CONTROL OF DIRECTIONALITY IN MYCOBACTERIOPHAGE L5 INTEGRASE-MEDIATED SITE-SPECIFIC RECOMBINATION

by

John August Lewis

B.S. University of Pittsburgh, 1994

Submitted to the Graduate Faculty of

Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2001

UNIVERSITY OF PITTSBURGH

FACULTY OF ARTS AND SCIENCES

This dissertation was presented

by

John August Lewis

It was defended on

October 15, 2001

and approved by

Dr. Roger Hendrix

Dr. James Pipas

Dr. Jeff Brodsky

Dr. Saleem Khan

Dissertation Advisor: Dr. Graham F. Hatfull

**CONTROL OF DIRECTIONALITY IN MYCOBACTERIOPHAGE L5
INTEGRASE-MEDIATED SITE-SPECIFIC RECOMBINATION**

John August Lewis, Ph.D.

University of Pittsburgh, 2001

Control of directionality in integrase-mediated site-specific recombination reactions is achieved by an architectural change fashioned by a class of accessory proteins know as recombination directionality factors (RDFs).   In the mycobacteriophage L5 system, no RDF had been previously identified. In the course of this work, the gene, *36*, was identified using *in vivo* screens and shown to play this role. The protein was over expressed using an *E. coli* expression system and then used to create an *in vitro* excision reaction assay. Initial work was done in characterizing the excision reaction including supercoiling and host factor requirements. Further analysis has shown that the protein binds specifically to a region within the left side of *attP* and *attR.* Once bound, complexes are formed that inhibit integration in the case of *attP* and stimulate excision when *attR* is present. Additional work was done to increase the utility of the L5 based integrating vector system, which has been commonly used to create stably integrated single copy transformants. The system lacked an effective means to recover DNA for high copy replication or curing of strains containing inserted DNA. With the identification of the L5 *xis* gene, we have been able to develop systems that allow the recovery of DNA into *E. coli* and curing strains of the integrated DNA. The third part of this work involved the characterization of the RDF class of proteins. Since this class of proteins is composed of a diverse group of small proteins, no

previous attempt had been made to characterize them. Extensive data mining yielded a collection of 63 putative or known RDFs. Further analysis of sequence data, chemical characteristics and other known properties argues that this class of proteins has evolved from multiple ancestral origins.

FORWARD

It is nearly impossible to become a successful scientist without having a great mentor. Graham Hatfull is such a leader who has a passion for science that is only outpaced by his commitment to his family. He has created a work environment that is a balance between research and an enjoyable social atmosphere that would make anyone enjoy science. Furthermore, he has taught me how to evaluate my research and to make sure that I complete the picture before I draw conclusions. I must also thank Graham for all the time he has spent helping me to refine my writing style. My committee, Roger Hendrix, Jim Pipas, Jeff Brodsky and Saleem Khan, have been extremely helpful in guiding my research and giving useful suggestion throughout my graduate career.

The people that I worked with made it enjoyable to be in the lab. In particular, I want to thank Lori Bibb and Curt Wadsworth who were always willing to listen to me babbling about science, my life or one of my obscure viewpoints. Carol Pena did much of the work of the work characterizing the L5 integration reaction that was integral in the development of my project and she also helped by establishing many of the protocols that I used in my studies. I would also like to thank Bill Jacobs, Marty Pavelka, Stoyan Bardarov at Albert Einstein College of Medicine and Tanya Parish and Neil Stoker at the London School of Hygiene and Tropical Medicine. I want to thank Gene Ney for proof reading this manuscript.

I must also thank my wife Ann, who gave up a lot to move here from NYC. She has been extremely supportive of my research and all the times I have had to stay late at the lab. She has done such a great job with our lovely baby girl, Grace, whose arrival

v

has made it a challenge to complete the revisions on this dissertation. I also want to let both of them know how much I love them.

TABLE OF CONTENTS

LIST OF TABLES

I. INTRODUCTION

I.A. Bacteriophage

Bacteriophages are viruses that are obligate parasites of bacteria cells. A phage can be characterized by its potential lifestyle choices, so there are two different classes of phages: temperate and lytic. While both types are able to appropriate the host cell's resources in order to replicate the viral DNA, form new phage particles and lyse the cell, temperate phages have a second means of existence, lysogeny (Figure 1A). In the lysogenic state, the phage genome is repressed and replicated by the host DNA replication machinery. The lysogen can subsequently be induced to enter the lytic growth phase.

In many phages, lysogeny is accompanied by site-specific recombination between the phage genome and a site in the host DNA. Integration of the DNA occurs at the phage attachment site, *attP*, and the bacterial attachment site, *attB*, and its products are two new sites *attR* and *attL*, which are the substrate sites for excision (Figure 1B). The control of integration and excision of the viral DNA must be accurately coordinated with the life cycle of the phage; abnormal recombination can result in loss of the episome. This is accomplished by having substrates and products that have different requirements for recombination.

I.B. Mycobacteriophage L5

Mycobacteriophage L5 is a temperate bacteriophage that infects several species of the *Mycobacterium* genus (Hatfull and Sarkis, 1993; Fullner and Hatfull, 1997). The

1

Figure 1. Temperate bacteriophage lifestyle choices

(A) The first step in the life cycle of a bacteriophage is infection, when the phage injects its DNA into the bacterial cell. At this point, there are two different pathways that the phage can take. The one is lytic growth in which the phage genome is actively expressed and the bacterial resources are used, so the phage can replicate itself and lyse the cell. The other choice, lysogeny, requires the repression of the phage genes and is often accompanied by integration of the phage DNA into the host genome. The resulting prophage can be induced to lytic growth.

(B) Lysogeny is often associated with an integration of the phage DNA into the host chromosome. The phage attachment site, *attP*, and the bacterial attachment site, *attB*, recombine to produce two new sites, *attL* and *attR*.

A

Bacteriophage

Bacterial Cell

attB

attP

Infection

Lysogeny

Lytic Growth

attL   attR

Induction

B

attP

attB

attR   attL
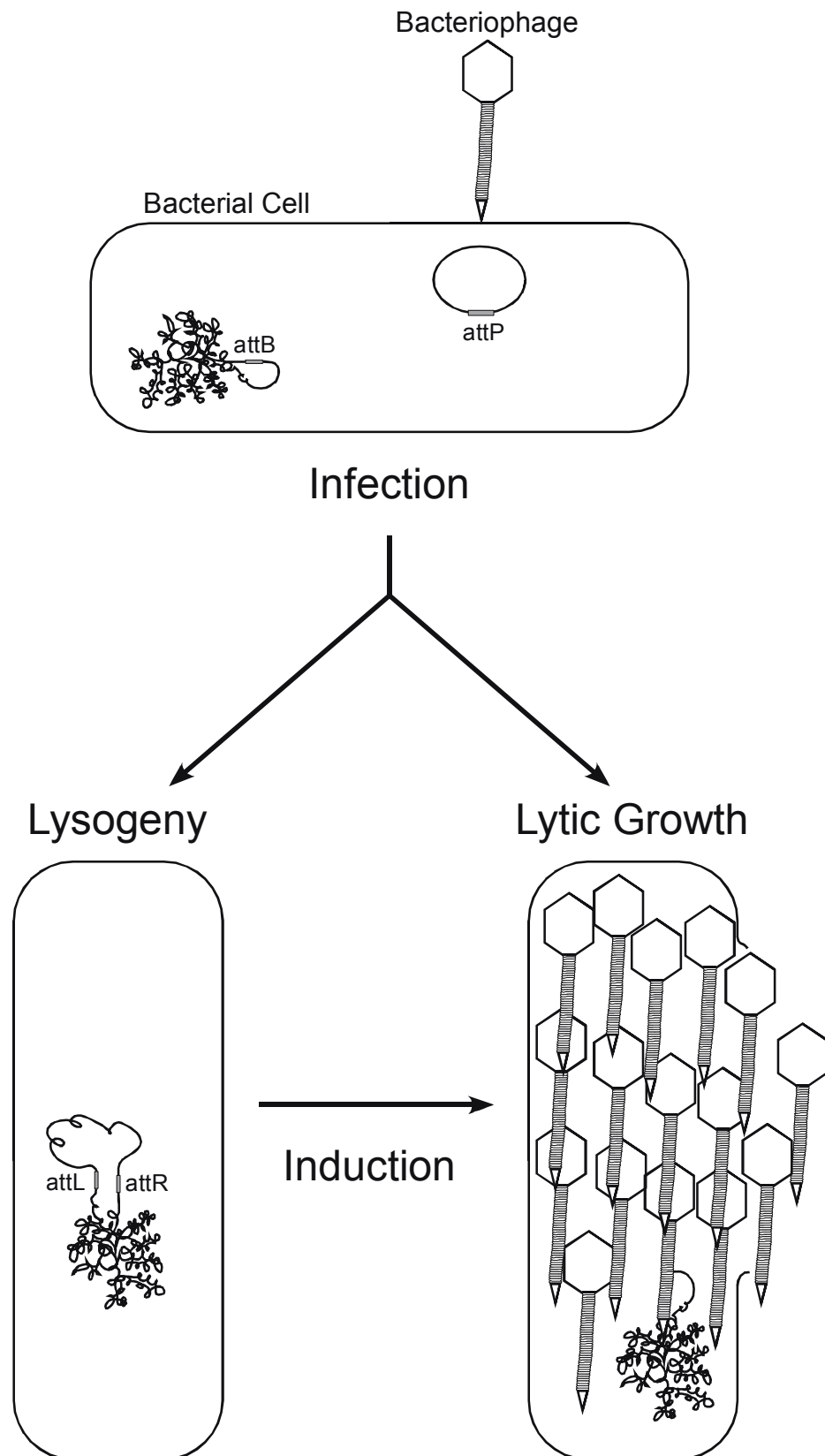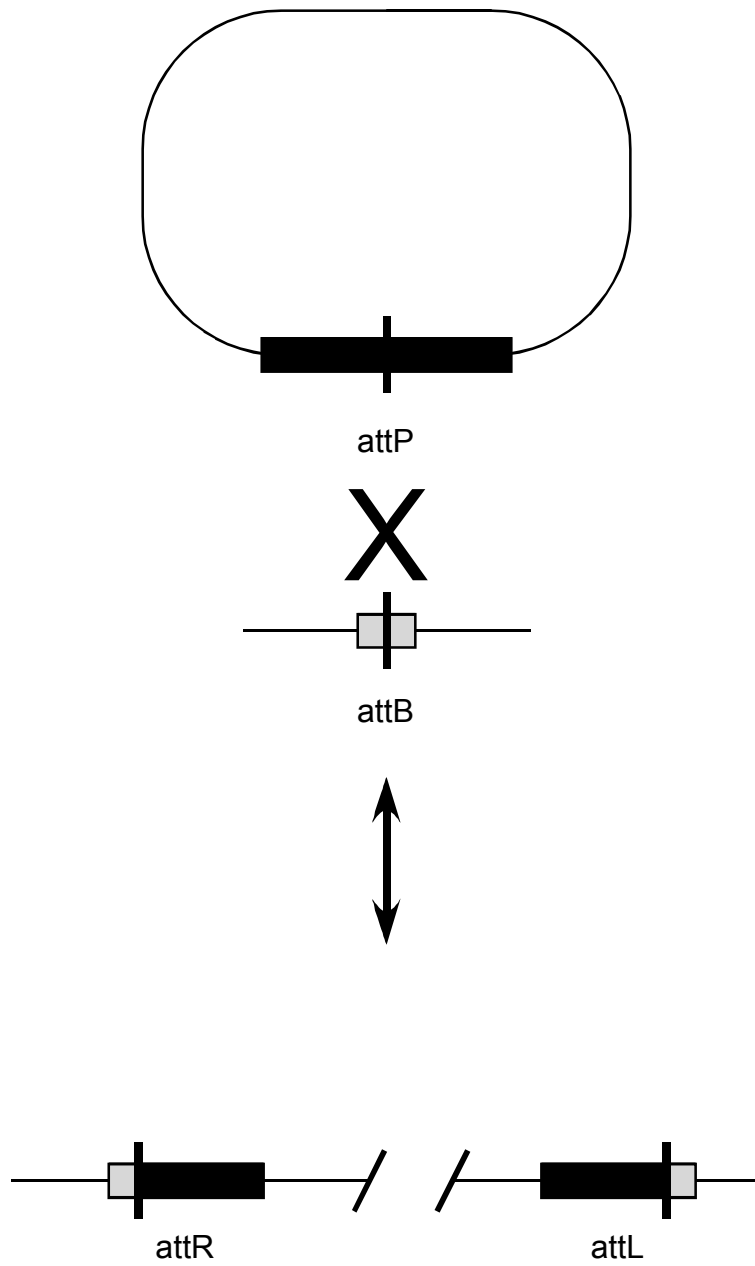
52,297 bp linear, double-stranded DNA genome was sequenced and 89 genes were identified; three of which are tRNAs (Figure 2) (Hatfull and Sarkis, 1993). Its genome is organized in two nearly equal arms that are separated by the phage attachment site (*attP*). All the genes in either arm are encoded on the same strand but on the opposite strands in the other arm, with the exception of two genes (*int* and *g34.1*). With this arrangement, transcription for both arms proceeds towards *attP* (or divergently towards the *attL* or *attR* when integrated). The function of many of the genes is unknown, but it appears as if the left arm genes are involved in virion assembly and structure. The right arm is composed of many DNA proliferation genes and other metabolic genes; it is also the location of the phage repressor gene, *g71*.

