *oriT*s suggest that these palindromes might fold into hairpins recognized by the IS91 transposase.

(iv) Integrons

Integrons are natural recombination platforms able to stockpile, shuffle and differentially express gene cassettes. Discovered by virtue of their importance in multiple antibiotic resistances, they were later identified in 10% of sequenced bacterial chromosomes, where they can contain hundreds of cassettes (Boucher, Koenig et al. 2007). The cassettes are generally single ORFs framed by *attC* recombination sites (Recchia and Hall 1995). When expressed, the integron integrase can recombine *attC* sites leading to excision of a circular cassette. Such a cassette can then be integrated at a primary recombination site named *attI*. *attC* recombination sites have been shown to be recognized and recombined by the integrase only as hairpins (Figure 14) (Bouvier, Demarre et al. 2005; Mazel 2006). A surprising feature

of *attC* hairpins is their huge polymorphism. Their stem length ranges from 54 to 80 bp and their loop length from 3 to 80 bp. Highly conserved mismatches known to be involved in hairpin recognition by the integrase are also present (Bouvier, Demarre et al. 2005; Bouvier, Ducos-Galand et al. 2009) (see the part "Strand selectivity").

III.2.e Other hairpin DNA: phage packaging, retrons, etc.

(i) Single-stranded phage packaging

The single-stranded filamentous phages (f1, fd, M13, Ike) contain IRs that can fold into hairpins. We have already described the hairpins involved in complementary strand synthesis, but the largest hairpin identified on these genomes is the packaging signal (PS) recognized in translocation of ssDNA into the virion capsid. This hairpin is probably recognized by the phage transmembrane protein pI and determines the orientation of DNA within the particle (Russel, Linderoth et al. 1997). Both the structure and sequence determinants of the PS-hairpin are required for its function (Russel and Model 1989).

(ii) Retrons

Retrons are DNA sequences found in the genomes of a wide variety of bacteria (Lampson, Inouye et al. 2005). They code for a reverse transcriptase similar to that produced by retroviruses and other types of retro-elements. They are responsible for synthesis of an unusual satellite DNA called msDNA (multicopy single-stranded DNA). msDNA is a complex of DNA, RNA and probably protein. It is composed of a small single-stranded DNA linked to a small single-stranded RNA molecule folded together into a secondary structure. msDNA is produced in many hundreds of copies per cell (Lampson, Inouye et al. 2005). Whether msDNA are selfish elements or play a role in the cell remains to be discovered.

III.3 Protein / hairpin recognition

All the hairpins described above have evolved different ways to realize their biological function. Most of them are recognized by proteins. Either they subvert host proteins that normally target other DNA substrates, or in some cases, proteins have

evolved that specifically recognize hairpin DNA features. I will briefly review here the mechanisms of protein / hairpin recognition.

III.3.a Mimicry: subverting the host proteins

Some of the hairpins described in the literature have evolved to mimic the "natural" target of the protein they interact with. The PAS sequences of single-stranded phages mimic Y-forks that are recognized by PriA. The *sso* of RCR plasmids, the F*rpo* hairpin and the filamentous phages priming hairpins all mimic promoters recognized by the host RNAP. The M13-A hairpin mimics a natural *dnaA* box and the CTX *attP* recombination site mimics the *V. cholerae dif* sites recognized by XerCD.

There is a noteworthy difference between hairpins like the CTX *attP* site, where mimicry is clear-cut, and the variety of hairpins recognized by RNAP. The latter indeed display an impressive diversity of structures and sequences. Although elements of the *ssoA* class present a large hairpin with near-consensus -35/-10 boxes (Kramer, Khan et al. 1997), other *sso* classes like *ssoU* present much more complex structures with several hairpins and -35/-10 boxes harder to recognize (Kramer, Espinosa et al. 1999). Another structural variation is that used by the filamentous phages. Here, a double hairpin acts as the recognition site with the -35 box on one stem-loop and the -10 box on the other (Higashitani, Higashitani et al. 1997). The fact that they are all recognized by RNAP suggests poor specificity of RNAP binding to hairpin DNA. The few common features of all these sequences are the widespread presence of mismatches in the hairpins and the fact that they do not work as promoters in dsDNA form, but bind RNAP very strongly when single-stranded (in some cases even more strongly than strong double-stranded promoters (Higashitani, Higashitani et al. 1997). These observations are consistent with the fact that the sigmaA and sigma70 of *B. subtilis* and *E.coli,* respectively, bind strongly to ssDNA containing promoter -10 sequences (Huang, J. et al. 1997). The mismatches that often span the -10 box could be there to ease access for RNAP and increase hairpin-promoter activity. High activity might be required by single-stranded molecules which need to synthesize their complementary strand promptly before triggering the SOS response of the host, as was observed for phages defective in complementary strand synthesis (Higashitani, Higashitani et al. 1992).

In all these cases, mimicry of dsDNA is not perfect: to different extents, mismatches are present in the hairpins. These mismatches are probably, in some cases, necessary for maintenance of long IR *in vivo,* as discussed above. But do they have a role in and an impact upon hairpin recognition? CTX might be the only mimicry case in which imperfection has a clear function: mismatches are essential for the irreversibility of single-stranded phage integration (Val, Bouvier et al. 2005).

III.3.b Protein recognition of hairpin features

Other systems have evolved proteins recognizing special features of hairpin DNA. This is the case for integron integrase IntI, for IS200/IS605 family transposase TnpA, for mobilizable plasmid relaxases (TrwC etc.), for N4 virion RNAP and probably for strand displacement replication proteins RepB. The features that make a hairpin structurally different from dsDNA are essentially: (1) the bottom of the stem, which can be either a Y-fork or a Holliday junction depending on whether the hairpin forms on ssDNA or as a cruciform; (2) the loop which is single-stranded; and (3) extrahelical bases and bulges produced by mismatches between the IRs.

The crystal structure of the interaction between IntI, N4 vRNAP, TnpA, TrwC and their cognate hairpins has been obtained (Guasch, Lucas et al. 2003; Ronning, Guynet et al. 2005; MacDonald, Demarre et al. 2006; Gleghorn, Davydova et al. 2008). All four highlight different mechanisms of recognition. IntI binds as a dimer to the stem of the hairpin and specifically recognizes two extrahelical bases. A central bulge in the stem also seems to be important for formation of a recombination synapse involving 4 IntI monomers. N4 vRNAP presents a base-specific interaction with the single-stranded loop of the hairpin and fits the stem structure through interaction with the phosphate and sugar backbone. TnpA binds the stem primarily through contact with the phosphate backbone, but also shows a base-specific interaction with the bases of the loop and, importantly, with an extrahelical T in the middle of the stem. Finally, the TrwC interaction is somewhat different from the others, since it binds not only to the hairpin structure, but also to the ssDNA 3' to the stem-loop, where the nicking site is present. The binding to the ssDNA part is base-specific, whereas the interaction with the hairpin occurs essentially through contact with the DNA backbone (Guasch, Lucas et al. 2003).

III.3.c Strand selectivity

Whether it be during phage complementary strand synthesis, at the sso of RCR plasmids or during conjugation, only one DNA strand is available. In these cases, the question of strand selectivity is not physiologically relevant. However, when both DNA strands are free to fold into hairpins, erroneous recognition of one strand over the other may be problematic. Indeed, an inverted repeat, once folded, generates the same hairpins on the top and bottom strands, except for the loop and eventual bulges and extrahelical bases. Still, in all the processes in which a protein recognizes hairpin features, strand selectivity has been observed: the protein recognizes one strand and not the other. In light of the hairpin/protein interactions described above, it is easy to understand how proteins discriminate between the two strands. They all show base-specific interactions with bases either in the loop, at the single-stranded base of the stem or with extrahelical bases. Any of these interactions can account for strand selectivity. Some of these systems appear to have good reason to process one strand and not the other. The N4 virion needs to initiate transcription in the right direction. Recombination of the wrong strand for integron cassettes would lead to their integration in the wrong direction, where they could not be transcribed. Finally, if a different strand of IS*608* is recognized at each end of the transposon, this would lead to the junction of the top strand with the bottom strand, a configuration that cannot be processed further and is likely to be lethal. Therefore, one strand had to be chosen and the other strongly discriminated against.

III.3.d On the origins of folded DNA binding proteins

While, in many examples described above, one can see that hairpins evolved to subvert the host machinery, in other instances, proteins evolved to specifically and sometimes exclusively recognize hairpin structures. This is the case for the RCR Rep proteins, the relaxases of conjugative elements, the transposase of IS*608*, the integron integrases and phage N4 vRNAP. Where do these proteins come from and what pushed them to recognize ssDNA rather than dsDNA?

(i) RCR Rep proteins, relaxases and IS608 transposase

Interestingly, the IS*608* transposase as well as conjugative relaxases have been found to be structurally similar to RCR Rep proteins (Ronning, Guynet et al. 2005). All of these proteins have in common the use of a tyrosine residue to covalently bind DNA. The Rep proteins belong to a vast superfamily spanning eubacteria, archae and eukaryotes (Ilyina and Koonin 1992). The superfamily is characterized by two sequence motifs: an HUH motif (histidine-hydrophobic residue-histidine) presumed to ligate a Mg2+ ion and required for nicking, and a YxxxY motif where the tyrosines (Y) bind the DNA covalently, with one of the tyrosine being optional. All these proteins thus probably have a common ancestor, ancient enough to account for the diversity of their functions and their spread among the kingdoms of life. The ability to bind hairpin DNA might have been an important feature in early stages of life when single-stranded DNA might have been more widely present. In this instance, the relaxases of conjugative plasmids obviously need to recognize ssDNA features to process the ssDNA in the recipient cell. Recombination of ssDNA by the IS*608* transposase is probably a way to target mobile elements and to ensure their spread. Finally, the reason why RCR plasmid Rep proteins would recognize hairpins rather than the more stable dsDNA is probably that origins of replications need to be strongly negatively coiled to unwind the double helix, and under these conditions, hairpins can be the most stable conformation of DNA.

(ii) Integron integrases

Integron integrases (IntI) are also tyrosine recombinases covalently binding DNA. However, they are not related to the Rep protein superfamily. The closest relatives to integron integrases are the XerCD proteins. However, IntI proteins carry an additional domain, compared to XerCD. This domain is involved in binding of the extrahelical bases of the *attC* hairpins that are essential for strand selectivity (MacDonald, Demarre et al. 2006; Bouvier, Ducos-Galand et al. 2009). It would be tempting to speculate that integrons diverged from a single-stranded CTX-like phage which already used XerCD to recombine hairpin DNA. This special feature of ssDNA recombination would then have been selected to form an evolving recombination platform, thanks to its ability to sense both stressful conditions and the occurrence of horizontal gene transfer.

(iii) N4 vRNAP

N4 vRNAP is an evolutionarily highly divergent member of the T7 family of RNAPs (Davydova, Kaganman et al. 2009). N4 vRNAP and T7 RNAP recognize their promoter with similar domains and motifs. However, N4 vRNAP recognizes a hairpin, whereas T7 RNAP recognizes dsDNA. The difference lies in the domain interacting with the hairpin loop. It displays substantial architectural complexity and base-specific interactions for N4 vRNAP, whereas the same domain in its counterpart just fits an AT-rich DNA sequence without base recognition (Cheetham and Steitz 1999). The reason why the N4 phage has evolved to transcribe several genes only from cruciform promoters is unclear. It is likely a way for the virion to sense the coiling state of DNA in the cell, which is known to be modified during the cell cycle and is particularly negative during the SOS stress response (Majchrzak, Bowater et al. 2006).

III.4 Single-stranded DNA, stress and horizontal transfer

We have seen how a variety of hairpins have been selected to be recognized by host proteins, especially in single-stranded phages and plasmids. The single-stranded nature of DNA during transfer of mobile elements drove the evolution of secondary structures able to hijack the host cell machinery (see part III.3.a). The use of host priming proteins, host RNAP or even host recombinases enables single-stranded phages not to bring additional proteins with them and still be processed into a replicative form. Similarly, when a quick reaction is required upon horizontal transfer, ssDNA hairpins are the best elements for driving the response, as exemplified by the hairpin promoters present on several conjugative plasmids.

Figure 23 ssDNA: at the crossroads of horizontal gene transfer, the SOS response and genetic rearrangements. 1) Conjugation, transformation, phage infection and environmental stress lead to the production of ssDNA in the cell. 2) The RecA proteins binds ssDNA and triggers the self-cleavage of LexA (brown circles). 3) The SOS regulon is derepressed, recombinases are expressed (orange triangles), and DNA coiling is modified. 4) Increased supercoiling leads to cruciform formation. 5) tes><year>2004</year><pub-dates><date>May 4</date></pub-dates></dates><accession-num>15118093</accession-num><urls><related-uced

All in all, hairpin formation is most likely to occur in the presence of ssDNA in the cell, which is the key element that triggers the SOS response (Figure 23). Several pathways can lead to the production of large amounts of ssDNA. This happens, for example, when the cell tries to replicate damaged DNA, causing replication forks to stall (Walker 1996). Another source of ssDNA comes from DNA intake by horizontal gene transfer and phage infection. For instance, conjugative transfer of R plasmids - conjugative plasmids carrying multiple resistances - has been shown to induce the SOS stress response in the recipient cell, except when an anti-SOS factor is encoded by the plasmid (*psiB*, already mentioned in the part "Hairpins

and transcription") (Z. Baharoglu, submitted for publication). Interestingly, the expression of these anti-SOS genes is under control of ssDNA promoters, i.e. of hairpin substrates. In the case of integrons, the control of the integrase expression by the SOS is probably a way for integrons to "know" when potential substrates are present in the cell and to recombine them. It is also certainly a way to ensure that rearrangements occur only when the bacteria is stressed.

The relation between horizontal gene transfer and genetic rearrangements was further strengthened when we recently observed in the lab that the induction of SOS during conjugative transfer of R plasmids results in induction of the integrase, allowing genome rearrangements in the recipient bacterium (Z. Baharoglu, submitted for publication). Furthermore, integrons are often found on conjugative plasmids and may well take advantage of the single-stranded transfer to acquire cassettes and spread horizontally.

Finally, not only does the SOS response promote genetic rearrangements, but it also induces horizontal gene transfer. It is known, for instance, that stress can induce competence in some bacteria (Claverys, Prudhomme et al. 2006) (Figure 23). Another effect of SOS induction is the derepression of genes involved in the single-stranded transfer of integrating conjugative elements (ICEs), such as SXT from *V.cholerae*, which is a ~100 kb ICE that transfers and integrates the recipient bacteria's genome, conferring resistance to several antibiotics (Beaber, Hochhut et al. 2004). Moreover, different ICEs are able to combine and create their own diversity in a RecA-dependent manner (i.e. using homologous recombination, which is also induced by SOS) (Garriss, Waldor et al. 2009; Wozniak, Fouts et al. 2009). As for R plasmids, SXT transfer was observed to induce SOS in *V.cholerae*. Finally, some lysogenic phages are also known to induce their lytic phase under stressful conditions (Galkin, Yu et al. 2009). One might thus see the use of ssDNA hairpins by integrons and other recombination systems as a mechanism for evolving: diversity is generated under stressful conditions.

III.5 Conclusion on hairpin DNA functions in the cell

The use of DNA hairpins in biological processes is ubiquitous in prokaryotes and their viruses. How do these hairpins arise from duplex DNA? Numerous cellular processes lead to the formation of ssDNA, notably replication and the mechanisms of horizontal gene transfer, but also DNA damage and repair. Furthermore, the implication of cruciform DNA has been demonstrated at the RCR *dso* and for N4 phage promoters. Nevertheless, functions associated with cruciforms do not seem to be widely spread due to the slow kinetics of cruciform formation. However, cruciforms might play a role in special cases, but the difficulty of probing them in vivo makes these events underestimated. In eukaryotes, cruciform binding proteins have recently been identified and are suggested to play a major role in genome translocation (Kurahashi, Inagaki et al. 2006) and replication initiation (Zannis-Hadjopoulos, Yahyaoui et al. 2008).

Not surprisingly, single-stranded phages have been found to use DNA hairpins at almost every step of their life cycle: complementary strand synthesis, replication, integration into the host chromosome and packaging. But hairpins play a role in the replication of a much larger number of elements, probably including the origin of replication of *E.coli*.

A striking feature is the opportunism of single-stranded DNA in subverting host machinery. The three different mechanisms of complementary strand synthesis have evolved hairpins directing priming by three different host proteins (DnaG, PriA, RNAP) in three different ways. Another example of opportunistic use of host machinery is the CTX phage which integrates *V. cholerae* chromosome I through a hairpin mimicking the XerCD recombination site. Also, the variety of hairpins recognized by the RNAP, either for replication priming or for transcription leads to the perception of ssDNA as evolutionarily very flexible.

Finally, the evolution of functions involving ssDNA is deeply intertwined with horizontal gene transfer, response to stress and genome plasticity. Horizontal gene transfers lead to ssDNA production and involve functions requiring hairpins. Together with stresses that also generate ssDNA, they activate the SOS response and trigger systems involved in genome plasticity, some of which use hairpin DNA, such as IS*608* or integrons. To close the loop, the SOS response can trigger more horizontal

transfer, notably through activation of natural transformation, ICE conjugation and lysogenic phages.

The cases discussed above illustrate at least three different families of proteins in which specific hairpin binding activities have independently evolved. It thus seems quite easy both for proteins to evolve hairpin binding activity and for hairpins to evolve in such a way that they can exploit host proteins. Hairpin recognition can be seen as a way for living systems to expand the repertoire of information storage in DNA beyond the primary base sequence. These hairpin recognition examples illustrate how DNA can carry information via its conformation. Finally, this review is probably not exhaustive, as new functions in which folded DNA plays a role most likely remain to be discovered.

Results

My PhD work consisted in two main projects. The first one, in collaboration with Dr. Céline Loot was to better understand the mechanism of integron recombination, and in particular to identify the pathways by which the *attC* recombination site could fold into a ssDNA hairpin. The second project was to assess the possibility of using the integron recombination machinery to generate genetic combinations for biotechnological purpose. I present here the two articles published during my PhD and describe further the results obtained.

Cellular pathways controlling integron cassette site folding

It is now admitted that the recombination site of integron cassettes (*attC*) recombines as a folded ssDNA hairpin. However, the most stable form of DNA in the cell is the double-stranded B-helix described by Watson & Crick. I reviewed in the introduction the pathways that can lead to the production of ssDNA in the cell, or the extrusion of hairpin as cruciform DNA.

It was at first not clear what pathway would *attC* sites use to fold. This work was initiated by Dr Céline Loot who had constructed a series of *attC* site derivatives with different VTS and stem lengths (Figure 2 of the paper). The initial idea was to see if there was a limitation to the *attC* site length and if this limitation could tell us something about the conditions in which *attC* sites fold. The observation was made that although all sites recombine at similar frequencies when delivered through conjugation (i.e. on ssDNA), they differ greatly when the *attC* site was carried by a replicative plasmid (Figure 4 of the paper). These first observations lead to the conclusion that *attC* site folding was not the limiting step of the recombination process when it was delivered on ssDNA. However we needed to understand what was limiting the recombination of some sites in replicative conditions.

At first there seemed to be no good explanation. Larger sites would tend to recombine at lower frequencies than smaller sites, but the correlation was very weak. For instance our largest site (VCR180) recombined only ~20 time less than our best site (*attC*aadA7) while intermediate sites like the wild type $VCR_{2/1}$ recombined ~320 time less frequently than *attC*aadA7. A better correlation was observed when only the

length of the loop of the hairpin (the VTS) was considered, but the correlation remained weak and this parameter could not explain why some sites with identical VTS length would differ by almost two logs in recombination frequency. We then realized that some sites tended to fold into alternative structures that did not form the integrase attachment sites. We thus formulated the hypothesis that "parasite" structures could compete with the proper hairpin fold and hinder recombination. We were able to verify this hypothesis experimentally through the construction of 2 pairs of variants (VCR97a/b and VCR116a/b). The (a) and (b) version of the sites only differed by a few base pair substitutions in the loop, but these substitutions allow or not the folding of a predominant parasite structure. The versions of the sites presenting the alternative structures recombined respectively 17 fold and 38 fold less than their counterpart (Figure 3B of the paper). Surprisingly the wild type VCR$_{2/1}$ site presented a similar strong parasite. These non-recombinogenic structures were thus not artificially introduced when constructing *attC* sites derivatives but may have a physiological relevance.

When the presence of parasite structures was accounted for, an effect of the VTS length could still be observed and the two parameters satisfactorily explained ~80% of the differences between the sites. However, this did not give much answer about the way *attC* sites fold. Indeed a larger VTS could account for lower recombination frequencies following two mechanisms: it could make a site less likely to fold on ssDNA formed during replication, and it could also be the consequence of cruciform extrusion. A larger VTS means that more DNA needs to be melted in order for the inverse repeats to anneal.

An assay was thus designed to see if the position of the *attC* bottom strand either on the lagging or leading strand of replication would influence recombination. If this was not the case, cruciform extrusion would surely be the predominant pathway for *attC* folding. This experiment showed that *attC* recombination was 4 times to 10 times more frequent on the lagging strand template than in the opposite orientation (Figure 3C of the paper). This clearly meant that when the site is on the template for the lagging strand synthesis, folding and recombination occurs mostly on the ssDNA produced by replication. However, there were still a significant number of recombination events occurring on the leading strand. Furthermore, when we look at the orientation of natural chromosomal integrons, they are all oriented so that the

bottom strand of the *attC* sites is on the leading strand of replication. It is very unlikely that the replication forks leave enough ssDNA on the leading strand for hairpins of 60bp or more to fold. There can thus be only two ways to account for the folding of the *attC* sites on the leading strand. Either, they fold as cruciform structures, or they fold on ssDNA gaps on the leading strand. Such ssDNA gaps are known to occur when the replication forks encounters a lesion on the leading strand and bypasses it, priming replication again only ~1kb downstream (see the introduction part III.1.a) (Pages and Fuchs 2003).

We decided to assess the possibility of cruciform extrusion with another experiment. We transformed suicide-plasmids carrying an *attC* site in recipient cells where the only way for them to be maintained is to recombine. The transformed DNA is fully double-stranded, and is not replicated. Cruciform extrusion is thus *a priori* the only way for the *attC* site to fold and recombine. We observed recombination events for all tested sites, with a frequency of up to 10^{-3} (recombined plasmids / transformed plasmids) suggesting that cruciform extrusion may explain a significant part of recombination events (Figure 6 of the paper). This result was quite unexpected, as previous work had shown that cruciform extrusion is very unlikely as soon as the inverted repeat is imperfect (Benham, Savitt et al. 2002). We were thus surprised to see recombination events in this assay, notably for sites with large VTS such as the WT $VCR_{2/1}$ site. However, our assay enables to detect cruciform formation events occurring at low frequencies, which previous techniques could not achieve. The fact that recombination occurs with cruciform structure was further confirmed by looking at the effect of supercoiling on recombination. Relaxed plasmid, as expected, recombined less than supercoiled plasmids.

One of the main conclusions of this work is that cruciform formation is probably much more likely to occur than previously thought, and this even for imperfect inverted repeats. Furthermore, we provide the first example of a recombination system able to use cruciform DNA as a substrate, and were able to decipher the ssDNA formation pathways responsible for *attC* site folding.

96

Cellular pathways controlling integron cassette site folding

Céline Loot[1,2,3], David Bikard[1,2,3],
Anna Rachlin[1,2,4] and Didier Mazel[1,2,*]

[1]Institut Pasteur, Unité Plasticité du Génome Bactérien, Paris, France
and [2]CNRS, URA2171–Génétique des génomes, Paris, France

By mobilizing small DNA units, integrons have a major function in the dissemination of antibiotic resistance among bacteria. The acquisition of gene cassettes occurs by recombination between the *attI* and *attC* sites catalysed by the IntI1 integron integrase. These recombination reactions use an unconventional mechanism involving a folded single-stranded *attC* site. We show that cellular bacterial processes delivering ssDNA, such as conjugation and replication, favour proper folding of the *attC* site. By developing a very sensitive *in vivo* assay, we also provide evidence that *attC* sites can recombine as cruciform structures by extrusion from double-stranded DNA. Moreover, we show an influence of DNA superhelicity on *attC* site extrusion *in vitro* and *in vivo*. We show that the proper folding of the *attC* site depends on both the propensity to form non-recombinogenic structures and the length of their variable terminal structures. These results draw the network of cell processes that regulate integron recombination.

Introduction

Integrons are genetic elements commonly found in bacteria from diverse phyla and environments (Mazel, 2006). They are defined as gene capture systems, which incorporate exogenous open reading frames and convert them to functional genes by ensuring their correct expression (Hall and Collis, 1995). The integron platform consists of three major elements: an *intI* gene coding a site-specific recombination enzyme belonging to the tyrosine-recombinase family (Azaro and Landy, 2002) called an integrase, a primary recombination site (*attI*), and an outwards oriented promoter (P_c), which directs transcription of captured gene cassettes (Hall

*Corresponding author. Département Génomes et Génétique, Institut Pasteur, Unité Plasticité du Génome Bactérien, 25 rue du Dr Roux, Paris F-75015, France. Tel.: +33 1 40 61 32 84; Fax: +33 1 45 68 88 34;
E-mail: mazel@pasteur.fr
[3]These authors contributed equally to this work
[4]Present address: McLean Hospital, Behavioral Genetics Lab, MRC 215, Mill St Belmont, MA 02478, USA

and Collis, 1995). Gene cassettes generally contain a single gene and an imperfect inverted repeat at the 3′ end called an *attC* site. Recombination events between the *attI* and *attC* sites, leading to the insertion of gene cassettes in the platform, are the most common and efficient reactions performed by integron-associated integrases (Collis *et al*, 2001).

The length of natural *attC* sites varies from 57 to 141 bp. They include two regions of inverted homology, R″–L″ and L′–R′, that are separated by a central region that is highly variable in length and sequence (Stokes *et al*, 1997) (Figure 1A). Contrasting with their sequence heterogeneity, the *attC* sites display a strikingly conserved palindromic organization (Hall *et al*, 1991; Stokes *et al*, 1997; Rowe-Magnus *et al*, 2003) that can form secondary structures through the self-pairing of DNA strands (Figure 1B). On folding, single-stranded (ss) *attC* sites present an almost canonical core site consisting of R and L boxes separated by an unpaired central segment, two or three extrahelical bases (EHB) and a variable terminal structure (VTS) (Bouvier *et al*, 2009). The VTSs vary in length among the various *attC* sites from three predicted unpaired nucleotides such as for $attC_{aadA7}$ to a complex branched secondary structure in the larger sites such as the VCRs (*Vibrio cholerae attC* sites; Figure 1B).

It has been shown using a DNA-binding assay that IntI1 binds strongly and specifically to the bottom strand (bs) of ss *attC* DNA (Francia *et al*, 1999). *In vivo*, it has also been shown that only the bs of the *attC* site is used as a substrate during the integration of gene cassettes (Bouvier *et al*, 2005), thereby creating a pseudo-Holliday Junction (pHJ) between the ss *attC* and the double-stranded (ds) *attI* site. It is currently believed that this pHJ is resolved through host processes, but these have yet to be identified.

Many questions about the specific nature of these unique gene capture systems still remain unanswered. Specifically, we do not yet know how and when the *attC* sites fold inside the cell, or the identity of the factors influencing these processes.

It has long been clear that structures more complex than the canonical double helix B-form DNA are biologically important. These include unpaired or mismatched bases, triplex DNA, Z-form DNA, hairpin loops, cruciforms and Holliday junctions (Bacolla and Wells, 2004). DNA secondary structures have been detected in both prokaryotes and eukaryotes and can be assembled essentially by two pathways: from an ssDNA by generating a hairpin structure or from dsDNA by extrusion of cruciform structures (Figure 1C).

The first pathway requires ssDNA, which can be produced during conjugation, natural transformation, viral infection, replication, transcription and DNA repair. Here, we examined the effect of ss availability mediated by two cellular processes considered as the principal sources of ssDNA in bacterial cells, conjugation and replication (Figure 1C). During conjugation, the transferred DNA is in essence single-stranded, whereas during replication, the lagging strand contains a
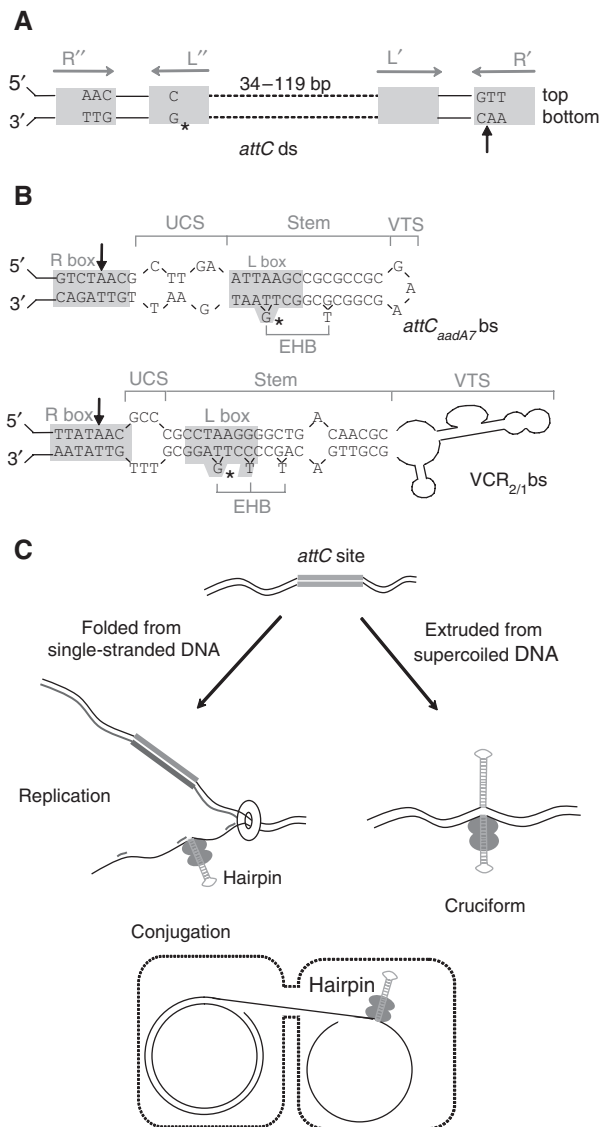
Figure 1 *attC* recombination sites and model for the *attC* folding. (**A**) Schematic representation of a double-stranded (ds) *attC* site. Inverted repeats R″, L″, L′ and R′ are indicated by grey boxes. The dotted lines represent the variable central part. The conserved nucleotides are indicated. Asterisks (*) show the conserved G nucleotides, which generate extrahelical bases (EHB) in the folded *attC* site bottom strand (bs). The black arrow shows the cleavage point. (**B**) Secondary structures of the *attC$_{aadA7}$* and VCR$_{2/1}$ sites bottom strands (bs). Structures were determined by the UNAFOLD online interface at the Institut Pasteur. The four structural features of *attC* sites, namely, the unpaired central segment (UCS), the EHB, the stem and the variable terminal structure (VTS) are indicated. Black arrows show the cleavage points. Primary sequences of the *attC* sites are shown (except for the VTS of the VCR$_{2/1}$ site). (**C**) Cellular processes possibly allowing the *attC* site folding. The different possible cellular pathways allowing proper folding of the *attC* site bottom strand are shown: from single-stranded DNA during replication and conjugation, and cruciform extrusion from supercoiled double-stranded DNA. IntI monomers are represented as grey ovals.

region of ssDNA reflecting the length of an Okazaki fragment (1–2 kb). These may, therefore, provide an opportunity for the formation of DNA secondary structure. Our results show that the availability of ssDNA in the cell during conjugation and/or during replication (lagging strand) could favour the

folding, and thus recombination, of *attC* sites of various lengths. However, the results could not be fully explained by these ssDNA production pathways, suggesting that other processes are likely involved. We thus considered the possibility that *attC* sites could be extruded from double-strand DNA into a cruciform structure (Figure 1C). Indeed, inverted repeats (perfect or imperfect) have the potential to form branched structures called cruciforms, in which inter-strand base pairing within the symmetric region is replaced by intra-strand base pairing (Courey, 1999). The formation of a cruciform by an inverted repeat involves a great deal of structural disruption as it requires a complete reorganization in base pairing. Furthermore, superhelicity is expected to directly influence cruciform extrusion. We tested this possibility as a second *attC*-folding pathway, and found that the *attC* sites could recombine from DNA essentially under ds form, after extrusion of the cruciform structure in a superhelicity-dependent manner. We confirmed these results by *in vitro* cruciform detection. We also showed that two parameters are implicated in the proper folding of the bs of the *attC* site: the length of the VTS and the propensity of the *attC* site to form a non-recombinogenic structure. Our results suggest that the contribution of these different processes varies as a function of the length and the sequence of the *attC* sites.

These results show that ssDNA structures, be they generated from replication, conjugation or extruded from dsDNA, can be recruited for specific processes such a site-specific recombination. The interplay between these cellular processes governs folding of *attC* sites, and certainly allows regulation of integron recombination by the host.

Results

Influence of single-strand DNA availability during conjugation on attC folding

We have earlier shown using a conjugation-based assay that the *attC* sites, contrarily to the *attI* sites, recombine as a folded structure generated from the bs of the *attC* site (Bouvier *et al*, 2005) (Figure 1B).

Natural *attC* sites sizes vary from 57 to 141 nt. To determine whether these size limits result from the constraints linked to the folding of ssDNA, we made a series of 21 *attC* site derivatives (Figure 2; Supplementary Figure S1 and Supplementary Table S1 and S2). Starting from the VCR$_{2/1}$ (the 123 bp signature *attC* site of the *V. cholerae* superintegron) (Mazel *et al*, 1998), we increased the length of the stem and/or VTS (up to sites of 180 nt). Inversely, we serially deleted the central part of the VCR$_{2/1}$ site to make shorter derivatives (down to 56 nt). Three wild-type *attC* sites were also tested: *attC$_{aadA7}$*, *attC$_{ereA2}$* and *attC$_{oxa2}$*. We first tested them in the suicide-conjugation assay we earlier developed (Bouvier *et al*, 2005). This assay uses conjugation, which proceeds exclusively through ssDNA transfer, to deliver the *attC* site in ss form to a recipient cell expressing the IntI1 integrase and carrying the *attI1* recombination site. The *attC* site provided by conjugation is carried on pSW plasmid that cannot replicate in the recipient (Demarre *et al*, 2005). This system permits the delivery of DNA in ss form to provide a substrate for recombination. We verified that all the constructed pSW::*attC* plasmids were transferred at similar rates ($\sim 3 \times 10^{-1}$) using a recipient strain able to sustain pSW replication.
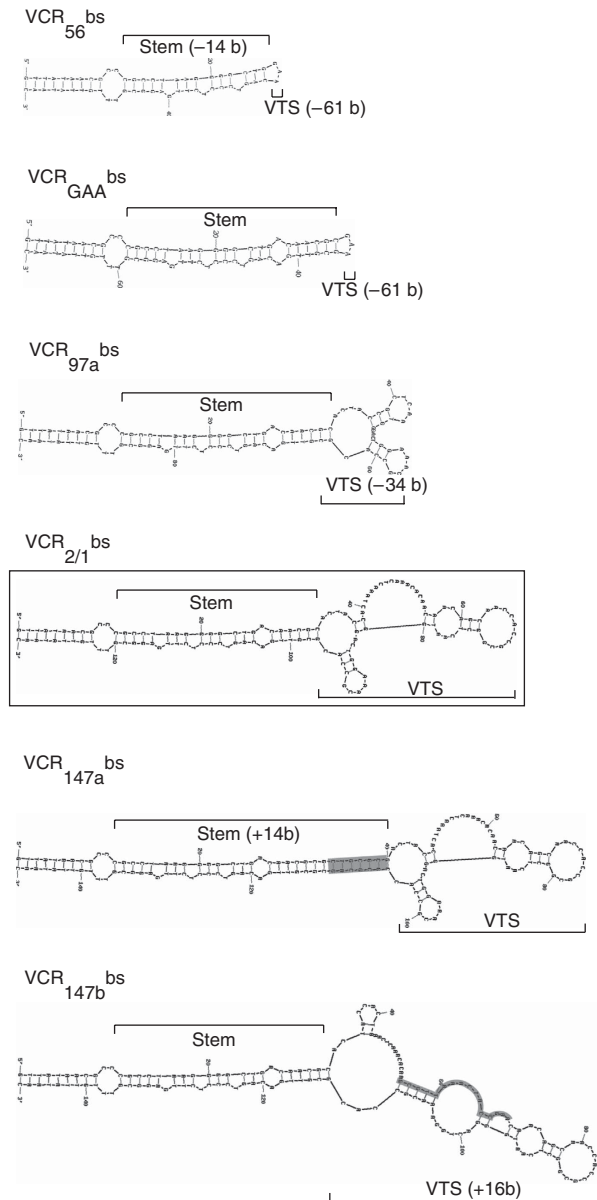
Figure 2 Predicted secondary structures of several constructed $VCR_{2/1}$ site derivatives. Secondary structures were determined using the UNAFOLD online interface of the Institut Pasteur. The *attC* sites are classified according to their size (smallest to largest). The natural $VCR_{2/1}$ site is boxed. Modifications made from the natural $VCR_{2/1}$ site in the stem and/or in the VTS are indicated.

The results are shown in Figure 3A. Recombination frequencies (see Supplementary Table S3) are plotted as a function of the probability of the *attC* site to fold into a recombinogenic site. We define a recombinogenic site as forming the R box, as well as having the G16 EHB of the L box (see Additional materials). The UNAfold software was used to estimate the probability to form active sites (Zuker, 2003). We observed that the recombination frequencies tended to drop for sites with very low probabilities to fold properly. Nevertheless, most of the *attC* sites displayed similar recombination frequencies when delivered by conjugation, regardless of their size or VTS length (Figure 2; Supplementary Figure S1 and Supplementary Table S3). This was in agreement with earlier observations (Bouvier *et al*, 2009).
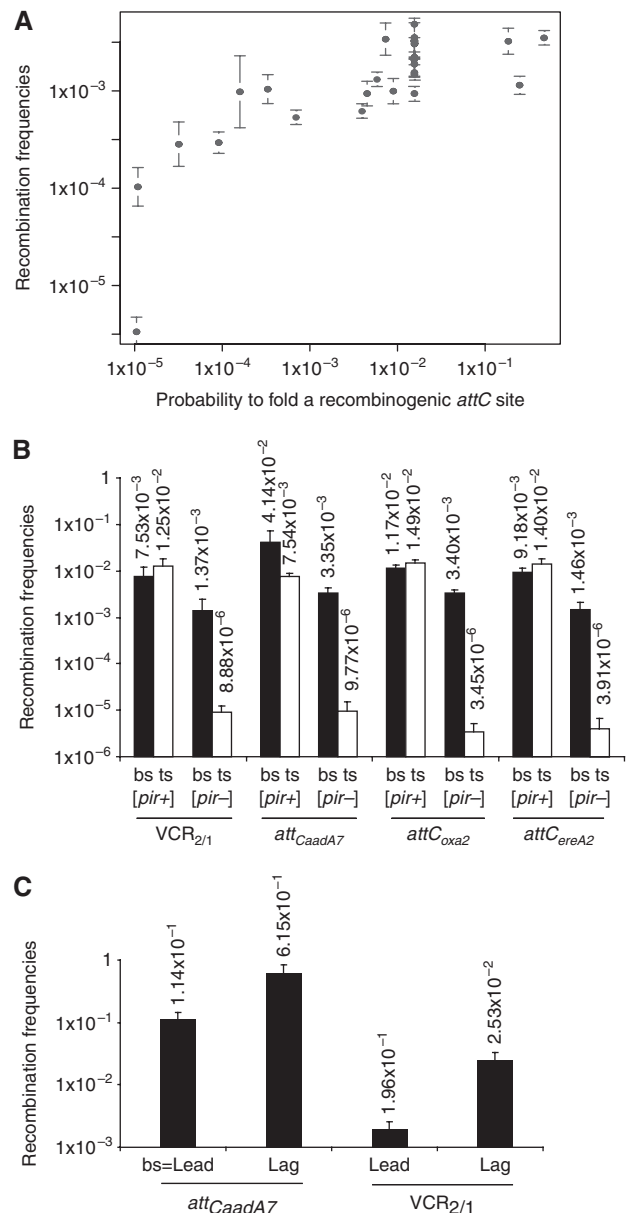
Figure 3 Influence of processes delivering single-stranded DNA on *attC* recombination. (**A**) Recombination frequencies of the different *attC* sites in the 'suicide-conjugation' assay as a function of their probability to fold a recombinogenic *attC* site (see Additional materials: suicide-conjugation assay). The UNAFOLD software was used (see Additional materials). (**B**) Recombination frequencies of four natural *attC* sites after conjugation in either a replication permissive *pir* + or non-permissive *pir*− recipient (see Additional materials: suicide-conjugation assay). Black and white columns correspond, respectively, to the recombination frequencies established when the bottom strand (bs) or the top strand (ts) is injected by conjugation. (**C**) Recombination frequencies of the $VCR_{2/1}$ and $attC_{aadA7}$ when the recombinogenic bs is carried on the lagging (lag) or leading (lead) strand of the replicating molecule (see Materials and methods: recombination assay with unidirectional-replicative substrate). Error bars show s.d.

These results suggested that if an upper limit to the length of *attC* sites exists, it is not constrained by the availability of ssDNA during conjugation.

In a second set of experiments, we compared the recombination frequencies of the *attC* sites when the bs or the top

strand (ts) was injected into recipient cells that either could or could not sustain their replication. Four natural *attC* sites were tested: $VCR_{2/1}$, $attC_{aadA7}$, $attC_{oxa2}$ and $attC_{ereA2}$. They were all cloned into the pSW::*attC*-B and pSW::*attC*-T plasmids, which permit the delivery of the bs and the ts, respectively, through conjugation (Bouvier *et al*, 2005). As earlier observed, when the ts of the *attC* site was injected into a *pir*− recipient strain, we obtained a much lower recombination frequency (Figure 3B). This confirmed that the *attC* sites are essentially ss in this assay. If they were not, one would expect similar recombination frequencies for the ts and the bs (Bouvier *et al*, 2005). We then performed the same experiment using a *pir*+ recipient cell that permitted the replication of the pSW::*attC* substrate once transferred. In these conditions, we observed for both pSW::*attC*-B and pSW::*attC*-T plasmids an equivalent high efficiency of recombination (Figure 3B). This suggested that replication can induce recombination at a similar or higher frequency than conjugation, bringing the recombination frequencies of the two transferred strands in this assay to the same level.

We conclude that recombination can happen both during the delivery of ssDNA (the bs) by conjugation and from other processes involving a replicating molecule. Recombination may indeed occur during replication-mediated DNA melting and/or during the extrusion of the *attC* site in a cruciform structure from dsDNA.

Influence of single-strand DNA availability during replication on attC folding

Replication is a process that transiently produces ssDNA. This in turn could regulate folding of the *attC* site (Figure 1C). On the basis of the differences in the dynamics between the lagging and leading strands during DNA replication (Wolfson and Dressler, 1972), proper bs *attC* folding is probably favoured by its localization on the lagging strand in which large regions of ssDNA are available (the length of the Okazaki fragments: 1–2 kb; Trinh and Sinden, 1991). If there are differences in recombination frequency based on which strand the *attC* is on, this would support the theory that the bs can be folded independently of the ts as a hairpin structure.

To test this hypothesis, we inserted an *attC* site (either $attC_{aadA7}$ or $VCR_{2/1}$) in both orientations into a unidirectional-replicating pTSC plasmid (Phillips, 1999) (Supplementary Table S1), so that the bs of the *attC* site is either on the leading or on the lagging strand. For the purpose of the experiment, the origin of replication is thermosensitive (oripSC101ts). These pTSC::*attC* plasmids were introduced into strain UB5201 with two others plasmids, one containing the *attI* site (pSU38Δ::*attI1*) and the other carrying the *intI1* gene (pBAD::*intI1*). Assays were performed at 30°C in the presence of arabinose ensuring the expression of the integrase gene. Recombination events between *attI* and *attC* sites were then selected on plates at 42°C with the pTSC::*attC* plasmid resistance marker (Cm). At 42°C, the temperature-sensitive pTSC::*attC* plasmids are unable to replicate and, therefore, cells containing these plasmids do not grow on Cm, unless there was a recombination event producing a cointegrate between the pTSC::*attC* and pSU38Δ::*attI1* plasmids. The results are shown in Figure 3C. We obtained for pSW::$attC_{aadA7}$ and pSW::$VCR_{2/1}$ a 4.2- and 12.1-fold increase in recombination, respectively, when the bs of these

attC sites corresponds to the lagging strand. As a control, the same experiment was performed on a bidirectional-replicating plasmid. As expected, we did not observe any significant differences between the two orientations for the *attC* site (data not shown).

Remarkably, there was still a high rate of recombination when bs of $attC_{aadA7}$ and $VCR_{2/1}$ were carried on the leading strand (1.14×10^{-1} and 1.96×10^{-3}, respectively; Figure 3C). Very little ssDNA is produced on the leading strand, making hairpins unlikely to fold. This led us to suspect another pathway.

Finally, these results show that ss production in both conjugation and replication influences and favours folding of the *attC* site, but other mechanisms are implicated. Specifically, we then studied the ability of the *attC* sites to extrude from dsDNA as cruciform structures able to recombine (Figure 1C).

Influence of double-strand and single-strand DNA availability on attC folding

We used all earlier tested *attC* sites derivatives (Figure 2; Supplementary Figure S1) in the replicative condition assay. Contrarily to the conjugation assay, *attC* sites are provided on a replicative plasmid and are thus mostly ds over the cell cycle, being only transiently ss during replication. In these conditions, the various sites display markedly different recombination frequencies (Supplementary Table S3). We observe a correlation between the VTS length and the recombination frequency (Figure 4A); *attC* sites containing a minimal VTS (3 nt) recombined at a frequency ranging from 1.01×10^{-1} (VCR_{GC}) to 3.53×10^{-1} ($attC_{aadA7}$), whereas *attC* sites containing a larger VTS recombined at a lower frequency, from 4.37×10^{-5} (VCR_{147c}) to 1.01×10^{-2} (VCR_{116b}). A negative effect of the VTS length on recombination is in agreement with the hypothesis of cruciform formation. Indeed, to go from a ds state to a cruciform, an *attC* site needs to melt at least the length of its VTS (see Discussion). The energy to melt this region can thus be considered as the energy of activation of cruciform formation, and thus is directly correlated to the probability of forming a cruciform after the Arrhenius equation (Supplementary Figure S2). Therefore, the larger the VTS is, the lower the probability of folding into a cruciform, which is what we observed. Furthermore, we found that the propensity of the *attC* site to form non-recombinogenic secondary structures directly impacts the recombination frequency. This is illustrated in Figure 4B, which represents the most favourable structure for the two 'a' and 'b' versions of VCR_{97} (see ΔGc and ΔG). The 'a' and 'b' versions only differ by a few bases substitutions in the VTS, but these point mutations have an impact on the formation of improperly folded structures. Indeed, the most favourable structure formed by VCR_{97a} differs greatly from the recombinogenic one and is most likely not an active substrate for recombination. On the other hand, the most stable structure for VCR_{97b} is very similar to that of the recombinogenic site. It might even be recognized by the integrase, which could help reach the active conformation. Not surprisingly, VCR_{97a} has a 20-fold lower recombination frequency than VCR_{97b}. The same effect of non-recombinogenic structures can also explain the 40-fold difference between the VCR_{116a} and VCR_{116b} (Supplementary Figure S1 and Table S3).
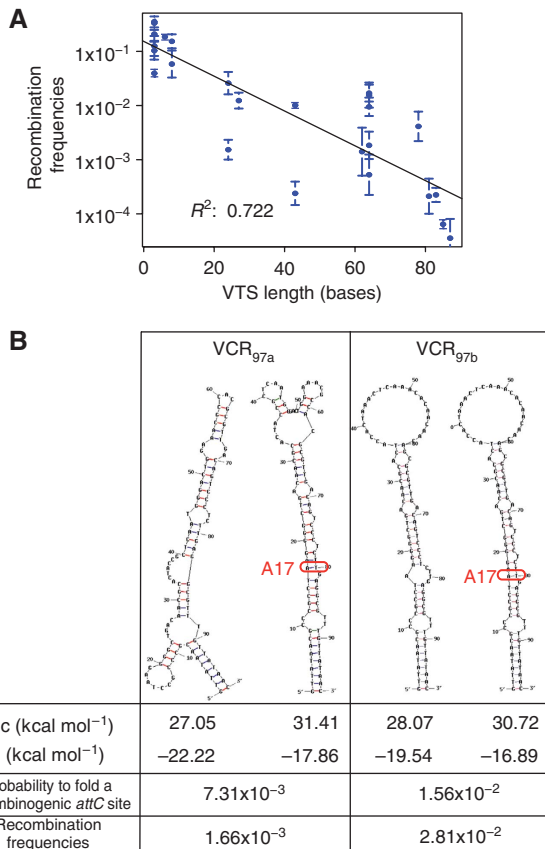
A

B

	VCR$_{97a}$		VCR$_{97b}$	
ΔGc (kcal mol^{-1})	27.05	31.41	28.07	30.72
ΔG (kcal mol^{-1})	−22.22	−17.86	−19.54	−16.89
Probability to fold a recombinogenic *attC* site	7.31x10^{-3}		1.56x10^{-2}	
Recombination frequencies	1.66x10^{-3}		2.81x10^{-2}	

Figure 4 *attC* recombination, in 'replicative' recombination conditions. (**A**) Recombination frequencies of the different *attC* sites in the 'replicative' assay as a function of their VTS length (see Additional materials: recombination assay with a replicative double-stranded substrate). Error bars show s.d. (**B**) Secondary structures of VCR97a and VCR97b as predicted by UNAFOLD. The free energy of single-strand *attC* site folding (ΔG) and of cruciform extrusion (ΔGc), the probability to fold a recombinogenic *attC* site and its measured recombination frequencies are indicated for each site. The A17, whose probability to bind the corresponding T is used as *a proxi* for the probability of the *attC* site to fold in a recombinogenic structure, is highlighted. For the calculation of the probability to fold a recombinogenic *attC* site and the free energy of cruciform formation, see Additional materials.

It thus seems that two main parameters are instrumental for the recombination frequency of *attC* sites in these conditions. First, the longer the VTS is, the lower the recombination frequency. Second, the accumulation of non-recombinogenic (improperly folded) sites dilutes the number of recombinogenic (properly folded) sites available for recombination. Note that these two parameters are not independent, as the longer is the VTS, the higher is the chance to form stable non-recombinogenic structures. An analysis of covariance performed with these two regressors (the length of the VTS and the probability of folding properly) explains 82.5% (R^2) of the experiment variance with a *P*-value of 4.72×10^{-9}, confirming the implication of these two parameters.

This, combined with the fact that high recombination frequencies can be obtained in replicative conditions when the bs of the *attC* site is on the leading strand of replication, strengthens the hypothesis that recombination can occur with *attC* sites that are extruded as cruciform structures.

Detection of *attC* cruciform structures

To test whether the *attC* sites could be directly extruded as cruciform structures from a dsDNA molecule, we performed complementary *in vitro* and *in vivo* analysis.

In vitro mapping of S1 nuclease-sensitive sites. Inverted repeat sequences in natural plasmids and phages have been shown to be centrally hyper-sensitive to cleavage by single-strand selective nucleases (Lilley, 1980; Panayotatos and Wells, 1981). Indeed, when folded into hairpins, they exhibit not only ssDNA in their loop, but also potential bulges of their stem. S1 nuclease, which cleaves ssDNA, was earlier used to probe the formation of cruciform structures *in vitro* (Noirot *et al*, 1990). We carried out the same kind of experiments to detect *in vitro* formation of *attC* cruciform structures (Figure 5A). The method consists of treating supercoiled plasmid DNA containing the *attC* sites with S1 nuclease. This is followed by a restriction digest with an enzyme having a single site in the molecule, which produces pairs of fragments arising from molecules linearized by S1 nuclease. Restriction enzymes used and expected band sizes (corresponding to the cruciform detection) are mentioned in Figure 5B. This assay showed that cruciform formation was occurring at detectable rates from the *attC*$_{aadA7}$ site, but not from the VCR$_{2/1}$ site (Figure 5C). These results coincided with the fact that *attC*$_{aadA7}$ has a very short VTS (3 nt) relative to that of VCR$_{2/1}$ (61 nt), allowing cruciform extrusion at a much higher frequency. To confirm this observation, we performed the same experiment, but using a mutant derivative of *attC*$_{aadA7}$, the *attC*$_{aadA7Mut3}$ site and a mutant derivative of VCR, the VCR$_{GAA}$ site (Figure 5A and C). The mutations in the *attC*$_{aadA7Mut3}$ disrupt the base pairing of the upper part of the stem and have been shown to decrease the recombination by >100-fold (Bouvier *et al*, 2005). As expected, we failed to detect bands indicating cruciform extrusion of this *attC*$_{aadA7}$ site derivative. On the contrary, VCR$_{GAA}$, which contains a small VTS, ensured cruciform extrusion at a higher level than the natural VCR site. These results coincide with the high recombination frequency of *attC*$_{aadA7}$ (3.53×10^{-1}) and VCR$_{GAA}$ (2.16×10^{-1}) in 'replicative' conditions (Supplementary Table S3).

To strengthen our hypothesis of *attC*$_{aadA7}$ site cruciform extrusion, we determined the precise boundaries of the S1 nuclease-generated fragments from a representative sample by sequencing. Figure 5D shows the distribution of the cleavage sites within the *attC*$_{aadA7}$ site and its neighbouring sequences. Among 91 sequenced clones, 73% (66/91) of the S1-cleavage sites clearly localize into *attC*$_{aadA7}$ at/near expected structural features, drawing a high-resolution structure map of the cruciform. The 25 remaining clones revealed S1-cleavage sites on both sides of the *attC*$_{aadA7}$ site up to a distance of 40 bases and could be explained by non-*attC*-specific extrusions of nearby inverted repeats (Supplementary Figure S4).

Recombination from non-replicative ds plasmid. In the 'replicative' assay described above, the *attC* site is carried by a dsDNA plasmid able to replicate and, therefore, to transiently produce ssDNA. To precisely study the ability of the *attC* site to recombine as a cruciform structure, we developed an assay ensuring the delivery of the *attC* site exclusively from dsDNA. To this end, Pir-dependent plasmids containing the *attC* sites were introduced by transformation in *pir−* deficient strains.
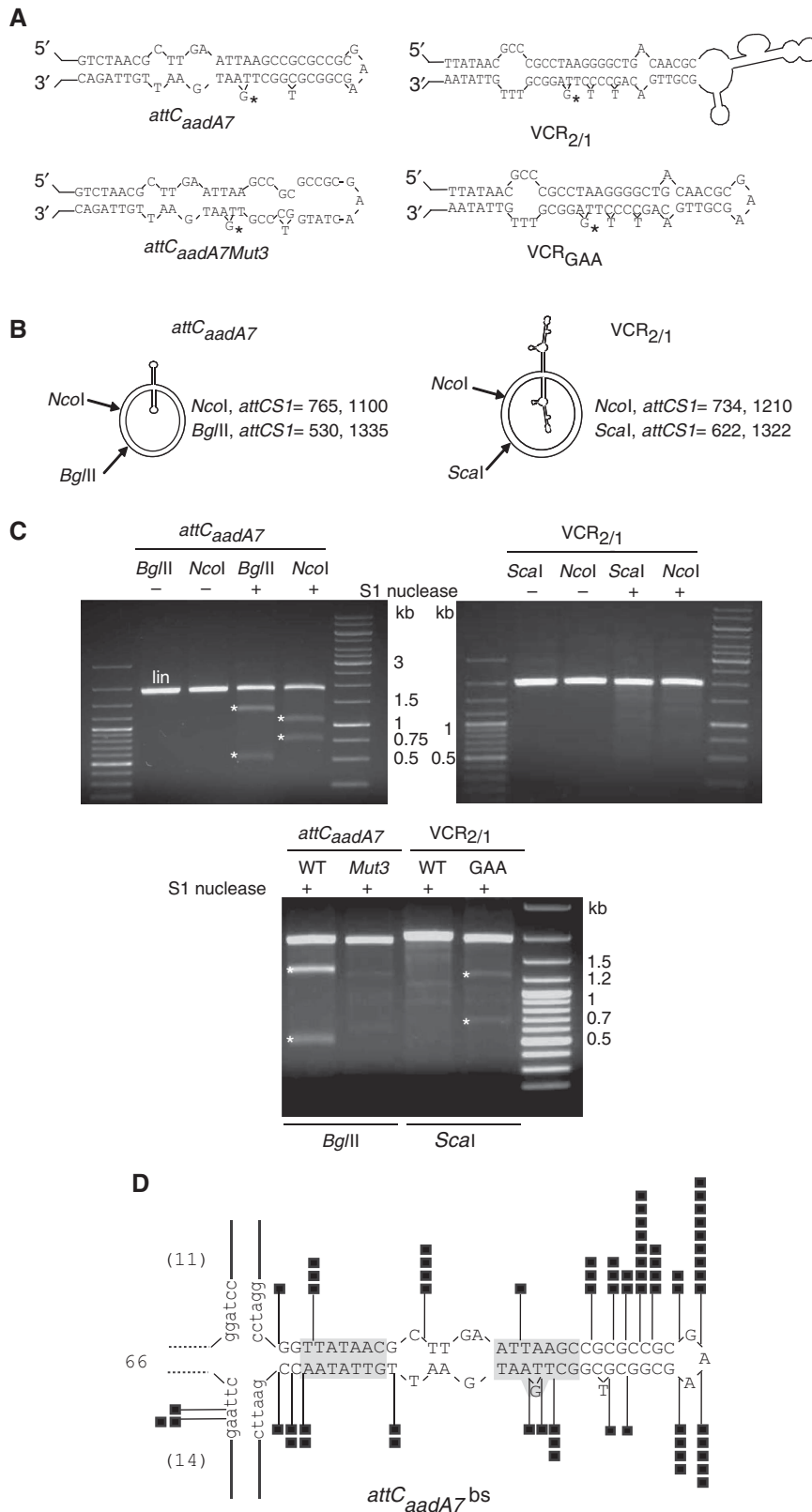
A

B

C

D

Figure 5 *In vitro* extrusion of the *attC* sites secondary structures from double-stranded DNA. (**A**) Schemes of the folded natural and mutants *attC* sites used in the S1 nuclease assay. (**B**) Schemes of the plasmids with extruded *attC_aadA7* and VCR_2/1 sites and the expected S1 fragments sizes (in bp) after cleavage by the different restriction enzymes. (**C**) *In vitro* mapping of S1 nuclease-sensitive sites. S1 nuclease-treated (+) or -untreated (−) plasmids were subjected to restriction digest (enzymes are indicated above each lane) and submitted to electrophoresis. The bands (white asterisks) smaller than the linear monomer (lin) result from S1 nuclease action and thus of cruciform extrusion. Kb, marker DNA. (**D**) Precise mapping of S1-cleavage position on the *attC_aadA7*-folded site. Each square corresponds to one S1-cleavage position in the sequenced representative sample.
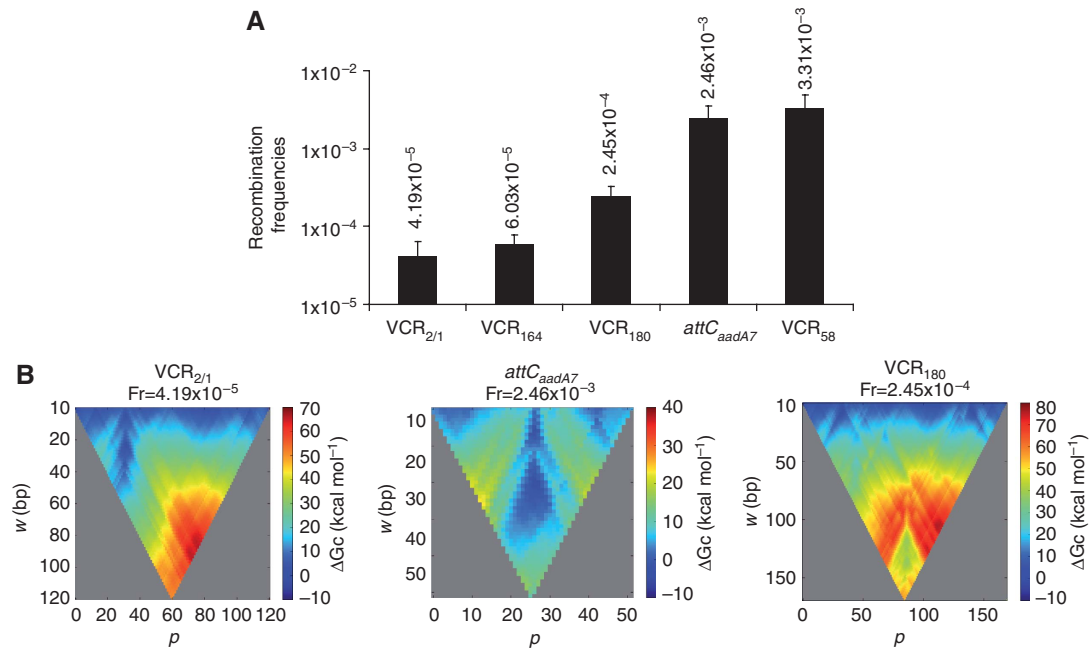
Figure 6 *In vivo* extrusion of the *attC* sites secondary structures from double-stranded DNA. (**A**) Recombination frequencies of different *attC* sites after transformation of *attC*-containing plasmids in non-replicative permissive recipient cells (see Materials and methods: recombination assay with a non-replicative substrate). Error bars show s.d. (**B**) Energy landscape of cruciform formation. The frequency of recombination (Fr) for each *attC* site is indicated. Each colour dot represents the free energy of cruciform formation (ΔGc in kcal/mol) for a portion of the *attC* site of length w (bp) at the position p (from favourable in blue to unfavourable in red). The tip of the triangle is the free energy of the whole *attC* site. As *attC* sites are symmetrical, favourable energies that are not along the centre of the triangle represent favourable non-recombinogenic structures. *attC*$_{aadA7}$ folds in a favourable recombinogenic site (blue colour from the middle of the base of the triangle). VCR$_{2/1}$ folds in a favourable non-recombinogenic structure (blue colour shifted from the middle). VCR$_{180}$ presents neither favourable recombinogenic nor favourable non-recombinogenic structures.

These plasmids can only be maintained on recombination with the *attI* site carried by a Pir-independent replicon. To establish a recombination frequency, we transform in parallel a *pir*+ permissive strain (UB5201-Pi) with the same samples of pSW-*attC* plasmids preparation. Note that the transformation efficiency of both UB5201 and UB5201-Pi strains are determined beforehand and used to adjust the recombination frequency (see Materials and methods). We found that all tested *attC* sites could lead to detectable cointegrate formation through recombination with the *attI* site (Figure 6A). These results strongly suggested that *attC* sites substrate could be formed by cruciform extrusion from dsDNA. Nevertheless, we cannot exclude the presence of nicked/damaged molecules in our plasmid preparation. Those molecules could allow *attC* site folding from the ssDNA generated during their repair and explain the obtained recombination events. However, as we failed to detect these nicked molecules by electrophoretic gel analysis, we concluded that, if they exist, these molecules represent a very minor part of the supercoiled plasmid preparation. If this minor part accounts for the majority of the recombination events, we should observe a much higher recombination frequency when all molecules are damaged. To test this hypothesis, we transformed the same cellular setup with identical quantities of either nicked or supercoiled molecules containing the *attC*$_{aadA7}$ site (see Additional materials; Supplementary Table S1 and S2). We modified the plasmids to introduce a single Nb-Bts1 endonuclease site either at 33 or 296 nt away from the *attC* site, and in the two orientations at each locations. This endonuclease only cuts one strand,

producing plasmids carrying a nick on either the bs or ts, depending on the orientation of the endonuclease site. The nicked molecules showed recombination frequencies similar to the supercoiled plasmid (Supplementary Figure S3). As none of these nicked molecules presented a higher frequency of recombination than the supercoiled ones, these results confirmed that the presence of a minor part of damaged/nicked molecules could not account for the frequency of recombination obtained in this assay. It is yet to be determined how these nicked molecules recombine. The most likely explanation is that nicked plasmids are rapidly ligated (Heitman *et al*, 1989), allowing the introduction of supercoils and recombination through *attC* site cruciform extrusion. On the other hand, we observed variations in the efficiency of recombination depending on the *attC* sites. It can certainly be explained by the formation of non-recombinogenic structures. This is illustrated by comparison of the recombination frequencies and the energy landscapes of cruciform formation (ΔGc) from VCR$_{2/1}$ and *attC*$_{aadA7}$ (Figure 6B). An improperly folded structure for the natural VCR$_{2/1}$ site is clearly visible (blue colour shifted from the middle of the base of the triangle) and correlates with its relatively low recombination frequency. In contrast, the most efficient site, *attC*$_{aadA7}$, presents a very favourable energy landscape to fold as a recombinogenic site (blue colour from the middle). The VCR$_{180}$ site presents neither, favourable energy landscape to fold a recombinogenic site, nor improperly folded structure, and recombines to an intermediate frequency of recombination. Here again, both the length of the VTS and the propensity to fold into non-recombinogenic
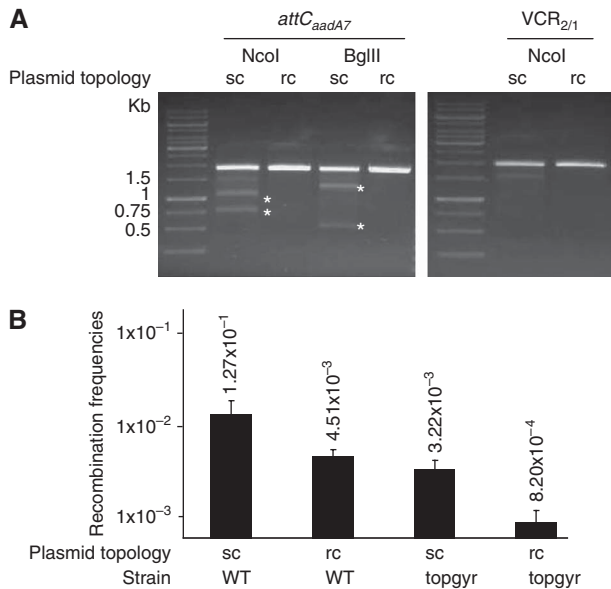
A

B

Figure 7 Influence of superhelicity on *attC* folding. (**A**) Mapping of S1 nuclease-sensitive sites of topoisomers products. Plasmid topological status is indicated by: sc, supercoiled and rc, relaxed by topoisomerase I treatment. The estimated band sizes for the extrusion of *attC*$_{aadA7}$ and VCR$_{2/1}$ in the pSW plasmid are given in Figure 5B. Kb, marker DNA. (**B**) Effect of topoisomerase I activity on the recombination frequency of the *attC*$_{aadA7}$ site in 'non-replicative' recombination assay. The assay was performed in two genetic backgrounds, wild type (WT) and *topA10 gyrB266* (topgyr, topoisomerase I and DNA gyrase-deficient strain) and using supercoiled and relaxed *attC*-containing plasmids.

structures seem to explain the recombination frequencies of the *attC* sites.

These results confirmed that substrate *attC* sites could be formed by cruciform extrusion from dsDNA and, therefore, lead us to test the effect of superhelicity on the recombination of *attC* sites under dsDNA form.

Influence of superhelicity on integron recombination

Cruciform extrusion requires the opening of the DNA double helix to allow intra-strand base pairing. The free energy held by the supercoiled molecule in the form of torsional underwinding is required for stabilization of cruciforms (Lilley *et al*, 1985). We first tested the effect of supercoiling on *attC* cruciform extrusion *in vitro*. For these, we used topoisomerase I, which catalyses the relaxation of negatively supercoiled DNA by introducing single-strand breaks that are subsequently religated. Relaxed DNA cannot stabilize cruciform structures. As expected, we observed an inhibitory effect of topoisomerase I treatment on *attC*$_{aadA7}$ cruciform extrusion *in vitro* (Figure 7A).

Second, to investigate the influence of supercoiling on the *attC* site folding *in vivo*, we used the earlier sample of negatively supercoiled and relaxed plasmids to transform a JTT1 WT strain containing the *attI* site and expressing the integrase. As the JTT1 strain is a *pir*− deficient strain, the Pir-dependent plasmids containing the *attC* sites cannot be selectively maintained without recombination with the *attI* site carried by the Pir-independent replicon. To establish a recombination frequency, we transformed in parallel a *pir*+ permissive strain (UB5201-Pi) with the same negatively

supercoiled and relaxed plasmids samples. Here again, the transformation efficiency of the both JTT1 and UB5201-Pi strains were determined beforehand and used to adjust the final ratio and normalize the results (see Materials and methods).

Negatively coiled plasmids appeared to recombine 2.8 times better than plasmids that are relaxed (Figure 7B). However, in this experiment, the topoisomerases and gyrases of the WT strain could act on the transformed plasmids by changing their supercoiling state before they might be recombined. Therefore, we repeated the experiment in the SD7 strain, a *topA10 gyrB266* derivative of JTT1, which, because of its mutations in topoisomerase I and DNA gyrase, has a lower intracellular level of negative supercoil density (Napierala *et al*, 2005). Calculation of the average super-helical density ($-\sigma_{av}$) of the pUC19 reference plasmid isolated from JTT1 and SD7 strains were 0.057 and 0.049, respectively (Napierala *et al*, 2005). In SD7, supercoiled plasmids recombined 3.9 times more efficiently than the relaxed ones (Figure 7B). Thus, superhelicity directly affects integron recombination, supporting the cruciform extrusion pathway for *attC* site recombination.

Discussion

DNA secondary structures

Extensive studies of DNA secondary structure during the past decades have shown that DNA is a dynamic molecule whose structure depends on the underlying nucleotide sequence and is influenced by the environment and the overall DNA topology. Several non-B-DNA structures have been described (Z-DNA, triplex DNA, unpaired DNA bases and hairpin or cruciform structures), which can be formed under physiological conditions. Circumstantial evidence suggests that secondary structures may serve functions in processes such as transcription regulation (Hartvig and Christiansen, 1996; Dai and Rothman-Denes, 1999), conjugation (Guasch *et al*, 2003), initiation of replication (Khan, 2005) or replication slippage (d'Alencon *et al*, 1994; Bierne *et al*, 1997). More recently, it has been shown that secondary structures are implicated in the recombination of genetic elements. Indeed, Xer recombinases can promote the direct integration of the (+)ssDNA genome of CTX into the ds *dif* site of *V. cholerae*. ssDNA substrates for integration can fold into a stem-loop structure, creating a small region of duplex DNA that is the target of site-specific recombinases (Val *et al*, 2005). Secondary structures are also implicated in the transposition of IS*608* of *Helicobacter pylori*, in which the TnpA transposase recognizes and cleaves only the top strand of the IS*608* ends that have folded into hairpin structures (Ton-Hoang *et al*, 2005; Guynet *et al*, 2008).

In the integron recombination process, the *attC* site is recognized by the integrase as an ss-folded substrate (Bouvier *et al*, 2005, 2009; MacDonald *et al*, 2006; Frumerie *et al*, 2010). To adopt this structural state, the DNA double helix needs the opening of inter-strand base pairing. As analysis of the integrase protein sequence failed to identify any helicase domain, we investigated this hypothesis that folding of the *attC* site could either be spontaneous and energy driven in supercoiled DNA, or could be driven by host factors. In this study, we tested the contribution of two pathways that could be involved: hairpin formation linked to

single-strand availability and/or a cruciform extrusion from supercoiled dsDNA.

For these, we used three different *in vivo* approaches, which differ significantly in terms of *attC* copy numbers and replicative status of the DNA. The 'suicide-conjugation assay' delivers by conjugation one copy of ss *attC*-containing plasmid (non-replicative) in the recipient cell. The 'non-replicative assay' delivers by transformation, one or more copies of the *attC*-containing non-replicative plasmid in the recipient cell. In the 'θ-replicative assay', a large number of replicative copies of the *attC*-carrying plasmid is contained in the recipient cell. Note, that the experimental procedures of these approaches are different and that it would be unwise to compare their respective recombination efficiencies.

Single-strand availability

Single-strand DNA production is obviously the most straightforward process allowing the folding of DNA into secondary structures. ssDNA is central for most examples of horizontal gene transfer. During natural transformation of bacteria, one DNA strand is taken up into the cytoplasm, whereas its complementary strand is degraded. During conjugation, ssDNA is unwound from the duplex plasmid and transferred into the recipient bacterium. During infection, filamentous phages such as M13, MV-L51 or φX174 are known to inject ssDNA. Inside the cell, there are essentially three processes that create ssDNA, replication, repair and transcription. During transcription, the size of the ssDNA is limited by the maximal size of the transcription bulge (25 nt) (Gamper and Hearst, 1982), and is likely not implicated in *attC* folding, as this size is too small. During DNA repair, the processing of double-strand breaks by the RecBCD complex is a significant source of RecA nucleofilaments, (i.e. the assemblage of RecA monomers on ssDNA) (Spies *et al*, 2005). Replication seems very appropriate to favour DNA secondary structures. During this process, after the melting of dsDNA by the replication machinery, an asymmetric fork is created in which one of the two strands (the lagging strand) provides a large quantity of ssDNA. Published observations suggest that secondary structures are easily made when carried on the lagging strand (Trinh and Sinden, 1991).

We carefully analysed the impact of conjugation and θ-replication processes on the folding of the *attC* site bs and on the integron recombination. We showed that conjugation ensures the folding of *attC* sites containing a larger VTS (e.g. VCR) or a shorter VTS (e.g. *attC_{aadA7}*) with even efficiency. We also showed that when carried on the lagging strand of the replicated DNA, the *attC* bs is recombined at a higher rate, showing that the availability of ssDNA impacts the recombination frequency of *attC* sites. These results show that replication can regulate the efficiency of integron recombination. We, therefore, examined the orientation of *attC* sites on all the chromosomal integrons encountered in sequenced bacterial genomes. We observed that the bs of all *attC* sites were located on the leading strand (Supplementary Figure S5). This specific orientation of *attC* sites could limit cassette rearrangements in chromosomal integrons. Nevertheless, it has been shown that DNA damage can uncouple the replication of the leading and lagging strand forming a partially ds molecule with an ss region of about 1 kb on the leading strand (Pages and Fuchs, 2003; Wang, 2005; Langston and O'Donnell, 2006). So, in this precise situation, which can be associated with stress conditions, we cannot exclude a high frequency of integron recombination even if the *attC* site bs is carried by the leading strand, thus ensuring a rapid adaptation of the stressed cells.