Five promoters have been identified in the phage genome (Nesbit *et al.*, 1995). Three of these ($P_{71\ 1-3}$) are located immediately upstream of *g71*. Along with these three, there is a fourth promoter in the right arm (P_{Left}) which is located at the right end and drives transcription to the left. The final promoter that has been identified (P_{Int}) is associated with *g33*, the integrase gene. In addition to these, there must be at least one promoter (P_{Right} ?) that drives transcription rightward in the left arm though none has been identified.

Mycobacteriophage L5 uses a novel repressoion mechanism. The phage repressor proteins, gp71, bind to approximately 30 asymmetric 13 bp sites throughout the genome (Nesbit *et al.*, 1995). Repression of transcription is highly biased according to the orientation of the sites. The repressor has been shown to interfere with transcription only when bound in one direction and even tandem repeats placed in the opposite orientation do not effectively reduce transcription (Brown *et al.*, 1997). Though the sites are distributed throughout the genome, there is a clustering of them around the P_{Left} and P_{71} promoters (Figure 2) (Nesbit *et al.*, 1995) (Brown *et al.*, 1997). While binding to those sites near the promoter prevents transcription initiation, other sites are

Figure 2. L5 genome map

The L5 genome is arranged into 89 genes including 3 tRNAs. The genes on the left arm are transcribed rightward and the genes on the right arm are transcribed leftward. The phage repressor binding sites, stoperators, are distributed throughout the genome and are indicated by arrows. Orientation of the stoperators has been designated by a plus or minus.

Left Arm

tRNAs

1 (?) 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 19.1 20 21 22 23 24 25 26 27 28 29 30 31 32

+

P?

cos

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

attP

33
Integrase

minor
tail protein

Bacteriophage r1t
structural
protein

head
assembly
protein

major
head
subunit

major
tail
subunit

minor
tail subunits

lysis
function?

Right Arm

34.1

P_{71}

Non-essential region

P_{left}

attP

+ + + + + + + + + + + + + + +++ + +++ -

cos

26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52

35 36 37 38 39 40 41 42 43 43.1 44 45 46 47 48 49 50 51 52 53 53.1 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 68.1 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89

Excisionase

DNA Pol

DAPB-DAP
intergenic
protein

B12-dependent
ribonucleoside
reductase

Glutaredoxin
similarity

DNA
Primase

T4 Endo VII
similarity

DNA
Helicase

Repressor
function

involved in transcriptional termination (Brown *et al.*, 1997). This system is thought to be responsible for the silencing of the bacteriophage genome during lysogeny.

I.C. Mycobacteriophage D29

Mycobacteriophage D29 is a lytic bacteriophage that infects mycobacteria. Despite its smaller size, 49,136 bp which encodes 77 genes including 5 tRNA genes (Ford *et al.*, 1998), the D29 genome is organized in a similar fashion to L5, and many of the proteins are related at the sequence level. A notable area of difference is the region of L5 which encodes the phage repressor. This area is missing in D29 and it appears as if this is a recent deletion, as the phage contains putative repressor binding sites throughout the genome and an intact integration system (Peña, C.E. *et al.*, 1998). Furthermore, D29 is subject to repression by the L5 repressor. D29 is thus likely to be a recently-derived lytic derivative of an L5-like ancestor.

I.D. Site-Specific recombination

The rearrangement of DNA via recombination can cause the inversion, integration, excision, resolution or reorganization of genetic material and is one of the important means of creating diversity in life. While homologous recombination is a mechanism that nearly all organisms use to shuffle gene combinations, it requires large regions of similar DNA and nucleoprotein complex containing many proteins (for review see(Bishop and Schiestl, 2000; Sonoda *et al.*, 2001). Conversely, site-specific recombination requires specific sequences on both DNA molecules and depending on the orientation of the sites, can lead to inversion, excision or integration. There are two families of proteins that catalyze site-specific recombination reactions: 1) serine recombinases and 2) tyrosine integrases. Both use their respective active site residue as a nucleophile to initiate strand cleavage and produce a covalently linked protein-DNA intermediate.

I.E. Bacteriophage Lambda site-specific recombination

I.E.i Components for recombination

The integration in bacteriophage Lambda is catalyzed by a tyrosine integrase (Int) and occurs between a 240 bp *attP* site and a 25 bp *attB* site (Mizuuchi and Mizuuchi, 1980; Mizuuchi and Mizuuchi, 1985). While Int is responsible for all catalytic activity, a host-encoded protein, IHF, is also required for integration (Nash and Robertson, 1981).   A third phage-encoded protein, Xis, mediates specificity of directionality (Abremski and Gottesman, 1982). An additional host factor, Fis, has been identified that can stimulate excision in conjunction with Xis (Ball and Johnson, 1991a).

I.E.ii. Requirements for in vitro recombination

The study of the Lambda recombination system has been greatly assisted by the development of an *in vitro* recombination assay.   Under proper conditions, efficient recombination is observed between *attP* and *attB* in the presence of two proteins, Int and IHF (Nash and Robertson, 1981). In the integrative reaction, the *attP* molecule must be supercoiled, but a linear or supercoiled *attB* DNA can be used (Mizuuchi *et al.*, 1978). The excision reaction can be achieved by adding a third protein, Xis (Abremski and Gottesman, 1982). However, excision does not show a supercoiling requisite, although supercoiling of either substrate, *attL* or *attR*, stimulates the reaction when protein concentrations are limiting (Abremski and Gottesman, 1979).   Furthermore, the reactions require $MgCl_2$ or the polyamine, spermadine (Nash and Robertson, 1981). Interestingly, these reactions occur independently of any high-energy cofactors (Nash and Robertson, 1981).

I.E.iii. The integrase protein

Lambda integrase is a member of the tyrosine integrase family of recombinases, which are also distant relatives of topoisomerases (Esposito and Scocca, 1997). Several integrases including Lambda's have been shown to be composed of several discrete domains by limited proteolysis (Figure 3A) (Moitoso de Vargas *et al.*, 1988). Two of these domains show distinct binding valences and interact with the different binding sites in *attP*; the domain at the amino-terminus interacts with arm-type sites (Moitoso de Vargas *et al.*, 1988). An RHR triad and a tyrosine at the active site characterize the conserved catalytic domain in the carboxy-terminus of the protein. Mutational analysis has shown that the RHR triad is required for catalysis and binding to the core (Parsons *et al.*, 1988; Han *et al.*, 1994; Arciszewska and Sherratt, 1995) and suicide substrates have identified the tyrosine as responsible for attacking the DNA (Pabo *et al.*, 1979). The structure of the Lambda catalytic domain has been determined by X-ray crystallography (Figure 3B) (Kwon *et al.*, 1997). The RHR residues form a basic pocket that is likely to interact with the catalytic site, but the tyrosine (a phenylalanine residue in the crystal structure) is on a flexible arm that places it about 20 Å from the RHR triad (Kwon *et al.*, 1997). The location of this residue is consistent with the observation that the tyrosine functions in trans with a triad from an adjacent Int molecule (Han *et al.*, 1993). However, other data suggests that the tyrosine acts in cis (Nunes-Duby *et al.*, 1994), so it is also possible that the flexible arm allows the tyrosine to act in both configurations.

I.E.iv. Attachment sites

The specificity of the phage recombination system is dependent on an array of binding sites present in *attP* (Figure 4). The 240 bp lambda *attP* has a short core with dyad symmetry and a 7 bp overlap region; this is the area in which strand exchange occurs (Mizuuchi *et al.*, 1981; Tirumalai *et al.*, 1998). In addition to the core sites, Int can

Figure 3. Domain organization of tyrosine integrases

(A) The tyrosine integrase is a multi-domain protein that has a common catalytic region at the carboxy-terminus. The catalytic region can be divided into three domains: core-type DNA recognition, catalysis, and protein - protein interaction. Most integrase proteins (from phages HP1, λ and L5 shown in figure) have a larger N-terminal domain that specifically recognizes arm-type sites which is involved in specificity and control of directionality. A subset of these recombinases (phage P1 Cre and 2μ plasmid Flp) has only a small amino-terminal domain that binds DNA non-specifically. A legend for the color-coding of the domains has been included in the figure.

(B) The structure of the carboxy-terminal domain of the Lambda integrase has been solved by X-ray crystallography (Kwon *et al.*, 1997). The RHR triad (in Blue) forms a basic pocket that is believed to interact with the catalytic site, but the active site tyrosine (in Red) is on a flexible arm that would allow it to act in trans with an adjacent integrase molecule or in cis. This figure was generated with RasMol using coordinates from the Protein Data Bank (1AE9).

A

Interactions with Core

HP1 Int

λ Int

L5 Int

P1 Cre

2μ Flp

Domains:

Core Recognition

Catalysis

Protein-Protein Interaction

Arm-Type Binding

Non-Specific DNA Binding

B

Catalytic Tyr

RHR Triad
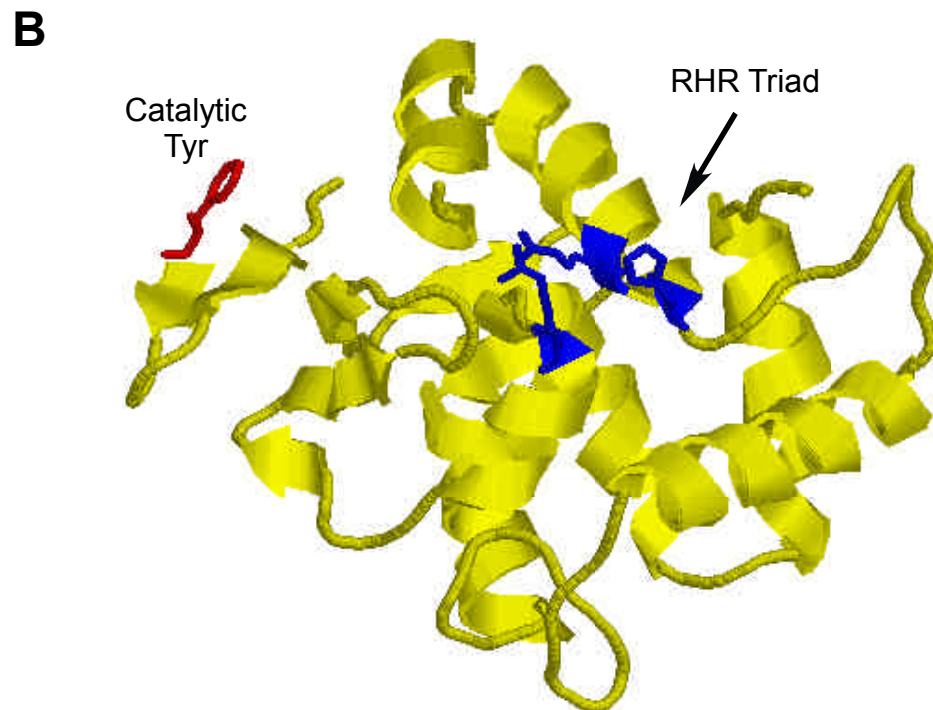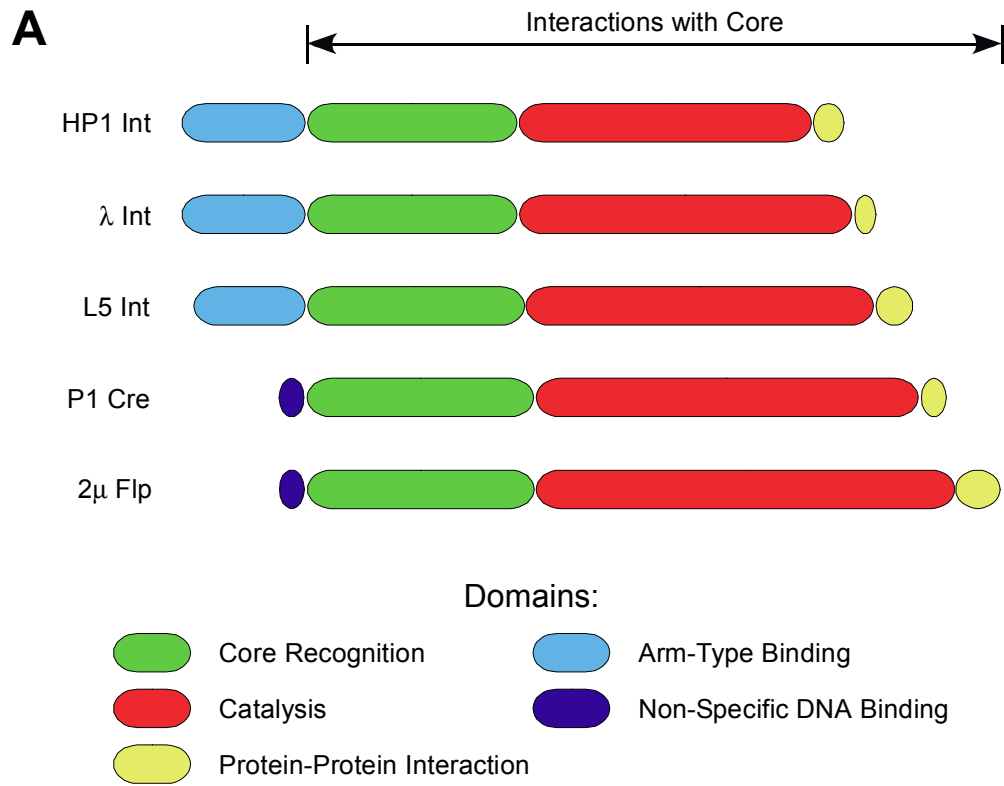
Figure 4. Recombination sites for tyrosine recombinases

Much of the specificity of site-specific recombination is determined by the presence of binding sites in the DNA substrate. In the simplest form (P1 Cre/loxP and 2μ plasmid Flp/frt) the required DNA can be limited to just core-type sites. The phage attachment sites (*attP*) shown here are typical of the greater complexity of binding sites required in phage systems. An array of arm-type sites is necessary for specificity and directional control.   For the systems that have been studied, locations of accessory protein interactions have also been included. A legend has been included in the figure to indicate which type of site is represented.

-150   -100   -50   0   50   100   150

14

P1 loxP

C   C'

2μ frt

C1   C2   C'

P2 attP

P1   P2   H   C   C'   P'1   P'2

P22 attP

P1   P2   P3   H   C   C'   H'   P'1   P'2

λ attP

P1   H1   P2   X1   X2   H2   C   C'   H'   P'1   P'2   P'3

F

L5 attP

P1   P2   P3   C   C'   P4   P5   100 bp   P6   P7

Binding Sites: core-type   Xis   IHF

arm-type   Fis

DNase I Protection: mIHF
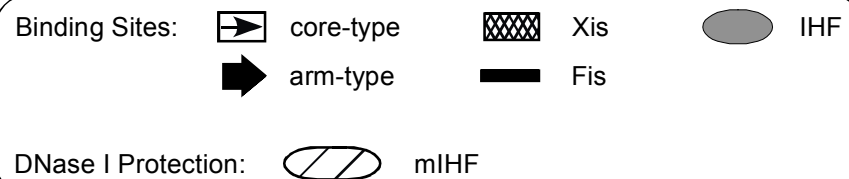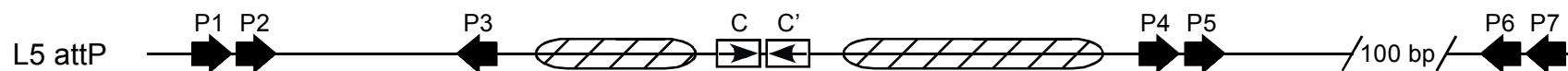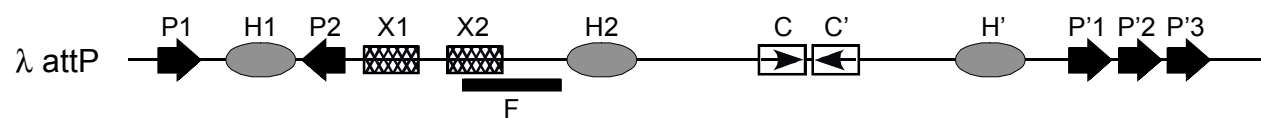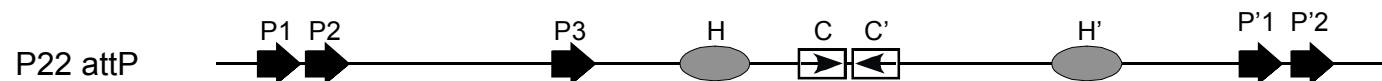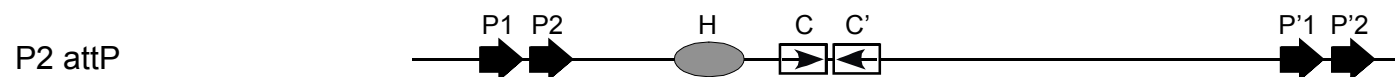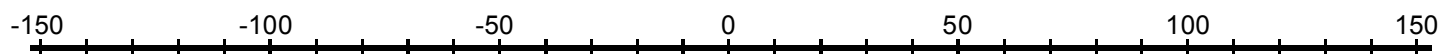
also bind 5 arm-type sites: P1 and P2 to the left of core and P'1 – P'3 to the right (Mizuuchi *et al.*, 1981). Besides the Int binding sites, *attP* also contains three IHF binding sites, H1 and H2 to the left and H' to the right (Robertson and Nash, 1988). While these are all of the sites required for integration, two Xis binding sites (X1 and X2) to the left of the core are required for excision (Yin *et al.*, 1985). A fourth protein, Fis, has been shown to further stimulate excision in the presence of Xis by binding to a site to the left of core (Numrych *et al.*, 1991). The bacterial attachment site, *attB*, is only 25 bp and contains just the core region (Mizuuchi and Mizuuchi, 1985).

I.E.v. Role of IHF

Integrative recombination has been shown to require a host-encoded protein, IHF (Nash and Robertson, 1981). This protein is a hetero-dimer encoded by the *E. coli himA* and *hip* genes (Kikuchi *et al.*, 1985). Binding of IHF creates a sharp bend in the DNA (Robertson and Nash, 1988), which is quite evident in the structure as determined by X-ray crystallography (Figure 5A) (Rice *et al.*, 1996). The bend introduced by IHF is required to bring the arm and core together to allow an Int mediated, intramolecular bridge (Figure 5B) (Moitoso de Vargas *et al.*, 1989). The requirement for IHF can be alleviated by using an alternate DNA bending protein, HU, or by introducing a poly A tract in the same region (Goodman *et al.*, 1992).

I.E.vi. The intasome

The Lambda intasome is a nucleoprotein complex that contains *attP*, Int and IHF. In the intasome, only four of the five arm-type sites are bound by Int; P2 is not occupied (Bauer *et al.*, 1986; Winoto *et al.*, 1986) (Thompson *et al.*, 1987a). Furthermore, IHF binds to all three sites and effectively bends the DNA. Another requirement for intasome formation is the supercoiling of the *attP* DNA (Richet *et al.*, 1986). The cooperative interactions of the proteins and DNA supercoiling generate a wrapped architecture in

Figure 5. Role of IHF in complex formation

(A) The structure of IHF bound to DNA has been determined by X-ray crystallography (Rice *et al.*, 1996). The binding of IHF to DNA produces an 180^0 bend at the site. The two subunits of IHF are shown in red and blue. This figure was generated with RasMol using coordinates from the Protein Data Bank (1IHF). (B) The bending caused by IHF is required to bring the arm and core together to allow an Int-mediated bridge.

A
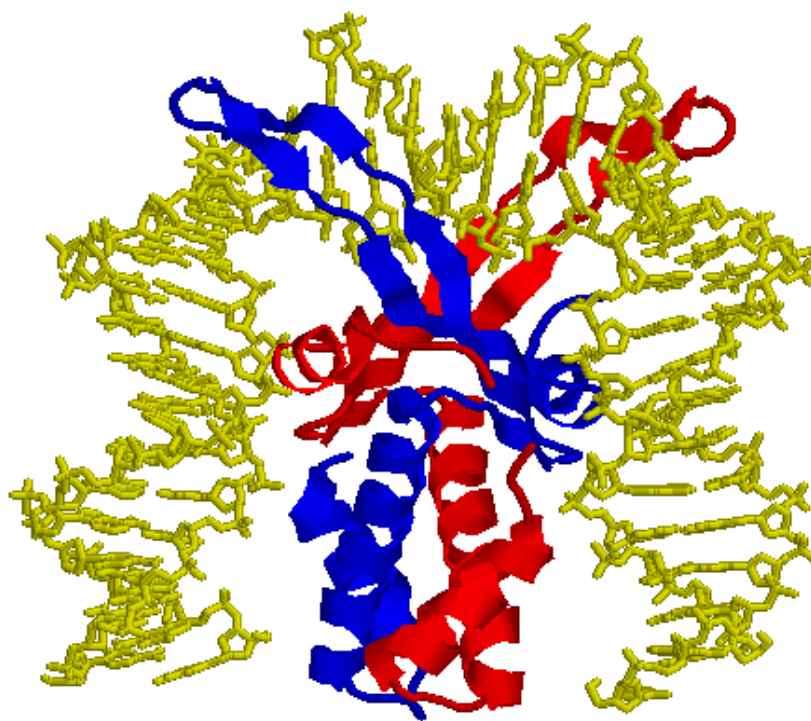
B
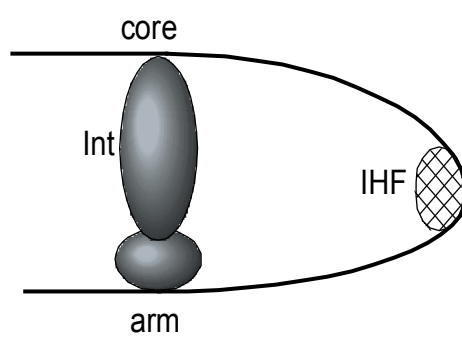
which the two integrase molecules form intramolecular bridges between arm-type sites and the core. Two other Int molecules are present but only the arm-type binding valence is satisfied.

I.E.vii. Synapsis

Synapsis of *attP* and *attB* follows intasome formation in the assembly of a recombinagenic complex. The *attB* DNA does not form a stable complex with Int or IHF. Instead, it is captured by interaction with a previously formed intasome (Richet *et al.*, 1988). All of the proteins necessary for recombination are present in the intasome. The structure of the lambda synaptic complex has not been determined, but the structure of a synaptic complex for the P1 tyrosine integrase, Cre, has been solved (Guo *et al.*, 1997). The P1 recombination system differs from the Lambda system because the substrate DNA, loxP, lacks arm-type sites (Figure 4), and the N-terminal domain that binds arm-type sites in the Lambda system is not present in the Cre protein (Figure 3A). However, the catalytic domain is related to Lambda Int, so the model for synapsis can be applied to the organization of the Lambda architecture. The structure of the P1 complex shows four integrase molecules bound to DNA in a planar arrangement with a sharp bend within the core overlap region (Figure 6A) (Guo *et al.*, 1997). The model suggested by this structure would be a planar arrangement of the DNA that requires little rearrangement of the core-type binding sites to allow recombination (Figure 6B).

I.E.viii. Chemisty of catalysis

Once a synaptic complex has been assembled, the cleavage is initiated by a nucleophilic attack of the 'top strands' by the active site tyrosine (Figure 7) (Craig and Nash, 1983). The resulting phosophotyrosyl intermediates are cut by another nucleophilic attack initiated by the 5'-OH of the cleaved strands to ligate the DNA. At this point, the DNA is in the form of a Holiday junction and can be resolved by cutting

Figure 6. Structure of tyrosine recombinase synapsis

(A) The structure of Cre integrase proteins bound to DNA was determined by X-ray crystalography (Guo *et al.*, 1997). It shows a planar arrangement of the DNA with a sharp bend near the middle of the cores. The top and side views of the structure of Cre bound to loxP, shown here, are a reproduction of a figure by Guo (Guo *et al.*, 1997).

(B) The pathway of recombination involves a planar arrangement of the two substrates, followed by a Holiday junction intermediate, and finally, resolution to yield products. Cartoons representing these three steps are shown.