Double-strand extrusion

DNA sequences that possess two-fold symmetry may re-organize their base pairing to form cruciform structures, in which there is local intra-strand hydrogen bonding (Lilley, 1980). Nevertheless, cruciforms are intrinsically less stable than the unbranched duplex DNA from which they are derived. Supercoiling provides free energy that may be used to stabilize unstable structural polymorphs. In particular, numerous *in vivo* methods have indicated that the superhelical density varies between −0.025 and −0.05 (Zheng *et al*, 1991). These values may be too low for cruciform formation. However, many factors (transcription, growth conditions, stress, topoisomerase I…) may transiently increase the local superhelical density to a critical level sufficient for cruciform extrusion (see review; Pearson *et al*, 1996). Indeed, evidence exists for the formation of cruciforms *in vivo* and implicates these non-B-DNA structures in various cell functions (see review; Pearson *et al*, 1996). Here, we studied *attC* folding as a cruciform structure and the implication of supercoiling in integron recombination. We monitored *in vitro* cruciform formation by detection of changes in nuclease sensitivity caused by the formation of these structures. We also presented an *in vivo* study that shows the ability of *attC* sites to extrude from dsDNA as cruciform structures. This assay consists of the transformation of supercoiled pSW::*attC* plasmids into a recipient strain in which they cannot replicate. The *attC* sites are carried by dsDNA and would mostly recombine after cruciform extrusion. In these conditions, we obtained significant recombination frequencies for all the tested *attC* sites. As expected, we observed a correlation between recombination frequency of the *attC* site as a cruciform structure and the length of the VTS. The propensity to form non-recombinogenic structures also seems to influence the formation of the proper cruciform. Nevertheless, it is quite surprising that sites with large VTS recombine at all. Those sites have a very unfavourable energy of cruciform formation even in highly supercoiled DNA, and spontaneous transitions to a cruciform state would be expected to occur with a much smaller probability than the observed recombination frequencies. This suggests that host proteins can favour cruciform formation in a process that is yet to be identified. It is important to note that in the conjugation assay, these two parameters (VTS size and presence of non-recombinogenic structures) do not seem to influence integron recombination. This is probably because of the fact that the *attC* sites are delivered as ssDNA and that, in these conditions, they have ample time to fold and be captured by the integrase even though they might have large VTS and non-recombinogenic structures.

We also studied the influence of superhelicity on *attC* folding. The dynamic balance between the activities of DNA gyrase and DNA topoisomerase I maintains the level of supercoiling in *Escherichia coli* (Pruss and Drlica, 1986; Lodge *et al*, 1989). Therefore, by using topoisomerase I and gyrase-deficient strains, as well as *in vitro* topoisomerase I-treated plasmids, we showed a significant effect of supercoiling on proper *attC* folding and recombination.

Until now, extrusions from only perfect palindromes have been observed *in vivo* (Pearson *et al*, 1996). Extrusion from an imperfect palindrome has only been observed in AT-rich sequences *in vitro* using two-dimensional gel electrophoresis (Benham *et al*, 2002). It has been shown that imperfections have major effects on the overall energetics of cruciform extrusion and on the course of this transition.

It had been earlier shown that intra-strand base pairing aptitude (i.e. the palindromic structure), not primary sequence, conditions the *attC* site recombination (Bouvier *et al*, 2005, 2009). As modification of the structural properties of the *attC* site directly affects the recombination efficiency, *attC* site folding is likely the limiting step in the recombination process. Consequently, the frequencies obtained in the transformation assay (recombination from ds plasmid) suggest a low probability of *attC* cruciform formation. This corroborates the fact that cruciform structures were never observed for imperfect palindromes that are not particularly AT rich.

We propose a stabilizing effect of the integrase, which could capture *attC* sites on their extrusion and recombine them efficiently. We obtained preliminary results that support this hypothesis (C Loot, V Parissi, D Bikard and D Mazel, in preparation) and we are currently exploring the parameters of this stabilization process. This type of stabilizing effect on cruciform formation has been earlier observed. For instance, in a study of the effect of S1 nuclease on cruciform extrusion, Singleton and Wells (1982) obtained data supporting the fact that the nuclease may exert a transient stabilizing effect on cruciform formation. Similarly, Noirot *et al* (1990) showed the ability of the initiator RepC protein to enhance cruciform extrusion from the pT181 origin of replication.

In this study, we chose to develop an *in vivo* assay (recombination from non-replicative ds plasmid), which, contrary to the other classical assays earlier used, allowed us to observe low probability *attC* site cruciform extrusion (from 10^{-3} down to 10^{-5}), and directly showed the implication of cruciform structures in integron recombination. Moreover, the fact that only the *attC* bs can recombine allowed us to separate hairpin formation occurring on the lagging strand from hairpin formation in cruciform structures (see Figure 3C).

About the constraints of the natural sites

Sites such as *attC*$_{aadA7}$ seem evolutionarily optimized as they display very favourable folding. On the contrary, the VCR sites display large VTS and non-recombinogenic structures hindering recombination.

In addition to those constraints exerted on the *attC* sites for their efficient folding, we found that their propensity to accumulate mutations could also affect their upper length limit (141 bp for *attC*$_{qacE}$; Stokes *et al*, 1997). Indeed, during the construction of the VCR$_{180}$ site, we obtained a much higher proportion of mutations than with the smaller sites. Long palindromes pose a threat to genome stability by hindering passage of the replication fork. It is known that cells have evolved a post-replicative mechanism for the elimination and/or repair of large DNA secondary structures using the SbcCD endonuclease (Leach, 1994). Thus, for the larger sites, if they can fold well, they would hinder replication and thus be unstable. Conversely, large sites that fold poorly would have recombination frequencies too low to be selected.

Folded *attC* sites as sensors of environmental stress

Cruciforms have lower thermodynamic stability than regular duplex DNA. They have been observed only in negatively supercoiled molecules, in which the unfavourable free energy of formation is offset by the topology of the torsionally stressed molecule. This can be a disadvantage, as cruciform structures can be observed only in relatively large supercoiled DNA circles, and are destabilized when a break is introduced at any location. In *E. coli*, superhelicity has been shown to vary considerably during cell growth and to change in different growth conditions (Balke and Gralla, 1987; Jaworski *et al*, 1991). The level of superhelicity can also vary between bacterial strains. Indeed, the average supercoil density of a pBR322 reporter plasmid extracted from mid-log cultures of *Salmonella* is 13% lower ($\sigma = -0.060$) than that from *E. coli* ($\sigma = -0.069$) (Champion and Higgins, 2007).

In addition, biological processes such as transcription may generate domains of supercoiling on circular DNA (Liu and Wang, 1987). On the other hand, changes in supercoiling in response to external and/or internal stimuli could have a significant function in the formation and stability of cruciform *attC* sites. Indeed, analysis of topology of reporter plasmids isolated from SOS+ and SOS− strains revealed higher levels of negative supercoiling in strains with the constitutively expressed SOS network, suggesting a link between the induction of bacterial SOS repair and changes in DNA topology (Majchrzak *et al*, 2006).

Integron recombination could not only be controlled by all processes implicated in the variation of DNA topology, but also we showed that it is controlled by the processes, which generate ssDNA (conjugation, replication…). Interestingly, ssDNA can also be produced in response to external stimuli such as environmental stress. For example, recently, it has been shown that in *Streptococcus pneumoniae*, a Gram+ bacterium, competence and, therefore, single-strand production is induced by an antibiotic stress response (Prudhomme *et al*, 2006). In *Bacillus subtilis*, the competence state has been found to be required for the cell to revert point mutations in auxotrophic alleles when grown on minimal medium (Robleto *et al*, 2007). These are two examples in which ssDNA production is triggered in response to stress conditions.

This novel concept for regulating the recombination of gene cassettes in integrons seems to be of considerable importance, as filamentous phages (Val *et al*, 2005) and insertion sequences (Ton-Hoang *et al*, 2005) also use this mobilization mechanism. It was recently shown that integron-mediated recombination is integrated with the SOS response (Guerin *et al*, 2009), which in turn is activated by the production of the substrate for integron-mediated recombination—ssDNA. The use of this unconventional form of DNA as substrate allows another level of regulation in the integron recombination process. The results presented here confirm the position of integrons as an integrated adaptative system and strengthen the model in which folded ssDNA can 'serve' as sensor of the environmental stress triggering bacterial adaptation.

Materials and methods

Bacterial strains and media
See Supplementary data.

Plasmids

Plasmids used in these studies are described in Supplementary Table S1. Primers were obtained from Sigma-Aldrich (France) and are listed in Supplementary Table S2.

In vivo recombination assays

Suicide-conjugation assay and recombination assay with a replicative ds substrate are described in Supplementary data.

Recombination assay with unidirectional-replicative substrate. In this assay, the two natural $attC_{aadA7}$ and $VCR_{2/1}$ sites were used to analyse the effect of replication on integron recombination. To this end, we constructed two unidirectional replicating pTSC plasmids carrying the $attC$ sites in either of the two orientations. A thermosensitive origin was chosen (oripSC101ts). Each pTSC::$attC$ plasmid was introduced in a UB5201 strain containing the pBAD::IntI1 plasmid and the pSU38Δ::$attI1$ plasmid. The transformed cells were grown for 6 h at 30°C (to allow pTSC::$attC$ replication) in the presence of the respective antibiotics: Cm (pTSC::$attC$ marker), Ap (pBAD::IntI1 marker) and Km (pSU38Δ ::$attI1$ marker). The integrase was expressed by addition of 0.2% arabinose. Then, cells were plated at 42°C on Cm so that only the cells containing recombined pTSC::$attC$ could grow. A total of 1% glucose was added to the plates to repress the pBAD promoter and prevent residual recombination events. The integration activity was calculated as the ratio of cells expressing the Cm^R marker to the total number of Ap^R Km^R clones.

Recombination assay with a non-replicative substrate. This assay supplies the $attC$ site on a ds plasmid that cannot replicate once introduced into the recipient cell by transformation. For these, we transformed a $pir-$ strain (UB5201) containing the pBAD::IntI1 and the pSU38Δ::$attI1$ plasmids with 200 ng of the pSW::$attC$ plasmids. Competent cells were prepared in the presence of 0.2% arabinose to allow integrase expression. Transformants were selected on Cm^R (the pSW::$attC$ marker). As pSW::$attC$ cannot replicate in the UB5201 strain, Cm^R clones correspond to $attC \times attI$ recombination events. To establish the recombination activity, we in parallel transformed a $pir+$ strain (UB5201-Pi), which allows the replication of the pSW::$attC$ plasmids. Transformants were selected for Cm^R (the pSW::$attC$ marker). The recombination activity corresponds to the ratio of Cm^R clones obtained in $pir-$ conditions to those obtained in $pir+$ conditions. Note that the efficiency of transformation of each strain was determined beforehand and used to adjust the final ratio and normalize the results. To study the superhelicity effect, we performed the same assay by transforming the JTT1 and SD7 strains with the same quantity (200 ng) of supercoiled and relaxed (topoisomerase I-treated) plasmids.

Topoisomers production

See Supplementary data.

In vitro detection of cruciform

Potential cruciform loops on the pSW::$attC_{aadA7}$, pSW::$attC_{aadA7Mut3}$, pSW::$VCR_{2/1}$ and pSW::VCR_{GAA} plasmids were detected by S1 nuclease sensitivity and digested with NcoI or BglII for the pSW::$attC_{aadA7}$ and pSW::$attC_{aadA7Mut3}$ plasmids and NcoI or ScaI for the pSW::$VCR_{2/1}$ and pSW::VCR_{GAA} plasmids (see Supplementary data).

Mapping of S1-cleavage position

To precisely map the S1-cleavage positions in the $attC_{aadA7}$-folded site, we purified the two fragments arising from molecules cleaved by S1 nuclease and BglII. A C nucleotides tail was added to the 3′ terminus of purified fragments using the Terminal transferase (TdT) (5′RACE Invitrogen kit, version 2.0). TdT is active on both ssDNA (protruding and recessing 3′ends) and dsDNA. The C-tailed products were then used as templates for PCR using the Abridged Anchor primer (containing a stretch of G nucleotides) and either the SW23beg primer for the larger band, or the SW23end primer for the smaller band. Amplified products were then blindly cloned using the TOPO TA cloning kit (Invitrogen). In total, 91 resulting clones were sequenced by using the MFD primer (see Supplementary data).

Analysis of recombination events and point localization

See Supplementary data.

Calculation of the probability to fold a recombinogenic attC site from ssDNA and of the free energy of cruciform formation

See Supplementary data.

Supplementary data

Supplementary data are available at *The EMBO Journal* Online (http://www.embojournal.org).

Conflict of interest

The authors declare that they have no conflict of interest.

References

Azaro MA, Landy A (2002) Chapter 7—λ integrase and the λ Int family. In *Mobile DNA II*, Craig NL, Craigie R, Gellert M, Lambowitz AM (eds) pp 118–148. Washington, DC: ASM Press

Bacolla A, Wells RD (2004) Non-B DNA conformations, genomic rearrangements, and human disease. *J Biol Chem* **279**: 47411–47414

Balke VL, Gralla JD (1987) Changes in the linking number of supercoiled DNA accompany growth transitions in Escherichia coli. *J Bacteriol* **169**: 4499–4506

Benham CJ, Savitt AG, Bauer WR (2002) Extrusion of an imperfect palindrome to a cruciform in superhelical DNA: complete determination of energetics using a statistical mechanical model. *J Mol Biol* **316**: 563–581

Bierne H, Vilette D, Ehrlich SD, Michel B (1997) Isolation of a dnaE mutation which enhances RecA-independent homologous recombination in the Escherichia coli chromosome. *Mol Microbiol* **24**: 1225–1234

Bouvier M, Demarre G, Mazel D (2005) Integron cassette insertion: a recombination process involving a folded single strand substrate. *EMBO J* **24**: 4356–4367

Bouvier M, Ducos-Galand M, Loot C, Bikard D, Mazel D (2009) Structural features of single-stranded integron cassette attC sites and their role in strand selection. *PLoS Genet* **5**: e1000632

Champion K, Higgins NP (2007) Growth rate toxicity phenotypes and homeostatic supercoil control differentiate Escherichia coli from Salmonella enterica serovar Typhimurium. *J Bacteriol* **189**: 5839–5849

Collis CM, Recchia GD, Kim MJ, Stokes HW, Hall RM (2001) Efficiency of recombination reactions catalyzed by class 1 integron integrase IntI1. *J Bacteriol* **183**: 2535–2542

Courey AJ (1999) Analysis of altered DNA structures: cruciform DNA. *Methods Mol Biol* **94**: 29–40

d'Alencon E, Petranovic M, Michel B, Noirot P, Aucouturier A, Uzest M, Ehrlich SD (1994) Copy-choice illegitimate DNA recombination revisited. *EMBO J* **13**: 2725–2734

Dai X, Rothman-Denes LB (1999) DNA structure and transcription. *Curr Opin Microbiol* **2**: 126–130

Demarre G, Guerout AM, Matsumoto-Mashimo C, Rowe-Magnus DA, Marlière P, Mazel D (2005) A new family of mobilizable suicide plasmids based on the broad host range R388 plasmid (IncW) or RP4 plasmid (IncPα) conjugative machineries and their cognate E. coli host strains. *Res Microbiol* **156**: 245–255

Francia MV, Zabala JC, de la Cruz F, Garcia-Lobo JM (1999) The IntI1 integron integrase preferentially binds single-stranded DNA of the attC site. *J Bacteriol* **181**: 6844–6849

Frumerie C, Ducos-Galand M, Gopaul DN, Mazel D (2010) The relaxed requirements of the integron cleavage site allow predictable changes in integron target specificity. *Nucleic Acids Res* **38:** 559–569

Gamper HB, Hearst JE (1982) A topological model for transcription based on unwinding angle analysis of *E. coli* RNA polymerase binary, initiation and ternary complexes. *Cell* **29:** 81–90

Guasch A, Lucas M, Moncalian G, Cabezas M, Perez-Luque R, Gomis-Ruth FX, de la Cruz F, Coll M (2003) Recognition and processing of the origin of transfer DNA by conjugative relaxase TrwC. *Nat Struct Biol* **10:** 1002–1010

Guerin E, Cambray G, Sanchez-Alberola N, Campoy S, Erill I, Da Re S, Gonzalez-Zorn B, Barbe J, Ploy MC, Mazel D (2009) The SOS response controls integron recombination. *Science* **324:** 1034

Guynet C, Hickman AB, Barabas O, Dyda F, Chandler M, Ton-Hoang B (2008) *In vitro* reconstitution of a single-stranded transposition mechanism of IS608. *Mol Cell* **29:** 302–312

Hall RM, Brookes DE, Stokes HW (1991) Site-specific insertion of genes into integrons: role of the 59-base element and determination of the recombination cross-over point. *Mol Microbiol* **5:** 1941–1959

Hall RM, Collis CM (1995) Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Mol Microbiol* **15:** 593–600

Hartvig L, Christiansen J (1996) Intrinsic termination of T7 RNA polymerase mediated by either RNA or DNA. *EMBO J* **15:** 4767–4774

Heitman J, Zinder ND, Model P (1989) Repair of the *Escherichia coli* chromosome after *in vivo* scission by the EcoRI endonuclease. *Proc Natl Acad Sci USA* **86:** 2281–2285

Jaworski A, Higgins NP, Wells RD, Zacharias W (1991) Topoisomerase mutants and physiological conditions control supercoiling and Z-DNA formation *in vivo*. *J Biol Chem* **266:** 2576–2581

Khan SA (2005) Plasmid rolling-circle replication: highlights of two decades of research. *Plasmid* **53:** 126–136

Langston LD, O'Donnell M (2006) DNA replication: keep moving and don't mind the gap. *Mol Cell* **23:** 155–160

Leach DR (1994) Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *Bioessays* **16:** 893–900

Lilley DM (1980) The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. *Proc Natl Acad Sci USA* **77:** 6468–6472

Lilley DM, Gough GW, Hallam LR, Sullivan KM (1985) The physical chemistry of cruciform structures in supercoiled DNA molecules. *Biochimie* **67:** 697–706

Liu LF, Wang JC (1987) Supercoiling of the DNA template during transcription. *Proc Natl Acad Sci USA* **84:** 7024–7027

Lodge JK, Kazic T, Berg DE (1989) Formation of supercoiling domains in plasmid pBR322. *J Bacteriol* **171:** 2181–2187

MacDonald D, Demarre G, Bouvier M, Mazel D, Gopaul DN (2006) Structural basis for broad DNA specificity in integron recombination. *Nature* **440:** 1157–1162

Majchrzak M, Bowater RP, Staczek P, Parniewski P (2006) SOS repair and DNA supercoiling influence the genetic stability of DNA triplet repeats in *Escherichia coli*. *J Mol Biol* **364:** 612–624

Mazel D (2006) Integrons: agents of bacterial evolution. *Nat Rev Microbiol* **4:** 608–620

Mazel D, Dychinco B, Webb VA, Davies J (1998) A distinctive class of integron in the *Vibrio cholerae* genome. *Science* **280:** 605–608

Napierala M, Bacolla A, Wells RD (2005) Increased negative super-helical density *in vivo* enhances the genetic instability of triplet repeat sequences. *J Biol Chem* **280:** 37366–37376

Noirot P, Bargonetti J, Novick RP (1990) Initiation of rolling-circle replication in pT181 plasmid: initiator protein enhances cruciform extrusion at the origin. *Proc Natl Acad Sci USA* **87:** 8560–8564

Pages V, Fuchs RP (2003) Uncoupling of leading- and lagging-strand DNA replication during lesion bypass *in vivo*. *Science* **300:** 1300–1303

Panayotatos N, Wells RD (1981) Cruciform structures in supercoiled DNA. *Nature* **289:** 466–470

Pearson CE, Zorbas H, Price GB, Zannis-Hadjopoulos M (1996) Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. *J Cell Biochem* **63:** 1–22

Phillips GJ (1999) New cloning vectors with temperature-sensitive replication. *Plasmid* **41:** 78–81

Prudhomme M, Attaiech L, Sanchez G, Martin B, Claverys JP (2006) Antibiotic stress induces genetic transformability in the human pathogen *Streptococcus pneumoniae*. *Science* **313:** 89–92

Pruss GJ, Drlica K (1986) Topoisomerase I mutants: the gene on pBR322 that encodes resistance to tetracycline affects plasmid DNA supercoiling. *Proc Natl Acad Sci USA* **83:** 8952–8956

Robleto EA, Yasbin R, Ross C, Pedraza-Reyes M (2007) Stationary phase mutagenesis in *B. subtilis*: a paradigm to study genetic diversity programs in cells under stress. *Crit Rev Biochem Mol Biol* **42:** 327–339

Rowe-Magnus DA, Guerout AM, Biskri L, Bouige P, Mazel D (2003) Comparative analysis of superintegrons: engineering extensive genetic diversity in the vibrionaceae. *Genome Res* **13:** 428–442

Singleton CK, Wells RD (1982) Relationship between superhelical density and cruciform formation in plasmid pVH51. *J Biol Chem* **257:** 6292–6295

Spies M, Dillingham MS, Kowalczykowski SC (2005) Translocation by the RecB motor is an absolute requirement for {chi}-recognition and RecA protein loading by RecBCD enzyme. *J Biol Chem* **280:** 37078–37087

Stokes HW, O'Gorman DB, Recchia GD, Parsekhian M, Hall RM (1997) Structure and function of 59-base element recombination sites associated with mobile gene cassettes. *Mol Microbiol* **26:** 731–745

Ton-Hoang B, Guynet C, Ronning DR, Cointin-Marty B, Dyda F, Chandler M (2005) Transposition of ISHp608, member of an unusual family of bacterial insertion sequences. *EMBO J* **24:** 3325–3338

Trinh TQ, Sinden RR (1991) Preferential DNA secondary structure mutagenesis in the lagging strand of replication in *E. coli*. *Nature* **352:** 544–547

Val ME, Bouvier M, Campos J, Sherratt D, Cornet F, Mazel D, Barre FX (2005) The single-stranded genome of phage CTX is the form used for integration into the genome of *Vibrio cholerae*. *Mol Cell* **19:** 559–566

Wang TC (2005) Discontinuous or semi-discontinuous DNA replication in *Escherichia coli*? *Bioessays* **27:** 633–636

Wolfson J, Dressler D (1972) Regions of single-stranded DNA in the growing points of replicating bacteriophage T7 chromosomes. *Proc Natl Acad Sci USA* **69:** 2682–2686

Zheng GX, Kochel T, Hoepfner RW, Timmons SE, Sinden RR (1991) Torsionally tuned cruciform and Z-DNA probes for measuring unrestrained supercoiling at specific sites in DNA of living cells. *J Mol Biol* **221:** 107–122

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415

ADDITIONAL MATERIALS

Bacterial strains and media

Bacterial strains used in this study are DH5α (Laboratory collection), Π1, β2163 (Demarre et al, 2005), UB5201 (Martinez and de la Cruz, 1990) and UB5201-Pi (Bouvier et al, 2005). JTT1 (gal-25, λ⁻, pyrF287, fnr-1, rpsL195 (StrR), iclR7, trpR72) and SD7 (JTT1 topA10 gyrB226) have been described by Pruss et al, (1982).

Escherichia coli strains were grown in Luria Bertani (LB) at 37°C or 30°C (for the thermosensible origin of replication). Antibiotics were used at the following concentrations: ampicillin (Ap), 100 µg/ml, chloramphenicol (Cm), 25 µg/ml, kanamycin (Km), 25 µg/ml. Thymidine (Thy) and diaminopimelic acid (DAP) were supplemented when necessary to a final concentration of 0.3mM. Glucose and L-arabinose were added at respectively 10 and 2mg/ml final concentration.

DNA procedures

Standard techniques were used for DNA manipulation and cloning (Sambrook et al, 1989). Restriction and DNA-modifying enzymes were purchased from New England Biolabs and Roche. DNA was isolated from agarose gels using the QIAquick gel extraction kit (Qiagen). Plasmid DNA was extracted using the miniprep or midiprep kits (Macherey-Nagel, Qiagen).

PCR were performed with the Taq DNA polymerase (Promega) according to the manufacturer's instructions. PCR products were purified using the QIAquick PCR purification kit (Qiagen). 1% agarose electrophoresis gels were used. The sequence of each constructed *attC* site was verified using an ABI BigDye Terminator v.3.1 sequencing kit and an ABI Prism 3100 Capillary GeneticAnalyzer (Applied Biosystem).

Plasmid constructions

pSW::VCRs_construction procedure

VCR mutant sites were constructed by the annealing of complementary partially overlapping primers. After annealing, the primers' end reconstitutes the *EcoRI* and *BglI* enzyme restriction sites. These products are ligated into the pSW23T plasmid linearized by *Eco*RI/*Bgl*I. Π1, a [Pir+] DH5α derivative that requires Thy to grow in MH medium, was used as a cloning strain.

pTSC::*attC*s construction procedure

attC$_{aadA7}$ and VCR$_{2/1}$ were reconstituted by the annealing of the two complementary partially overlapping primers. After annealing, the products are ligated into the pTSC29 plasmid linearized by SmaI. DH5α was used as a cloning strain.

Nicked pSW23T::*aadA7* construction procedure

The pSW23T::*aadA7* nicked plasmids were constructed by the annealing of two complementary primers reconstituting the Nt-BtsI cleavage site. We used two sets of primers which reconstitutes after annealing either MfeI or NaeI enzyme restriction sites. These products are ligated into the pSW23T::aadA7 plasmid linearized by MfeI or NaeI. The two orientations for each cloning site are selected; Π1, a [Pir+] DH5α derivative that requires Thy to grow in MH medium, was used as a cloning strain.

The nicking of each of the 4 constructions is performed by digestion of the Nt-BtsI restriction enzyme (Biolabs). The efficiency of the reaction was controlled thanks to the differential migration of the supercoiled and nicked molecules on electrophoresis gel.

Topoisomers production

We treated the pSW::*attC* substrates with the topoisomerase I protein which catalyze the relaxation of negatively supercoiled DNA. 1µg of supercoiled DNA plasmids (pSW::*attC*$_{aadA7}$ and pSW::VCR$_{2/1}$) was incubated at 37°C in a volume of 100µL with 10 units of Topoisomerase I (Biolabs) for 2h using the suggested buffer. The reactions were stopped with EDTA and proteinase K treatment and DNA was purified using the QIAquick PCR purification kit (Qiagen).

***In vitro* detection of cruciform**

Potential cruciform loops were detected by S1 nuclease sensitivity. For these, we prepared plasmids isolated from exponentially growing cells. Indeed, it has previously been shown that the level of supercoiling of plasmids is lower in starved cells (stationary phase) than during the exponential bacterial grown (Balke and Gralla, 1987). 1 µg of supercoiled plasmid DNA (pSW::*attC*$_{aadA7}$ and pSW::VCR$_{2/1}$) were incubated at 37°C in a volume of 100µL with 50 units of S1 nuclease (Fermentas) for 45 min using the suggested buffer. The reactions were stopped with EDTA and proteinase K treatment. DNA was purified using the QIAquick PCR purification kit (Qiagen) and digested with NcoI or BglII for pSW::*attC*$_{aadA7}$ and NcoI or ScaI for pSW::VCR$_{2/1}$.

Analysis of recombination events and point localization

Note for all the recombination assays, recombination frequencies correspond to the average of at least three independent trials. Recombination events were checked by Polymerase chain reaction (PCR) using the GoTaq Flexi DNA (Promega) on eight randomly chosen clones per experiment. MFD/SW23begin were used for analysis of *attI* x *attC* co-integrates formation. The recombination point was precisely determined by sequencing with SW23beg (*attI* x *attC* co-integrates), MRV (*attI* x *attC* co-integrates obtained with the pTSC plasmids). Sequences were

verified using an ABI BigDye Terminator v.3.1 sequencing kit and an ABI Prism 3100 Capillary GeneticAnalyzer (Applied Biosystem). Primers were obtained from Sigma-Aldrich (France) and are listed in Table S2.

In vivo recombination assay

<u>Suicide conjugation assay</u>

This conjugation assay was based on that of Biskri et *al.* 2005 and was previously implemented in Bouvier et *al.* 2005 (Biskri et al, 2005; Bouvier et al, 2005). Briefly, the *attC* sites provided by conjugation are carried on a suicide vector from the R6K- based pSW family that is known to use the Pir protein to initiate its own replication. This plasmid also contains an RP4 origin of transfer (*oriT*RP4). The orientation of the oriT sequence determines which of the two strands is transferred. The donor strain β2163 carries an RP4 integrated in its chromosome which requires DAP to grow in rich medium and sustains pSW replication through the expression of a chromosomally integrated *pir* gene. The recipient strain UB5201, which contains the pBAD::*intI1* [Ap^R] (expressing the IntI1 integrase) and the pSU38Δ::*attI1* [Km^R] (carrying the *attI1* site), is devoid of a *pir* gene and therefore cannot sustain replication of the suicide vector. The only way for the pSW vector to be maintained in the recipient cell is to form a co-integrate by *attC* x *attI* recombination. The recombination activity is calculated as the ratio of transconjugants expressing the pSW marker [Cm^R] to the total number of recipient clones [Ap^R, Km^R].

We also perform the same assay but using a pir+ recipient cell (UB5201-Pi) insuring the pSW::*attC* replication once transferred. In this case, after the overnight conjugation, cells are resuspended into 2 ml of LB followed by total DNA plasmid extraction. The obtained DNA is used to transform the DH5α pir- cells. The recombination activity is calculated as the ratio of cells expressing the pSW::*attC* marker [Cm^R] to the total number of Km^R clones.

<u>Recombination assay with a replicative double-stranded substrate</u>

This assay allows supplying the *attC* site on double strand replicative plasmid. Three plasmids, pBAD::*IntI1*, pSU38::*attI1* and pSW::*attC*, harboring the different *attC* site derivatives were transformed into a pir+ cell (UB5201-Pi). This strain allows the pSW::*attC* replication (see the description of the pSW plasmid family above). After overnight growth in the presence of appropriate antibiotics and 0.2% arabinose to allow the intI1 expression, cells were harvested and total plasmid DNA extracted. This was then introduced by transformation into the DH5α pir- cell. Transformants were selected for Cm^R (the pSW::*attC* marker). As pSW::*attC* cannot replicate in the pir- DH5α strain, Cm^R clones correspond to *attC* x *attI* recombination events. Recombination activity is calculated as the ratio of Cm^R to Km^R transformants. In order to control the implication of the *attC* site as a single-stranded form in this replicative test, we constructed an *attC* site lacking the double-stranded R'' box (see Table S1 and S2). In this *attC* site, the R'/R'' box of the folded *attC* site is affected. As expected, we obtained a very low frequency of recombination (1.25×10^{-6} +/- 5.06×10^{-7}).

Calculation of the probability to fold a recombinogenic *attC* site from single-stranded DNA

UNAFOLD software was used to compute the probability to form active *attC* sites from single-stranded DNA. We consider that to be folded properly, *attC* sites need to form the R and L boxes (Figure 1). A proper L box is characterized by the presence of the extrahelical G16. If we constrain the proper pairing of A17, we can observe that the most energetically favorable fold is by far the proper fold (data not shown). Based on the assumption that we are in an equilibrium state, this means that the majority of the molecules that pair the A17 properly should have this proper fold. We computed the probability to pair the A17 properly, using the hybrid-ss function of UNAFold.

Calculation of the free energy of cruciform formation

The free energy of cruciform formation (ΔGc) was computed as the sum of four terms: the energy to melt the double stranded DNA (ΔGdb), the folding energy of the bottom and top strands (ΔGbot and ΔGtop), and the energy contributed by the change in superhelicity (ΔGs). ΔGdb, ΔGbot and ΔGtop were computed using UNAFOLD. ΔGs was computed according to JF Marko and ED Siggia (Marko and Siggia, 1995) for a superhelix density of -0.06 and a plasmid size of 2kb.

	Plasmids name and description
p929	pSU38Δ::attI1, ori_{p15A} [Km^R] Biskri et *al.*, 2005
p3938	pBAD::intI1, ori$_{ColE1}$ [Ap^R], Demarre et al, 2007
p7523	pTSCaadA7 (ori-), fwd/rev aadA7 fragment in pTSC29 digested, *oripSC101ts* [Cm^R] (this study)
p7546	pTSCaadA7 (ori+), fwd/rev aadA7 fragment in pTSC29 digested,), *oripSC101ts* [Cm^R] (this study)
p7545	pTSCVCR (ori-), fwd1/rev1 and fwd2/rev2 VCR fragment in pTSC29 digested, *oripSC101ts* [Cm^R] (this study)
p7544	pTSCVCR (ori+), fwd1/rev1 and fwd2/rev2 VCR fragment in pTSC29 digested, *oripSC101ts* [Cm^R] (this study)
p4136	pSW23T::aadA7 (B), $oriT_{RP4}$, $oriV_{R6K}$ [Cm^R] (Bouvier et al., 2005)
p7945	pSW23T::aadA7 (T), $oriT_{RP4\,INV}$, $oriV_{R6K}$ [Cm^R] (this study)
p4192	pSW23T::aadA7 Mut3 (B), $oriT_{RP4}$, $oriV_{R6K}$ [Cm^R] (Bouvier et al., 2005)
p1880	pSW23T::VCR$_{2/1}$ (B), $oriT_{RP4}$, $oriV_{R6K}$ [Cm^R] (Biskri et al, 2005)
p2656	pSW23T::VCR$_{2/1}$ (T), $oriT_{RP4\,INV}$, $oriV_{R6K}$ [Cm^R] (Bouvier et al, 2005)
p3615	pSW23T::ereA2 (B), $oriT_{RP4}$, $oriV_{R6K}$ [Cm^R] (Bouvier et al., 2009)
p4392	pSW23T::ereA2 (T), $oriT_{RP4\,INV}$, $oriV_{R6K}$ [Cm^R] (Bouvier et al, 2009)
p3616	pSW23T ::oxa2 (B), $oriT_{RP4}$, $oriV_{R6K}$ [Cm^R] (Bouvier et al, 2009)
p4390	pSW23T :: oxa2 (T), $oriT_{RP4\,INV}$, $oriV_{R6K}$ [Cm^R] (Bouvier et al, 2009)
p6823	pSW23T::VCR56 (B), EcoRI/BglII fwd/rev VCR56 fragment in pSW23T digested (this study)
p6824	pSW23T::VCR58 (B), EcoRI/BglII fwd/rev VCR58 fragment in pSW23T digested (this study)
p4893	pSW23T::VCR-GAA (B), EcoRI/BglII fwd/rev VCR-GAA fragment in pSW23T digested (this study)
p4191	pSW23T::VCR-TTC (B), EcoRI/BglII fwd/rev VCR-TTC fragment in pSW23T digested (this study)
p7329	pSW23T::VCR-GC (B) EcoRI/BglII fwd/rev VCR-GC fragment in pSW23T digested (this study)
p7330	pSW23T::VCR-TA (B) EcoRI/BglII fwd/rev VCR-TA fragment in pSW23T digested (this study)
p7332	pSW23T::VCR97a (B), EcoRI/BglII fwd/rev VCR-97a fragment in pSW23T digested (this study)
p7527	pSW23T::VCR97b (B), EcoRI/BglII fwd/rev VCR-97b fragment in pSW23T digested (this study)
p7589	pSW23T::VCR100 (B), EcoRI/BglII fwd/rev VCR-100 fragment in pSW23T digested (this study)
p7333	pSW23T::VCR116a (B), EcoRI/BglII fwd1/rev1 and fwd2/rev2 VCR-116a fragment in pSW23T digested (this study)
p7528	pSW23T::VCR116b (B), EcoRI/BglII fwd1/rev1 and fwd2/rev2 VCR-116b fragment in pSW23T digested (this study)
p7114	pSW23T::VCRb (B), EcoRI/BglII fwd1/rev1 and fwd2/rev2 VCRb fragment in pSW23T digested (this study)
p6730	pSW23T::VCR139 (B), EcoRI/BglII fwd1/rev1 and fwd2/rev2 VCR-139 fragment in pSW23T digested (this study)
p6731	pSW23T::VCR147a (B), EcoRI/BglII fwd1/rev1 and fwd2/rev2 VCR-147a fragment in pSW23T digested (this study)
p6990	pSW23T::VCR147b (B), EcoRI/BglII fwd1/rev1 and fwd2/rev2 VCR-147b fragment in pSW23T digested (this study)
p7591	pSW23T::VCR147c (B), EcoRI/BglII fwd1/rev1 and fwd2/rev2 VCR-147c fragment in pSW23T digested (this study)
p6938	pSW23T::VCR147d (B), EcoRI/BglII fwd1/rev1 and fwd2/rev2 VCR-147d fragment in pSW23T digested (this study)
p6814	pSW23T::VCR164 (B), EcoRI/BglII fwd1/rev1 and fwd2/rev2 VCR-164 fragment in pSW23T digested (this study)
p7684	pSW23T::VCR180 (B), EcoRI/BglII fwd1/rev1, fwd2/rev2 and fwd3/rev3 VCR-180 fragment in pSW23T digested (this study)
p7781	pSW23T::VCRΔR'' (B), EcoRI/BglII fwd1/rev1 VCRΔR'' fragment in pSW23T digested (this study)
p8426	pSW23T::aadA7 (B), bottom strand cleaved, MfeI fwd/rev Nb-BtsI (MfeI) fragment in pSW23T digested (this study)
p8427	pSW23T::aadA7 (B), top strand cleaved, MfeI fwd/rev Nb-BtsI (MfeI) fragment in pSW23T digested (this study)
p8428	pSW23T::aadA7 (B), bottom strand cleaved, NaeI fwd/rev Nb-BtsI (NaeI) fragment in pSW23T digested (this study)
p8429	pSW23T::aadA7 (B), top strand cleaved, NaeI fwd/rev Nb-BtsI (NaeI) fragment in pSW23T digested (this study)

Table S1: Plasmids used and constructed in this study

Table S2: Primers used in this study

Sequences are given in 5' → 3' direction.