A

Cleaving subunit

17' 3'

5'

E

A

95 Å

5'

5'

17

3'

5'

3'

17

5'

Cleaving subunit

17' 5'

3'

105 Å

20   20   20   20

A   A

E   E

17'   17'

M   M

17   17

N   N

341   341

341   341
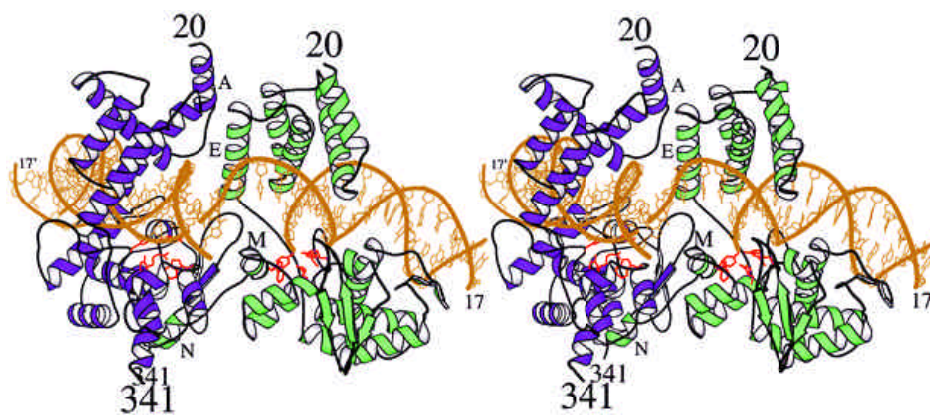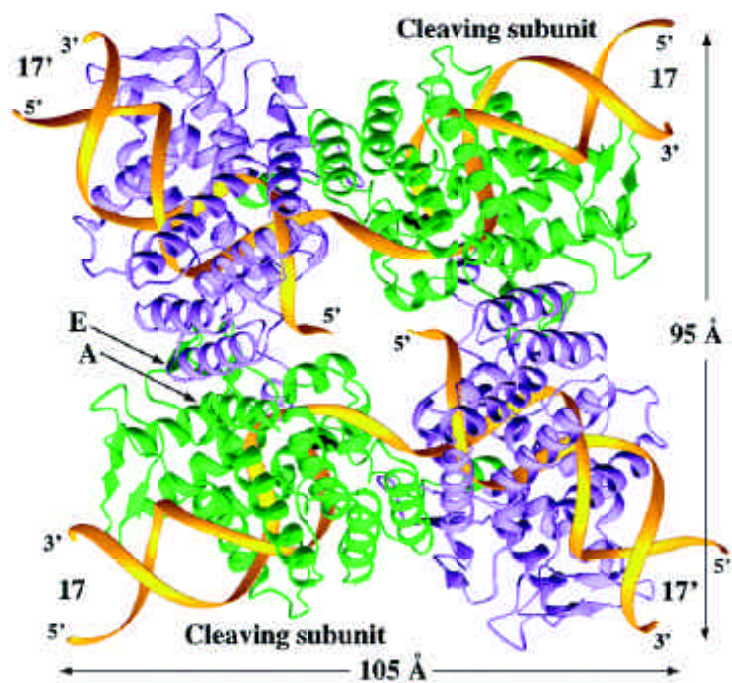
B
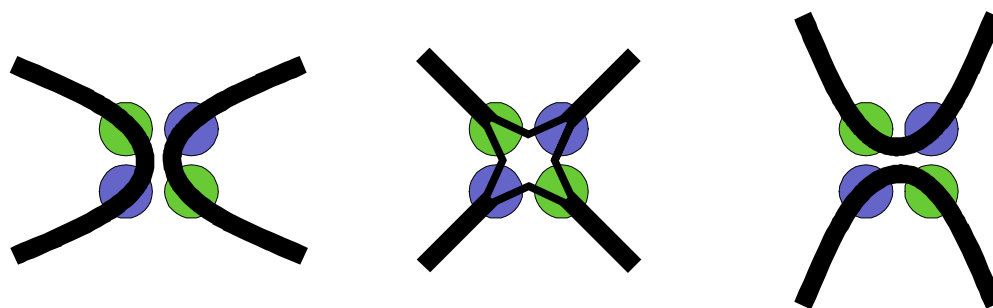
Figure 7. Chemistry of Lambda integration

The tyrosine integrase-mediated recombination reaction is initiated by a nucleophilic attack of the DNA's phosphodiester linkage by the recombinase's tyrosyl hydroxyl group. This produces a covalent linkage between the 3′ phosphate of the DNA and the tyrosine of the recombinase. The release of the protein is accompanied with the ligation of the DNA strands by a nucleophilic attack of the phosphotyrosyl intermediate by the 5′ hydroxyl of a cleaved DNA strand.
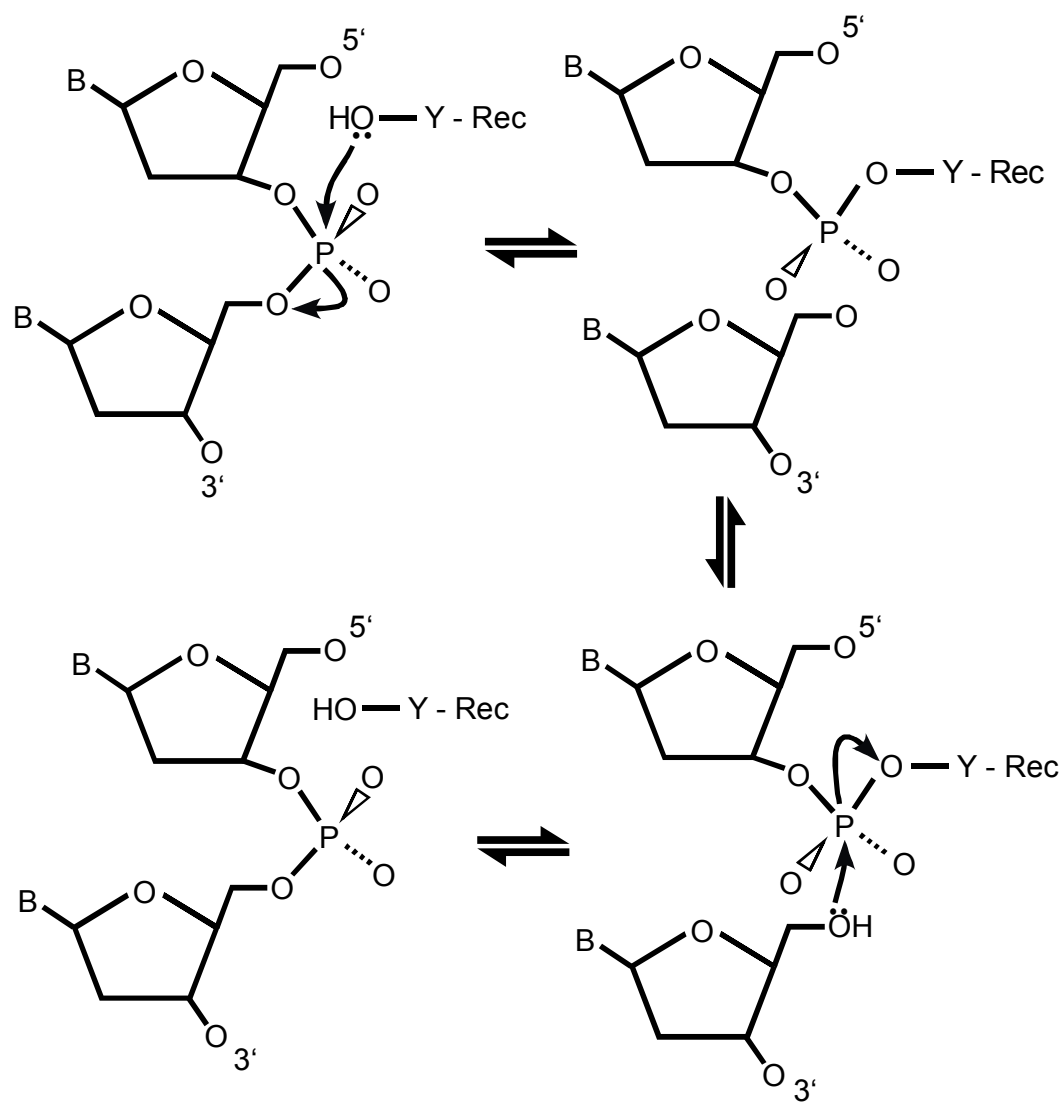
and ligating the bottom strands using the same mechanism of nucleophililc attack (Figure 6B). The Holiday junction intermediate also can be resolved to reproduce the substrate if the second round of cleavage is on the top strand (Franz and Landy, 1995)

I.E.ix. Control of directionality

While the recombination of *attP* and *attB* in integrative recombination only requires Int and IHF, the products, *attL* and *attR*; are not efficiently combined with these two proteins alone. The primary difference between these sets of DNA is the arrangement of the arm-type sites (Figure 8). For integration, all of the arm-type sites flank the core on one DNA molecule, *attP*. However, *attL* and *attR* each have sites on one side of their core that were derived from the corresponding side of *attP*. An additional protein, Xis, is required to mediate excision by binding to two sites in *attR*, bending the DNA and promoting complex formation (Better *et al.*, 1983; Bushman *et al.*, 1984); *attL* DNA complexes do not require Xis (Better *et al.*, 1983). Furthermore, a different subset of the arm-type and accessory sites is used for excisive recombination. Binding to P1 and H1 in *attR* and P'3 in *attL* is not required (Bushman *et al.*, 1985; Bauer *et al.*, 1986; Winoto *et al.*, 1986; Kim *et al.*, 1990). The use of only three arm-type sites indicates that the amino-terminal domain of one of the four Int molecules is not bound to an arm-type site, and can be used to predict a model of the synaptic complex (Figure 9) (Bushman *et al.*, 1985).

A fourth protein, Fis, has been shown to enhance excision by cooperatively binding with Xis (Thompson *et al.*, 1987b; Numrych *et al.*, 1990). Stimulation of excision by 20-fold *in vitro* and 200-fold *in vivo* has been observed (Thompson *et al.*, 1987b; Ball and Johnson, 1991a). However, Fis has also been shown to have a moderate stimulation of lysogen formation *in vivo* (~20-fold) (Ball and Johnson, 1991b) but no detectable effect *in vitro* (Thompson *et al.*, 1987b). Fis appears to stimulate both integration and excision, but the effects are more pronounced in excision.

Figure 8. Arrangement of λ recombination sites

The phage attachment site (*attP*) for the λ recombination system has a core region with dyad symmetry; there is also an array of sites that flank the core. The N-terminal domain of the integrase recognizes the arm-type sites P1 and P2 to the left of core and P'1, P'2 and P'3 to the right. IHF binds and bends the DNA at the IHF sites H1, H2 and H'. Excision and inhibition of integration requires two additional sites to the left of the core: Xis binding sites X1 and X2. A fourth protein, Fis, is not absolutely required but binds site F to stimulate recombination. The bacterial attachment site, *attB*, has only the core binding sites. The substrates for excision, *attR* and *attL*, has the same array of sites, but they are now split between the two attachment junctions. This difference in location of accessory sites prevents the recombination of *attR* and *attL* by Int and IHF alone.