A) Primers used to generate the pSW::*attC* derivatives plasmids.

attC	sites	Sequences
VCR$_{56}$	rev	GGGCTGACAACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG
	fwd	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTTGTCAGCCCCTT
VCR$_{58}$	rev	GGGCTGCGAAGCAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG
	fwd	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGCTTCGCAGCCCCTT
VCR$_{GAA}$	rev	GGGCTGACAACGCCTTGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG
	fwd	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCAAGGCGTTGTCAGCCCCTT
VCR$_{TTC}$	rev	GGGCTGACAACGCGAAGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG
	fwd	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCTTCGCGTTGTCAGCCCCTT
VCR$_{TA}$	rev	GGGCTGACAACGCTAAAAAATGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG
	fwd	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCATTTTTTAGCGTTGTCAGCCCCTT
VCR$_{GC}$	rev	GGGCTGACAACGCGCCCCCGGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG
	fwd	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCCGGGGGCGCGTTGTCAGCCCCTT
VCR$_{97a}$	rev	GGGCTGACAACGCACTACCGCTCAATGGGACTGGAAACGCCACGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG
	fwd	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGTGGCGTTTCCAGTCCCATTGAGCGGTAGTGCGTTGTCAGCCCCTT
VCR$_{97b}$	rev	GGGCTGACAACGCACTACCACTAAACTCAAACACAACAACAGCGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG
	fwd	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGCTGTTGTTGTGTTTGAGTTTAGTGGTAGTGCGTTGTCAGCCCCTT
VCR$_{100}$	rev	GGGCTGACAACGCACTACCACTAAACTCAAACACAACAACAGCCCGGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG
	fwd	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCCGGGCTGTTGTTGTGTTTGAGTTTAGTGGTAGTGCGTTGTCAGCCCCTT
VCR$_{116a}$	rev1 rev2	GACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG GGGCTGACAACGCACTACCACTAAACTCAAACACAACAACTCAATGGGACTGGAAACGCCACGCGTT
	fwd1 fwd2	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGTGGCGTTTCCAGTCC CATTGAGTTGTTGTGTTTGAGTTTAGTGGTAGTGCGTTGTCAGCCCCTT
VCR$_{116b}$	rev1 rev2	AATGGGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG GGGCTGACAACGCACTACCACTAAACTCAAACACAACAACAGCAACCACCGCGGCTC
	fwd1	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCCCATTGAGCCGCGGTGG

	fwd2	TTGCTGTTGTTGTGTTTGAGTTTAGTGGTAGTGCGTTGTCAGCCCCTT
VCRb	**rev1** **rev2**	GAAACGCCGCGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG GGGCTGACAACGCGCTACCACTAAACTCAAACACAACAACAGCAACCACCGCGGCTCAATGGGACTG
	fwd1 **fwd2**	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGGGGCGTTTCCAGTCCCATTGAGCCGCGGTGG TTGCTGTTGTTGTGTTTGAGTTTAGTGGTAGCGCGTTGTCAGCCCCTT
VCR139	**rev1** **rev2**	ACCAGCGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG GGGCTGACAACGCGCTGACTACCACTAAACTCAAACACAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCC
	fwd1 **fwd2**	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGCTGGTGGCGTTTCCAGTCCCATTGAGC CGCGGTGGTTGCTGTTGTTGTGTTTGAGTTTAGTGGTAGTCAGCGCGTTGTCAGCCCCTT
VCR147a	**rev1** **rev2**	ACCGCGCAGCGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG GGGCTGACAACGCGCTGCGCGACTACCACTAAACTCAAACACAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCC
	fwd1 **fwd2**	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGCTGCGCGGTGGCGTTTCCAGTCCCATTGAGC CGCGGTGGTTGCTGTTGTTGTGTTTGAGTTTAGTGGTAGTCGCGCAGCGCGTTGTCAGCCCCTT
VCR147b	**rev1** **rev2**	GAAACGCCACGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG GGGCTGACAACGCACTACCACTAAACTCAAACACAAGCGTTAAACTACCGAACAACAGCAACCACCGCGGCTCAATGGGACTG
	fwd1 **fwd2**	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGTGGCGTTTCCAGTCCCATTGAGCCGCGGTGG TTGCTGTTGTTCGGTAGTTTAACGCTTGTGTTTGAGTTTAGTGGTAGTGCGTTGTCAGCCCCTT
VCR147c	**rev1** **rev2**	GACTGGAAACGCCACGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG GGGCTGACAATTTACTACCACTAAACTCAAACACAAGCGTTAAACTACCGAACAACAGCAACCACCGCGGCTCAATGG
	fwd1 **fwd2**	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGTGGCGTTTCCAGTCCCATTGAGCC GCGGTGGTTGCTGTTGTTCGGTAGTTTAACGCTTGTGTTTGAGTTTAGTGGTAGTAAATTGTCAGCCCCTT
VCR147d	**rev1** **rev2**	GAAACGCCACCGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG GGGCTGACAACGCGACTACCACTAAACTCAAACACAACGTTAAACTACCGACAACAGCAACCACCGCGGCTCAATGGGACTG
	fwd1 **fwd2**	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGGTGGCGTTTCCAGTCCCATTGAGCCGCGGTGG TTGCTGTTGTCGGTAGTTTAACGTTGTGTTTGAGTTTAGTGGTAGTCGCGTTGTCAGCCCCTT
VCR147e	**rev1** **rev2**	GAAACGCCACGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG GGGCTGACAACGTACTACCACTAAACTCAAACACAAGCGTTAAACTACCGAACAACAGCAACCACCGCGGCTCAATGGGACTG
	fwd1 **fwd2**	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGTGGCGTTTCCAGTCCCATTGAGCCGCGGTGG TTGCTGTTGTTCGGTAGTTTAACGCTTGTGTTTGAGTTTAGTGGTAGTACGTTGTCAGCCCCTT
VCR147f	**rev1** **rev2**	GACTGGAAACGCCACGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG GGGCTGACAACTTACTACCACTAAACTCAAACACAAGCGTTAAACTACCGAACAACAGCAACCACCGCGGCTCAATGG
	fwd1 **fwd2**	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGTGGCGTTTCCAGTCCCATTGAGCC GCGGTGGTTGCTGTTGTTCGGTAGTTTAACGCTTGTGTTTGAGTTTAGTGGTAGTAAGTTGTCAGCCCCTT
VCR164	**rev1** **rev2**	ACGCGGCGTCGCGCGCAGCGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG GGGCTGACAACGCGCTGCGCGCGCGCCGCACTACCACTAAACTCAAACACAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCC

	fwd1	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGCTGCGCGCGACGCCGCGTGGCGTTTCCAGTCCCATTGAGC
	fwd2	CGCGGTGGTTGCTGTTGTTGTGTTTGAGTTTAGTGGTAGTGCGGCGCGCGCGCAGCGCGTTGTCAGCCCCTT
VCR₁₈₀	**rev1** **rev2** **rev3**	CCACGCGGTGTCGGCGGTGTCGCGCGCAGCGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGATCTG ACTACCACTAAACTCAAACACAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACG GGGCTGACAACGCGCTGCGCGCGCGCCGCCGCACCGC
	fwd1 **fwd2** **fwd3**	AATTCAGATCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGCTGCGCG CGACGCCGCCGACACCGCGTGGCGTTTCCAGTCCCATTGAGCCGCGGTGGTT GCTGTTGTTGTGTTTGAGTTTAGTGGTAGTGCGGTGCGGCGGCGCGCGCGCAGCGCGTTGTCAGCCCCTT
VCR₋ᴿ''	**rev1** **rev2**	ACGCGTTGACAGTCCCTCTTGAGGCGTTTCAGATCTG GGGCTGACAACGCACTACCACTAAACTCAAACACAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCC
	fwd1 **fwd2**	AATTCAGATCTGAAACGCCTCAAGAGGGACTGTCAACGCGTGGCGTTTCCAGTCCCATTGAGC CGCGGTGGTTGCTGTTGTTGTGTTTGAGTTTAGTGGTAGTGCGTTGTCAGCCCCTT

B) Primers used to generate the pTSC derivative plasmids

Name		Sequences
VCR	**VCR rev1**	ACGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAACAGA
	VCR rev2	CCGGTTATAACGCCCGCCTAAGGGGCTGACAACGCACTACCACTAAACTCAAACACAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCC
	VCR fwd1	TCTGTTATAACAAACGCCTCAAGAGGGACTGTCAACGCGTGGCGTTTCCAGTCCCATTGAGC
	VCR fwd2	CGCGGTGGTTGCTGTTGTTGTGTTTGAGTTTAGTGGTAGTGCGTTGTCAGCCCCTTAGGCGGGCGTTATAACCGG
attC*aadA7*	***attC*₍aadA7₎ rev**	TGCCTAACGCTTGAATTAAGCCGCGCCGCGAAGCGGCGTCGGCTTGAATGAATTGTTAGGCA
	***attC*₍aadA7₎ fwd**	TGCCTAACAATTCATTCAAGCCGACGCCGCTTCGCGGCGCGGCTTAATTCAAGCGTTAGGCA

C) Primers used to generate the nicked pSW::*aadA7* derivatives plasmids

Name		Sequences
Nb-BtsI (MfeI)	**rev**	AATTGGGCACTGCGCTAGCC
	fwd	AATTGGCTAGCGCAGTGCCC
Nb-BtsI (NaeI)	**rev**	GGCGGCACTGCGCTAGCGCC
	fwd	GGCGCTAGCGCAGTGCCGCC

D) Primers used to confirm the *attC* x *attI* insertion and the *attC* x *attC* deletion events and to map the S1 cleavage sites.

Name	Sequences
MFD	CGCCAGGGTTTTCCCAGTCAC
MRV	AGCGGATAACAATTTCACACAGGA
Sw23beg	CCGTCACAGGTATTTATTCGGCG
Sw23end	CCTCACTAAAGGGAACAAAAGCTG
AAP	GGCCACGCGTCGACTAGTACGGGIIGGGIIGGGIIG

Table S3: Description of the used *attC* sites

Sequences of the bottom strand of each *attC* sites are presented. Frequencies of recombination of the *attC* derivative sites during the suicide conjugation assay (Conj) (see also Figure 3A) and during the replicative assay (Rep) (see also Figure 4A) are indicated. The VTS size of the *attC* sites and the probability to fold a recombinogenic *attC* site (Pfold) are also indicated. b=base

attC$_{aadA7}$	**Conj**=3.35×10^{-3} **Rep**=3.53×10^{-1}	**VTS size**=3b	**Pfold**=0.184	
TGCCTAACGCTTGAATTAAGCCGCGCCGCGAAGCGGCGTCGGCTTGAATGAATTGTTAGGCA				
VCR$_{56}$	**Conj**=3.08×10^{-3} **Rep**=3.96×10^{-1}	**VTS size**=3b	**Pfold**=0.0157	
GTTATAACGCCCGCCTAAGGGGCTGGAACAGTCCCTCTTGAGGCGTTTGTTATAAC				
VCR$_{58}$	**Conj**=3.08×10^{-3} **Rep**=1.08×10^{-1}	**VTS size**=3b	**Pfold**=0.0156	
GTTATAACGCCCGCCTAAGGGGCTGCGAAGCAGTCCCTCTTGAGGCGTTTGTTATAAC				
attC$_{ereA2}$	**Conj**=1.46×10^{-3} **Rep**=3.39×10^{-1}	**VTS size**=3b	**Pfold**=0.247	
CGCATAACGCGCTGATCACCGGCGGTTGAAAACCGTCCGGTGGATTGGCAGGTTATGCG				
attC$_{oxa2}$	**Conj**=3.40×10^{-3} **Rep**=1.85×10^{-1}	**VTS size**=6b	**Pfold**=0.47	
CGCCCAACGTTGAAGTAACCGGCGCTGCGCGGTTTTATCGCGCAGCGTCCGAGTTGACTGCCGGGTTGGGCG				
VCR$_{GAA}$	**Conj**=3.04×10^{-3} **Rep**=2.16×10^{-1}	**VTS size**=3b	**Pfold**=0.0156	
GTTATAACGCCCGCCTAAGGGGCTGACAACGCCTTGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC				
VCR$_{TTC}$	**Conj**=1.93×10^{-3} **Rep**=1.34×10^{-1}	**VTS size**=3b	**Pfold**=0.0156	
GTTATAACGCCCGCCTAAGGGGCTGACAACGCGAAGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC				
VCR$_{TA}$	**Conj**=3.47×10^{-3} **Rep**=1.35×10^{-1}	**VTS size**=8b	**Pfold**=0.0156	
GTTATAACGCCCGCCTAAGGGGCTGACAACGCTAAAAAATGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC				
VCR$_{GC}$	**Conj**=2.38×10^{-3} **Rep**=1.01×10^{-1}	**VTS size**=8b	**Pfold**=0.0156	
GTTATAACGCCCGCCTAAGGGGCTGACAACGCGCCCCCCGGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC				
VCR$_{97a}$	**Conj**=9.65×10^{-4} **Rep**=1.66×10^{-3}	**VTS size**=24b	**Pfold**=0.00731	
GTTATAACGCCCGCCTAAGGGGCTGACAACGCACTACCGCTCAATGGGACTGGAAACGCCACGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC				
VCR$_{97b}$	**Conj**=9.46×10^{-4} **Rep**=2.81×10^{-2}	**VTS size**=24b	**Pfold**=0.0156	
GTTATAACGCCCGCCTAAGGGGCTGACAACGCACTACCACTAAACTCAAACACAACAACAGCGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC				
VCR$_{100}$	**Conj**=1.56×10^{-3} **Rep**=1.29×10^{-2}	**VTS size**=27b	**Pfold**=0.0156	
GTTATAACGCCCGCCTAAGGGGCTGACAACGCACTACCACTAAACTCAAACACAACAACAGCCCGGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC				
VCR$_{116a}$	**Conj**=3.53×10^{-3} **Rep**=2.65×10^{-4}	**VTS size**=43b	**Pfold**=0.0045	
GTTATAACGCCCGCCTAAGGGGCTGACAACGCACTACCACTAAACTCAAACACAACAACTCAATGGGACTGGAAACGCCACGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC				
VCR$_{116b}$	**Conj**=3.32×10^{-3} **Rep**=1.01×10^{-2}	**VTS size**=43b	**Pfold**=0.0156	

GTTATAACGCCCGCCTAAGGGGCTGACAACGCACTACCACTAAACTCAAACACAACAACAGCAAC
CACCGCGGCTCAATGGGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC

VCR$_a$ **Conj**=1.89×10^{-3} **Rep**=1.11×10^{-3} **VTS size**=64b **Pfold**=0.000159

GTTATAACGCCCGCCTAAGGGGCTGACAACGCACTACCACTAAACTCAAACACAACAACAGCAAC
CACCGCGGCTCAATGGGACTGGAAACGCCACGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC

VCR$_b$ **Conj**=1.02×10^{-3} **Rep**=2.09×10^{-3} **VTS size**=62b **Pfold**=0.00896

GTTATAACGCCCGCCTAAGGGGCTGACAACGCGCTACCACTAAACTCAAACACAACAACAGCAAC
CACCGCGGCTCAATGGGACTGGAAACGCCGCGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC

VCR$_{139}$ **Conj**=2.13×10^{-3} **Rep**=9.96×10^{-3} **VTS size**=64b **Pfold**=0.0155

GTTATAACGCCCGCCTAAGGGGCTGACAACGCGCTGACTACCACTAAACTCAAACACAACAACAG
CAACCACCGCGGCTCAATGGGACTGGAAACGCCACCAGCGCGTTGACAGTCCCTCTTGAGGCGTTT
GTTATAAC

VCR$_{147a}$ **Conj**=5.65×10^{-3} **Rep**=1.77×10^{-2} **VTS size**=64b **Pfold**=0.0156

GTTATAACGCCCGCCTAAGGGGCTGACAACGCGCTGCGCGACTACCACTAAACTCAAACACAACA
ACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCCACCGCGCAGCGCGTTGACAGTCCCTCTTG
AGGCGTTTGTTATAAC

VCR$_{147b}$ **Conj**=3×10^{-4} **Rep**=2.42×10^{-4} **VTS size**=81b **Pfold**=0.0000912

GTTATAACGCCCGCCTAAGGGGCTGACAACGCACTACCACTAAACTCAAACACAAGCGTTAAACTA
CCGAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCCACGCGTTGACAGTCCCTCTTG
AGGCGTTTGTTATAAC

VCR$_{147c}$ **Conj**=3.47×10^{-6} **Rep**=4.37×10^{-5} **VTS size**=87b **Pfold**=0.0000105

GTTATAACGCCCGCCTAAGGGGCTGACAATTTACTACCACTAAACTCAAACACAAGCGTTAAACTA
CCGAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCCACGCGTTGACAGTCCCTCTTG
AGGCGTTTGTTATAAC

VCR$_{147d}$ **Conj**=1.08×10^{-3} **Rep**=5.63×10^{-3} **VTS size**=78b **Pfold**=0.000331

GTTATAACGCCCGCCTAAGGGGCTGACAACGCGACTACCACTAAACTCAAACACAACGTTAAACTA
CCGACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCCACCGCGTTGACAGTCCCTCTTGA
GGCGTTTGTTATAAC

VCR$_{147e}$ **Conj**=3.11×10^{-4} **Rep**=2.23×10^{-4} **VTS size**=83b **Pfold**=0.0000319

GTTATAACGCCCGCCTAAGGGGCTGACAACGTACTACCACTAAACTCAAACACAAGCGTTAAACTA
CCGAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCCACGCGTTGACAGTCCCTCTTG
AGGCGTTTGTTATAAC

VCR$_{147f}$ **Conj**=1.11×10^{-4} **Rep**=6.20×10^{-5} **VTS size**=85b **Pfold**=0.0000109

GTTATAACGCCCGCCTAAGGGGCTGACAACTTACTACCACTAAACTCAAACACAAGCGTTAAACTA
CCGAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCCACGCGTTGACAGTCCCTCTTG
AGGCGTTTGTTATAAC

VCR$_{164}$ **Conj**=3.82×10^{-3} **Rep**=2.06×10^{-3} **VTS size**=64b **Pfold**=0.0156

GTTATAACGCCCGCCTAAGGGGCTGACAACGCGCTGCGCGCGCGCCGCACTACCACTAAACTCAAA
CACAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCCACGCGGCGTCGCGCGCAGCGC
GTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC

VCR$_{180}$ **Conj**=1.51×10^{-3} **Rep**=1.72×10^{-2} **VTS size**=64b **Pfold**=0.0156

GTTATAACGCCCGCCTAAGGGGCTGACAACGCGCTGCGCGCGGCCGCCGCACCGCACTACCACTAA
ACTCAAACACAACAACAGCAACCACCGCGGCTCAATGGGACTGGAAACGCCACGCGGTGTCGGCG
GCGTCGCGCGCAGCGCGTTGACAGTCCCTCTTGAGGCGTTTGTTATAAC

Figure S1: Proposed secondary structures of the used *attC* sites

Secondary structures were determined using the UNAFOLD online interface of the Institut Pasteur. G:C and A:T base pairs are marked by red and blue dashes respectively. The 5'and 3'ends are indicated and the bases (b) are numerated. The *attC* sites are classified according to their size (smallest to largest). The natural *attC* sites (WT: Wild Type) are indicated. The modifications made from the wild type VCR site are described for all the VCR derivatives.
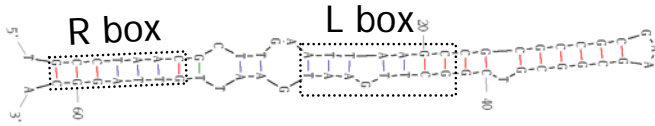
$attC_{aadA7}$bs (WT)
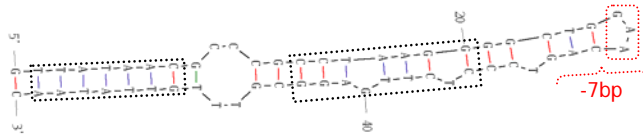
VCR_{56}bs (substitution of the VTS by 3b) (7bp deletion in the stem)
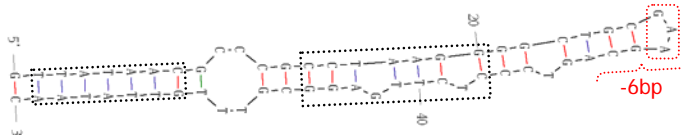
VCR_{58}bs (substitution of the VTS by 3b) (6bp deletion in the stem)
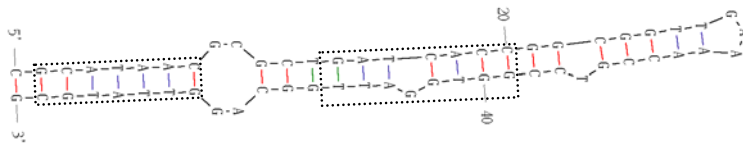
$attC_{ere2}$bs (WT)
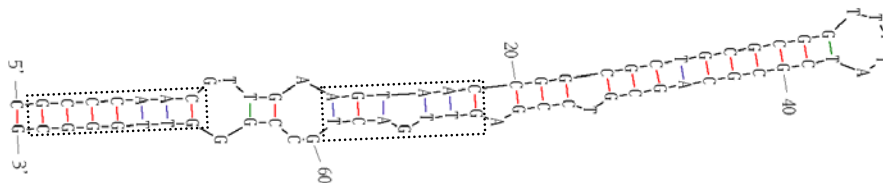
$attC_{oxa2}$bs (WT)
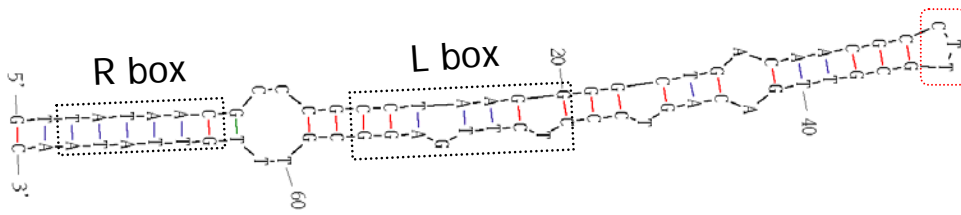
VCR$_{GAA}$bs
(substitution of the VTS by 3b)
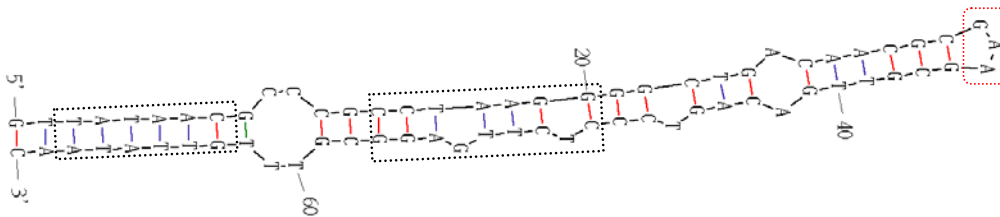
VCR$_{TTC}$bs
(substitution of the VTS by 3b)
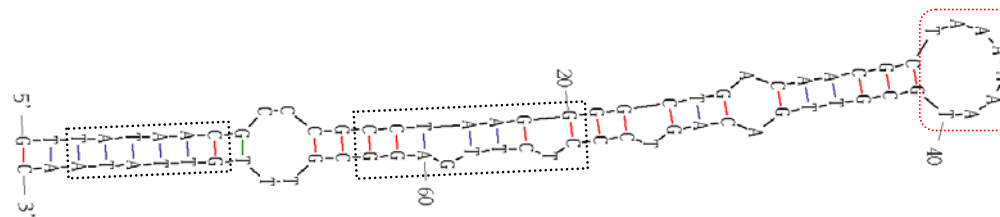
VCR$_{TA}$bs
(substitution of the VTS by a stretch of TA)
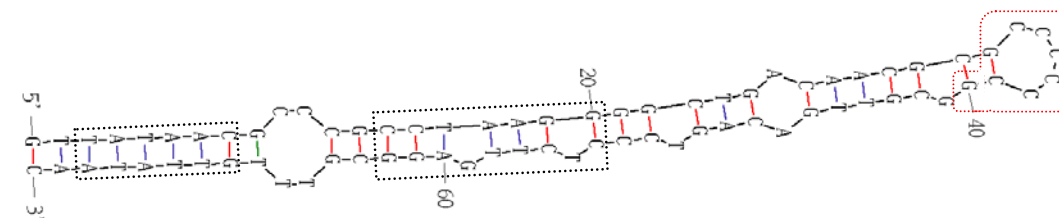
VCR$_{GC}$bs
(substitution of the VTS by a stretch of GC)
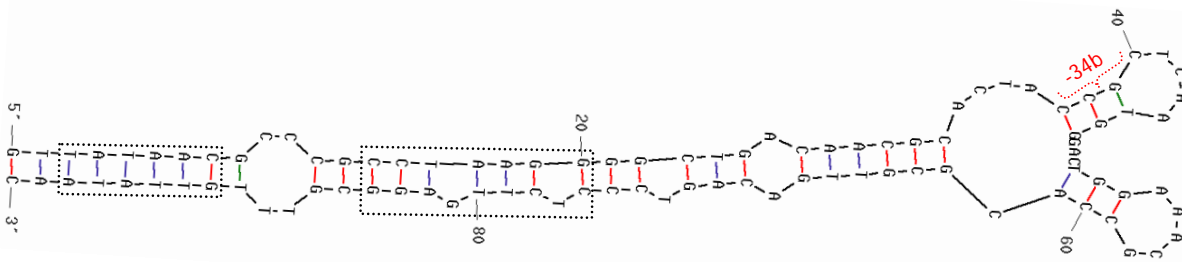
VCR$_{97a}$bs
(34b deletion in the VTS)
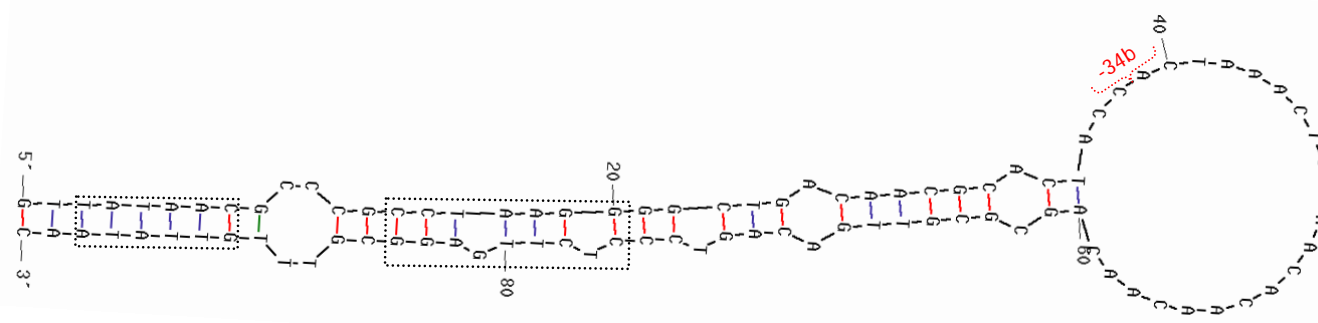
VCR$_{97b}$bs
(34b deletion in the VTS)
(modification of a part of the 30 remaining bases)
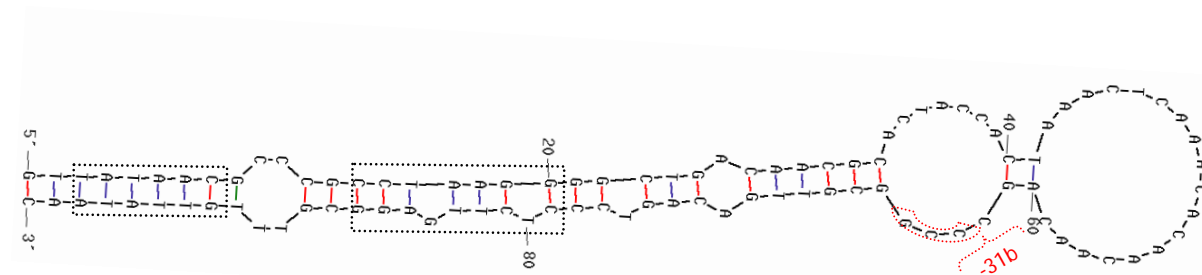
VCR$_{100}$bs
(31b deletion in the VTS)
(modification of 3b among the 33 remaining bases)

VCR$_{116a}$bs
(15b deletion in the VTS)

VCR$_{116b}$bs
(15b deletion in the VTS)

VCR$_{a}$bs (WT)