A
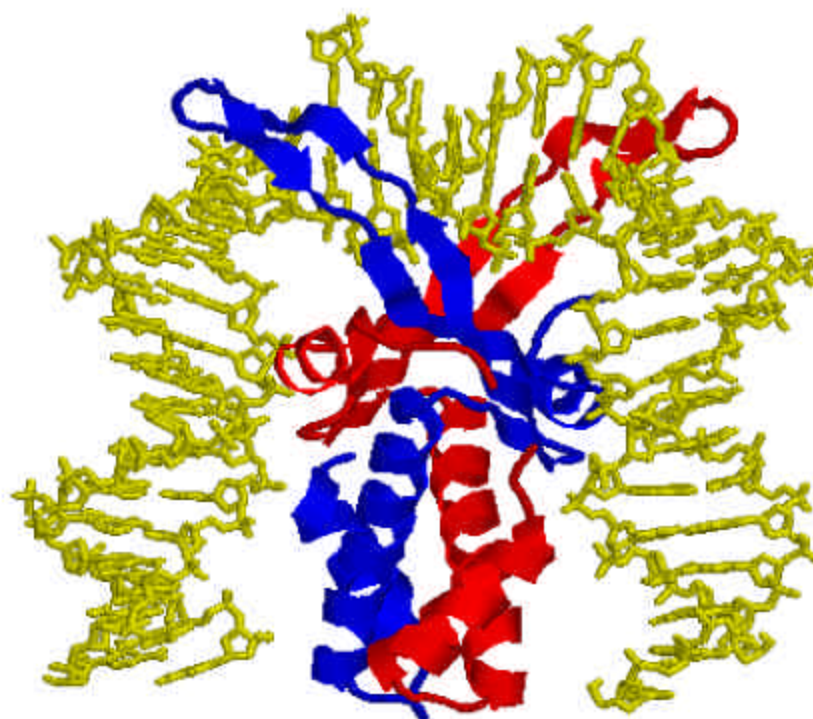
B
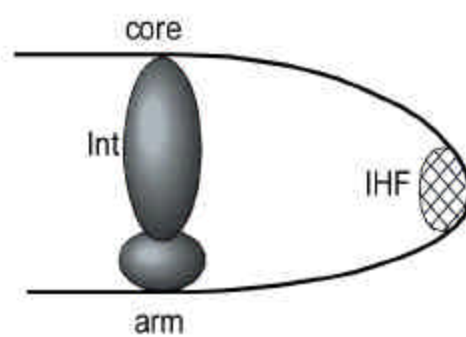
Figure 9. Lambda excision model

The proposed model for Lambda excision requires a bending of both substrate DNAs mediated by IHF and bending of *attR* by Xis, which allows the integrase to form a bridge between a core- and arm-type site. A proposed model for an excisive complex has the P'1, P'2 and P2 arm-type sites and all four core-type sites occupied. This model does not account for the interactions of Fis, which have been shown to stimulate excision. This model is a reproduction of a figure by Bushman (Bushman *et al.*, 1985).
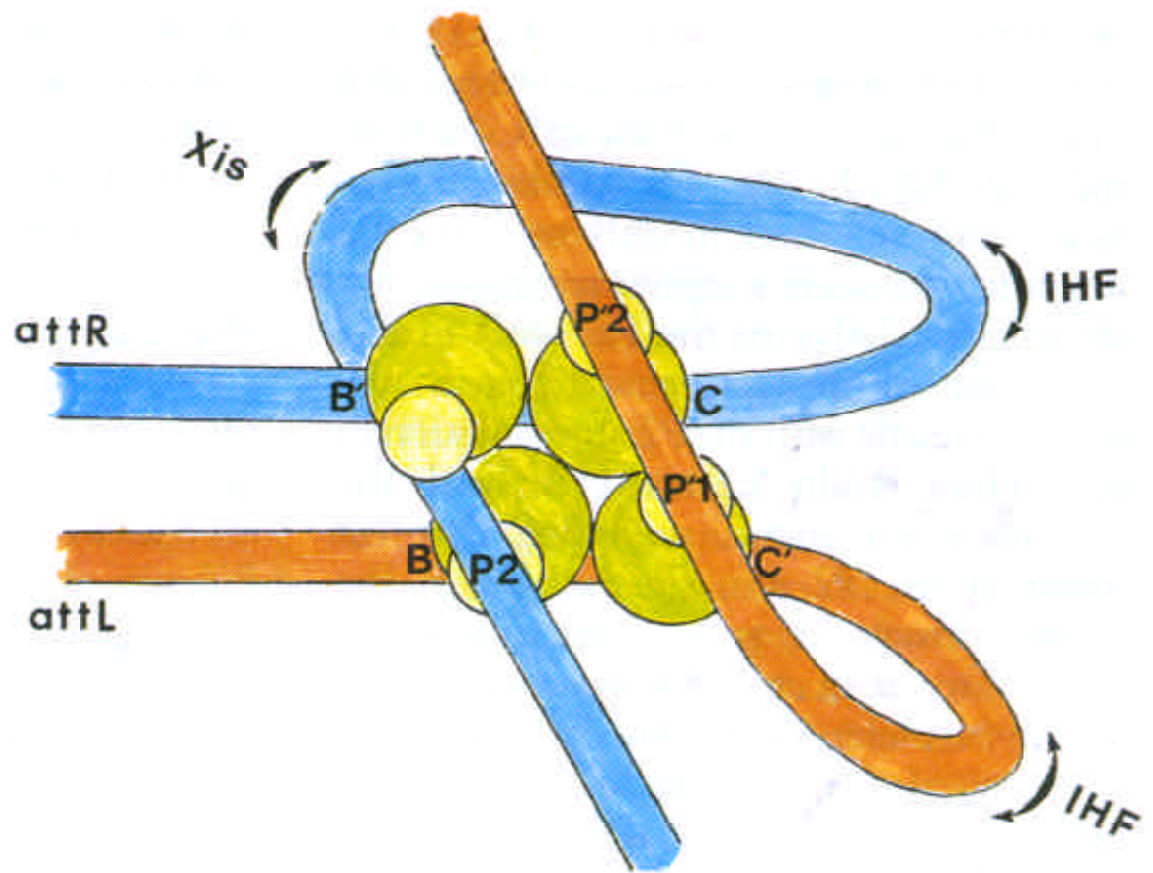
The inhibition of integration is achieved by the Xis protein binding to *attP* DNA. The same subset of sites used in excision is occupied in an inhibited *attP* complex (Thompson *et al.*, 1987a; Moitoso de Vargas and Landy, 1991). Furthermore, the exclusion of Int from the necessary arm-type sites is not by direct competitive binding, but by a change in the architecture of the DNA complex (Moitoso de Vargas and Landy, 1991).

I.E.x. Regulation of recombination

In the Lambda system, much work has been done to determine how the expression of the proteins involved in site-specific recombination are coordinated with the phage's lifecycle. The same regulator proteins are involved in both regulation of recombination and of the growth state of the phage. Upon infection, the protein that is integral in determining the fate of the virus is CII. This protein is sensitive to host encoded proteases that are believed to reflect the condition of the cell. In healthy cells, the proteases degrade CII and lytic growth is set in motion by production of the anti-terminator protein, N. If CII is not inactivated, it positively regulates the expression of the phage repressor gene, CI, and a lysogenic state is entered.

Entry into lysogeny must be accompanied by expression of Int but not Xis to allow integration of the phage DNA. The CII protein that is involved in production of the phage repressor also promotes transcription from a site within the *xis* gene, so only the downstream *int* gene is transcribed. During lytic growth, the Xis protein inhibits integration, and the presence of Int is not desired. The P_L promoter is responsible for transcription of the *xis* gene but can only achieve this in the presence of the anti-terminator, N. If this were the only means of regulation, both Int and Xis would be made from this transcript, but a retro-regulation is also involved in limiting Int production. The presence of N allows the transcript to proceed through a terminator downstream of *int* and through a region called *sib*. This sequence is target by RNase III

that degrades the mRNA, and since the *int* gene is closer to the site, it is degraded earlier. The effect is a reduction of mRNA encoding Int and thus a lowering of its expression. The induction of a prophage into lytic growth requires expression of both Int and Xis, which is achieved by the same method of transcription used in lytic growth. The anti-terminator, N, allows transcription to proceed through the *int* and *xis* genes from the P$_L$ promoter, but due to a change in relative location of the *sib* site and its absence from the mRNA, it is not targeted for degradation. The *sib* site and *int* are on opposite sides of *attP*, and during lytic growth the *attP* site is intact, so *sib* is located downstream of the *int* gene. In an integrated prophage, the attachment site is split, so the bacterial genome is downstream of the *int* gene, and the *sib* site is located at the opposite end of the prophage. This complex regulator system assures that recombination of the phage DNA is tightly coordinated with viral lifecycle.

I.F. Alternate methods of directional control

I.F.i. Other tyrosine recombinases

The phage recombination systems have distinguishable products and substrates that can be recognized and recombined by a different set of proteins. The lambda mechanism utilizes an accessory protein that stimulates one of the directions and often inhibits the other direction (Lewis and Hatfull, 2000). These recombination directionality factors (RDFs) play an architectural role in determining the recombination specificity. In other systems, a different recombinase is expressed for each direction. In the *E. coli* switch for type 1 fimbriae, the FimB and FimE recombinases are each used for only one direction (Gally *et al.*, 1996). However, the biological role of some recombination systems dictates that the substrate and product must have identical sequence features. This is the case with the Xer recombinases, which are involved in recombining identical sites in dimeric chromosomes, so an alternate mechanism must

be used. Part of the specificity of the ColE1 Xer recombination complex, in which the recombinase only binds to a minimal core (~30 bp), is achieved by using a series of sites that flank the core that are bound by accessory proteins, ArgR and PepA (Colloms *et al.*, 1997). Furthermore, these proteins help to align the complex only if specific topological requirements are met (Colloms *et al.*, 1997).

I.F.ii. Serine recombinases

The DNA-invertases and resolvases (Inv/Res) share a common domain at the amino-terminus that contains the catalytic core and utilizes a serine as the catalytic residue (Figure 10A). Due to the nature of these systems, the sequences of the substrates and products must be identical. DNA topology is the feature that allows for specificity in choosing substrate molecules (Stark *et al.*, 1989). However, this is not the case for all serine recombinases.

In recent years, a new class of proteins has been identified that has a catalytic domain which is related to the Inv/Res family, but which has a large C-terminal domain of unknown function (Figure 10B) (Carrasco *et al.*, 1994; Christiansen *et al.*, 1994; Thorpe and Smith, 1998). This new class of proteins has been termed serine integrases (Int-S) or large serine recombinases. Presumably, the catalytic mechanism of these systems must be similar to those of other serine recombinases (Breuner *et al.*, 2001). In the best-studied system, φC31, the minimal sites of recombination are small (< 50 bp) (Thorpe *et al.*, 2000), but the mechanism of recombination has not been determined. Furthermore, no accessory proteins are necessary for an *in vitro* recombination reaction (Thorpe and Smith, 1998).

Differences between the mechanisms of the Int-S family and the Inv/Res family might include the specificity of recombination or the directional control. For several of the large serine recombinases such as XisF (Ramaswamy *et al.*, 1997), TP901-1 (Breuner *et al.*, 1999) and Rv1586c (L. Bibb and G.F. Hatfull unpublished observations), an

Figure 10. Alignment of serine recombinases

The N-terminal catalytic domain of serine recombinases is conserved between invertases, resolvases and large serine recombinases. An alignment (A) of the amino-terminal segment of the protein sequences for the Salmonella Hin invertase (gi:118766), Mu Gin invertase (gi: 9633542), γ-δ resolvase (gi:135945), Tn3 resolvase (gi:135947), Anabaena XisF (gi:1075645), TP901-1 integrase (gi:1246426), *M. tuberculosis* Rv1586c (gi:2117251), and φC-31 integrase (gi:3947502) shows conserved residues in red and similar residues in green. A cartoon representation (B) of the same proteins has been included to highlight the differences in the linear organization of the domains in invertases and resolvases versus large serine recombinases or serine integrases. The small carboxy-terminus of the invertases and resolvases contain a helix-turn-helix DNA binding motif, while in contrast, the large serine recombinases contain a lare carboxy-terminal domain of unknown function. In both panels, the catalytic serine residue has been indicated with an arrow.

A
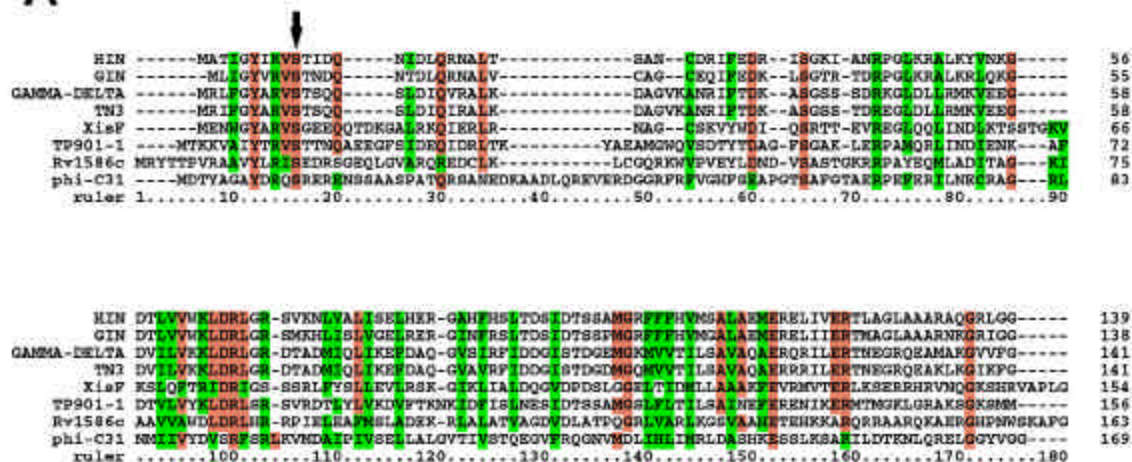
B
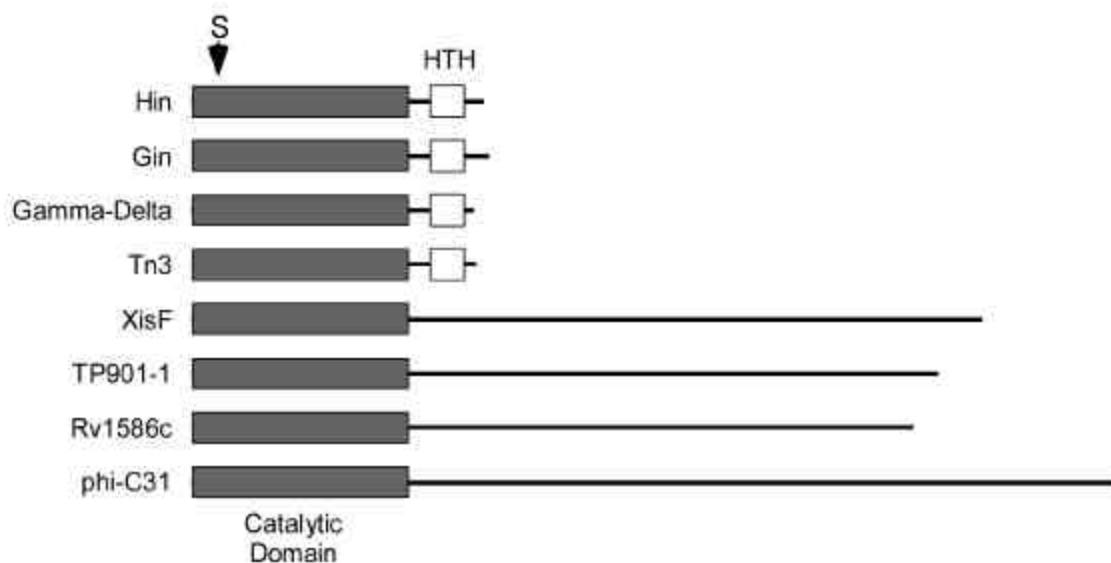
accessory protein has been identified that mediates directional control. This differs from the resolvase and invertase reactions, which rely heavily on topology as a mechanism of directionality. Despite having the same mechanism of catalysis, alternate methods of directionality are utilized.