stem

VTS

VCR$_{b}$bs

(2b substitution in the VTS
Reducing it by 4b and
extending the stem by 2bp)
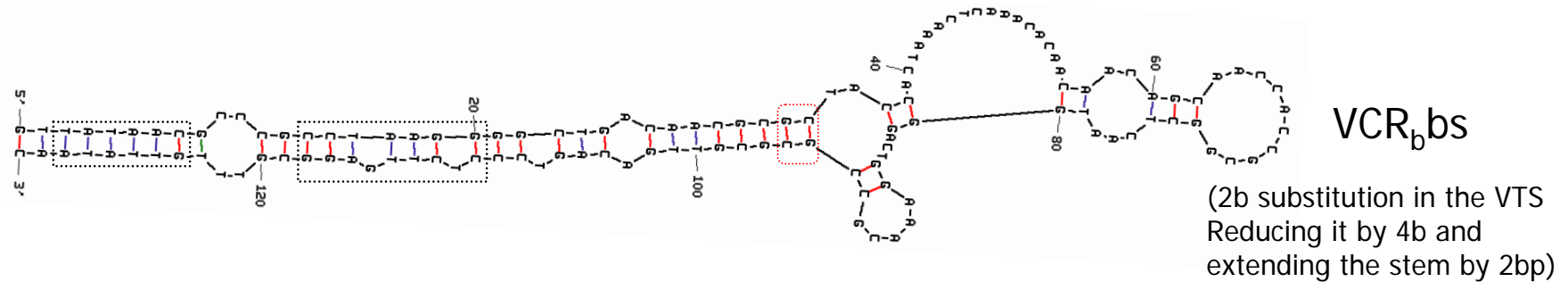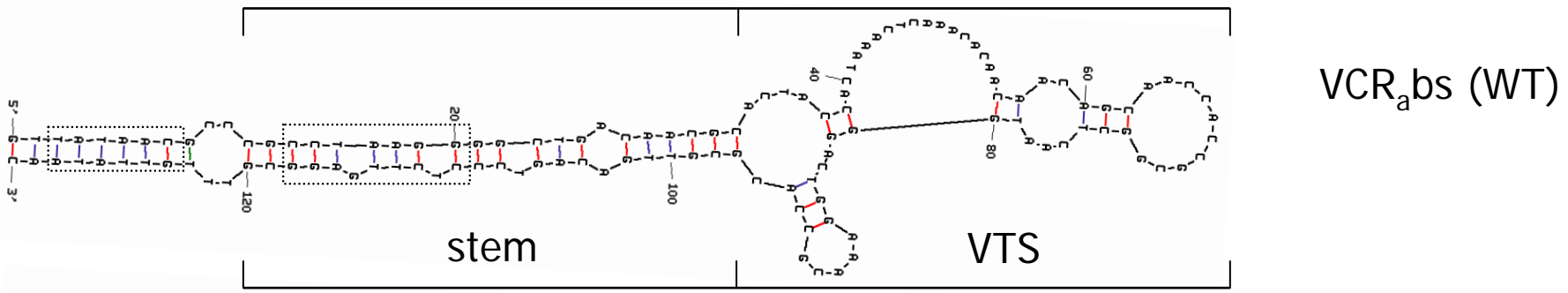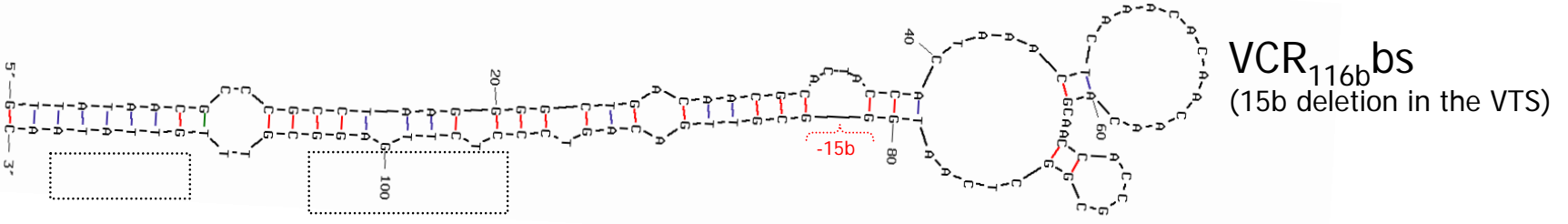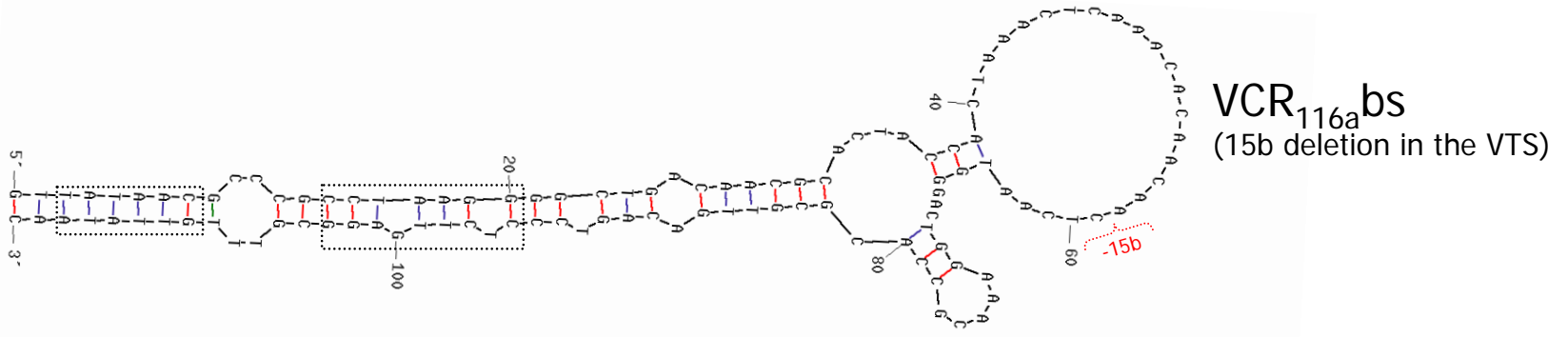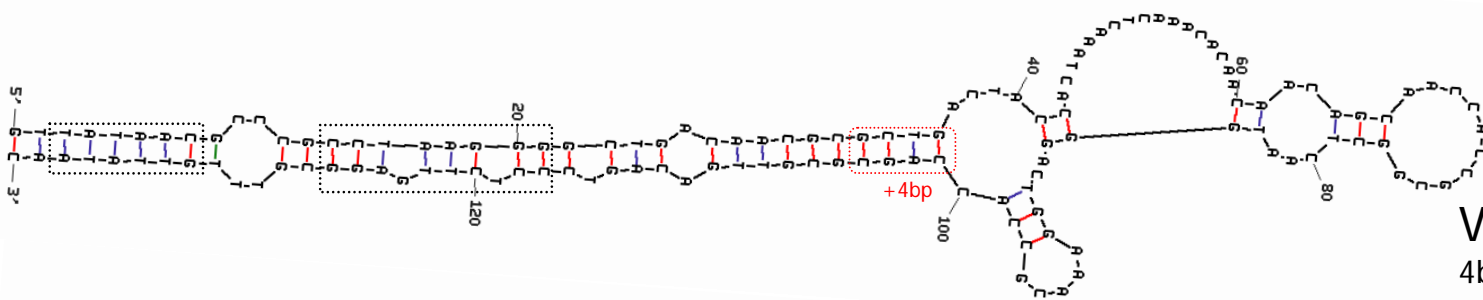
VCR$_{139}$bs
4bp addition in the stem
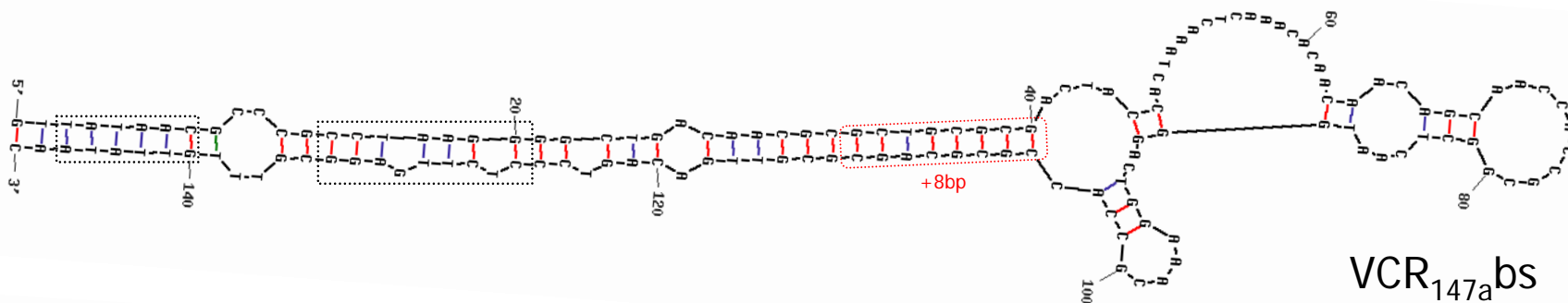
VCR$_{147a}$bs
8bp addition in the stem
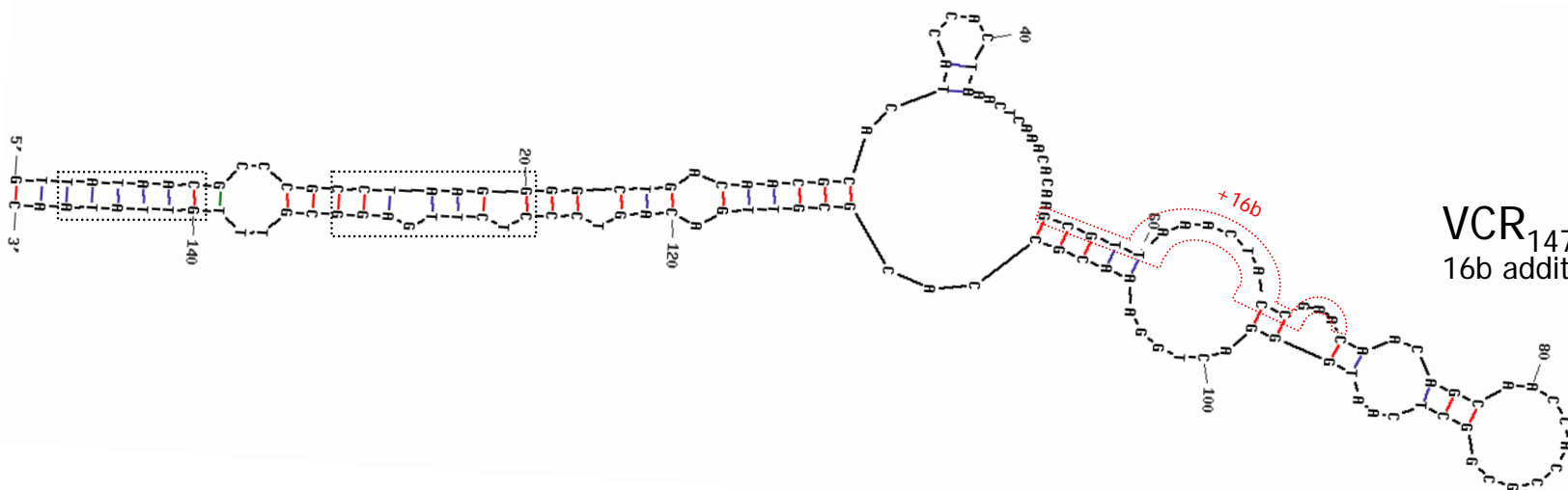
VCR$_{147b}$bs
16b addition in the VTS
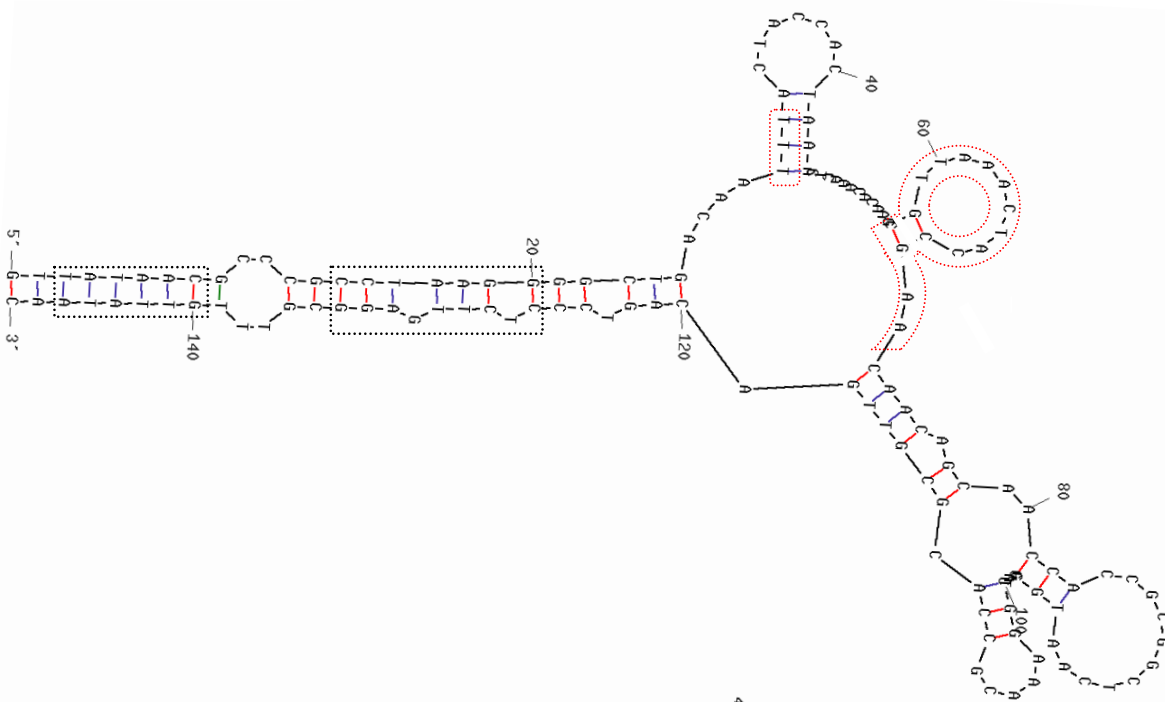
VCR$_{147c}$bs
3b substitution in the stem
reducing it by 14b
16b addition in the VTS
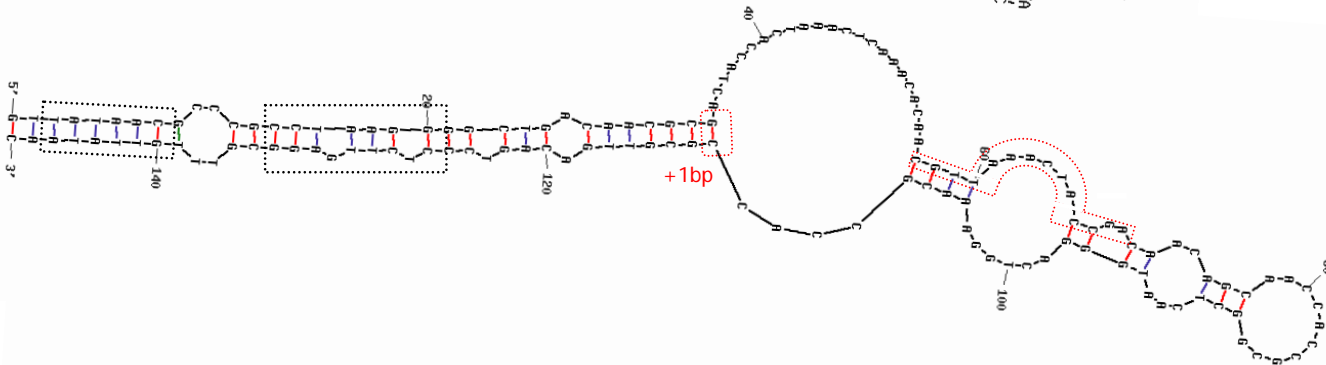
VCR$_{147d}$bs
1b addition in the stem
extending it by 1bp
14b addition in the VTS
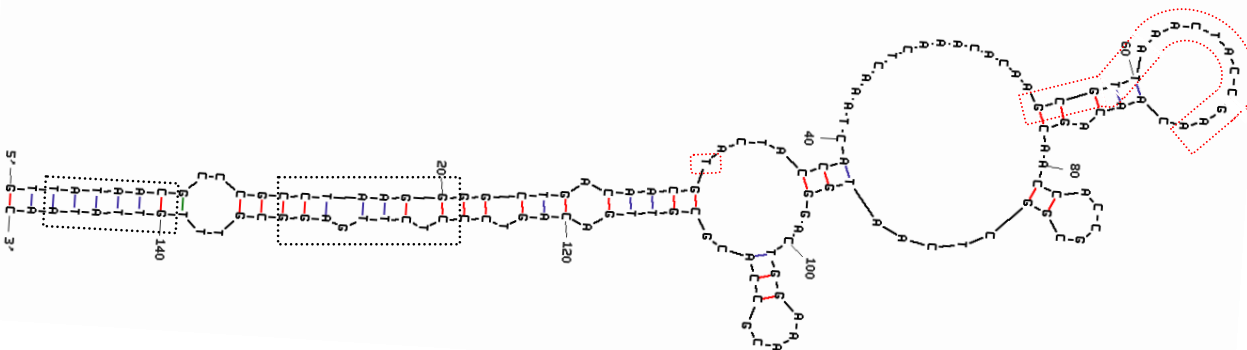
VCR$_{147e}$bs
1b substitution in the stem
reducing it by 2pb
16b addition in the VTS
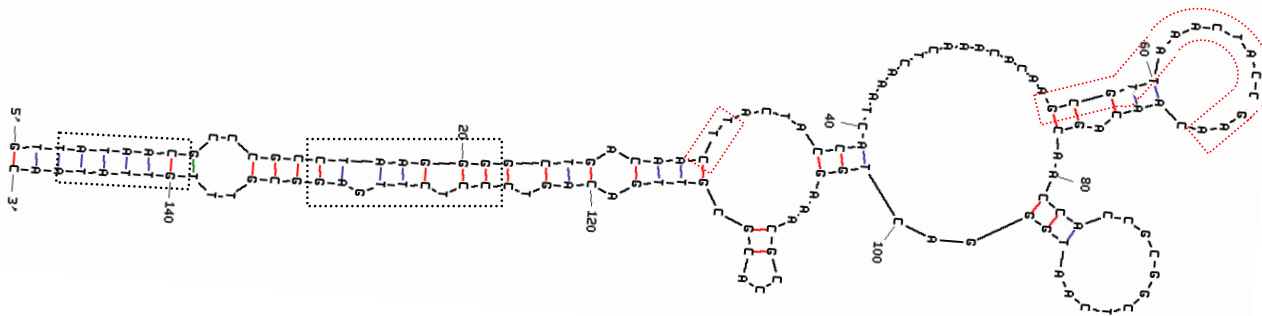
VCR$_{147f}$bs
3b substitution in the stem
reducing it by 4 bp
16b addition in the VTS
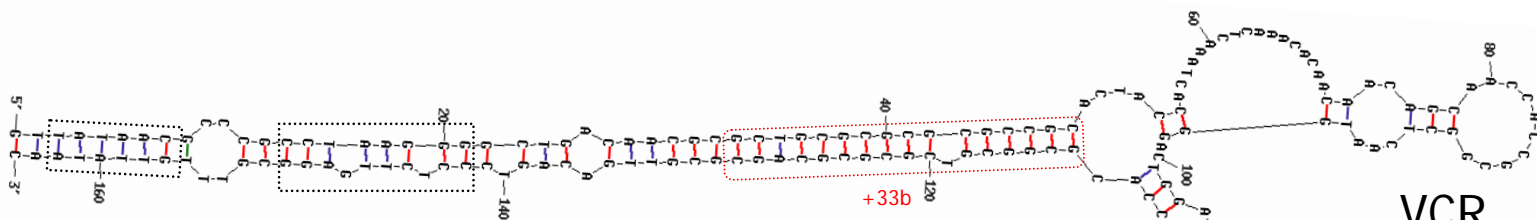
VCR$_{164}$bs
33b addition in the stem

VCR$_{180}$bs
49b addition in the stem
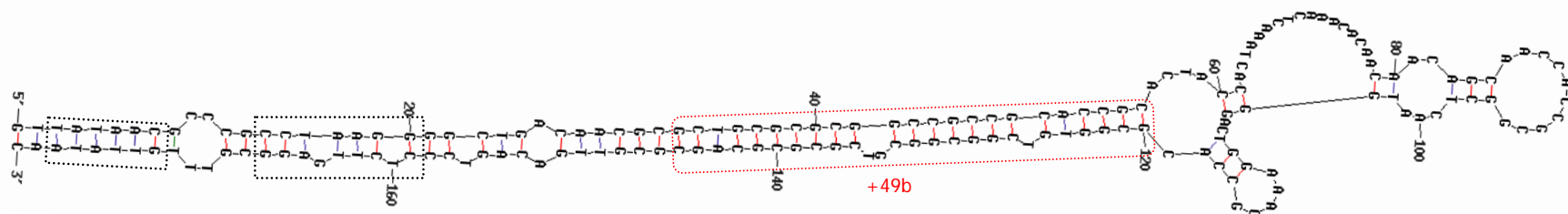
Figure S2
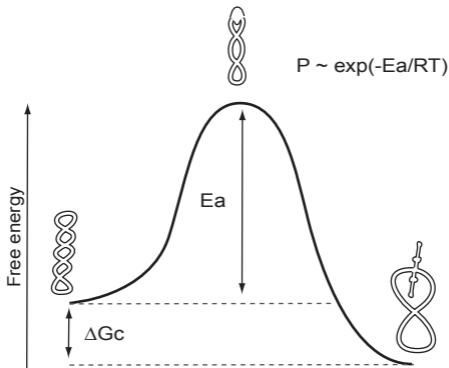
Energy Path of cruciform formation in a supercoiled DNA molecule.
In supercoiled DNA, melting at the dyad of a palindrome sequence
can lead to intrastrand base-pairing and initiates cruciform formation.
Branch migration elongating the cruciform is then energy driven.
The energy required to realize the dyad melting can be viewed as the
activation energy of cruciform formation and is directly linked to the
probability of cruciform formation following the Arrhenius equation.

Figure S3

Plasmid topology	ccc	oc	ccc	oc	ccc	oc	ccc	oc
Cleavage position (from the *attC* site)	338 nt		348 nt		43 nt		53 nt	
Cleaved strand	BOT		TOP		BOT		TOP	

Bar values: 3.92×10^{-3}, 1.49×10^{-3}, 1.04×10^{-3}, 3.58×10^{-3}, 1.85×10^{-3}, 1.00×10^{-3}, 8.74×10^{-4}, 1.52×10^{-3}

Y-axis: Recombination frequencies
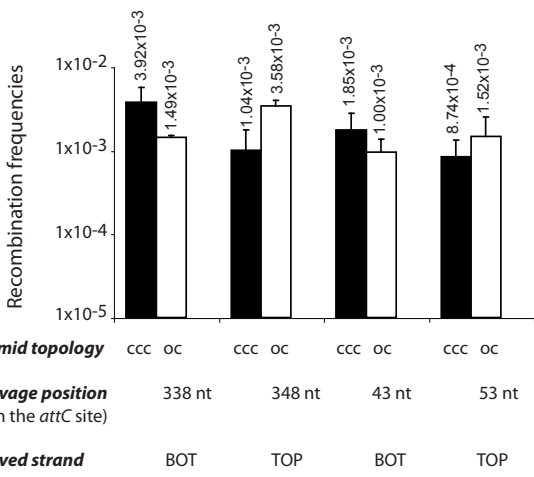
Effect of strand cleavage on recombination frequencies

The figure shows the recombination frequencies of different *attC*-containing plasmids (supercoiled or nicked) after transformation in non permissive recipient cell (see supplementary information). Error bars show standard deviations.

ccc, supercoiled plasmid; **oc,** open circular (nicked) plasmid;
BOT, bottom strand; **TOP,** top strand; **nt,** nucleotides.

Figure S4
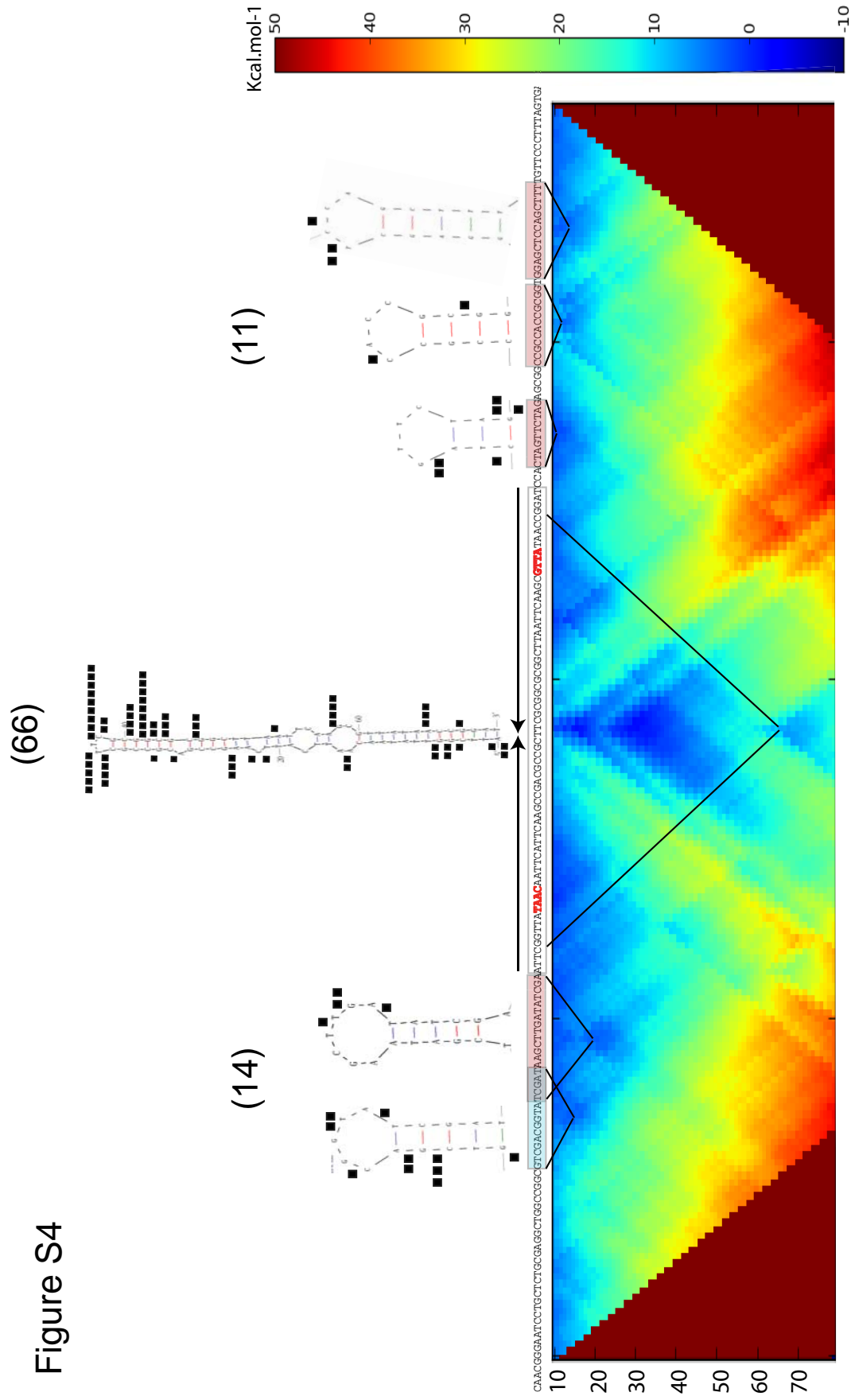
The color graph represent the energy of cruciform formation along the sequence. On the y-axis is the size of the sequence window considered, and on the x-axis its position. Hairpins where S1 cleavage sites were found are represented above the graph and their position on the graph shown with black lines. Black squares by the hairpins represent the exact cleavage positions. Note that the large majority of S1 cleavage sites (66/91) are specific to the *attC* site (the central hairpin).

Figure S5
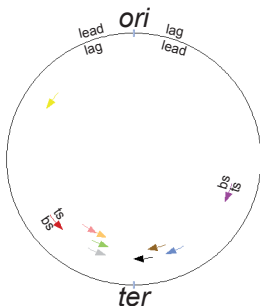
→	*Pseudoalteromonas haloplanktis*	→	*Photobacterium profundum*
→	*Vibrio cholerae*	→	*Vibrio vulnificus CMCP6*
→	*Vibrio fisheri*	→	*Vibrio harveyi*
→	*Vibrio vulnificus YJ016*	→	*Vibrio parahaemolyticus*
→	*Vibrio splendidus*	→	*Treponema denticola*

Location of the chromosomal integrons in sequenced bacterial genomes and relative orientation of the *attC* sites.

The position of the arrows along the circle represents the relative location of the different chromosomal integrons present in bacterial genomes. Each bacterial strain is identified by a specific color. Genomes analysis were made using the MAGE (MAgnifying GEnome) platform developped at the Genoscope (Evry, France, http://www.genoscope.cns.fr/agc/mage).

The arrows orientation indicates the cassettes orientation.

The origin *(ori)* and terminus *(ter)* regions of replication, and the respective leading (lead) and lagging (lag) strands of the two replichores are indicated.

ts, top strand; bs, bottom strand.

The Synthetic Integron: an in vivo genetic shuffling device

Being able to construct complex genetic circuits that work at first trial is the ultimate goal of synthetic biologist. However we have seen how intricate and hard this task can be. Genetic tools that generate large numbers of designs are thus extremely useful. Integrons are recombination systems naturally efficient at combining genetic elements to find adaptive solutions. They thus seemed a tool of choice to generate genetic combination for engineering purpose as well.

When this project was initiated, little was known about the dynamics of recombination in an integron cassette array. Would a synthetic integron be able to shuffle gene cassettes, or would the cassette be lost? Would the frequency of recombination be high enough to generate large number of arrangements in a cell culture?

In order to answer these questions, we needed a simple system with a phenotype easy to select. The system also needed to be composed of enough genes to demonstrate the combinatory power of integrons in rearranging them. The tryptophan biosynthesis operon came as a reasonable choice. It consists in 5 genes whose proteins catalyze 7 reaction steps from chorismate to tryptophan. All the arrangements of 5 genes represent 120 combinations, and it is easy to select for tryptophan production simply by growing cells on minimum media. However, the first trials were a failure. It took several months to delete the natural tryptophan operon, to associate the 5 *trp* genes and a reporter *lacZ* with *attC* sites, and to build several arrays with the 6 cassettes in different orders. But once those arrays were constructed, it appeared that they all conferred prototrophy to a *trp-* strain without the need of any rearrangements. There was thus no way to select for recombination events, and their frequency was too low to be detected by PCR screens.

Regulatory cassettes were thus included in the integrons: two transcriptional terminators were added to the array, one at the beginning and one in the middle, with the purpose of stopping the leaking gene expression. A promoter cassette was also added at the end of the array. However, to our surprise, the plasmid on which this array was constructed still conferred prototrophy to a *trp-* strain Suspecting this to be

due to the number of copies of the *trp* genes when carried on the plasmid, we integrated the array in the chromosome to reduce it to the chromosomal level; expecting that the strain would remain auxotroph for tryptophan.. This worked, and we finally had a way to select for recombination events. However this was not the end of our struggle as the first recombination experiments only yielded a single solution: the deletion of the transcriptional terminators. I had not spent so much time on these constructions just to show that integrons could delete cassettes in an array! Fortunately, when we attempted to screen more of the recovered prototrophs, we were able to detect more extensive rearrangements. The rest of the work was then mainly to realize lots of PCR screens in order to find enough combinations to be able to tell something meaningful about the recombination dynamics in the array, and the number of arrangements generated.

We also developed other methods to generate combinations with the integron machinery. In particular, we could show that it was possible to deliver cassettes through conjugation and to integrate them at an *attI* site placed in the chromosome. This method provides an easy way to deliver cassettes to an integron platform. Furthermore, it offers the possibility to increase the combinatorial power of the assay through the delivery of several cassette arrays from different donor strains simultaneously.

The synthetic integron: an *in vivo* genetic shuffling device

David Bikard[1,2], Stéphane Julié-Galau[1,2], Guillaume Cambray[1,2] and Didier Mazel[1,2],*

[1]Institut Pasteur, Unité Plasticité du Génome Bactérien, Département Génomes et Génétique and
[2]CNRS, URA2171, F-75015 Paris, France

ABSTRACT

As the field of synthetic biology expands, strategies and tools for the rapid construction of new biochemical pathways will become increasingly valuable. Purely rational design of complex biological pathways is inherently limited by the current state of our knowledge. Selection of optimal arrangements of genetic elements from randomized libraries may well be a useful approach for successful engineering. Here, we propose the construction and optimization of metabolic pathways using the inherent gene shuffling activity of a natural bacterial site-specific recombination system, the integron. As a proof of principle, we constructed and optimized a functional tryptophan biosynthetic operon in *Escherichia coli*. The *trpA-E* genes along with 'regulatory' elements were delivered as individual recombination cassettes in a synthetic integron platform. Integrase-mediated recombination generated thousands of genetic combinations overnight. We were able to isolate a large number of arrangements displaying varying fitness and tryptophan production capacities. Several assemblages required as many as six recombination events and produced as much as 11-fold more tryptophan than the natural gene order in the same context.

INTRODUCTION

Synthetic biology aims to engineer useful novel functions in existing organisms (1–3). Recent efforts have shown how one can couple mathematical modelling with the precise characterization of libraries of genetic elements to rapidly construct synthetic networks with predictable functions (4). However, unpredictable interactions will probably remain a significant challenge to genetic engineering for some time to come. One way to circumvent this problem might be to admit the limits of our predictive abilities and to test large numbers of random designs. This approach has already proved successful for the directed evolution of proteins. Indeed, accurate prediction of protein folding and function from primary sequence remains a challenge (5). Nevertheless, methods such as gene shuffling have enabled the improvement of protein stability and performance as well as changes in their reaction and substrate specificity (6). These methods have mainly focused on mimicking and enhancing natural recombination to promote diversity. In contrast to rational design, directed evolution does not require a comprehensive knowledge of the system being implemented (7). We believe the same principles can be applied to the engineering of larger genetic systems and metabolic pathways. Combinatorial approaches have recently been used for promoter engineering (8), or for the modification of intergenic regions in synthetic metabolic pathways with the purpose of improving the balance of gene expression (9). Along these lines, a new method called MAGE allows the combinatorial mutagenesis of a few base pairs at multiple genomic loci to improve endogenous metabolic pathways (10). In this work, we used the unique recombination properties of the integron machinery to enable the rapid and large-scale generation of combinations *in vivo*.

Integrons were first discovered owing to their involvement in multiple antibiotics resistance phenotypes. They were later identified in the genomes of ~10% of sequenced bacteria (11) and can represent a significant proportion of these genomes (e.g. 3% in *Vibrio cholera*; 12). Integrons are composed of a tyrosine recombinase (the integrase, IntI), a primary recombination site (*attI*), and an array of gene cassettes (13). Cassettes generally consist of promoterless ORFs flanked by *attC* recombination sites. In some *vibrionaceae*, there can be more than 200 such cassettes. Upon stress (14), the integrase is expressed and can recombine *attC* sites, leading to the excision of circular cassettes that can be further integrated at the *attI* site. These recombination events lead to deletions and rearrangements in the cassette array as well as the capture of new cassettes by lateral gene transfer (15).

They represent extremely powerful evolutionary devices, whose success is exemplified by their ubiquitous spread associated with multiple antibiotics resistances. Here, we designed three different setups based on the integron recombination parts to test the recombining power of this genetic system. We chose to assay the assembly of a known anabolic pathway, tryptophan biosynthesis, involving the consecutive action of multiple enzymes, and selected clones with increased production of the final metabolite.

MATERIALS AND METHODS

Culture conditions

Escherichia coli strains were grown in Luria Bertani broth (LB) or 63B1 minimal medium with glucose 0.4% (MM63B1) at 37°C. Antibiotics were used at the following concentrations: ampicillin (Ap), 100 µg/ml, chloramphenicol (Cm), 25 µg/ml. Diaminopimelic acid (DAP) was supplemented when necessary to a final concentration of 0.3 mM and tryptophan to a final concentration of 50 µg/ml. Chemicals were obtained from Sigma-Aldrich (France).

Plasmids

pSW plasmids were constructed as follow. First, the *attP* recombination site of the lambda phage was PCR amplified and cloned at the SacI site of the previously published pSW23T (16). The multiple cloning site was then replaced with the BioBrick strandard restriction sites through a PCR with oligos o936 and o937 (see Supplementary Table S4 for the list of oligos). The tryptophan operon genes were PCR amplified from the genome of *E. coli* MG1655 with a first set of primers (trp(A-E)-F and trp(A-E)-R) framing them with common 3′- and 5′-ends. A second set of primers (pos(1-6)-F and pos(1-6)-R) was then used to add the *attC*$_{aadA7}$ site on the 3′-end and either BioBrick standard restriction sites or BglI restriction sites allowing the directional assembly of the cassette array in successive steps leading to pSWlib, pSW-BA and pSW-CED (see Supplementary Table S5 for the description of plasmids). The BglI and BioBrick restriction sites present in the *trp* genes were deleted trough site-directed mutagenesis with primers o1002–o1007. The *attI1* site was entered to the registry of standard biological parts as BBa_J99002 and *attC*$_{aadA7}$ as BBa_J99001. BBa_J23100 is a constitutive promoter and BBa_B0015 a transcriptional terminator. Biobricks details can be found at partsregistry.org). The *cat* [CmR] gene of pSWlib was replaced by the *aadA7* [SpecR] gene to give pSWKspec as described by Demarre *et al.* (16). J23100 and *attI1* were cloned in pSWKspec through BioBrick standard assembly, giving p7421.

Strains

The tryptophan operon was deleted in the TG1 strain, with the method described by Chaveroche *et al.* (17) giving the strain TG1Δ*trp*::km. The deletion of the tryptophan operon was then P1 transduced into MG1655recA::Tn10 [RecA was supplied from the plasmid pCY579 (18)] and the kanamycin resistance was subsequently excised trough FRT recombination mediated by the pCP20 plasmid, resulting in strain ω7814 (see Supplementary Table S6 for the description of strains).

pSWlib and p7421 were integrated into the *attB* site of the ω7814 chromosome through lambda recombination mediated by plasmid pTSA29-CXI (19) to give ω7830 and ω7902, respectively.

Chromosomal recombination assay

Overnight cultures of ω7842 were diluted to the 1/100 in LB medium with arabinose 0.2% and grown overnight. These were then plated on MM63B1 and LB agar. Prototroph frequencies were established as the ratio of the number of clones on MM63B1 over LB. The 788 clones were screened for cassette integration at the *attI* site. Twenty-nine positive clones and 30 negative clones were further analysed to determine their precise gene order.

Plasmidic recombination assay

Overnight cultures of ω7661-int were diluted to the 1/100 in LB medium with arabinose 0.2% and grown overnight. Plasmid extractions were realized with Macherey–Nagel NucleoPlasmid kits and transformed into electrocompetent ω7814 strains. Transformants were plated on MM63B1 and LB + Cm. Prototrophs were analysed through series of PCRs.

Conjugation assays

Overnight cultures of the donor(s) strain(s) (ω7893 or ω8066 + ω8067) and recipient strain (ω7902-int) were diluted to the 1/100 and grown in LB + DAP (300 µg/ml) and LB + arabinose 0.2%, respectively to OD$_{600}$ = 0.6. Two millilitre of the donor cells were mixed with 4 ml of the recipient cells and filtered onto 0.22 µm filters which were incubated overnight on LB + DAP Petri dishes. The bacteria on the filter were then resuspended in LB and plated on MM63B1 or LB. Prototroph frequencies were established as the ratio of the number of clones on MM63B1 over LB. The cassettes orders in the recovered prototrophs were established through series of PCRs.

Tryptophan production and growth rate

We performed here biological measures of tryptophan production relying on a co-culture between a tryptophan producing strain and an auxotroph reporter strain. The growth of the reporter strain is indicative of the presence of tryptophan in the medium and thus of the tryptophan production by the other strain. Overnight cultures of the prototroph strains obtained from the recombination assays were diluted to the 1/100 in MM63B1 together with ω8072 in 96 wells Corning culture plates (1/1 ratio). Fluorescence (em:560 nm, ex:600 nm) and OD600 were measured in a Tecan infinite200 plate reader. Tryptophan production was

assessed as the difference in fluorescence between the sample and the control well (DB8072 only) divided by the sample OD_{600}. Growth rates were measured in MM63B1 and MM63B1 + tryptophan in a Tecan infinite200 plate reader.

RESULTS

Tryptophan biosynthesis involves seven reactions from chorismate as a starting point (20). In *E. coli*, five proteins are involved in this pathway (TrpA-E), with TrpE and TrpC each carrying two catalytic domains. In a first assay, the genes for each of these proteins were 'packaged' into identically designed cassettes in which the gene of interest was preceded by the same ribosome binding site (BBa_B0030) and followed by a well described 64 bp *attC* recombination site [$attC_{aadA7}$ (21)]. These artificial cassettes were assembled downstream of an *attI* primary recombination site in an arbitrary order along with four 'regulatory' cassettes to form the starting library plasmid pSWlib. Two of the regulatory cassettes carried strong transcriptional terminators and were primarily intended to prevent the initial construct from producing tryptophan. One cassette carried the reporter gene *lacZα* and the last one a constitutive promoter (see Figure 1 for the detailed arrangement). The entire pSWlib plasmid was integrated into the chromosome of ω7814, an *E. coli* strain that is deleted for the whole-tryptophan

operon (see 'Materials and Methods' section). As expected, the resulting strain remained auxotrophic for tryptophan. To induce rearrangements that might lead to expression of the tryptophan biosynthetic pathway, the integrase IntI1 gene was expressed from an arabinose-inducible P_{BAD} promoter located on a plasmid (pBAD-IntI1). After an overnight culture with arabinose, tryptophan prototrophs were recovered at a frequency of 3.4×10^{-3}.

Rearrangements were then analysed through series of PCR reactions giving the exact integron gene order in the prototrophic strains. It appeared from these data that most (28/30) rearrangements only involved deletion events removing both transcriptional terminators. We thus screened a larger number of clones for cassette integration at the *attI* site and found that 3.7% (29/788) showed more extensive rearrangements. Their precise gene order was determined, and can all be explained by one or several *attC* × *attC* excisions followed by *attC* × *attI* integrations, as well as cassettes duplications in some cases (Table 1). The frequency of such events—hereafter referred to as reordering events—in the unselected population is thus on the order of 10^{-4}. Since a single reordering event only allows a small number of different gene arrangements to be attained, we also attempted to assess the frequency of multiple reordering and duplication events. These numbers are hard to determine precisely since they would require a very large number of PCR screens. Among the recovered
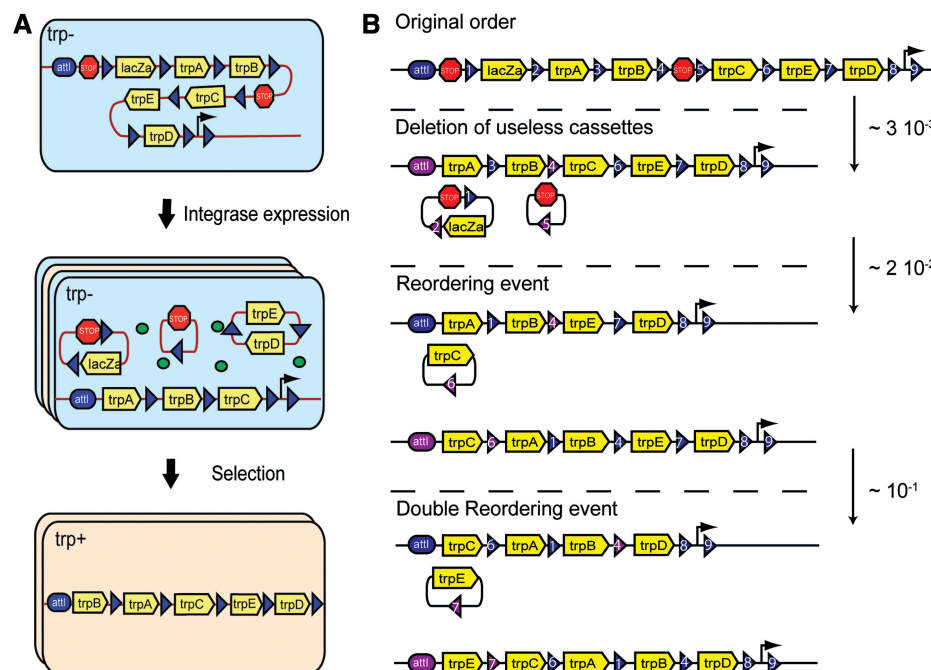

Figure 1. Original cassette arrangement and recombination events leading to tryptophan production. The *trp* cassettes were assembled in an arbitrary order together with two strong transcriptional terminators preventing the initial construct from producing tryptophan. Another cassette carries the gene for the LacZα peptide to act as a reporter that could be assayed by colony color, and another cassette carries a constitutive promoter that should give increased expression of any downstream genes in the array. No promoter was placed upstream of the *attI*. Expression of the cassettes is likely driven by leaking expression from the upstream pSWlib RP4 region. $attC_{aadA7}$ recombination sites are represented by blue triangles (pink after they underwent recombination). (**A**) Integron cassettes on the chromosome are shuffled through deletion and integration mediated by the integrase (green circles) in a *trp*- strain which is then selected for prototrophy. Cells with blue or orange background colors are tryptophan auxotrophs and prototrophs, respectively. (**B**) Frequencies of recombination. $attC_{aadA7}$ are numbered according to their initial order.

Table 1. Possible recombination history for some of the recovered prototrophs

Cassettes order	Possible recombination history
t | Z | A | B | t | C | E | D | p |	Original order
A | B | C | E | D | p |	attI × attC2 / attC4 × attC5
B | C | Z | A | E | D | p |	attI × attC1 / attC4 × attC5 / attC2 × attC6 -> attI
E | Z | A | B | C | D | p |	attI × attC1 / attC4 × attC5 / attC6 × attC7 -> attI
B | C | A | B | C | E | D |	attI × attC2 / attC4 × attC5 / attC8 × attC9 / attC3 × attC6 dup-> attI
C | E | D | Z | A | B | t | p |	attI × attC1 / attC5 × attC8 -> attI
E | A | B | C | E | D | p |	attI × attC2 / attC4 × attC5 / attC6 × attC7 dup-> attI
D | A | B | C | E | t | Z |	attC4 × attC5 / attC8 × attC9 / attC2 × attC7 -> attI / attC2 × attC8 -> attI
B | A | C | E | D |	attI × attC2 / attC4 × attC5 / attC8 × attC9 / attC7 × attC8 -> attI
E | C | A | B | D | p |	attI × attC2 / attC4 × attC5 / attC5 × attC6 -> attI / attC5 × attC7 -> attI

Letters represent integron cassettes and are abbreviations as follow. t, BioBrick terminator BBa_B0015; p, BioBrick promoter BBa_J23100 ('Materials and Methods' section); A, *trpA*; B, *trpB*; C, *trpC*; D, *trpD*; E, *trpE*; Z, *lacZα*. The arrow means that the excised cassette was subsequently integrated at the attI site. The *attC* sites are numbered following their order in the original cassettes array. Note that the last line of the table represents the recombination history depicted in Figure 1B. We consider here that integration events occur preferentially at the attI site as previously described (32).

prototrophs, 2.4% (19/788) displayed single reordering events, 0.5% (4/788) double reordering events and 0.8% (6/788) events involving the duplication of at least one cassette. We can thus estimate the order of magnitude for the frequency of double reordering events in the unselected population to be 10^{-5}. It is noteworthy that this frequency is higher than the product of the frequencies of two single reordering events, suggesting that once a recombination event has occurred, a second one is more likely to happen. We can also note that none of the genotyped arrangements placed the promoter cassette upstream of the *trp* genes. Their expression is likely driven by the pL1 promoter of the oriRP4 region (22) upstream of the *attI*.

Since single reordering events are more frequent than multiple ones, the types and number of gene combinations actually obtained in a culture are highly dependent on the original gene order. In order to determine the number of unique combinations that retain the five *trp* genes and that can be obtained by a defined number of reordering and deletion events, we numerically generated all combinations and counted the unique ones. Single reordering events of the nine initial cassettes can yield 36 unique combinations and 342 if we include the deletion of non-essential cassettes. Double reordering events can yield 904 unique combinations and 5022 if we include deletions. We want to assess the expectation of the number of unique combinations that we can obtain in a culture were C cells realize randomly one of the N possible combinations. If we make the approximation that all combinations have the same probability, this problem is known as the 'coupon collector's problem' which asks the question of the number of sample trials (C), with replacement, required to collect a complete set of N coupons. We numerically simulated this experiment 10 000 times and computed the estimators of the expectation and of the standard deviation. In a 1 ml overnight culture of 10^9 cells, we should have an average of $C = 10^4$ cells realising one of the $N = 5022$ possible double reordering events. The expectation is to have 4337 (± 20) unique combinations in 1 ml of overnight culture. Nevertheless, it seems that some rearrangements are more probable than others.

For instance, the excision of the *trpB* cassette and its integration at the *attI* recombination site occurred in 6 of the 19 single reordering events isolated experimentally.

In order to assess the functional significance of the selected genotypes, we then measured the tryptophan production and generation time for 10 arrangements. The sequence of the *trpA-E* genes in the parental strain was verified. Since the frequency at which we recovered prototroph strains is much higher than the point mutation frequency of *E. coli*, we can assume that the phenotype of the different arrangements is explained by cassette order only. We also constructed and tested the combination mimicking the natural gene order. The measured tryptophan productions ranged from 4-fold less to a 2.8-fold more than the original strain carrying the natural *trp* operon (MG1655recA; Figure 2). Hence, both gene order and gene copy number in the operon have a drastic effect on tryptophan production. Interestingly, the wild-type order (EDCBA) has one of the lowest production levels. We also observed that growth rate could be affected by the cassettes arrangement, with an up to 40% decrease of growth rate compared to the parental MG1655recA strain, without correlation to the trp production level (Figure 2).

Overall, these results demonstrate that a synthetic integron can efficiently be used to generate functional gene combinations from a library of independent candidate gene cassettes. In order to improve the flexibility of the system presented above, we devised and tested two alternative methods that allow one to carry out the rearrangement and selection steps in different genetic backgrounds and enhance the delivery of cassettes. Having the synthetic integron on the chromosome presents some disadvantages in those cases one wants to generate combinations in one strain and select for good solutions in another genetic background. To address this concern, we assessed the potential of a synthetic integron carried by a plasmid (Figure 3B). The original cassette array was cloned into a low copy number plasmid [BBa_pSB4C5 (23)] and recombination was induced during an overnight culture. Plasmid DNA was recovered and transformed into a tryptophan auxotroph, and transformants where
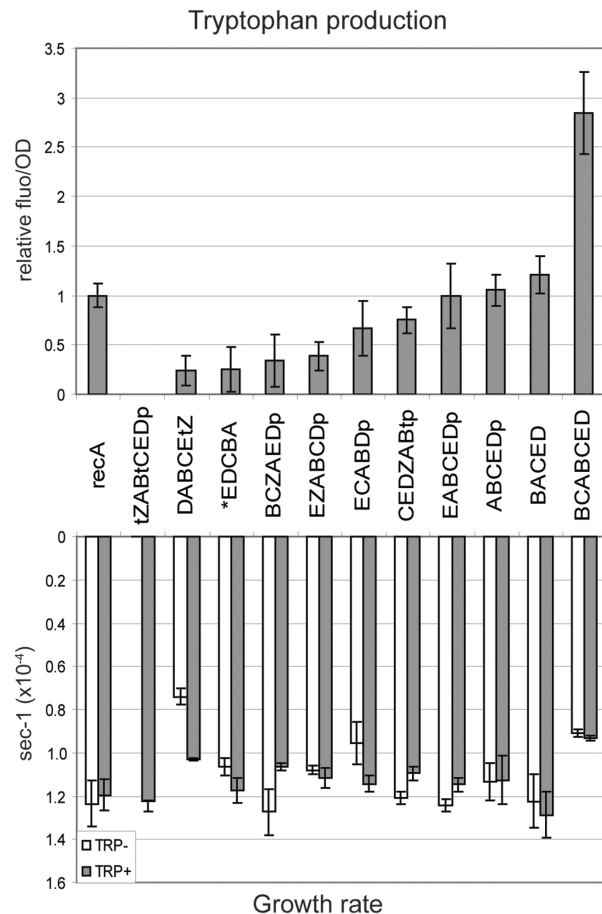
Figure 2. Tryptophan production and growth rates of some arrangements. Letters represent integron cassettes and are abbreviations as follow. t, BioBrick terminator BBa_B0015; A, *trpA*; B, *trpB*; C, *trpC*; D, *trpD*; E, *trpE*; Z, *lacZα*; p, BioBrick promoter BBa_J23100. Tryptophan production was measured in a biological assay involving a co-culture with an auxotroph reporter strain ('Material and Methods' section). The growth rates of the different arrangements were measured in minimum media supplemented or not with tryptophan. The WT gene order is highlighted with a star. There is no obvious relation between the cassettes order the level of tryptophan production and the growth rates measured. Data are means from three independent experiments; error bars show SD.

plated on minimal medium to select for plasmids carrying functional *trp* operons. The proportion of prototrophs among the transformants was 2.2×10^{-4}. Subsequent analysis revealed that most of them contained multiple plasmids each carrying different genes of the tryptophan pathway. Nevertheless, out of the 96 colonies screened, we were able to identify six clones carrying all the genes in a single plasmid. They were all in different combinations, and three of them carried duplications of one gene or more (see Supplementary Table S1).

Because the capacity to test a large number of candidate genes, and the ease of including new genes of interest within an existing scaffold, are of prime importance, we considered the possibility of delivering integron cassettes through conjugation. This procedure presents the possibility of building large arrays of cassettes in independent cloning strains and delivering them for chromosomal

integration and gene shuffling in a recipient bacteria. In a first assay, we used the pSWlib suicide plasmid. pSWlib was delivered through conjugation into a tryptophan auxotroph carrying an *attI* site on the chromosome and expressing the integrase from the pBAD-IntI1 plasmid. The transconjugants were selected for growth on minimal medium. Being unable to replicate in the recipient cell, the pSWlib can only be maintained if it integrates at the *attI* site. Alternatively, the plasmid can be lost by recombination of excised *trp* cassettes into the recipient chromosome. The frequency of recipient cells becoming prototrophs was 2.3×10^{-5} (see Supplementary Table S2 for details on the recovered arrangements). Delivering the cassettes through conjugation also offers the possibility of delivering different cassette arrays at the same time, increasing the combinatorial power of the assay. In a second assay, the five *trp* genes were thus split between two donor plasmids pSW-BA and pSW-CED. Two donor strains each carrying a different plasmid where used in a conjugation assay with the recipient cell described above (Figure 3C). Prototrophs were recovered at a frequency of 5×10^{-6}. We determined the gene combinations obtained in eight prototroph colonies through series of PCRs. All arrangements were different from each other. The pSW-BA and pSW-CED were integrated in the chromosome of the recipient strain either through $attI \times attI$ recombination or $attI \times attC$ recombination. Further recombination events or deletion of the suicide pSW vector were also identified (see Supplementary Table S3).

DISCUSSION

We demonstrated here for the first time, the ability to efficiently generate large number of genetic combinations and arrangements *in vivo* using site-specific recombination. The functional arrangements we isolated in the chromosomal assay required from two to six recombination events, leading to cassette loss, reordering events and duplications. The generated operons varied both in growth rate and tryptophan production capacities in an uncorrelated manner. The effect on the strain fitness is presumably due to the generation of a misbalance of the operon enzymes expression leading to toxic effects and non optimal allocation of resources. It is for instance known that indole (the product of TrpA) has oxidative toxic effects (24). Besides, an excessive strain on the chorismate pool could deplete the biosynthetic pathways of amino-acids and metabolites such as tyrosine, phenylalanine, ubiquinone and tetrahydrofolate. Similar effects are very likely to occur in any attempt to implement any synthetic metabolic pathways, but are mostly unpredictable; even for the present case where we used *E. coli* genes in *E. coli* and though the tryptophan operon is one of the best studied biosynthesis pathways. This new method should thus find applications in metabolic engineering where rational decisions about candidate genes in a pathway, gene order and gene regulation are hard to make.
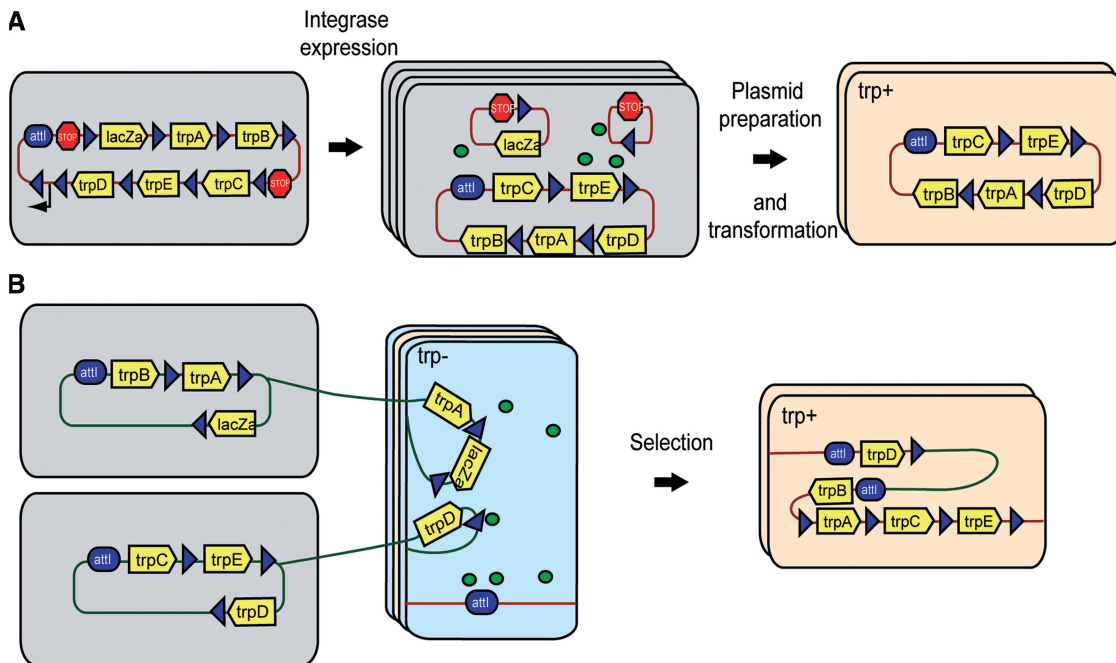
Figure 3. Alternative experimental setups. (**A**) Integron cassettes are shuffled directly on a plasmid in a cloning strain. They are then extracted and transformed in the *trp*⁻ selection strain. (**B**) Integron cassettes are delivered through conjugation into the *trp*⁻ strain where they are recombined on the chromosome. Color code is the same as in figure one, except for cell with a different genetic background, which are colored in grey.

Among the setups we tested, we show that conjugation is a powerful way to deliver cassettes in a chromosomal platform of a recipient strain. Simultaneous conjugation from several donor strains could be an extremely practical way to combine ready-made elements. One could consider the possibility of having plasmid libraries of various biobricks (regulatory elements, genes, etc.) that could be easily reused for the combinatorial synthesis of new systems and pathways. Other obvious areas of applications could include the shuffling and recombination of protein domains, the study of larger chromosomal re-arrangements and the random design of regulatory networks. One could indeed easily construct libraries of promoters and transcription factor cassettes that could be randomly combined using our method. A synthetic integron could also simply be used as a 'landing platform' for consecutive and targeted integrations of genetic elements in a host of interest using conjugation.

The question of the production of combinations in large numbers is central in both biotechnology and synthetic biology. The recently described MAGE method (10), illustrates how important is this challenge. Although both our synthetic integron and MAGE allow the generation of thousands of combinations, they have very different outputs. Whereas the synthetic integron can generate random arrangements of large exogenous genetic elements, the MAGE method permits to target mutations of a few base pairs at several genomic loci simultaneously. Both methods could advantageously be combined to manipulate at the same time genetic elements arrangement and their sequence. The only constraint of these approaches is the availability of a screen powerful enough to discriminate good solutions in a large population. This problem is being tackled by the recent developments of ultrahigh-throughput screening technologies, notably using microfluidics (25,26).

For some applications, one could also see the necessity of packaging genetic elements of interest in integron cassettes as a limitation. However, *attC* recombination sites are remarkably flexible in sequence. It has indeed been shown that they recombine as folded single stranded DNA and that recombination is mostly driven by structural features of the stem-loop and not by primary sequence (27–29). This opens the possibility of creating *attC* sites 'à la carte', and to use them as protein linkers, for instance.

For the moment, the most time consuming step in the utilization of this new approach probably is the assembly of large integron cassette arrays. Nevertheless, new methods in DNA assembly, such as SLIC (30) or DNA-assembler (31), promise to overcome this hurdle. Such developments, together with the exponential progress in DNA synthesis, are empowering synthetic biology in unprecedented ways, changing the speed at which we will increase our knowledge of living systems and our ability to manipulate them.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Endy,D. (2005) Foundations for engineering biology. *Nature*, **438**, 449–453.
2. Elowitz,M.B. and Leibler,S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**, 335–338.
3. Gardner,T.S., Cantor,C.R. and Collins,J.J. (2000) Construction of a genetic toggle switch in Escherichia coli. *Nature*, **403**, 339–342.
4. Ellis,T., Wang,X. and Collins,J.J. (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotech.*, **27**, 465–471.
5. Lee,D., Redfern,O. and Orengo,C. (2007) Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.*, **8**, 995–1005.
6. Kolkman,J.A. and Stemmer,W.P. (2001) Directed evolution of proteins by exon shuffling. *Nat. Biotechnol.*, **19**, 423–428.
7. Cambray,G. and Mazel,D. (2008) Synonymous genes explore different evolutionary landscapes. *PLoS Genet.*, **4**, e1000256.
8. Cox,R.S., Surette,M.G. and Elowitz,M.B. (2007) Programming gene expression with combinatorial promoters. *Mol. Syst. Biol.*, **3**, 145.
9. Pfleger,B.F., Pitera,D.J., Smolke,C.D. and Keasling,J.D. (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.*, **24**, 1027–1032.
10. Wang,H.H., Isaacs,F.J., Carr,P.A., Sun,Z.Z., Xu,G., Forest,C.R. and Church,G.M. (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, **460**, 894–898.
11. Boucher,Y., Koenig,J.E., Stokes,H.W. and Labbate,M. (2007) Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol.*, **15**, 301–309.
12. Mazel,D. (2006) Integrons: agents of bacterial evolution. *Nat. Rev. Microbiol.*, **4**, 608–620.
13. Recchia,G.D. and Hall,R.M. (1995) Gene cassettes: a new class of mobile element. *Microbiology*, **141**, 3015–3027.
14. Guerin,E., Barbé,J., Cambray,G., Ploy,M., Sanchez-Alberola,N., Campoy,S., Erill,I., Da Re,S., Gonzalez-Zorn,B. and Mazel,D. (2009) The SOS response controls integron recombination. *Science*, **324**, 1034.
15. Rowe-Magnus,D., Guerout,A. and Mazel,D. (2002) Bacterial resistance evolution by recruitment of super-integron gene cassettes. *Mol. Microbiol.*, **43**, 1657–1669.
16. Demarre,G., Guérout,A.M., Matsumoto-Mashimo,C., Rowe-Magnus,D.A., Marlière,P. and Mazel,D. (2005) A new family of mobilizable suicide plasmids based on broad host range R388 plasmid (IncW) and RP4 plasmid (IncPalpha) conjugative machineries and their cognate Escherichia coli host strains. *Res. Microbiol.*, **156**, 245–255.
17. Chaveroche,M.K., D'Enfert,C. and Ghigo,J.M. (2000) A rapid method for efficient gene replacement in the filamentous fungus Aspergillus nidulans. *Nucleic Acids Res.*, **28**, E97.
18. Cronan,J.E. (2003) Cosmid-based system for transient expression and absolute off-to-on transcriptional control of Escherichia coli genes. *J. Bacteriol.*, **185**, 6522–6529.
19. Valens,M., Penaud,S., Rossignol,M., Cornet,F. and Boccard,F. (2004) Macrodomain organization of the Escherichia coli chromosome. *EMBO J.*, **23**, 4330–4341.
20. Crawford,I. (1989) Evolution of a biosynthetic pathway: the tryptophan paradigm. *Ann. Rev. Microbiol.*, **43**, 567–600.
21. Biskri,L., Bouvier,M., Guérout,A., Boisnard,S. and Mazel,D. (2005) Comparative study of class 1 integron and Vibrio cholerae superintegron integrase activities. *J. Bacteriol.*, **187**, 1740–1750.
22. Ziegelin,G., Pansegrau,W., Strack,B., Balzer,D., Kröger,M., Kruft,V. and Lanka,E. (1991) Nucleotide sequence and organization of genes flanking the transfer origin of promiscuous plasmid RP4. *DNA Seq. J. DNA Seq. Mapping*, **1**, 303–327.
23. Shetty,R.P., Knight,T.F. and Endy,D. (2008) Engineering BioBrick vectors from BioBrick parts. *J. Biol. Eng.*, **2**, 5.
24. Garbe,T.R., Kobayashi,M. and Yukawa,H. (2000) Indole-inducible proteins in bacteria suggest membrane and oxidant toxicity. *Arch. Microbiol.*, **173**, 78–82.
25. Agresti,J.J., Antipov,E., Abate,A.R., Ahn,K., Rowat,A.C., Baret,J., Marquez,M., Klibanov,A.M., Griffiths,A.D. and Weitz,D.A. (2010) Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proc. Natl Acad. Sci. USA*, **107**, 4004–4009.
26. Brouzes,E., Medkova,M., Savenelli,N., Marran,D., Twardowski,M., Hutchison,J.B., Rothberg,J.M., Link,D.R., Perrimon,N. and Samuels,M.L. (2009) Droplet microfluidic technology for single-cell high-throughput screening. *Proc. Natl Acad. Sci. USA*, **106**, 14195–14200.
27. Bouvier,M., Demarre,G. and Mazel,D. (2005) Integron cassette insertion: a recombination process involving a folded single strand substrate. *EMBO J.*, **24**, 4356–4367.
28. Frumerie,C., Ducos-Galand,M., Gopaul,D.N. and Mazel,D. (2010) The relaxed requirements of the integron cleavage site allow predictable changes in integron target specificity. *Nucleic Acids Res.*, **38**, 559–569.
29. Bouvier,M., Ducos-Galand,M., Loot,C., Bikard,D. and Mazel,D. (2009) Structural features of single-stranded integron cassette attC sites and their role in strand selection. *PLoS Genet.*, **5**, e1000632.
30. Li,M.Z. and Elledge,S.J. (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat. Methods*, **4**, 251–256.
31. Shao,Z., Zhao,H. and Zhao,H. (2009) DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.*, **37**, e16.
32. Collis,C.M., Recchia,G.D., Kim,M., Stokes,H.W. and Hall,R.M. (2001) Efficiency of recombination reactions catalyzed by class 1 integron integrase IntI1. *J. Bacteriol.*, **183**, 2535–2542.

Supplementary Tables

The synthetic integron: an *in vivo* genetic shuffling device

David Bikard, Stéphane Julié-Galau, Guillaume Cambray, Didier Mazel

clone	Genotype
1	_ A | B | C | E | D | p |
2	_ D | A | B | C | E |
3	_ B | C | A | C | E | D | p |
4	_ C | E | A | B | C | E | D |
5	_ C | A | B | C | E | D | p |
6	_ C | Z | A | B | E | D | p |

Table S1. Cassettes arrangements conferring prototrophy on pSB4C5 plasmids
The letter code is the same as in Table 1. Vertical bars represent *attC* recombination sites.
Horizontal bars represent *attI* recombination sites.
Original order is _ t \| Z \| A \| B \| t \| C \| E \| D \| p \|

clone	Genotype
1	_ A | B | C | E | D | pSW _ t | Z |
2	_ A | B | C | E | D | pSW _ t |
3	_ A | B | C | E | D | pSW _ t | Z |
4	_ A | B | C | E | D | pSW _ t |
5	_ A | B | C | E | D | pSW _ t | Z |
6	_ B | A | pSW _ C | E | D |
7	_ B | C | E | D | p | pSW _ A |
8	_ A | B | C | E | D | pSW _ t | Z |

Table S2. Cassettes arrangements of the recovered prototrophs from the conjugation assay with one donor strain.
The letter code is the same as in Table 1. Vertical bars represent *attC* recombination sites.
Horizontal bars represent *attI* recombination sites. The delivered plasmid is pSWlib: pSW _ t \| Z \| A \| B \| t \| C \| E \| D \| p \|

clone	Genotype
1	_ B | A | C | E | D | pSW _
2	_ C | B | A | Z | pSW E | D | pSW _
3	_ B | A | C | E | D | pSW _
4	_ D | pSW _ B | A | C | E |
5	_ D | pSW _ C | E | A | Z | pSW _ B |
6	_ D | pSW _ C | E | Z | pSW _ B | A |
7	_ B | D | pSW _ C | E | A | Z | pSW _
8	_ E | D | pSW _ C | pSW _ B | A | Z |

Table S3. Cassettes arrangements of the recovered prototrophs from the conjugation assay with two donor strains.
The letter code is the same as in Table 1. Vertical bars represent *attC* recombination sites.
Horizontal bars represent *attI* recombination sites. The delivered plasmids are pSW-BA: pSW _ B \| A \| Z \| , and pSW-CED: pSW _ C \| E \| D \|

number	name	sequence
936	pSWK_AttP-R	AGTACTCTAGAAGCGGCCGCGAATTCTATCAAGCTTATCGATACCGTCGACG
937	pSWK_AttP-F	GCTTCTAGAGTACTAGTAGCGGCCGCTGCAGGCGGTGGAGCTCAAATCAAATAATG
938	lacZalpha-F	CCTCAGCTACACGTGCACTGATTAAAGAGGAGAAAATGACCATGATTACGGATTCACTGG
939	trpE-F	CCTCAGCTACACGTGCACTGATTAAAGAGGAGAAAATGCAAACACAAAAACCGACTCTCG
940	trpD-F	CCTCAGCTACACGTGCACTGATTAAAGAGGAGAAAATGGCTGACATTCTGCTGCTCGATA
941	trpC-F	CCTCAGCTACACGTGCACTGATTAAAGAGGAGAAAATGATGCAAACCGTTTTAGCGAAAA
942	trpA-F	CCTCAGCTACACGTGCACTGATTAAAGAGGAGAAAATGGAACGCTACGAATCTCTGTTTG
943	trpB-F	CCTCAGCTACACGTGCACTGATTAAAGAGGAGAAAATGACAACATTACTTAACCCCTATT
944	LacZalpha-R	GCCGCGAAGCGGCGTCGGCTTGAATGAATTGTTATAACCTTAATTCAGGCTGCGCAACTGTTGGGAA
945	trpA-R	GCCGCGAAGCGGCGTCGGCTTGAATGAATTGTTATAACCTAACTGCGCGTCGCCGCTTTCATC
946	trpB-R	GCCGCGAAGCGGCGTCGGCTTGAATGAATTGTTATAACCAGATTCGTAGCGTTCCATCAGATT
947	trpC-R	GCCGCGAAGCGGCGTCGGCTTGAATGAATTGTTATAACCTCATTGTTCCTTTCCTTAATATGC
948	trpD-R	GCCGCGAAGCGGCGTCGGCTTGAATGAATTGTTATAACCGTTTGCATCATTTACCCTCGTGCC
949	trpE-R	GCCGCGAAGCGGCGTCGGCTTGAATGAATTGTTATAACCATCAGAAAGTCTCCTGTGCATGAT
950	pos1-F	GCAGAATTCGCGGCCGCTTCTAGAGCCTCAGCTACACGTGCACTG
951	pos1-R	TTTCTGCAGGCCATTCAGGCGGTTATAACGCTTGAATTAAGCCGCGCCGCGAAGCGGCGTCGGCTTGAAT
952	pos2-F	TTTGAATTCGCCTGAATGGCCCTCAGCTACACGTGCACTG
953	pos2-R	TTTCTGCAGGCCAGTGAGGCGGTTATAACGCTTGAATTAAGCCGCGCCGCGAAGCGGCGTCGGCTTGAAT
954	pos3-F	TTTGAATTCGCCTCACTGGCCCTCAGCTACACGTGCACTG
955	pos3-R	TTTCTGCAGGCCATCGAGGCGGTTATAACGCTTGAATTAAGCCGCGCCGCGAAGCGGCGTCGGCTTGAAT
956	pos4-F	TTTGAATTCGCCTCGATGGCCCTCAGCTACACGTGCACTG
957	pos4-R	TTTCTGCAGGCCAGCCAGGCGGTTATAACGCTTGAATTAAGCCGCGCCGCGAAGCGGCGTCGGCTTGAAT
958	pos5-F	TTTGAATTCGCCTGGCTGGCCCTCAGCTACACGTGCACTG
959	pos5-R	TTTCTGCAGGCCATTCAGGCGGTTATAACGCTTGAATTAAGCCGCGCCGCGAAGCGGCGTCGGCTTGAAT
960	pos6-F	TTTGAATTCGCCTGAATGGCCCTCAGCTACACGTGCACTG
961	pos6-R	AGCCTGCAGCGGCCGCTACTAGTAGGTTATAACGCTTGAATTAAGCCGCGCCGCGAAGCGGCGTCGGCTTGAAT
1002	ΔBgl1TrpA-F	AAACGCCACTCTGCGCGCATTTGCGGCAGGTGTGACTCCG
1003	ΔBgl1TrpA-R	CGGAGTCACACCTGCCGCAAATGCGCGCAGAGTGGCGTTT
1004	ΔBgl1TrpD-F	GCCATGTTAATGCGCCTGCATGGACATGAAGATCTGCAAG
1005	ΔBgl1TrpD-R	CTTGCAGATCTTCATGTCCATGCAGGCGCATTAACATGGC
1006	ΔBgl1TrpE-F	TTCCGGCAACGGCGAAGCACTCCTGGCACTACTGGATAAC
1007	ΔBgl1TrpE-R	GTTATCCAGTAGTGCCAGGAGTGCTTCGCCGTTGCCGGAA

Table S4. Oligos used for the plasmids constructions

number	name	Genotype / Construction	
p3938	pBAD::intI1	oriColE1 [ApR]	Demarre et al, 2007
p970	pSW23t	pSW23::oriTRP4; cat[CmR]; oriVR6Kγ	Demarre et al, 2005
p4849	pSW23t-attP	λattP cloned into the SacI site of pSW23t	this work
p6682	pSWK-attP	p4849 with biobrick restriction sites	this work
p8283	pSWlib	pSWK-attP::attI1-BBa_B0015-attC-lacZα-attC-trpA-attC-trpB-attC-BBa_B0015-attC-trpC-attC-trpE-attC-trpD-attC-BBa_J23100-attC	this work
p7661	pSB4C5lib	pSB4C5::attI1-BBa_B0015-attC-lacZα-attC-trpA-attC-trpB-attC-BBa_B0015-attC-trpC-attC-trpE-attC-trpD-attC-BBa_J23100-attC	this work
p7366	pSWKspec-attP	oriTRP4; aadA7[SpecR]; oriVR6K; λattP	this work
p7421	pSWplt	pSWKspec-attP::J23100-attI1	this work
p8013	pSW-CED	pSWKspec-attP::attI1-trpC-trpE-trpD	this work
p8014	pSW-BA	pSWKspec-attP::attI1-trpB-trpA-lacZα	this work
p7204	BBa_J61002:: BBa_J23107	oriColE1 [ApR]; [mRFP1]	registry of standard biological parts

Table S5. Plasmids table

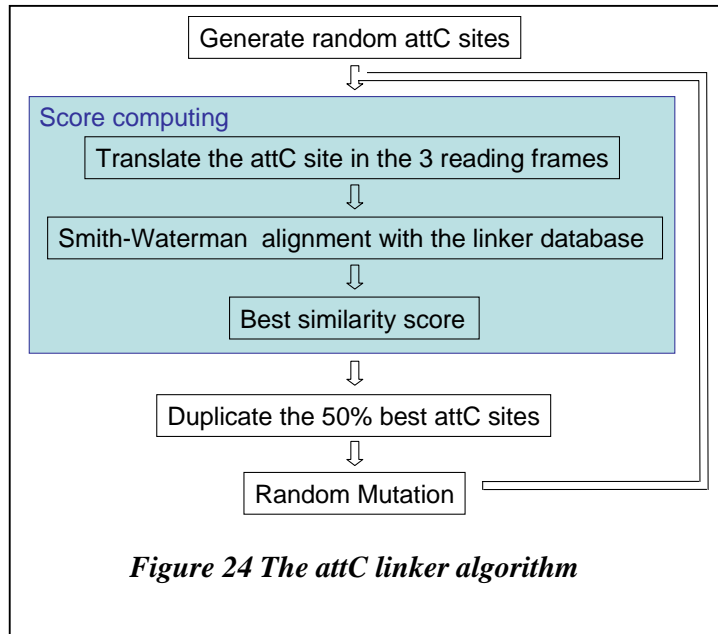| Name	genotype	source or reference
pi1	DH5α thyA::(erm-pir116) [EmR]	Demarre 2005
N2691	recA269::Tn10	R.G. Lloyd
ω7796	Δtrp::Km transduced into N2691	this work
ω7814	recA269::Tn10 Δtrp::frt	this work
ω7830	DB7814 attB::pSWlib	this work
ω7842	DB7830 pBAD::intI1	this work
ω7661-int	pSB4C5lib and pBAD::intI1 transformed in pi1	this work
ω7249	(F-) RP4-2-Tc::MuΔnic35 ΔdapA::(erm-pir) [KmR ErmR]	Babic et al. 2008
ω7850	recA269::Tn10 tranduced in ω7249	this work
ω7893	pSWlib transformed in ω7850	this work
ω8066	pSW-CED transformed in ω7850	this work
ω8067	pSW-AB transformed in ω7850	this work
ω7902	7814 attB::pSWplt	this work
ω7902-int	pBAD::intI1 transformed in ω7902	this work
ω8072	p7204 transformed in ω7814	this work

Table S6. Strains table

attC sites has linkers for protein domain shuffling

In this work we suggest that besides the engineering of metabolic pathways, synthetic integrons could be used to shuffle protein domains. It is now well established that in several cases, domains can be interchanged to produce new functional proteins (Grunberg and Serrano 2010). Several recent studies took advantage of the natural modularity of proteins to rewire signaling networks (Park, Zarrinpar et al. 2003; Yeh, Rutigliano et al. 2007; Bashor, Helman et al. 2008). A family of proteins that could be particularly interesting to engineer with a combinatorial approach include the non-ribosomal polypeptide and polyketide synthetases. Non-ribosomal polypeptides and polyketides are a vast family of molecules that often present pharmaceutical interest (Walsh 2004). They are synthesized by complex proteins composed of series of functional modules that each catalyzes the addition of a new monomer onto a growing polyketide/polypeptide chain. The order and the modular composition of the protein determine the output molecule (Khosla, Kapur et al. 2009). Functional chimeric proteins have been made that catalyze the formation of new compounds (Bedford, Jacobsen et al. 1996), and combinatorial approaches could certainly help discover many more interesting molecules.

However, changing the domain composition of a protein still involves a great deal of disruption, and testing several designs might be necessary to find working variants. Mutagenesis methods currently used for the production of chimeric protein libraries are reviewed in the introduction (part I.4.a.v). In cases where protein modules are well defined and where the presence of a large linker between the domains is not problematic, integrons might be an interesting alternative to existing methods. A synthetic integron enables to generate large number of variants directly in vivo, and a greater diversity of composition than with methods such as SISDC could likely be obtained; with fewer efforts too (Hiraga and Arnold 2003). Furthermore, the orientation of the cassette is preserved and no frameshift can be introduced, which is a great advantage in comparison to methods such as NRR (Bittker, Le et al. 2004).

To this end, the *attC* sites used need to be good protein linkers. In order to make such *attC* linker, we can take advantage of the flexibility of these recombination sites. We have seen in the introduction (part II.3.b) that almost only structural features
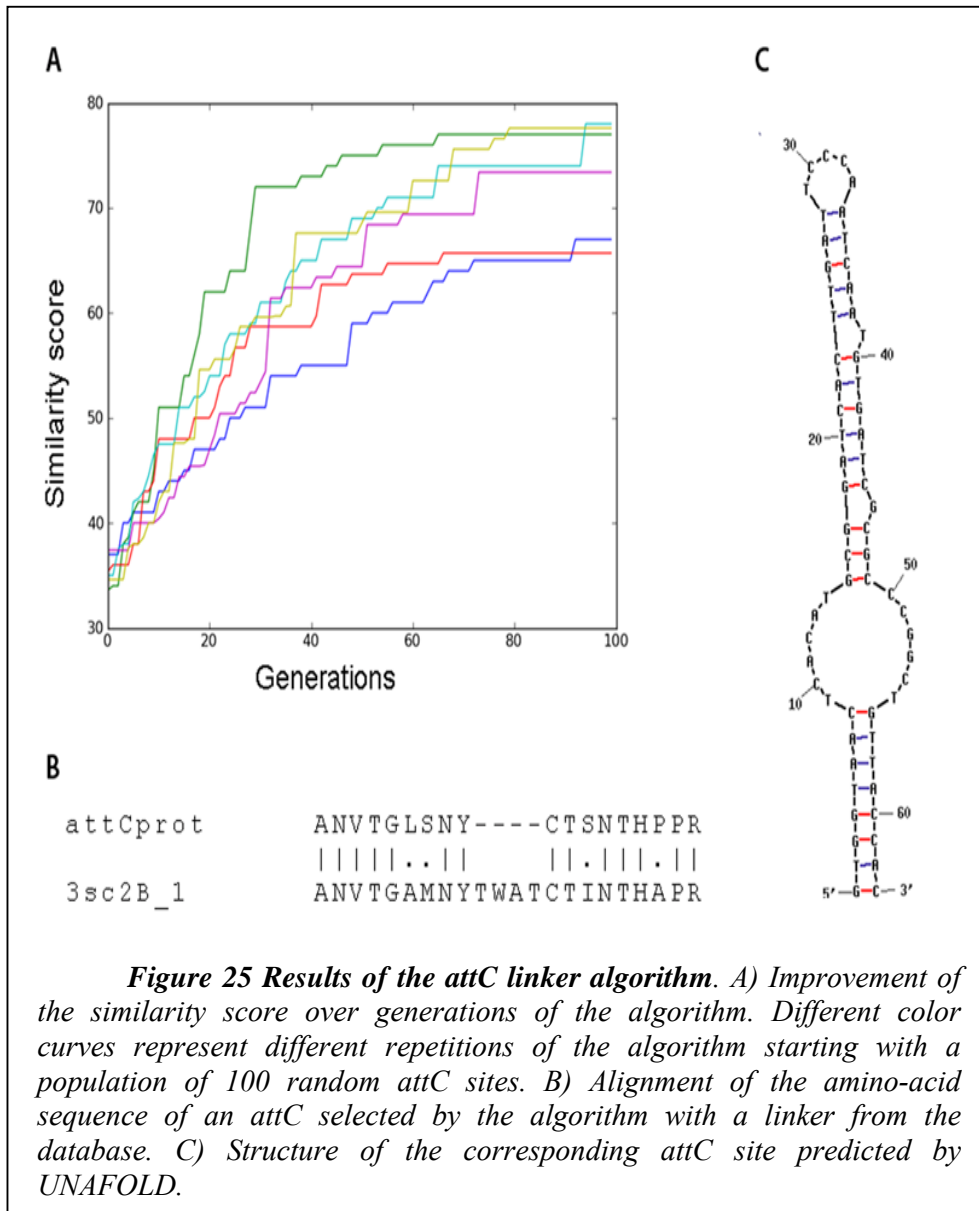
146

are required for *attC* recombination, which leaves ample room for modifications of the primary sequence. I have written an algorithm that consists in several rounds of random mutations and selection for improved *attC* sites. Mutations are realized so that the structure of the site is preserved, and the sites with the best homology scores to linkers retrieved from a database are selected (LinkerDB: www.ibi.vu.nl/programs /linkerdbwww) (Figure 24).

Figure 24 The attC linker algorithm

This algorithm allowed selecting many different *attC* sites (Figure 25). Further work would now be required to test the functionality of the selected sequences, both as proper protein linkers and as recombinogenic *attC* sites. The sequences generated are completely new, and might not all be functional. Experimental measurements will probably reveal new sites with altered functionality. The analysis of such sites will enable to still improve our model of *attC* site folding and recombination. In return, this better knowledge will allow writing a better scoring function for the algorithm, so that new functional sites can be created at first trial.