I.G. Biological restraints on mechanism of directionality

The levels of recombination control can vary from the simple cases that merely moderate the frequency of recombination, to the more complicated systems whose substrates and products have different specificities. The simplest mechanism of control is to moderate the level or timing of recombinase expression. However, this method plays no role in specificity and only limits the amount of recombination. The need to form a large nucleoprotein complex is a means of specificity, and in many cases, accessory proteins are involved in aligning the DNA substrates into an active synaptic complex (for review (Hallet and Sherratt, 1997). Additionally, many of the proteins involved show cooperative binding characteristics in establishing the architectures. For control of directionality, both DNA topology and sequence differences can be used to differentiate between substrates and products.

With a multitude of mechanisms available, the biological requirements of the recombination event are more influential in determining how strictly the reaction must be controlled and which mechanisms are utilized. The catalytic domain, serine or tyrosine, is not integral in determining the means by which specificity is controlled. A more important factor is the biological role of the recombination and how much flexibility is allowed. In some systems, such as resolving DNA dimers, in which recombination sequences must be identical, both tyrosine and serine recombinases can be used, and DNA topology is integral in controlling the reaction (Stark *et al.*, 1989; Sherratt *et al.*, 1995). In systems that require more rigid control, such as the integration and excision of the bacteriophage genome, which must be tightly coordinated with the

life cycle of the phage, asymetric sites and accessory proteins have evolved to regulate directionality. Once again, both serine and tyrosine recombinases have been identified that function with similar mechanisms for specificity.

I.H. Mycobacteriophage L5 Integration

I.H.i. Components

Bacteriophage L5 has an integration system that uses a tyrosine integrase that is homologous to Lambda integrase, but the two proteins are distant relatives and the sites are quite different in sequence. The two proteins involved in L5 integration are the integrase (Int), which is encoded by L5 gene *33*, and the host encoded mycobacterial integration host factor (mIHF). The mIHF protein is a specific DNA binding protein that creates a bend in the DNA (Pedulla and Hatfull, 1998). Despite the functional similarities of IHF and mIHF, the proteins are unrelated. Furthermore, mIHF has not been shown to bind *attP* DNA specifically, but several mutations in the region where mIHF binds are inhibitory to integration (M. Kahlenberg and G.F. Hatfull, unpublished). Integration occurs between a 252 bp *attP* and a 29 bp *attB* to produce *attL* and *attR* which are the substrates for excision (Peña *et al.*, 1996; Peña *et al.*, 1997).

I.H.ii. Attachment sites

The *attP* site has a core region with dyad symmetry separated by an asymmetric stretch of 7bp. Flanking the core are three arm-type sites to the left (P1-P3) and four to the right (P4-P7) (Figure 11A and B). Only four of these sites, P1/P2 and P4/P5 are necessary for integrative recombination (Peña *et al.*, 1997). The *attB* site is located in the tRNA-Gly gene of the host, so a 43 bp common core that is identical between the *attP* and *attB* sites is necessary to regenerate the 3' end of the tRNA upon integration of the phage (Lee and Hatfull, 1993), but only 29 bp are necessary for recombination (Peña *et*

Figure 11. The L5 recombination sites

The L5 attachment site (*attP*) contains a core with dyad symmetry and 7 arm-type sites (P1-P7). Sites P1-P3 are to the left of the core, while P4-P7 are to the right. The bacterial attachment site (*attB*) only contains the core. The *attR* and *attL* sites both contain the core and arm-type sites from the left or right of *attP*, respectively. An additional feature shared amongst the sites is an identical 45 bp common core that is the 3′ end of the tRNA-gly gene. This sequence is necessary so that when the phage integrates into the bacterial tRNA, the gene is reconstructed with the phage DNA. The sequence for part of the *attP* site has been included; P6 and P7 arm-type sites have been left out, as they are not necessary for integration. The sequence of *attP* (A) has been included with the core and arm-type sites labeled. A schematic representation of all four sites (B) shows the rearrangement of arm-type sites as a result of integration.

A TGCGCTGAGTCTTCAGCGATCCCCATCCGCGACGTGCCAACTAGGTCTCCTCTCGTCGTGAACAA
ACGCGACTCAGAAGTCGCTAGGGGTAGGCGCTGCACGGTTGATCCAGAGGAGAGCAGCACTTGTT
　R　Q　T　K　L　S　G　W　G　R　R　A　M　-NH$_2$

P$_{int}$

GGCTACCGGGTTGCAACTCCTGTGCAACTCTCAGGCTTCAACGCGCTTCTACGACCTGCAATTTC
CCGATGGCCCAACGTTGAGGACACGTTGAGAGTCCGAAGTTGCGCGAAGATGCTGGACGTTAAAG
P1　　　　　P2

TTTCCACTTAGAGGATGCAGCCGAGAGGGGGTAAAAACCTATCTTGACCGGCCCATATGTGGTCG
AAAGGTGAATCTCCTACGTCGGCTCTCCCCCATTTTTGGATAGAACTGGCCGGGTATACACCAGC
P3

GCAGACACCCATTCTTCCAAACTAGCTACGCGGGTTCGATTCCCGTCGCCCGCTCCGCTGGTCAG
CGTCTGTGGGTAAGAAGGTTTGATCGATGCGCCCAAGCTAAGGGCAGCGGGCGAGGCGACCAGTC
Common Core

AGGGTGTTTTCGCCCTCTGGCCATTTTTCTTTCCAGGGGTCTGCAACTCTTGTGCGACTCTTCTG
TCCCACAAAAGCGGGAGACCGGTAAAAAGAAAGGTCCCCAGACGTTGAGAACACGCTGAGAAGAC
P4　　　　　P5

ACCTGGGCATACGCGGTTGCAACGCATCCCTGATCTGGCTACTTTCGATGCTGACAAACGAATAG
TGGACCCGTATGCGCCAACGTTGCGTAGGGACTAGACCGATGAAAGCTACGACTGTTTGCTTATC

B

P1/P2　　P3　　　Core　　　P4/P5　　P6/P7
attP

Core
attB

Core　　　P4/P5　　P6/P7
attL

P1/P2　　P3　　Core
attR

al., 1996). In addition to the integase binding, *attP* contains some specificity for mIHF binding (M. Kahlenberg and G.F. Hatfull, unpublished).

I.H.iii. Assembly pathway intermediates

There are multiple architectures that have been identified in the assembly pathway of an active synaptic complex (Figure 12) (Pena *et al.*, 2000). An *attP*-containing complex, the intasome, has been identified which has integrase molecules bridging between the P4/P5 sites and the core; mIHF is also required to form the bend between the sites. A second potential precursor, synaptic complex 2, has been observed that contains both *attP* and *attB* joined by intermolecular bridging of integrase. The P1/P2 arm-type sites are involved in binding the integrase molecules that recruit and are bound to the *attB* DNA. However, this complex does not require mIHF and thus does not have the proper bends in the DNA to allow for the proper alignment required for integration. It is not clear if either of these complexes is an obligate intermediate or if the assembly can proceed through either pathway.

I.H.iv. Synapsis

The most complete synaptic complex that has been observed (synaptic complex 1) has all 4 arm-type sites (P1/P2 and P4/P5) occupied, and both cores sites (*attB* and *attP*) are present. The most likely conformation for this complex, based on observed intermediates, is the P4/P5 sites bridged to the *attP* core and the P1/P2 sites bridged to the *attB* core. The complex that has been captured could be in an "open" conformation that might require a conformational change to activate catalysis. Though no structural data exists for the L5 synaptic complex, it presumably would be similar to those of homologous proteins that have been examined, such as Cre. Therefore, the *attB* and *attP* core must lie in a planar conformation with a sharp bend in the center. Upon initial cleavage, a covalent intermediate will form, and after ligation of the first strands, a

Figure 12. L5 recombination complexes

Integration by the L5 recombination system requires the formation of nucleoprotein complexes. The first step in the pathway is one of two complexes: the intasome or synaptic complex 2. The intasome contains only *attP* DNA and Int and mIHF proteins. The Int is bound to the core and the P4 and P5 arm-type sites; mIHF bends the DNA to stabilize the complex. Synaptic complex 2 contains both *attP* and *attB* and only the Int protein, which forms a bridge between the P1 and P2 arm-type sites of *attP* and the core of *attB*. It is not clear if *attB* is recruited before or after the intasome conformation. Synaptic complex 1 contains *attP*, *attB* and both proteins, mIHF and Int. Both cores and the four arm-type sites (P1/P2 and P4/P5) are occupied. While this complex contains all the components necessary for recombination, it might require some rearrangement to become active. After integration, the *attL* DNA remains in a complex that resembles the intasome; *attR* is only found as free DNA. A legend has been included to identify the proteins and binding sites.

Intasome
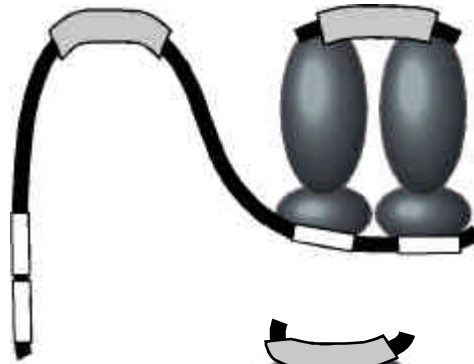
Synaptic Complex II

Synaptic Complex I

attL complex

Legend:

Integrase

mIHF

core

arm-type site

Holiday junction intermediate should result. Neither the covalent intermediate nor Holiday junction has been trapped or observed in the L5 integration reaction.

I.H.v. Control of Directionality

The proteins necessary for integration are not sufficient to recombine the products of integration, *attR* and *attL*. The primary difference from the substrates and products is the organization of arm-type sites (Figure 11B); the cores differ by only several bases. In studies of the integration reaction, an *attL* complex was identified that mimics the intasome (Figure 12) (Pena *et al.*, 2000). However, after recombination *attR* DNA was only detected in an unbound state. The mechanism of excision requires some means to bring *attR* into a complex and bring the two DNA molecules into a synaptic complex. Presumably, an RDF exists in the L5 genome that will mediate this complex formation.

I.I Host Species

I.I.i. Mycobacterium tuberculosis

Though Mycobacteriophage L5 can infect several species, the most well known and must relevant to human life is probably *M. tuberculosis.* This pathogen is the causative agent for tuberculosis (TB) and currently infects nearly one-third of the world's population (Kochi, 1991). Though not all of those infected show active TB, it is still the leading cause of death in adults worldwide. While the discovery and use of antibiotics seemed to have curbed the disease, the increase in multi-drug resistant strains and the AIDS epidemic have brought TB back into the forefront. The World Health Organization reports that the leading killer for AIDS patients is tuberculosis.