Figure 25 Results of the attC linker algorithm. *A) Improvement of the similarity score over generations of the algorithm. Different color curves represent different repetitions of the algorithm starting with a population of 100 random attC sites. B) Alignment of the amino-acid sequence of an attC selected by the algorithm with a linker from the database. C) Structure of the corresponding attC site predicted by UNAFOLD.*

Discussion

The discovery of ssDNA as a recombination substrate in a number systems (CTX phage, IS*608*, Integrons) raised important questions, some of these still unanswered (Bouvier, Demarre et al. 2005; Ronning, Guynet et al. 2005; Val, Bouvier et al. 2005; Barabas, Ronning et al. 2008). I will first discuss the implications of recombining ssDNA in the light of what we have learned, in particular about the way IRs can fold in the cell. Then, we will see that the evolution of ssDNA as a substrate for integron recombination can only be understood in the broader context of horizontal gene transfer and stress response. Finally I will discuss the use of integrons as a mutagenesis tool for directed evolution in comparison to other available techniques and propose further potential applications.

I. Recombining ssDNA

I.1 Hairpin formation: cruciform extrusion vs. single-stranded hairpin

How and under what conditions do ssDNA hairpins fold in the cell? Among the hairpins with biological functions reviewed in the introduction, some obviously fold from ssDNA (the *sso* of RCR and the single-stranded phage hairpins). On the other hand, there is consistent evidence that the N4 hairpin promoters and the hairpin of the RCR plasmids *dso* fold as cruciforms (Noirot, Bargonetti et al. 1990; Dai, Greizerstein et al. 1997). However, there are only a few cases of successful cruciform detection of natural IR *in vivo*. Indeed, most reported *in vivo* cruciform detection involved artificial conditions favoring hairpin extrusion: small loops, IR in AT-rich regions, perfect palindromes with AT-rich centers and GC-rich stems, topoisomerase mutants or salt shock to increase supercoiling (Zheng and Sinden 1988; Sinden, Zheng et al. 1991; Zheng, Kochel et al. 1991).

When investigating the conditions that can lead to integron *attC* site folding (Loot et al., 2010), we realized that these recombination sites are extremely good candidates for studying hairpin formation *in vivo*. Recombination events can only happen with folded *attC* sites and can be detected at very low frequencies.

Furthermore, only the bottom strand of the *attC* site is recognized by the integrase (Francia, Zabala et al. 1999; Bouvier, Demarre et al. 2005). This enables distinguishing recombination events occurring with hairpins formed during replication on the lagging strand template, from events occurring with hairpins extruding as cruciforms, or during other processes such as repair.

We found that *attC* hairpins fold much more frequently on the lagging strand template than through other processes. However, in natural chromosomal integrons the recombinogenic strand of *attC* sites is always found on the leading strand of replication. Under such conditions, *attC* hairpin can presumably only fold on ssDNA generated by repair or through cruciform extrusion. However, *attC* sites are very imperfect IRs with a central spacer sequence (the VTS) of up to 80bp. Such imperfections are known to hinder cruciform formation, and extrusion of imperfect hairpins had previously only been reported for very AT-rich IRs (Benham, Savitt et al. 2002). Nevertheless, transformation and recombination of non-replicative plasmids carrying *attC* sites enabled to show that *attC* sites can extrude cruciforms at rather low frequencies ($<10^{-3}$). Most surprisingly, *attC* sites with large VTS were also able to fold cruciform structures. It has been observed that integron cassettes are particularly AT-rich (Mazel 2006), which could favor *attC* site extrusion following a C-type mechanism (see the introduction part III.1.b).

Putting these results in perspective with what was already known about IRs folding *in vivo* gives the following picture. Large perfect IRs can presumably fold into cruciforms but are genetically unstable because of their propensity to hinder replication and be cleaved by SbcCD. Small perfect (or almost-perfect) IRs can fold into cruciforms only when their sequence and context allow it. The N4 promoters and pT181 plasmid origin of replication are examples of such IRs with biological functions. Imperfect IRs are genetically more stable regardless of their size, but fold into cruciforms only rarely. They could still be involved in biological functions that take place at low frequencies such as integrons or IS*608* recombination. Alternatively, imperfect IRs present in topologically constrained regions such as replication origins could also fold into cruciforms, which might be the case for the M13-A hairpin and for the *ssi* present in some origins of replication (see the introduction, part III.2.a). However, one should remember that these hairpins are specifically bound by cognate proteins that could stabilize cruciforms.

I.2 Hairpin recombination substrates as a sensors of environmental stress

The integron integrase is controlled by the SOS regulon which might be a way to detect when ssDNA and potential recombination substrates are entering the cell (Guerin, Cambray et al. 2009). Additionally, this also ensures that recombination occurs under stressful conditions, i.e. when innovation is needed. The study of *attC* site folding revealed another pathway which probably contributes to the same purpose. Indeed, the folding of *attC* sites into recombinogenic hairpins can occur when their bottom strand is on the template for the synthesis of the lagging strand. However, as mentioned above, all chromosomal integrons are oriented so that the bottom strand of the *attC* sites is on the leading strand. While ssDNA is produced on the lagging strand template under normal conditions, large amounts of ssDNA on the leading strand are synonymous with DNA damage. It would thus be interesting to investigate if DNA damages can lead to more *attC* folding and recombination, independently of the integrase regulation by the SOS. Another reason why this might be the case is that SOS response increases the negative supercoiling of DNA (Majchrzak, Bowater et al. 2006), and may thus promote the extrusion of cruciforms, enabling *attC* folding on dsDNA upon DNA damage or horizontal gene transfer.

I.3 Parasite structures and recombination control

When an IR forms a hairpin, several folds may be possible, depending on the perfection of the IR and on the presence of neighboring sequences that could capture the IR into other structures. Those different folds will have different probabilities of occurrence and will be able to interconvert into one another or not, depending on how much they differ. For IS*608* the recognized hairpins are small and almost perfect, parasite structures are thus unlikely. In contrast, the CTX phage *attP* site, folds into a much larger structure (~170bp) that could form parasites more easily, especially since the whole genome of 17kb is presumably single stranded upon entry in the cell. Although this possibility has not yet been studied, alternative structures could interfere with the proper fold of the *attP* site. Pointing in this direction, observation

154

were made that neighboring sequences of the *attP* site do affect its recombination frequency (Marie-Eve Val, personal communication).

The *attC* sites of integrons are very variable both in sequence and in size. The larger ones in particular may be subject to improper folding. One of the unexpected finding of our work on *attC* site was that a parasite structure, found in a natural *attC*, could strongly hinder recombination. However, only one of the 169 *V.cholerae El Tor attC* sites, the VCR$_{2/1,}$ was used in our assays. To make sure that we had not picked by chance an abnormal VCR, I realized a quick analysis of all the *V.cholerae attC* sites together with the *attC* sites of other chromosomal integrons. This analysis revealed a continuum of sites with varying probabilities to fold properly. However, while sites with a predominant parasite structure represent the majority of *V.cholerae attC* sites, they are almost absent in other strains like *X.campestris* or *V.vulnificus*.

Further experimental work would be required to know if the parasites found for the *attC* sites of *V.cholerae* hinder recombination in *V.cholerae* (we did the experiment in *E.coli* only). It is possible that host factors contribute to the proper folding and allow sites with parasites to recombine as well as sites without. If this is the case, parasites could be a mechanism to ensure that only members of the same species can recombine these cassettes efficiently?

However, if the parasite structures of VCRs do hinder recombination in *V.cholerae* as well, the selective forces that shape *attC* sites should be investigated in more details. Our experiments revealed that *attC* sites with parasites recombined less frequently in all conditions, excepted conjugation. Indeed, when the sites were delivered as ssDNA through conjugation, they all recombined with similar frequencies regardless of their VTS size and of the presence of alternative structures (unless those alternative structures were made artificially very strong). Alternatively, parasite structures may thus be a way to ensure that recombination occurs upon horizontal transfer only.

I.4 Recombination dynamics

In order to be maintained, an *attC* site has to find a balance between several phenomenons. First it must be functional, i.e. it must be recombined from time to time

and the associated gene must give a selective advantage to the host (Cambray 2009). Otherwise, either the host will be out-competed or the cassette will slowly drift and eventually disappear (Cambray 2009). Second, it must be stable, i.e. it must not be too easy to excise, and it must make sure that once excised it is reintegrated with the best possible chances. The IntI integrase does not reintegrate cassettes with a 100% chance; there is thus a selective pressure not to recombine too much. The balance between these two opposing selective forces might be achieved differently in different integrons, and influence for instance the number of cassettes that an integron can "store".

In order to understand this point better, further studies would be required to find the parameters that influence cassette reintegration or loss. In particular, cassettes are often characterized by the fact that the sequence immediately after the recombination point of the previous *attC* site is the inverse repeat of the sequence just before the next *attC* site (Figure 26). Because of these inverse repeats that frame the cassettes, once excised, the *attC* site of the circular intermediate has a stem 5-7bp longer than the sites before recombination (Rowe-Magnus, Guerout et al. 2003). There is so far no explanation for this organization, but it would be worth investigating if this longer stem might promote reintegration after excision. This hypothesis can hold only if the most frequent excision event is the excision of a single cassette, and not multiple cassettes simultaneously. Therefore the probability for distant *attC* sites to recombine instead of adjacent *attC* sites should also be assessed.
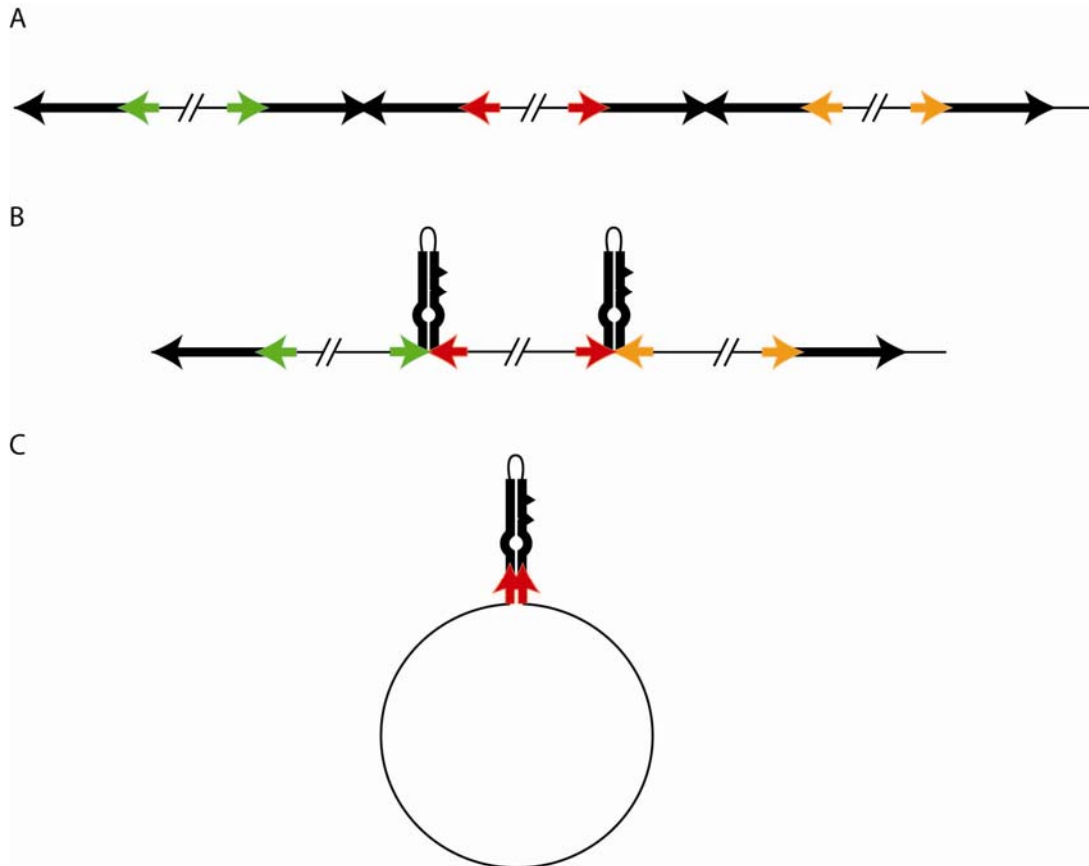
Figure 26 Inverse repeats at the base of the attC sites.

A) Organisation of the cassette array in an integron. attC sites are represented by two converging black arrows. The colored arrows represent the small inverse repeats that frame most integron cassettes. B) attC site folding leading, and excision of a single cassette (C). Note that the stem of the attC on the excised cassette is longer than the stem of the original attC sites.

Finally cassettes that recombine only when they are needed will be more successful than other cassettes because they minimize the chances to be lost while maximizing the chances to be useful to the cell and be selected. One way to achieve this could be to control recombination in the way riboswitches control translation, i.e. through changes in conformation of the *attC* site mediated by the binding of a regulatory molecule to an aptamer. It would thus be truly interesting to look for aptamers in *attC* sites, especially the ones where parasite structures hindering recombination have been identified.

I.5 Solving the junction

Besides the folding problem mentioned earlier, there could be mechanistic problems when a ssDNA substrate is recombined. If both recombination substrates are ssDNA hairpins, then a single strand exchange is sufficient to complete the recombination reaction. However, if one of the recombination partners is double-stranded, after the first strand exchange, a second strand exchange cannot solve the junction, but would form abortive and potentially lethal recombination products. This is the case for the recombination between the single-stranded CTX *attP* site and the double-stranded *dif* site, or between an integron single-stranded *attC* site and a double-stranded *attI* site. The second strand exchange would lead to the linearization of the molecule carrying the double-stranded substrate. Therefore, the isomerisation of the recombination synapse and the second strand exchange must be prohibited. For CTX, the mismatch in the *attP* core site probably plays this role. For integrons, the asymmetry of the complex and especially the binding of one IntI monomer to the *attC* extra-helical "T ", which pulls the catalytic tyrosine away from the phosphate links, makes a second strand exchange impossible (MacDonald, Demarre et al. 2006).

Still, the junction has to be resolved somehow. It was thus proposed for both integron recombination and CTX integration that replication is involved in this last step (Val, Bouvier et al. 2005; Mazel 2006). One of the consequences of this model is that recombination would be semi-conservative. After the passage of a replication fork, the recombination product would be formed, and the initial substrate of the partner carrying the double-stranded site would be reconstituted. Experiments are now being undertaken in the lab to assess the semi-conservative nature of integron recombination.

If replication seems a likely way to solve the pseudo-Hollyday junction, the precise mechanism remains to elucidate. Does the junction wait for a replication fork to arrive? Or is a replication complex directly assembled at the junction? A candidate host factor for this function could be the PriA protein. PriA can recruit a replication complex at branched DNA structures (notably Y-forks) which are structurally similar to the pseudo-Hollyday junction (Tanaka, Mizukoshi et al. 2007).

Finally, this unique mechanism of recombination may be crucial for cassette creation. Indeed, the mechanism by which new cassettes are created is still unknown.

The semi-conservative nature of recombination could account for cassette duplications. A cassette that is excised and reintegrated at the *attI* site of the conserved integron (deriving from the replication of the top strand) would be duplicated on this molecule. Duplication and divergence of cassettes may well account for a number of cassette creations. Large integrons notably contain duplicated cassettes; some of them present as much as 24 times and spread throughout the cassette array (Anne-Marie Guerout, personal communication). An alternative explanation for cassette creation involves the random insertion of cassettes at secondary recombination sites. If two such events occur nearby, the region in-between the integrated *attC* sites could be captured by an integron.

<div style="text-align:center; border:1px solid black; padding:10px;">

II. Integrons as a new tool for genetic engineering

</div>

Integron are powerful adaptation devices with a unique recombination mechanism. As synthetic biology expands, new tools are always welcome for the implementation of new functions. I demonstrated how integrons can be used to assemble and shuffle genetic elements of interest. I will discuss this point further as well as other potential applications of the integron recombination machinery.

II.1 Generation of random genetic circuits

Directed evolution relies on the construction of libraries of mutants that are screened for improved phenotypes. It enables to overcome the limitations of rational design in the engineering of ever more complex systems. Essentially, mutagenesis techniques can be classified in several categories depending on the kind of mutation they produce:

- Random point mutations can be introduced genome-wide in mutator strains or using chemicals (Stefan, Radeghieri et al. 2001).

- Random point mutations over a region up to 10kb (about the limit of what can currently be amplified through PCR) can be introduced through error-prone PCR (Lerner and Inouye 1994).

- Random or controlled mutagenesis of a defined small region (~10bp) can be achieved *in vitro* using synthetic oligos through cassette mutagenesis or site directed mutagenesis (Ruvkun and Ausubel 1981; Wells, Vasser et al. 1985).

- Random and combinatorial mutagenesis of multiple small regions (~10bp) *in vivo*, can be achieved with the MAGE method (Wang, Isaacs et al. 2009).

- Random or controlled mutagenesis of a <u>random</u> small region (within a defined region of up to a few kb) can be achieved in vitro by random insertion/deletion (RID) (Murakami, Hohsaka et al. 2002).

- Random recombination of homologous sequences genome-wide can be achieved in some bacterial strains through whole-genome shuffling (Zhang, Perry et al. 2002).

- Random recombination of homologous sequences of up to 10kb can be achieved through sexual PCR (Stemmer 1994).

- Random recombination of heterologous sequences can be achieved *in vitro* through various techniques of randomized assembly ligation and overlap extension PCR (Tsuji, Onimaru et al. 2001; Guet, Elowitz et al. 2002; Bittker, Le et al. 2004; Cox, Surette et al. 2007; Gertz, Siggia et al. 2009).

The Synthetic Integron is the first tool to achieve random recombination of heterologous sequences *in vivo*. It is thus interesting to put in contrast with the randomized assembly ligation methods which have successfully been used to achieve this *in vitro*. Randomized ligation is realized by mixing together several building blocks and ligating them. Assemblies of a given size can eventually be purified through electrophoresis. With this method random numbers of building blocks are ligated together in random order and can then be inserted into a plasmid. Alternatively, one may use homologous overhangs between the parts (often generated with restriction enzymes) in order to have more control over the assembly order. To this end, the final construct can be divided into a given number of units, (1, 2, …, N). Variants of each unit are constructed, and all variants are mixed together for the ligation reaction. Carefully chosen overhangs ensure that a variant of unit 1 is followed by a variant of unit 2, itself followed by a variant of unit 3 etc (Guet, Elowitz et al. 2002; Cox, Surette et al. 2007)(Hiraga and Arnold 2003).

Random assemblies of 7 units have been achieved in this way (Meyer, Hochrein et al. 2006), but it is not clear if this strategy can be scaled up to more units. However for constructs where larger homology sequences can be tolerated, method relying on homologous recombination could probably allow combining variants of more than 10 units (Li and Elledge 2007; Shao, Zhao et al. 2009). Gibson and colleagues have reported the simultaneous assembly of 25 fragments with 6kb overlaps in yeast, and suggested that with their method genetic pathways could be constructed in a combinatorial fashion such that each member in the combinatorial library has a different combination of gene variants (Gibson, Benders et al. 2008). Here, the limitation comes from the number of different elements that can simultaneously be transformed in a single yeast cell.

The Synthetic Integron does not allow distinguishing between unit types to ensure that a particular order is conserved. However, the size and number of the elements that can be combined is not constrained by an assembly method, and can potentially be much higher. Natural integrons containing more than 200 cassettes have indeed been isolated (Mazel 2006). Another advantage of integrons is that new elements can easily be added in an existing cassette array, whereas other methods require all building blocks to be present in the initial assembly reaction. Furthermore, with the integron method, combinations are very easily generated, and can directly be screened without the need of further cloning. Finally, another level of complexity can be added to the synthetic integron through conjugation, enabling to yield more combinations. In a conjugation assay where several donor strains deliver different cassette arrays, not only can the integron shuffle the integrated cassettes, but new combinations can also arise from the fact that different donor strains will randomly deliver their cassettes to the receptor cells. Moreover, conjugation is an easy, efficient and widespread method to deliver DNA, which might be handy for the engineering of bacterial strains where transformation does not work.

Consequently, both methods, randomized ligation and synthetic integron, will likely be used for different purposes. While randomized ligation was already successfully applied to the engineering of promoter and small regulatory networks, integrons may advantageously be used for the engineering of large metabolic pathway and the delivery of genetic elements through conjugation as we demonstrated with the tryptophan operon.

II.2 Recombination sites "a la carte"

The fact that integrons *attC* sites recombine as ssDNA hairpins offers the possibility to modify them extensively. They are recognized mostly through structural features and the only conserved sequence is the core "GTT/AAC". Else, the *attC* site is only defined by its hairpin structure with a bulge between the R and L box, and two or three extrahelical bases in the L box (Bouvier, Ducos-Galand et al. 2009). Furthermore, it was recently shown that even the GTT/AAC can accept a wide range of mutations, provided that the first nucleotide of the triplet is the same for both recombination partners (Frumerie, Ducos-Galand et al. 2010). I have developed an algorithm that takes advantage of these properties of *attC* sites to generate sequences

162

that would make good protein linkers. Many more *attC* sites could be generated with a similar method in order to fulfill other useful function. *attC* sites could likely be turned into promoter sequence, transcriptional terminator and even recombination sites recognized by other recombinases as well as the integrase.

Another application could be the construction of sites with predictable recombination frequencies. During our work on integron folding, I developed a statistical model that explains recombination frequency with two parameters: the VTS size and the probability to fold properly. This model could now be tested against other sites to verify its prediction capacities. It could then be used to design new sites with given recombination frequencies.

II.3 In vivo DNA assembly

Finally, integrons may simply be used to sequentially assemble large gene constructs directly *in vivo*. Assembling several DNA parts using site-specific recombination can be much quicker than traditional digestion/ligation methods as advertised by Invitrogen for their Gateway® cloning system. (Invitrogen now offers the possibility to assemble in a single step up to 4 DNA fragments using a recombination system based on the lambda integrase *in vitro*.)

Integron could be used to sequentially integrate DNA parts *in vivo* into an *attI* site previously placed at any position of interest. We demonstrated this possibility with the conjugation assay developed for the Synthetic Integron paper. We were able to integrate multiple parts delivered from several plasmids to an *attI* site placed on the chromosome. The part we want to add to the array can simply be cloned into an "entry" vector where it would be framed by two *attC* sites. The entry vector would be delivered through suicide conjugation into the engineered strain and integrated at the *attI* by the integrase expressed in *trans*. Integration events would be selected with the antibiotic resistance of the entry vector. Its excision could then be selected with a counter-selection marker like the *ccdB* toxin. In order to ensure that unwanted deletion or reordering events do not occur, notably with the cassette already present in the array, the recombination sites could be chosen among the variants with different core sequences described by Frumerie and colleagues (Frumerie, Ducos-Galand et al. 2010). For instance a site with the wild-type "GTT" core sequence is able to recombine with a "GGG" core sequence, while two "GGG" cannot recombine. If the

attC site associated with the "entry" vector has a "GTT" core site and all other sites are "GGG", then only the proper integration and excision events are supposed to occur.

References

A

Adhya, S. (1989). "Multipartite genetic control elements: communication by DNA loop." Annu Rev Genet **23**: 227-50.

Alper, H., K. Miyaoku, et al. (2005). "Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets." Nat Biotechnol **23**(5): 612-6.

Alper, H. and G. Stephanopoulos (2007). "Global transcription machinery engineering: a new approach for improving cellular phenotype." Metab Eng **9**(3): 258-67.

Althorpe, N. J., P. M. Chilley, et al. (1999). "Transient transcriptional activation of the Incl1 plasmid anti-restriction gene (ardA) and SOS inhibition gene (psiB) early in conjugating recipient bacteria." Mol Microbiol **31**(1): 133-42.

Alvarez-Martinez, C. E. and P. J. Christie (2009). "Biological diversity of prokaryotic type IV secretion systems." Microbiology and molecular biology reviews: MMBR **73**(4): 775.

Arai, K. and A. Kornberg (1981). "Unique primed start of phage phi X174 DNA replication and mobility of the primosome in a direction opposite chain synthesis." Proceedings of the National Academy of Sciences of the United States of America **78**(1): 69.

Arakawa, Y., M. Murakami, et al. (1995). "A novel integron-like element carrying the metallo-beta-lactamase gene *blaIMP*." Antimicrobial Agents and Chemotherapy **39**(7): 1612-5.

Arkin, A. P. and D. C. Youvan (1992). "Optimizing nucleotide mixtures to encode specific subsets of amino acids for semi-random mutagenesis." Biotechnology (N Y) **10**(3): 297-300.

Babic, A., A. B. Lindner, et al. (2008). "Direct Visualization of Horizontal Gene Transfer." Science **319**(5869): 1533.

Balke, V. L. and J. D. Gralla (1987). "Changes in the linking number of supercoiled DNA accompany growth transitions in Escherichia coli." J Bacteriol **169**(10): 4499-506.

Barabas, O., D. R. Ronning, et al. (2008). "Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection." Cell **132**(2): 208.

Barlow, R. S. and K. S. Gobius (2006). "Diverse class 2 integrons in bacteria from beef cattle sources." J Antimicrob Chemother **58**(6): 1133-8.

Barrick, J. E. and R. R. Breaker (2007). "The power of riboswitches." Sci Am **296**(1): 50-7.

Bashor, C. J., N. C. Helman, et al. (2008). "Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics." Science **319**(5869): 1539-43.

Bass, B. L. and T. R. Cech (1984). "Specific interaction between the self-splicing RNA of Tetrahymena and its guanosine substrate: implications for biological catalysis by RNA." Nature **308**(5962): 820-6.

Beaber, J. W., B. Hochhut, et al. (2004). "SOS response promotes horizontal dissemination of antibiotic resistance genes." Nature **427**(6969): 72-4.

Bedford, D., J. R. Jacobsen, et al. (1996). "A functional chimeric modular polyketide synthase generated via domain replacement." Chem Biol **3**(10): 827-31.

Benham, C. J., A. G. Savitt, et al. (2002). "Extrusion of an imperfect palindrome to a cruciform in superhelical DNA: complete determination of energetics using a statistical mechanical model." J Mol Biol **316**(3): 563-81.

Benzer, S. and S. P. Champe (1961). "AMBIVALENT rII MUTANTS OF PHAGE T4." Proc Natl Acad Sci U S A **47**(7): 1025-38.

Biskri, L. and D. Mazel (2003). "Erythromycin esterase gene ere(A) is located in a functional gene cassette in an unusual class 2 integron." Antimicrob Agents Chemother **47**(10): 3326-31.

Bittker, J. A., B. V. Le, et al. (2004). "Directed evolution of protein enzymes using nonhomologous random recombination." Proc Natl Acad Sci U S A **101**(18): 7011-6.

Bloom, J. D. and F. H. Arnold (2009). "In the light of directed evolution: pathways of adaptive protein evolution." Proc Natl Acad Sci U S A **106 Suppl 1**: 9995-10000.

Boucher, Y., J. E. Koenig, et al. (2007). "Integrons: mobilizable platforms that promote genetic diversity in bacteria." Trends in microbiology **15**(7): 301.

Boucher, Y., M. Labbate, et al. (2007). "Integrons: mobilizable platforms that promote genetic diversity in bacteria." Trends Microbiol **15**(7): 301-9.

Bouvier, M., G. Demarre, et al. (2005). "Integron cassette insertion: a recombination process involving a folded single strand substrate." Embo J **24**(24): 4356-67.

Bouvier, M., M. Ducos-Galand, et al. (2009). "Structural features of single-stranded integron cassette attC sites and their role in strand selection." PLoS Genet **5**(9): e1000632.

Boyd, E. F., S. Almagro-Moreno, et al. (2009). "Genomic islands are dynamic, ancient integrative elements in bacterial evolution." Trends Microbiol **17**(2): 47-53.

Brannigan, J. A. and A. J. Wilkinson (2002). "Protein engineering 20 years on." Nat Rev Mol Cell Biol **3**(12): 964-70.

Burrus, V. and M. K. Waldor (2004). "Shaping bacterial genomes with integrative and conjugative elements." Res Microbiol **155**(5): 376-86.

Cadwell, R. C. and G. F. Joyce (1992). "Randomization of genes by PCR mutagenesis." PCR Methods Appl **2**(1): 28-33.

Cambray, G. (2009). Evolutivité: le cas des intégrons et utilisation de sequences synonymes en evolution dirigée. Paris, Université Paris Diderot. **Doctorat:** 293.

Cambray, G. and D. Mazel (2008). "Synonymous genes explore different evolutionary landscapes." PLoS genetics **4**(11): e1000256.

Caramel, A. and K. Schnetz (1998). "Lac and lambda repressors relieve silencing of the Escherichia coli bgl promoter. Activation by alteration of a repressing nucleoprotein complex." Journal of molecular biology **284**(4): 875.

Carothers, J. M., J. A. Goler, et al. (2010). "Selecting RNA aptamers for synthetic biology: investigating magnesium dependence and predicting binding affinity." Nucleic Acids Res **38**(8): 2736-47.

Carr, K. M. and J. M. Kaguni (2002). "Escherichia coli DnaA protein loads a single DnaB helicase at a DnaA box hairpin." The Journal of biological chemistry **277**(42): 39815.

Carr, P. A. and G. M. Church (2009). "Genome engineering." Nature Biotechnology **27**(12): 1151.

Chalker, A. F., D. R. Leach, et al. (1988). "Escherichia coli sbcC mutants permit stable propagation of DNA replicons containing a long palindrome." Gene **71**(1): 201.

Champion, K. and N. P. Higgins (2007). "Growth rate toxicity phenotypes and homeostatic supercoil control differentiate Escherichia coli from Salmonella enterica serovar Typhimurium." J Bacteriol **189**(16): 5839-49.

Cheetham, G. M. and T. A. Steitz (1999). "Structure of a transcribing T7 RNA polymerase initiation complex." Science **286**(5448): 2305-9.

Chen, I. s., P. J. Christie, et al. (2005). "The ins and outs of DNA transfer in bacteria." Science (New York, N.Y.) **310**(5753): 1456.

Chilley, P. M. and B. M. Wilkins (1995). "Distribution of the ardA family of antirestriction genes on conjugative plasmids." Microbiology (Reading, England) **141**: 2157.

Claverys, J. P., B. Martin, et al. (2009). "The genetic transformation machinery: composition, localization, and mechanism." FEMS Microbiol Rev **33**(3): 643-56.

Claverys, J. P., M. Prudhomme, et al. (2006). "Induction of competence regulons as a general response to stress in gram-positive bacteria." Annu Rev Microbiol **60**: 451-75.

Coleman, N., S. Tetu, et al. (2004). "An unusual integron in Treponema denticola." Microbiology **150**(Pt 11): 3524-6.

Collins, J., G. Volckaert, et al. (1982). "Precise and nearly-precise excision of the symmetrical inverted repeats of Tn5; common features of recA-independent deletion events in Escherichia coli." Gene **19**(1): 139.

Collis, C. M., G. Grammaticopoulos, et al. (1993). "Site-specific insertion of gene cassettes into integrons." Molecular Microbiology **9**(1): 41-52.

Collis, C. M. and R. M. Hall (1992). "Gene cassettes from the insert region of integrons are excised as covalently closed circles." Mol Microbiol **6**(19): 2875-85.

Collis, C. M. and R. M. Hall (1995). "Expression of antibiotic resistance genes in the integrated cassettes of integrons." Antimicrob Agents Chemother **39**(1): 155-62.

Collis, C. M., M. J. Kim, et al. (2002). "Characterization of the class 3 integron and the site-specific recombination system it determines." J Bacteriol **184**(11): 3017-26.

Collis, C. M., M. J. Kim, et al. (2002). "Integron-encoded IntI integrases preferentially recognize the adjacent cognate attI site in recombination with a 59-be site." Mol Microbiol **46**(5): 1415-27.

Collis, C. M., G. D. Recchia, et al. (2001). "Efficiency of recombination reactions catalyzed by class 1 integron integrase IntI1." J Bacteriol **183**(8): 2535-42.

Courey, A. J. and J. C. Wang (1983). "Cruciform formation in a negatively supercoiled DNA may be kinetically forbidden under physiological conditions." Cell **33**(3): 817.

Cox, E. C. (1976). "Bacterial mutator genes and the control of spontaneous mutation." Annu Rev Genet **10**: 135-56.

Cox, R. S., M. G. Surette, et al. (2007). "Programming gene expression with combinatorial promoters." Molecular systems biology **3**: 145.

Cozzarelli, N. R. and J. C. Wang (1990). DNA Topology and Its Biological Effects, Cold Spring Harbor Laboratory Pr.

Crameri, A., G. Dawes, et al. (1997). "Molecular evolution of an arsenate detoxification pathway by DNA shuffling." Nat Biotechnol **15**(5): 436-8.

Crameri, A., S. A. Raillard, et al. (1998). "DNA shuffling of a family of genes from diverse species accelerates directed evolution." Nature **391**(6664): 288-91.

Cromie, G. A., C. B. Millar, et al. (2000). "Palindromes as substrates for multiple pathways of recombination in Escherichia coli." Genetics **154**(2): 513-22.

Dai, X., M. B. Greizerstein, et al. (1997). "Supercoil-induced extrusion of a regulatory DNA hairpin." Proc Natl Acad Sci U S A **94**(6): 2174-9.

Dai, X. and L. B. Rothman-Denes (1998). "Sequence and DNA structural determinants of N4 virion RNA polymeraseâ€"promoter recognition." Genes & Development **12**(17): 2782.

Dalbadie-McFarland, G., L. W. Cohen, et al. (1982). "Oligonucleotide-directed mutagenesis as a general and powerful method for studies of protein function." Proc Natl Acad Sci U S A **79**(21): 6409-13.

Das, B., J. Bischerour, et al. (2010). "Molecular keys of the tropism of integration of the cholera toxin phage." Proceedings of the National Academy of Sciences of the United States of America **107**(9): 4377.

Datta, S., C. Larkin, et al. (2003). "Structural Insights into Single-Stranded DNA Binding and Cleavage by F Factor TraI." Structure **11**(11): 1369.

Davydova, E. K., I. Kaganman, et al. (2009). "Identification of bacteriophage N4 virion RNA polymerase-nucleic acid interactions in transcription complexes." J Biol Chem **284**(4): 1962-70.

Dayn, A., S. Malkhosyan, et al. (1991). "Formation of (dA-dT)n cruciforms in Escherichia coli cells under different environmental conditions." J Bacteriol **173**(8): 2658-64.

Dayn, A., S. Malkhosyan, et al. (1992). "Transcriptionally driven cruciform formation in vivo." Nucleic Acids Res **20**(22): 5991-7.

de Lorenzo, V. and A. Danchin (2008). "Synthetic biology: discovering new worlds and new words. The new and not so new aspects of this emerging research field." EMBO Reports **aop**(current).

del Solar, G., R. Giraldo, et al. (1998). "Replication and control of circular bacterial plasmids." Microbiol Mol Biol Rev **62**(2): 434-64.

Demarre, G., C. Frumerie, et al. (2007). "Identification of key structural determinants of the IntI1 integron integrase that influence attC x attI1 recombination efficiency." Nucleic Acids Res **35**(19): 6475-89.

Densmore, D., T. H. Hsiau, et al. (2010). "Algorithms for automated DNA assembly." Nucleic Acids Res **38**(8): 2607-16.

Dillingham, M. S. and S. C. Kowalczykowski (2008). "RecBCD enzyme and the repair of double-stranded DNA breaks." Microbiology and molecular biology reviews: MMBR **72**(4): 642.

Dubnau, D. (1999). "DNA uptake in bacteria." Annu Rev Microbiol **53**: 217-44.

Dutra, B. E. and S. T. Lovett (2006). "Cis and trans-acting effects on a mutational hotspot involving a replication template switch." J Mol Biol **356**(2): 300-11.

Ellis, T., X. Wang, et al. (2009). "Diversity-based, model-guided construction of synthetic gene networks with predicted functions." Nat Biotech **27**(5): 465.

Elowitz, M. B. and S. Leibler (2000). "A synthetic oscillatory network of transcriptional regulators." Nature **403**(6767): 335.

Endy, D. (2005). "Foundations for engineering biology." Nature **438**(7067): 449.

Fluit, A. C. and F. J. Schmitz (2004). "Resistance integrons and super-integrons." Clin Microbiol Infect **10**(4): 272-88.

Francia, M. V., P. Avila, et al. (1997). "A hot spot in plasmid F for site-specific recombination mediated by Tn21 integron integrase." J Bacteriol **179**(13): 4419-25.

Francia, M. V., F. de la Cruz, et al. (1993). "Secondary-sites for integration mediated by the Tn21 integrase." Molecular Microbiology **10**(4): 823-8.

Francia, M. V. and J. M. Garcia Lobo (1996). "Gene integration in the Escherichia coli chromosome mediated by Tn21 integrase (Int21)." J Bacteriol **178**(3): 894-8.

Francia, M. V., A. Varsaki, et al. (2004). "A classification scheme for mobilization regions of bacterial plasmids." FEMS microbiology reviews **28**(1): 79.

Francia, M. V., J. C. Zabala, et al. (1999). "The IntI1 integron integrase preferentially binds single-stranded DNA of the *attC* site." Journal of Bacteriology **181**: 6844-6849.

Frank, R., W. Heikens, et al. (1983). "A new general approach for the simultaneous chemical synthesis of large numbers of oligonucleotides: segmental solid supports." Nucleic Acids Res **11**(13): 4365-77.

Frumerie, C., M. Ducos-Galand, et al. (2010). "The relaxed requirements of the integron cleavage site allow predictable changes in integron target specificity." Nucleic Acids Res **38**(2): 559-69.

Furlong, J. C., K. M. Sullivan, et al. (1989). "Localized chemical hyperreactivity in supercoiled DNA: evidence for base unpairing in sequences that induce low-salt cruciform extrusion." Biochemistry **28**(5): 2009.

Galkin, V. E., X. Yu, et al. (2009). "Cleavage of bacteriophage lambda cI repressor involves the RecA C-terminal domain." J Mol Biol **385**(3): 779-87.

Gamper, H. B. and J. E. Hearst (1982). "A topological model for transcription based on unwinding angle analysis of E. coli RNA polymerase binary, initiation and ternary complexes." Cell **29**(1): 81-90.

Gardner, T. S., C. R. Cantor, et al. (2000). "Construction of a genetic toggle switch in Escherichia coli." Nature **403**(6767): 339.

Garriss, G., M. K. Waldor, et al. (2009). "Mobile antibiotic resistance encoding elements promote their own diversity." PLoS Genet **5**(12): e1000775.

Gellert, M. and H. Nash (1987). "Communication between segments of DNA during site-specific recombination." Nature **325**(6103): 401-4.

Gertz, J., E. D. Siggia, et al. (2009). "Analysis of combinatorial cis-regulation in synthetic and genomic promoters." Nature **457**(7226): 215.

Gibson, D. G., G. A. Benders, et al. (2008). "One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome." Proc Natl Acad Sci U S A **105**(51): 20404-9.

Gilbert, W. (1978). "Why genes in pieces?" Nature **271**(5645): 501.

Gillings, M., Y. Boucher, et al. (2008). "The evolution of class 1 integrons and the rise of antibiotic resistance." J Bacteriol **190**(14): 5095-100.

Gillings, M. R., M. P. Holley, et al. (2005). "Integrons in Xanthomonas: a source of species genome diversity." Proc Natl Acad Sci U S A **102**(12): 4419-24.

Gleghorn, M. L., E. K. Davydova, et al. (2008). "Structural basis for DNA-hairpin promoter recognition by the bacteriophage N4 virion RNA polymerase." Molecular cell **32**(5): 707.

Goeddel, D. V., D. G. Kleid, et al. (1979). "Expression in Escherichia coli of chemically synthesized genes for human insulin." Proc Natl Acad Sci U S A **76**(1): 106-10.

Gonzalez-Perez, B., M. a. Lucas, et al. (2007). "Analysis of DNA processing reactions in bacterial conjugation by using suicide oligonucleotides." The EMBO journal **26**(16): 3847.

Gravel, A., N. Messier, et al. (1998). "Point mutations in the integron integrase IntI1 that affect recombination and/or substrate recognition." Journal of Bacteriology **180**(20): 5437-5442.

Grindley, N. D., K. L. Whiteson, et al. (2006). "Mechanisms of site-specific recombination." Annu Rev Biochem **75**: 567-605.

Grunberg, R. and L. Serrano (2010). "Strategies for protein synthetic biology." Nucleic Acids Res **38**(8): 2663-75.

Guasch, A., M. Lucas, et al. (2003). "Recognition and processing of the origin of transfer DNA by conjugative relaxase TrwC." Nat Struct Biol **10**(12): 1002-10.

Gueguen, E., P. Rousseau, et al. (2005). "The transpososome: control of transposition at the level of catalysis." Trends Microbiol **13**(11): 543-9.

Guerin, E., G. Cambray, et al. (2009). "The SOS response controls integron recombination." Science **324**(5930): 1034.

Guet, C. C., M. B. Elowitz, et al. (2002). "Combinatorial synthesis of genetic networks." Science (New York, N.Y.) **296**(5572): 1466.

Guynet, C., A. Achard, et al. (2009). "Resetting the site: redirecting integration of an insertion sequence in a predictable way." Molecular cell **34**(5): 612.

Guynet, C., A. B. Hickman, et al. (2008). "In vitro reconstitution of a single-stranded transposition mechanism of IS608." Mol Cell **29**(3): 302-12.

Hall, R. M., D. E. Brookes, et al. (1991). "Site-specific insertion of genes into integrons: role of the 59-base element and determination of the recombination cross-over point." Mol Microbiol **5**(8): 1941-59.

Hansson, K., O. Skold, et al. (1997). "Non-palindromic attI sites of integrons are capable of site-specific recombination with one another and with secondary targets." Molecular Microbiology **26**(3): 441-53.

Hansson, K., L. Sundstrom, et al. (2002). "IntI2 integron integrase in Tn7." J Bacteriol **184**(6): 1712-21.

Hatfield, G. W. and C. J. Benham (2002). "DNA topology-mediated control of global gene expression in Escherichia coli." Annual review of genetics **36**: 175.

Heidelberg, J. F., J. A. Eisen, et al. (2000). "DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*." Nature **406**: 477 - 483.

Heller, R. C. and K. J. Marians (2006). "Replication fork reactivation downstream of a blocked nascent leading strand." Nature **439**(7076): 557.

Higashitani, A., N. Higashitani, et al. (1997). "Minus-strand origin of filamentous phage versus transcriptional promoters in recognition of RNA polymerase." Proceedings of the National Academy of Sciences of the United States of America **94**(7): 2909.

Higashitani, N., A. Higashitani, et al. (1996). "Recognition mechanisms of the minus-strand origin of phage f1 by Escherichia coli RNA polymerase." Genes to cells: devoted to molecular & cellular mechanisms **1**(9): 829.

Higashitani, N., A. Higashitani, et al. (1992). "SOS induction in Escherichia coli by infection with mutant filamentous phage that are defective in initiation of complementary-strand DNA synthesis." J Bacteriol **174**(5): 1612-8.

Hiraga, K. and F. H. Arnold (2003). "General method for sequence-independent site-directed chimeragenesis." J Mol Biol **330**(2): 287-96.

Hirose, S. and K. Matsumoto (2005). Possible roles of DNA supercoiling in transcription. DNA conformation and transcription. T. Ohyama, Springer. **XII:** 138-143.

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.

Hochhut, B., J. Marrero, et al. (2000). "Mobilization of plasmids and chromosomal DNA mediated by the SXT element, a constin found in Vibrio cholerae O139." Journal of bacteriology **182**(7): 2043.

Honda, Y., H. Sakai, et al. (1991). "Functional division and reconstruction of a plasmid replication origin: molecular dissection of the oriV of the broad-host-range plasmid RSF1010." Proceedings of the National Academy of Sciences of the United States of America **88**(1): 179.

Honda, Y., H. Sakai, et al. (1988). "Two single-strand DNA initiation signals located in the oriV region of plasmid RSF1010." Gene **68**(2): 221.

Honda, Y., H. Sakai, et al. (1989). "RepB' is required in trans for the two single-strand DNA initiation signals in oriV of plasmid RSF1010." Gene **80**(1): 155.

Horton, R. M., H. D. Hunt, et al. (1989). "Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension." Gene **77**(1): 61-8.

Horwitz, M. S. and L. A. Loeb (1988). "An E. coli promoter that regulates transcription by DNA superhelix-induced cruciform extrusion." Science (New York, N.Y.) **241**(4866): 703.

Horwitz, M. S. and L. A. Loeb (1988). "An E. coli promoter that regulates transcription by DNA superhelix-induced cruciform extrusion." Science **241**(4866): 703-5.

Huang, X., F. J., et al. (1997). "sigma factor mutations affecting the sequence-selective interaction of RNA polymerase with -10 region single-stranded DNA." Nucleic acids research **25**(13): 2603.

Ilyina, T. V. and E. V. Koonin (1992). "Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria." Nucleic Acids Res **20**(13): 3279-85.

Isaacs, F. J., D. J. Dwyer, et al. (2006). "RNA synthetic biology." Nat Biotech **24**(5): 545.

Jaworski, A., N. P. Higgins, et al. (1991). "Topoisomerase mutants and physiological conditions control supercoiling and Z-DNA formation in vivo." J Biol Chem **266**(4): 2576-81.

Jenison, R. D., S. C. Gill, et al. (1994). "High-resolution molecular discrimination by RNA." Science **263**(5152): 1425-9.

Johansson, C., L. Boukharta, et al. (2009). "Mutagenesis and Homology Modeling of the Tn21 Integron Integrase IntI1." Biochemistry **48**(8): 1743–1753.

Johansson, C., M. Kamali-Moghaddam, et al. (2004). "Integron integrase binds to bulged hairpin DNA." Nucleic Acids Res **32**(13): 4033-43.

Jones, A. L., P. T. Barth, et al. (1992). "Zygotic induction of plasmid ssb and psiB genes following conjugative transfer of Incl1 plasmid Collb-P9." Molecular microbiology **6**(5): 605.

Jove, T., S. Da Re, et al. (2010). "Inverse correlation between promoter strength and excision activity in class 1 integrons." PLoS Genet **6**(1): e1000793.

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.

Kaguni, J. M. and A. Kornberg (1982). "The rho subunit of RNA polymerase holoenzyme confers specificity in priming M13 viral DNA replication." The Journal of biological chemistry **257**(10): 5437.

Katayama, T., S. Ozaki, et al. "Regulation of the replication cycle: conserved and diverse regulatory systems for DnaA and oriC." Nat Rev Microbiol **8**(3): 163-70.

Keasling, J. D. (2008). "Synthetic biology for synthetic chemistry." ACS chemical biology **3**(1): 64.

Kelley, W. L. (2006). "Lex marks the spot: the virulent side of SOS and a closer look at the LexA regulon." Mol Microbiol **62**(5): 1228-38.

Kersulyte, D., B. Velapatino, et al. (2002). "Transposable element ISHp608 of Helicobacter pylori: nonrandom geographic distribution, functional organization, and insertion specificity." J Bacteriol **184**(4): 992-1002.

Khalil, A. S. and J. J. Collins (2010). "Synthetic biology: applications come of age." Nature reviews. Genetics **11**(5): 367.

Khan, S. A. (2005). "Plasmid rolling-circle replication: highlights of two decades of research." Plasmid **53**(2): 126-36.

Khosla, C., S. Kapur, et al. (2009). "Revisiting the modularity of modular polyketide synthases." Curr Opin Chem Biol **13**(2): 135-43.

Kirby, J. and J. D. Keasling (2009). "Biosynthesis of plant isoprenoids: perspectives for microbial engineering." Annu Rev Plant Biol **60**: 335-55.

Kitamura, K., Y. Kinoshita, et al. (2002). "Construction of block-shuffled libraries of DNA for evolutionary protein engineering: Y-ligation-based block shuffling." Protein Eng **15**(10): 843-53.

Koenig, J. E., Y. Boucher, et al. (2008). "Integron-associated gene cassettes in Halifax Harbour: assessment of a mobile gene pool in marine sediments." Environ Microbiol **10**(4): 1024-38.

Koepsel, R. R. and S. A. Khan (1987). "Cleavage of single-stranded DNA by plasmid pT181-encoded RepC protein." Nucleic acids research **15**(10): 4085.

Kolkman, J. A. and W. P. Stemmer (2001). "Directed evolution of proteins by exon shuffling." Nature biotechnology **19**(5): 423.

Konieczny, I. (2003). "Strategies for helicase recruitment and loading in bacteria." EMBO Rep **4**(1): 37-41.

Kornberg, A. and T. A. Baker (1992). DNA Replication, W.H. Freeman & Company.

Kowalczykowski, S. C. (1994). "In vitro reconstitution of homologous recombination reactions." Experientia **50**(3): 204.

Kowalczykowski, S. C., D. A. Dixon, et al. (1994). "Biochemistry of homologous recombination in Escherichia coli." Microbiol Rev **58**(3): 401-65.

Kowalczykowski, S. C. and R. A. Krupp (1987). "Effects of Escherichia coli SSB protein on the single-stranded DNA-dependent ATPase activity of Escherichia coli RecA protein. Evidence that SSB protein facilitates the binding of RecA protein to regions of secondary structure within single-stranded DNA." J Mol Biol **193**(1): 97-113.

Kramer, M. G., M. Espinosa, et al. (1998). "Lagging strand replication of rolling-circle plasmids: specific recognition of the ssoA-type origins in different gram-positive bacteria." Proceedings of the National Academy of Sciences of the United States of America **95**(18): 10505.

Kramer, M. G., M. Espinosa, et al. (1999). "Characterization of a single-strand origin, ssoU, required for broad host range replication of rolling-circle plasmids." Molecular microbiology **33**(3): 466.

Kramer, M. G., S. A. Khan, et al. (1997). "Plasmid rolling circle replication: identification of the RNA polymerase-directed primer RNA and requirement for DNA polymerase I for lagging strand synthesis." The EMBO journal **16**(18): 5784.

Kramer, M. G., S. A. Khan, et al. (1998). "Lagging-strand replication from the ssoA origin of plasmid pMV158 in Streptococcus pneumoniae: in vivo and in vitro influences of mutations in two conserved ssoA regions." Journal of bacteriology **180**(1): 83.

Kreuzer, K. N. (2005). "Interplay between DNA replication and recombination in prokaryotes." Annual review of microbiology **59**: 43.

Kurahashi, H., H. Inagaki, et al. (2006). "Chromosomal translocations mediated by palindromic DNA." Cell Cycle **5**(12): 1297-303.

Labbate, M., R. J. Case, et al. (2009). "The integron/gene cassette system: an active player in bacterial adaptation." Methods Mol Biol **532**: 103-25.

Lambert, P. F., D. A. Waring, et al. (1986). "DNA requirements at the bacteriophage G4 origin of complementary-strand DNA synthesis." J. Virol. **58**(2): 450.

Lampson, B. C., M. Inouye, et al. (2005). "Retrons, msDNA, and the bacterial genome." Cytogenetic and genome research **110**(1-4): 491.

Langston, L. D. and M. O'Donnell (2006). "DNA replication: keep moving and don't mind the gap." Mol Cell **23**(2): 155-60.

Leach, D. R. (1994). "Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair." Bioessays **16**(12): 893-900.

Lee, S. Y., H. U. Kim, et al. (2009). "Metabolic engineering of microorganisms: general strategies and drug production." Drug Discov Today **14**(1-2): 78-88.

Lerner, C. G. and M. Inouye (1994). "Localized random polymerase chain reaction mutagenesis." Methods Mol Biol **31**: 97-112.

Levesque, C., S. Brassard, et al. (1994). "Diversity and relative strength of tandem promoters for the antibiotic-resistance genes of several integrons." Gene **142**(1): 49-54.

Li, M. Z. and S. J. Elledge (2007). "Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC." Nature Methods **4**(3): 251.

Lilley, D. M. (1985). "The kinetic properties of cruciform extrusion are determined by DNA base-sequence." Nucleic Acids Res **13**(5): 1443-65.

Liu, L. F. and J. C. Wang (1987). "Supercoiling of the DNA template during transcription." Proc Natl Acad Sci U S A **84**(20): 7024-7.

Liu, Y., V. Bondarenko, et al. (2001). "DNA supercoiling allows enhancer action over a large distance." Proc Natl Acad Sci U S A **98**(26): 14883-8.

Llosa, M., S. Bolland, et al. (1991). "Structural and functional analysis of the origin of conjugal transfer of the broad-host-range IncW plasmid R388 and comparison with the related IncN plasmid R46." Mol Gen Genet **226**(3): 473-83.

Lou, C., X. Liu, et al. "Synthesizing a novel genetic sequential logic circuit: a push-on push-off switch." Mol Syst Biol **6**: 350.

Low, K. B. (1972). "Escherichia coli K-12 F-prime factors, old and new." Bacteriological reviews **36**(4): 587.

Lynch, S. A. and J. P. Gallivan (2009). "A flow cytometry-based screen for synthetic riboswitches." Nucleic Acids Res **37**(1): 184-92.

MacDonald, D., G. Demarre, et al. (2006). "Structural basis for broad DNA specificity in integron recombination." Nature **440**: 1157-1162.

Majchrzak, M., R. P. Bowater, et al. (2006). "SOS repair and DNA supercoiling influence the genetic stability of DNA triplet repeats in Escherichia coli." J Mol Biol **364**(4): 612-24.

Marquez, C., M. Labbate, et al. (2008). "Recovery of a functional class 2 integron from an Escherichia coli strain mediating a urinary tract infection." Antimicrob Agents Chemother **52**(11): 4153-4.

Martin, C., J. Timm, et al. (1990). "Transposition of an antibiotic resistance element in mycobacteria." Nature **345**(6277): 739-743.

Martin, V. J., D. J. Pitera, et al. (2003). "Engineering a mevalonate pathway in Escherichia coli for production of terpenoids." Nat Biotechnol **21**(7): 796-802.

Martinez, E. and F. de la Cruz (1988). "Transposon Tn21 encodes a RecA-independant site-specific integration system." Molecular and General Genetics **211**: 320-325.

Martinez, E. and F. de la Cruz (1990). "Genetic elements involved in Tn*21* site-specific integration, a novel mechanism for the dissemination of antibiotic resistance genes." EMBO Journal **9**: 1275-1281.

Masai, H. and K. Arai (1996). "DnaA- and PriA-dependent primosomes: two distinct replication complexes for replication of Escherichia coli chromosome." Frontiers in bioscience: a journal and virtual library **1**: d48.

Masai, H. and K. Arai (1997). "Frpo: a novel single-stranded DNA promoter for transcription and for primer RNA synthesis of DNA replication." Cell **89**(6): 897.

Masai, H., N. Nomura, et al. (1990). "The ABC-primosome. A novel priming system employing dnaA, dnaB, dnaC, and primase on a hairpin containing a dnaA box sequence." J Biol Chem **265**(25): 15134-44.

Mazel, D. (2006). "Integrons: agents of bacterial evolution." Nat Rev Microbiol **4**(8): 608-20.

Mazel, D., B. Dychinco, et al. (1998). "A distinctive class of integron in the *Vibrio cholerae* genome." Science **280**(5363): 605-608.

McGlynn, P., A. A. Al-Deib, et al. (1997). "The DNA replication protein PriA and the recombination protein RecG bind D-loops." Journal of molecular biology **270**(2): 212.

Mejean, V. and J. P. Claverys (1984). "Use of a cloned DNA fragment to analyze the fate of donor DNA in transformation of Streptococcus pneumoniae." J Bacteriol **158**(3): 1175-8.

Mendiola, M. V., I. Bernales, et al. (1994). "Differential roles of the transposon termini in IS91 transposition." Proc Natl Acad Sci U S A **91**(5): 1922-6.

Messier, N. and P. H. Roy (2001). "Integron integrases possess a unique additional domain necessary for activity." J Bacteriol **183**(22): 6699-706.

Meyer, M. M., L. Hochrein, et al. (2006). "Structure-guided SCHEMA recombination of distantly related beta-lactamases." Protein Eng Des Sel **19**(12): 563-70.

Meyer, M. M., J. J. Silberg, et al. (2003). "Library analysis of SCHEMA-guided protein recombination." Protein Sci **12**(8): 1686-93.

Miao, D. M., Y. Honda, et al. (1993). "A base-paired hairpin structure essential for the functional priming signal for DNA replication of the broad host range plasmid RSF1010." Nucleic Acids Res **21**(21): 4900-3.

Mironov, A. S., I. Gusarov, et al. (2002). "Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria." Cell **111**(5): 747-56.

174

Monzingo, A. F., A. Ozburn, et al. (2007). "The structure of the minimal relaxase domain of MobA at 2.1 A resolution." <u>Journal of molecular biology</u> **366**(1): 165.

Mortier-Barriere, I., M. Velten, et al. (2007). "A key presynaptic role in transformation for a widespread bacterial protein: DprA conveys incoming ssDNA to RecA." <u>Cell</u> **130**(5): 824-36.

Mott, M. L. and J. M. Berger (2007). "DNA replication initiation: mechanisms and regulation in bacteria." <u>Nature reviews. Microbiology</u> **5**(5): 343.

Murakami, H., T. Hohsaka, et al. (2002). "Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs." <u>Nat Biotechnol</u> **20**(1): 76-81.

Murchie, A. I. and D. M. Lilley (1987). "The mechanism of cruciform formation in supercoiled DNA: initial opening of central basepairs in salt-dependent extrusion." <u>Nucleic Acids Res</u> **15**(23): 9641-54.

Nandi, S., J. J. Maurer, et al. (2004). "Gram-positive bacteria are a major reservoir of Class 1 antibiotic resistance integrons in poultry litter." <u>Proc Natl Acad Sci U S A</u> **101**(18): 7118-22.

Nelson, T. C. (1951). "Kinetics of genetic recombination in Escherichia coli." <u>Genetics</u> **36**(2): 162.

Nemergut, D. R., M. S. Robeson, et al. (2008). "Insights and inferences about integron evolution from genomic data." <u>BMC Genomics</u> **9**: 261.

Nesvera, J., J. Hochmannova, et al. (1998). "An integron of class 1 is present on the plasmid pCG4 from gram-positive bacterium *Corynebacterium glutamicum*." <u>FEMS Microbiology Letters</u> **169**(2): 391-395.

Neumann, H., K. Wang, et al. (2010). "Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome." <u>Nature</u> **464**(7287): 441-4.

Noirot, P., J. Bargonetti, et al. (1990). "Initiation of rolling-circle replication in pT181 plasmid: initiator protein enhances cruciform extrusion at the origin." <u>Proc Natl Acad Sci U S A</u> **87**(21): 8560-4.

Noirot, P., J. Bargonetti, et al. (1990). "Initiation of rolling-circle replication in pT181 plasmid: initiator protein enhances cruciform extrusion at the origin." <u>Proceedings of the National Academy of Sciences of the United States of America</u> **87**(21): 8560.

Nomura, N., H. Masai, et al. (1991). "Identification of eleven single-strand initiation sequences (ssi) for priming of DNA replication in the F, R6K, R100 and ColE2 plasmids." <u>Gene</u> **108**(1): 15.

Nunes-Duby, S. E., H. J. Kwon, et al. (1998). "Similarities and differences among 105 members of the Int family of site-specific recombinases." <u>Nucleic Acids Research</u> **26**(2): 391-406.

Ostermeier, M., J. H. Shim, et al. (1999). "A combinatorial approach to hybrid enzymes independent of DNA homology." <u>Nat Biotechnol</u> **17**(12): 1205-9.

Oussatcheva, E. A., J. Pavlicek, et al. (2004). "Influence of global DNA topology on cruciform formation in supercoiled DNA." <u>J Mol Biol</u> **338**(4): 735-43.

Pages, V. and R. P. Fuchs (2003). "Uncoupling of leading- and lagging-strand DNA replication during lesion bypass in vivo." <u>Science</u> **300**(5623): 1300-3.

Panayotatos, N. and R. D. Wells (1981). "Cruciform structures in supercoiled DNA." <u>Nature</u> **289**(5797): 466-70.

Panchenko, A. R., Z. Luthey-Schulten, et al. (1996). "Foldons, protein structural modules, and exons." <u>Proc Natl Acad Sci U S A</u> **93**(5): 2008-13.

Pandey, D. P. and K. Gerdes (2005). "Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes." Nucleic Acids Res **33**(3): 966-76.

Park, J. H., K. H. Lee, et al. (2007). "Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation." Proc Natl Acad Sci U S A **104**(19): 7797-802.

Park, S. H., A. Zarrinpar, et al. (2003). "Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms." Science **299**(5609): 1061-4.

Partridge, S. R., G. Tsafnat, et al. (2009). "Gene cassettes and cassette arrays in mobile resistance integrons." FEMS Microbiol Rev **33**(4): 757-84.

Pearson, C. E., H. Zorbas, et al. (1996). "Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication." J Cell Biochem **63**(1): 1-22.

Peisajovich, S. G. and D. S. Tawfik (2007). "Protein engineers turned evolutionists." Nat Meth **4**(12): 991.

Pfleger, B. F., D. J. Pitera, et al. (2006). "Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes." Nature biotechnology **24**(8): 1027.

Pitera, D. J., C. J. Paddon, et al. (2007). "Balancing a heterologous mevalonate pathway for improved isoprenoid production in Escherichia coli." Metab Eng **9**(2): 193-207.

Platt, J. R. (1955). "POSSIBLE SEPARATION OF INTERTWINED NUCLEIC ACID CHAINS BY TRANSFER-TWIST." Proceedings of the National Academy of Sciences of the United States of America **41**(3): 181.

Pruss, G. J. and K. Drlica (1989). "DNA supercoiling and prokaryotic transcription." Cell **56**(4): 521-3.

Rackham, O. and J. W. Chin (2005). "A network of orthogonal ribosome x mRNA pairs." Nat Chem Biol **1**(3): 159-66.

Rackham, O. and J. W. Chin (2006). "Synthesizing cellular networks from evolved ribosome-mRNA pairs." Biochem Soc Trans **34**(Pt 2): 328-9.

Ramirez, M. S., S. Pineiro, et al. (2010). "Novel insights about class 2 integrons from experimental and genomic epidemiology." Antimicrob Agents Chemother **54**(2): 699-706.

Recchia, G. D. and R. M. Hall (1995). "Gene cassettes: a new class of mobile element." Microbiology **141**: 3015-3027.

Recchia, G. D. and R. M. Hall (1995). "Plasmid evolution by acquisition of mobile gene cassettes: plasmid pIE723 contains the aadB gene cassette precisely inserted at a secondary site in the incQ plasmid RSF1010." Mol Microbiol **15**(1): 179-87.

Recchia, G. D., H. W. Stokes, et al. (1994). "Characterisation of specific and secondary recombination sites recognised by the integron DNA integrase." Nucleic Acids Res **22**(11): 2071-8.

Reddy, M. S., M. B. Vaze, et al. (2000). "Binding of SSB and RecA protein to DNA-containing stem loop structures: SSB ensures the polarity of RecA polymerization on single-stranded DNA." Biochemistry **39**(46): 14250-62.

Ronning, D. R., C. Guynet, et al. (2005). "Active site sharing and subterminal hairpin recognition in a new class of DNA transposases." Mol Cell **20**(1): 143-54.

Rowe-Magnus, D. A., A.-M. Guerout, et al. (2001). "The evolutionary history of chromosomal super-integrons provides an ancestry for multi-resitant integrons." Proceedings of the National Academy of Sciences of the United States of America **98**: 652-657.

Rowe-Magnus, D. A., A. M. Guerout, et al. (2003). "Comparative analysis of superintegrons: engineering extensive genetic diversity in the vibrionaceae." Genome Res **13**(3): 428-42.

Rowe-Magnus, D. A., A. M. Guerout, et al. (2002). "Bacterial resistance evolution by recruitment of super-integron gene cassettes." Mol Microbiol **43**(6): 1657-69.

Rowe-Magnus, D. A., A. M. Guerout, et al. (2001). "The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons." Proc Natl Acad Sci U S A **98**(2): 652-7.

Roy, R., A. G. Kozlov, et al. (2009). "SSB protein diffusion on single-stranded DNA stimulates RecA filament formation." Nature **461**(7267): 1092-7.

Russel, M., N. A. Linderoth, et al. (1997). "Filamentous phage assembly: variation on a protein export theme." Gene **192**(1): 23.

Russel, M. and P. Model (1989). "Genetic analysis of the filamentous bacteriophage packaging signal and of the proteins that interact with it." Journal of virology **63**(8): 3284.

Ruvkun, G. B. and F. M. Ausubel (1981). "A general method for site-directed mutagenesis in prokaryotes." Nature **289**(5793): 85-8.

Saiki, R. K., D. H. Gelfand, et al. (1988). "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase." Science **239**(4839): 487-91.

Saito, H., T. Minamisawa, et al. (2007). "Motif programming: a microgene-based method for creating synthetic proteins containing multiple functional motifs." Nucleic Acids Res **35**(6): e38.

Schleif, R. (2000). "Regulation of the L-arabinose operon of Escherichia coli." Trends Genet **16**(12): 559-65.

Schmidt-Dannert, C., D. Umeno, et al. (2000). "Molecular breeding of carotenoid biosynthetic pathways." Nat Biotechnol **18**(7): 750-3.

Schmidt, M. (2009). Synthetic Biology. Dordrecht, Springer Netherlands.

Schvartzman, J. B. and A. Stasiak (2004). "A topological view of the replicon." EMBO Rep **5**(3): 256-61.

Segal, H. and B. G. Elisha (1997). "Identification and characterization of an aadB gene cassette at a secondary site in a plasmid from Acinetobacter." FEMS Microbiol Lett **153**(2): 321-6.

Serrano, L. (2007). "Synthetic biology: promises and challenges." Mol Syst Biol **3**.

Shao, Z., H. Zhao, et al. (2009). "DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways." Nucleic acids research **37**(2): e16.

Shi, L., M. Zheng, et al. (2006). "Unnoticed spread of class 1 integrons in gram-positive clinical strains isolated in Guangzhou, China." Microbiol Immunol **50**(6): 463-7.

Shlyakhtenko, L. S., P. Hsieh, et al. (2000). "A cruciform structural transition provides a molecular switch for chromosome structure and dynamics." Journal of molecular biology **296**(5): 1169.

Shortle, D. (1983). "A genetic system for analysis of staphylococcal nuclease." Gene **22**(2-3): 181-9.

Sinden, R. R., S. S. Broyles, et al. (1983). "Perfect palindromic lac operator DNA sequence exists as a stable cruciform structure in supercoiled DNA in vitro but not in vivo." Proceedings of the National Academy of Sciences of the United States of America **80**(7): 1797.

Sinden, R. R., G. X. Zheng, et al. (1991). "On the deletion of inverted repeated DNA in Escherichia coli: effects of length, thermal stability, and cruciform formation in vivo." Genetics **129**(4): 991-1005.

Singh, J., M. Mukerji, et al. (1995). "Transcriptional activation of the Escherichia coli bgl operon: negative regulation by DNA structural elements near the promoter." Molecular microbiology **17**(6): 1085.

Sorum, H., M. C. Roberts, et al. (1992). "Identification and cloning of a tetracycline resistance gene from the fish pathogen Vibrio salmonicida." Antimicrob Agents Chemother **36**(3): 611-5.

Stark, B. C., R. Kole, et al. (1978). "Ribonuclease P: an enzyme with an essential RNA component." Proc Natl Acad Sci U S A **75**(8): 3717-21.

Stefan, A., A. Radeghieri, et al. (2001). "Directed evolution of beta-galactosidase from Escherichia coli by mutator strains defective in the 3'-->5' exonuclease activity of DNA polymerase III." FEBS Lett **493**(2-3): 139-43.

Stemmer, W. P. (1994). "Rapid evolution of a protein in vitro by DNA shuffling." Nature **370**(6488): 389-91.

Stokes, H. W. and R. M. Hall (1989). "A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons." Molecular Microbiology **3**(12): 1669-1683.

Stokes, H. W., C. L. Nesbo, et al. (2006). "Class 1 integrons potentially predating the association with tn402-like transposition genes are present in a sediment microbial community." J Bacteriol **188**(16): 5722-30.

Stoltenburg, R., C. Reinemann, et al. (2007). "SELEXâ€"A (r)evolutionary method to generate high-affinity nucleic acid ligands." Biomolecular Engineering **24**(4): 381.

Sudarsan, N., M. C. Hammond, et al. (2006). "Tandem riboswitch architectures exhibit complex gene control functions." Science **314**(5797): 300-4.

Sullivan, K. M. and D. M. Lilley (1987). "Influence of cation size and charge on the extrusion of a salt-dependent cruciform." J Mol Biol **193**(2): 397-404.

Sun, W. and G. N. Godson (1998). "Structure of the Escherichia coli primase/single-strand DNA-binding protein/phage G4oric complex required for primer RNA synthesis." Journal of molecular biology **276**(4): 689.

Swart, J. R. and M. A. Griep (1993). "Primase from Escherichia coli primes single-stranded templates in the absence of single-stranded DNA-binding protein or other auxiliary proteins. Template sequence requirements based on the bacteriophage G4 complementary strand origin and Okazaki fragment." The Journal of biological chemistry **268**(17): 12970.

Tanaka, T., T. Mizukoshi, et al. (2007). "Escherichia coli PriA protein, two modes of DNA binding and activation of ATP hydrolysis." The Journal of biological chemistry **282**(27): 19917.

Tatum, E. L. and J. Lederberg (1947). "Gene Recombination in the Bacterium Escherichia coli." Journal of bacteriology **53**(6): 673.

The, M. J. (1989). "Human insulin: DNA technology's first drug." Am J Hosp Pharm **46**(11 Suppl 2): S9-11.

Thomas, R. (1973). "Boolean formalization of genetic control circuits." J Theor Biol **42**(3): 563-85.

Tokuriki, N. and D. S. Tawfik (2009). "Protein Dynamism and Evolvability." Science **324**(5924): 203.

Topp, S. and J. P. Gallivan (2010). "Emerging applications of riboswitches in chemical biology." ACS chemical biology **5**(1): 139.

Trinh, T. Q. and R. R. Sinden (1991). "Preferential DNA secondary structure mutagenesis in the lagging strand of replication in E. coli." Nature **352**(6335): 544-7.

Tsuji, T., M. Onimaru, et al. (2001). "Random multi-recombinant PCR for the construction of combinatorial protein libraries." Nucleic Acids Res **29**(20): E97.

Tuerk, C. and L. Gold (1990). "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase." Science **249**(4968): 505-10.

Vaisvila, R., R. Morgan, et al. (1999). "The pacIR gene resides within a potential super-integron in the *Pseudomonas alcaligenes* genome." IXth international congress of bacteriology and applied microbiology, Sidney, 1999.

Vaisvila, R., R. D. Morgan, et al. (2001). "Discovery and distribution of super-integrons among Pseudomonads." Molecular Microbiology **42**: 587-601.

Val, M.-E., M. Bouvier, et al. (2005). "The single-stranded genome of phage CTX is the form used for integration into the genome of Vibrio cholerae." Molecular cell **19**(4): 559.

Val, M. E., M. Bouvier, et al. (2005). "The single-stranded genome of phage CTX is the form used for integration into the genome of Vibrio cholerae." Mol Cell **19**(4): 559-66.

Voigt, C. A., C. Martinez, et al. (2002). "Protein building blocks preserved by recombination." Nat Struct Biol **9**(7): 553-8.

Walker, G. C. (1996). "The SOS Response of Escherichia coli. Escherichia coli and Salmonella." Neidhardt, FC Washington DC American Society of Microbiology **1**: 1400-1416.

Walsh, C. T. (2004). "Polyketide and nonribosomal peptide antibiotics: modularity and versatility." Science **303**(5665): 1805-10.

Wang, H. H., F. J. Isaacs, et al. (2009). "Programming cells by multiplex genome engineering and accelerated evolution." Nature **460**(7257): 894.

Wang, H. H., F. J. Isaacs, et al. (2009). "Programming cells by multiplex genome engineering and accelerated evolution." Nature **460**(7257): 894-8.

Wang, J. C. and A. S. Lynch (1993). "Transcription and DNA supercoiling." Current opinion in genetics & development **3**(5): 764.

Wells, J. A., M. Vasser, et al. (1985). "Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites." Gene **34**(2-3): 315-23.

White, J. H. and W. R. Bauer (1987). "Superhelical DNA with local substructures. A generalization of the topological constraint in terms of the intersection number and the ladder-like correspondence surface." Journal of molecular biology **195**(1): 205.

Wickner, S. and J. Hurwitz (1975). "Association of phi X174 DNA-Dependent ATPase Activity with an Escherichia coli Protein, Replication Factor Y, Required for in vitro Synthesis of phi X174 DNA." Proceedings of the National Academy of Sciences **72**(9): 3342.

Wieland, M. and J. S. Hartig (2008). "Improved aptazyme design and in vivo screening enable riboswitching in bacteria." Angew Chem Int Ed Engl **47**(14): 2604-7.

Win, M. N. and C. D. Smolke (2008). "Higher-order cellular information processing with synthetic RNA devices." Science **322**(5900): 456-60.

Winkler, W., A. Nahvi, et al. (2002). "Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression." Nature **419**(6910): 952-6.

Wozniak, R. A., D. E. Fouts, et al. (2009). "Comparative ICE genomics: insights into the evolution of the SXT/R391 family of ICEs." PLoS Genet **5**(12): e1000786.

Xu, H., J. Davies, et al. (2007). "Molecular characterization of class 3 integrons from Delftia spp." <u>J Bacteriol</u>.

Yadav, V. G. and G. Stephanopoulos "Reevaluating synthesis by biology." <u>Curr Opin Microbiol</u> **13**(3): 371-6.

Yeh, B. J., R. J. Rutigliano, et al. (2007). "Rewiring cellular morphology pathways with synthetic guanine nucleotide exchange factors." <u>Nature</u>.

Yokobayashi, Y., R. Weiss, et al. (2002). "Directed evolution of a genetic circuit." <u>Proceedings of the National Academy of Sciences of the United States of America</u> **99**(26): 16587.

Zannis-Hadjopoulos, M., W. Yahyaoui, et al. (2008). "14-3-3 cruciform-binding proteins as regulators of eukaryotic DNA replication." <u>Trends in biochemical sciences</u> **33**(1): 44.

Zhang, Y.-X., K. Perry, et al. (2002). "Genome shuffling leads to rapid phenotypic improvement in bacteria." <u>Nature</u> **415**(6872): 644.

Zhao, H. (2007). "Directed evolution of novel protein functions." <u>Biotechnol Bioeng</u> **98**(2): 313-7.

Zhao, J., A. Bacolla, et al. (2010). "Non-B DNA structure-induced genetic instability and evolution." <u>Cell Mol Life Sci</u> **67**(1): 43-62.

Zheng, G., D. W. Ussery, et al. (1991). "Estimation of superhelical density in vivo from analysis of the level of cruciforms existing in living cells." <u>J Mol Biol</u> **221**(1): 122-9.

Zheng, G. X., T. Kochel, et al. (1991). "Torsionally tuned cruciform and Z-DNA probes for measuring unrestrained supercoiling at specific sites in DNA of living cells." <u>J Mol Biol</u> **221**(1): 107-22.

Zheng, G. X. and R. R. Sinden (1988). "Effect of base composition at the center of inverted repeated DNA sequences on cruciform transitions in DNA." <u>J Biol Chem</u> **263**(11): 5356-61.