The study of *M. tuberculosis* is limited by the slow growth rate and the fact that it is a human pathogen. The completion of the sequencing of the H37Rv and CSU93 strains has been a great assistance in studying the organism. However, the use of phage

and phage-derived tools is another powerful method that will ultimately help to get a more complete understanding of the bacteria and the disease.

I.I.ii. Mycobacterium bovis *BCG*

M. bovis bacille Calmette Guerin (BCG) was originally isolated as a virulent strain in the early 1900's (Calmette and Guerin, 1908). Due to the lack of efficient storage methods, the strain was preserved by passage through serial cultures. This extended growth outside of an animal has allowed numerous mutations to arise that have yielded an avirulent strain. BCG is closely related to *M. tuberculosis*, and is currently used as a vaccine strain to protect against TB. When the genomes are compared, there are several major deletions in BCG that have been detected, but there are also likely to be numerous smaller mutations that will only be found when the BCG genome has been sequenced (Mahairas *et al.*, 1996). The comparison of these two strains offers insights into the cause of virulence in *M. tuberculosis* and potential targets for treatment of TB. This is an extremely useful laboratory strain because of the low health risk to humans, though it still is a slow grower.

I.I.iii. Mycobacterium smegmatis

M. smegmatis is non-pathogenic fast-growing mycobacterium. These two properties make it extremely useful for laboratory study. The most commonly used laboratory strain, mc^{2}155, has a mutation that allows for high efficiency transformation (Snapper *et al.*, 1988). This organism is used extensively for preliminary studies, but due to metabolic and genomic differences, all results must be viewed with caution when trying to draw parallels to the slow-growers.

I.J. Phages as genetic tools

The use of bacteriophages as tools to study life dates back to 1915 when they were first identified independently by Frederick Twort and Felix d'Herelle. Later, phages were integral in showing that DNA contained genetic determinants. In current times, bacteriophage-derived tools are commonly used for many molecular and genetic techniques. Cosmid vectors are derived from phage λ; the Cre recombination systems are used extensively in many different organisms; and phage enzymes such as T4 DNA ligase are indispensable in the modern molecular biology lab.

Mycobacteriophage L5 offers a diverse set of tools that can be used to facilitate the study of mycobacteria (Hatfull, 1994). One of the most commonly used systems developed from L5 is an integrating vector system. A plasmid containing the L5 *attP* and *int* genes is capable of efficiently and stably transforming mycobacteria by integrating a single copy into the bacterial genome (Figure 13) (Lee *et al.*, 1991). An alternate method has been developed that uses two plasmids: one containing *attP* and another non-replicating, suicide plasmid expressing Int. The two-plasmid system transforms bacteria at a lower efficiency, but increases the stability of the integrated vector (Lewis and Hatfull, 2000). This integrating vector system allows the introduction of large DNA segments (~50 kb) and has the potential to be used as a means to modify the BCG vaccine strain to protect against other pathogens (Stover *et al.*, 1991; Hatfull *et al.*, 1994).

The largest deficiency of this technique is the inability to effectively remove the DNA once it had integrated. There are two occasions in which it is desirable to remove the vectors. One is to cure the strains of the inserted DNA to allow further study of these strains. The second is to allow recovery of the DNA into *E. coli* to allow amplification of the DNA for further molecular studies or to introduce it into other

Figure 13. Two methods for integrating vectors

The use of the L5 integrase and *attP* has proven to be an extremely useful tool for introducing DNA into mycobacterial cells. The most commonly used method has a single plasmid that contains both *int* and *attP*. This plasmid efficiently transforms bacteria and stably integrates a single copy of the DNA into the bacterial genome. A second method has been developed that transforms at a lower efficiency, but creates more stable integrants. In this system, two plasmids are cotransformed: one containing *attP* and one with the integrase. The plasmid with *attP* is stably integrated while the one with integrase is a non-replicating plasmid and is only transiently present to express the Int required for integration. The absence of the integrase in the transformed strains eliminates Int-mediated *attR* / *attL* recombination that occurs at a low frequency.

Integrase

attP

integrating vector

attL   attR

attP   Integrase

integrating vector   suicide plasmid
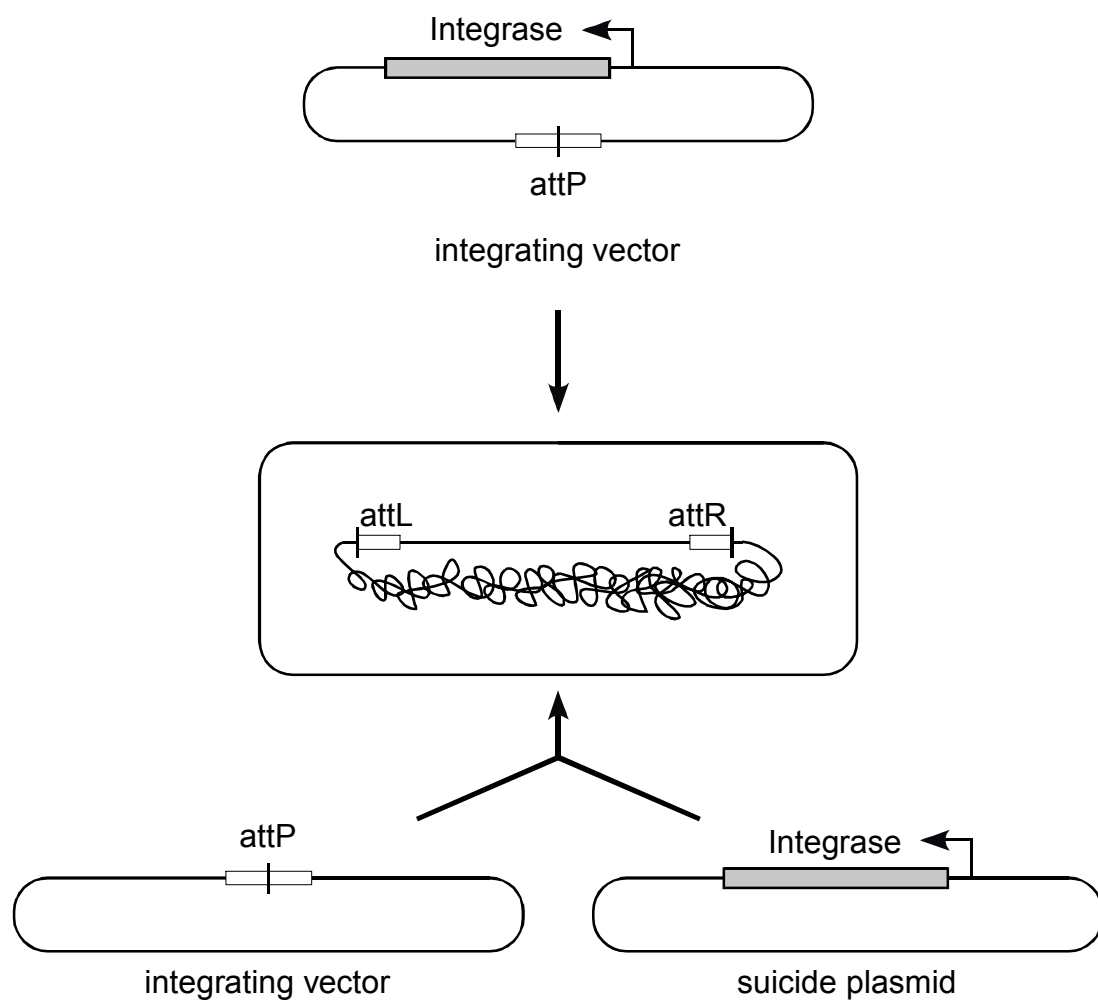
mycobacterial strains. The identification of the RDF is integral in developing techniques that will overcome these shortcomings.

I.K Specific Aims

I.K.i. Identification of the L5 RDF

For all integrase systems that have been studied, an accessory protein is required for directional control. Due to the small size and diversity of sequences, the L5 Xis has not been identified through sequence comparisons. Stability studies of different L5 derived integrating vectors gave a potential location for the *xis* gene (Lee *et al.*, 1991), and the first aim of this study was to identify the L5 gene through the use of a genetic assay.

I.K.ii. Excisive recombination in vitro

The L5 integration reaction has been extensively characterized by using efficent *in vitro* recombination assays. It is fair to assume that a similar analysis could be performed with the excision reaction. The initial step towards this goal was to over-express and purify L5 Xis. Once usable samples of the protein had been obtained, conditions and requirements for excisive recombination could be determined.

I.K.iii. Mechanism of directional control

Integrase-mediated site-specifc recombination requires the formation of nucleo-protein complexes. In order to understand the reaction and how the specificity of directional control is determined, these complexes must be examined. A variety of techniques such as native gel electrophoresis and DNase I footprinting are useful for identifying and characterizing complexes involved in excisive recombination. Comparison of these complexes to those identified for integration will shed light on how directionality is mediated.

I.K.iv. L5 integrating vector system

The L5 integrating vector system is already a powerful tool used to study mycobacteria. This method will be greatly enhanced by using the L5 Xis to remove the integrated DNA. Techniques will be developed to cure strains and to shuffle the DNA into *E. coli*.

I.K.v. Recombination Directionality Factors

Many RDFs have been identified through experimental techniques, but finding them using simple sequence similarity searches has been difficult because of their small size and sequence diversity. Furthermore, no systematic attempt has been made to identify or classify this group of proteins. The final aim of this project is to rigorously search sequence databases to identify new and previously identified RDFs. These RDFs will then be examined using a variety of bioinformatic tools to determine protein characteristics and to speculate on the phylogeny of this group of proteins.

II. MATERIAL AND METHODS

II.A Bacterial Strains

II.A.i. Escherichia coli

E. coli strains XL1-Blue (Stratagene) and DH5α were used for DNA propagation. Strains SC110 or GM2163 was used when unmethylated DNA was needed. Strain BL21(DE3)pLysS was used for protein expression (Novagen).

II.A.ii. Mycobacterium

Mycobacterium smegmatis mc^{2}155 is a high frequency transformation strain of *M. smegmatis* that was used for all *M. smegmatis* experiments (Snapper *et al.*, 1988). *M. bovis* BCG-C is a non-virulent strain that was used for all BCG work.

II.B. Growth of bacterial cultures

E. coli strains were grown in LB broth (Difco) or on LB plates containing 1.5% agar at 37°C. Antibiotics for selection in *E. coli* were added to the following concentrations: kanamycin (kan), 20 μg/ml; carbenicillin (Cb), 50 μg/ml; tetracylcine (tet), 10 μg/ml for liquid and 6.25 μg/ml for plates; and hygromycin (hyg) 100 μg/ml. *M. smegmatis* was grown in Middlebrook 7H9 broth or on Middlebrook 7H10 agar plates (Difco). *M. bovis* BCG was grown in Middlebrook 7H9 broth or on Middlebrook 7H11 agar plates (Difco). Media for mycobacterial growth were supplemented with 10% ADC, 1 mM CaCl$_2$, 50 μg/ml Cb and 10 μg/ml chlorohexamide (chx). Liquid medium and plates used for replica plating were supplemented with 0.05% - 0.1% Tween 80 to

prevent clumping. Antibiotics for selection in mycobacteria were added to the following concentrations: kan, 20 μg/ml; hyg, 100 μg/ml; tet, 0.5 μg/ml. Sucrose was added to 10% in 7H10 agar for SacB selection. For preparation of mycobacteria for phage infections, a saturated culture of *M. smegmatis* was diluted 1/100 and grown overnight in a baffled flask in medium lacking Tween 80. For BCG, dense cultures were pelleted and resuspended in 2 volumes of medium lacking Tween 80 and incubated overnight. Plating top agar was made using 7H9 medium with 0.7% agar and supplemented with 1% Glucose and 1 mM CaCl₂.

II.C. Recombinant plasmids and cosmids

II.C.i. *Recombinant plasmids and cosmids constructed by others*

II.C.i.a. *pCP plasmids*

Plasmid pCP10 is a hygromycin resistant (hygR) and carbencillin resistant (CbR) plasmid which replicates extrachromosomally in both *E. coli* and mycobacteria and was used purely as an extra-chromosomal plasmid control. It is a derivative of pYUB41Plasmid pCPΔr13 is derived from pMH94 (see below) by shortening the *attP* region to remove the P6/P7 arm type sites to leave 116 bases to the right of the core (Peña *et al.*, 1997). It is CbR and kanamycin resistant (kanR), and it replicates extrachromosomally in *E. coli* and integrates in mycobacteria. Plasmid pCP24, which is CbR, contains *attR* and *attL*, which was made by an *in vitro* recombination reaction between pMH94 and a Sac I fragment from pMH12.1 containing *attB*. The linear product was ligated at the Sac I site to generate the completed plasmid.

II.C.i.b. pGS74

Plasmid pGS74 is a reporter construct that contains the FFlux gene downstream of the hsp60 promoter with 10 tandem copies of the gp71 binding site between them (Brown *et al.*, 1997).

II.C.i.c. pMH plasmids

Plasmid pMH94 and pMH5 contain the *attP* and *int* region of L5 in a kanR and CbR backbone; the difference is the length of the insert (Lee *et al.*, 1991). Plasmid pMH5 contains 4.9 kb of L5 DNA which includes 11 ORFs (*32-42*) and the *attP*site; the shorter, 2.1 kb piece of DNA in pMH94 has only the *attP* and one intact gene, *int* (see Section III).

II.C.i.d. pYUB plasmids

The pYUB plasmids are kind gifts from the W. R. Jacobs Lab. The pYUB53 plasmid is exrtachromosomally replicating, TetR and KanR. Plasmid pYUB572 is a shuttle phasmid vector containing an *E. coli* origin of replication, *bla*, cos sites, and Pac I sites for cloning into phAE87. The pYUB657plasmid contains the counter selectable sucrose sensitivity marker, *sacB* (Pelicic *et al.*, 1996).

II.C.ii. Construction of pJL plasmids

II.C.ii.a. pJL5 to pJL8

The pJL5 to pJL8 series of plasmids are kanR, integrating plasmids that contain a sucrose sensitivity counter selectable marker, *sacB*. DNA containing the *sacB* gene was isolated as a Pst I fragment from pMP62 and inserted into either pMH5 or pMH94 partially digested with Pst I. Plasmids pJL5 and pJL6 are derivatives of pMH5 containing the gene in opposite orientations; pJL7 and pJL8 are the respective pMH94 derivatives.

II.C.ii.b pJL9 and pJL10

These plasmids, pJL9 and pJL10, are derivatives of pJL7 and pJL8 which contain additional L5 DNA to complete L5 gene *34.1*. It was constructed by ligating a Bgl II / Bcl fragment from pMH5 into pJL7 or pJL8 digested with Bgl II and Bam HI.

II.C.ii.c. pJL11

Plasmid pJL11, a hygromycin-resistant derivative of pJL7, was generated by removing the aph gene by partial digestion with Hind III and XbaI and inserting a Hind III - XbaI fragment containing the HygR gene from pLT91 (C. Peebles and G. Hatfull, unpublished observations).

II.C.ii.d. pJL12 to pJL19

The series of plasmids contains segments of L5 DNA from pMH5 inserted into vector pMV261 (Stover *et al.*, 1991) (see Figure 14B). These plasmids contain the following DNA segments: pJL12and pJL13, a Bgl II - Bcl I fragment (L5 co-ordinates 26,358 - 27,396) in the Bam HI site of pMV261 in forward and reverse directions respectively; pJL14 and pJL15, a Sal I fragment(L5 co-ordinates 27,081 - 27,829) in the Sal I site of pMV261in forward and reverse directions respectively; pJL16 andpJL17, a Sal I fragment (L5 co-ordinates 26,112 - 27,081) in the Sal I site of pMV261 in forward and reverse directions respectively; pJL18, an MscI - NheI fragment (L5 co-ordinates25,607 - 26,854) in the MscI - NheI sites of pMV261; pJL19, a Bcl I - MscI fragment (L5 co-ordinates 26,358 - 25,607) in the BamHI - HpaI sites of pMV261.

II.C.ii.e. pJL20

Inserting a PCR-derived fragment into pET21a generated the L5 gp36-overexpression plasmid, pJL20. For the PCR reaction, the primersL5-36a - `TTC CAT ATG CCC CCG AGA GCA TCC ATC` and L5-36b - `TTC CTC GAG CTA CTT GCC GAT TCG CAT C` were

used to amplify gene *36*. The product was digested with Nde I and Xho I and ligated into the same sites in pET21a (Stratagene).

II.C.ii.f. pJL22

Plasmid pJL22 was made by inserting the Bgl II - XhoI fragment from pJL20 (see below) into the BamHI and Sal I sites of pMV261, such that the only L5 gene present is *36*, which is fused to the translation initiation signals of pET21a (Stratagene) and the hsp60 promoter from pMV261.

II.C.ii.g. pJL27

Plasmid pJL27 is an ampR shuttle phasmid construction vector containing L5 *xis* downstream of the hsp60 promoter. The plasmid was constructed by placing an Aat II fragment from JL22 into pYUB572 digested with the same.

II.C.ii.h. pJL28 and pJL29

In order to lower and possibly regulate the expression of the L5 *xis* gene, it was cloned into a plasmid with temperature sensitive mutants of the L5 repressor gene, *71* (from L5c^{ts32} or L5c^{ts43}). Plasmids pJL28 and pJL29 contain L5 gene *71* with its own promoter and the L5 *xis* gene downstream. To further down regulate expression, ten copies of the gp71 binding site are placed in tandem just upstream of *xis*. The plasmids was constructed by means of a three way ligation using an Xba I - Bam HI fragment from pGS74 containing 10 copies of the gp71 binding site, an XbaI - Bam H segment from pMD159 or pMD162 (Donnelly-Wu *et al.*, 1993)(for pJL28 or pJL29, respectively) and the vector pJL22 cut with Xba I.

II.C.ii.i. pJL32 to pJL34

Plasmid pJL32 is a derivative of pYUB53 in which the gene conferring kanR has been deleted, so the plasmid is exclusively tetR. This was accomplished by digesting pYUB53 with Pst I and religating. The pJL33 and pJL34 plasmids are derivatives of pJL32 that contain the hsp60 promoter and *xis* from pJL22 in both directions. They were constructed by digesting pJL32 and pJL22 with AatII and ligating the segments.

II.C.ii.j. pJL35

Plasmid pJL35 is a derivative of the pJL11 integrating plasmid. The *int* gene was partially deleted by digesting with PstI and religating. The resulting plasmid is hygR and sucS, but lacks a functional copy of *int*.

II.C.ii.k. pJL36

To simplify cloning of a shuttle phasmid, both pJL22 and pJL27 were cut with Nhe I and Apa LI and joined to construct a kanR derivative of pJL27. This plasmid is a cosmid vector with the L5 *xis* gene downstream of the hsp60 promoter.

II.C.ii.l. pJL38

Plasmid pJL38 contains a minimal *attB* placed into pUC119. It was constructed using oligos RevAttB – `GGG CGC GGA TCC GC` and AttB-789N – `CGG GGT ACC CCA GAG CNN NAG TCT TCC AAA CTA GCT ACG CGG ATC CGC GCC C`. A double stranded *attB* was made by annealing the oligos in a solution containing 45 μM of each and 0.1 M NaCl. The mixture was boiled and slowly cooled to room temperature overnight; the RevAttB oligo was extended with Klenow polymerase. The dsDNA was digested with Bam HI and Asp718 and ligated into pUC119 cut with the same. Sequence analysis of clones that were competent in recombination identified pJL38 as having the wild-type sequence.

II.C.ii.m. pJL39

Plasmid pJL39 is a pCPΔR13 derivative in which an *in vitro* recombination of the *attP* with a Pvu II fragment from pJL38 containing *attB* produced *attR* and *attL*. The Pvu II ends of the linear product were ligated to complete the plasmid.

II.C.ii.n. pJL46 and pJL51 to pJL54

These plasmids all contain *attR* and/or *attL* derived from *in vitro* recombination. Plasmid pJL46 was constructed by recombining pJL38 with an EcoR I / Bam HI fragment from containing *attP*. The linear product was ligated to the polylinker from pHSX digested with Eco RI and Bam HI. Plamids pJL51 and pJL52 contain the *attL* fragment from pJL46 removed by EcoR I digestion and placed into pBluescript SK+ in either orientation. The pJL53 and pJL54 plasmids are the *attR* counterparts, only digested with Bam HI.

II.C.ii.o. pJL47 and pJL48

These plasmids are cosmid vectors that express both *int* and *xis*, each with an hsp60 promoter. They were constructed by cutting pJL36 with Bcl I and cutting the hsp60 promoter and the *xis* gene from pJL44 and the promoter and *int* gene from pJL45 with Bcl I and partially with Bgl II. These constructs both have the *xis* gene upstream of the *int* with the overall orientation of the insert in both directions.

II.D. Recombinant Phage

II.D.i. phAE87 phage

Shuttle phasmid phAE87 is a kind gift from S. Bardarov and W.R. Jacobs that was derived from a temperature-sensitive non-replicating strain of TM4 (Bardarov *et al.*, 1997).

II.D.ii. phJL phage

Three derivatives of phAE87 were constructed to express the L5 *xis* gene. phJL1 was made by cutting phAE87 and pJL36 with Pac I and joining them to create a phasmid that expresses Xis. The phJL2 and phJL3 phasmids contain both the *int* and *xis* genes from pJL47 and pJL48, respectively, and were constructed by a Pac I digest of each plasmid and the phAE87 phasmid.

II.E. PCR analysis of attachment sites and junctions

The PCR scheme used to identify the presence of attachment sites or junctions was the same as that previously described (Peña *et al.*, 1997). The *attB* primers (primer 1 and primer 2) produce a band of 361 bp, whereas the *attP* primers (primers 3 and 4) yield a band of 221 bp. When all four primers are used additional bands of 316 bp and 266 bp, corresponding to *attL* and *attR* respectively, are also generated. New primers were designed to amplifiy BCG *attB*. Primer JL24 - `CCC TTG ATG TCA GGC TGC TTC` and primer JL25 – `GCT CCT TGG TAG TCG ATA ACG` are used in place or primer 1 and primer 2 and produce nearly identical length products.

II.F. Protein expression and purification

The L5 Xis protein was overexpressed by inducing *E. coli* strain BL21(DE3)pLysS containing pJL20 with 1mM IPTG. Pelleted cells were frozen at -80º C then resuspended in TDE (20mM Tris pH 8, 1mM DTT, 1mM EDTA, 50mM NaCl) buffer containing 1mM phenylmethylsulfonyl flouride (PMSF). DNAse I and MgCl$_2$ were added to a final concentration of 20 μg/ml and 4mM respectively, and the extract was incubated for 20 min at 37º C. After clarification of the extract by centrifugation at 16,000 x g for 20 min., the lysate was boiled for 15 minutes; functional assays showed that boiling did not result in a significant reduction in excisionase activity. Collecting an ammonium sulfate cut between 55% and 80% saturated ammonium sulfate and

resuspending the pellet in HED50 further purified gp36. The HED50 buffer is 20 mM HEPES pH 8.0, 1 mM EDTA, 1 mM DTT and 50 mM NaCl. The proteins were then loaded onto a weak cation, carboxymethyl resin (CM) column using a Sprint BioCad chromatography system and eluted with a 50mM to 500mM NaCl gradient. Fractions eluting at approximately 200mM NaCl contained highly purified excisionase plus a single visible contaminating band. Fractions collected from the flow through material when loading this column were found to contain substantial amounts of excisionase and were used to reload the CM column. Excisionase eluted at the same salt concentration as previously but the fractions were free from contaminating proteins as viewed by Coomassie blue staining. An alternate method, which produces a greater yield than reloading the flow through, is to collect the fractions from the first column and load them onto a strong cation resin column (HS, Perseptive Biosystems). When washing with a 50 mM to 1M NaCl gradient, the Xis protein is eluted with little contamination at near 1 M salt. However, quite often, the fractions from the first CM column were used as these were still relatively pure. The protein was stored at – 20°C in 10 mM Tris pH 7.5, 1 mM EDTA, 1 mM DTT and 50 mM NaCl.

II.G. Protein gel electrophoresis

Due to the small size of gpXis, a three layer Tricine-SDS-PAGE system was used for electrophoretic separation of the proteins. A 49.5 % acrylamide/3 % bis-acrylamide stock solution was used to make a 4% stacking gel, a 10 % spacer gel and a 16.5% separating gel as previously described (Schägger and Jagow, 1987). The gels were stained using 0.25% Coomassie blue, 45% methanol and 9% acetic acid and destained using a 25% methanol, 7% acetic acid solution.

II.H. *In vitro* excision recombination assay

An *in vitro* excision assay was developed using similar approaches to those described for the L5 *in vitro* integration reaction (Lee and Hatfull, 1993). The standard reaction included 20mM Tris (pH 7.5), 10mM EDTA (pH 8.0), 25mM NaCl, 10mM Spermidine, 1mM DTT, 1 mg/ml BSA, 0.021pmol of pCP24 DNA, 0.74pmol of purified Int, 1.1pmol of purified mIHF and 13.4 pmol of purified Xis. Proteins were diluted using protein dilution buffer containing 10 mM Tris pH 7.5, 1 mM DTT and 1 mg/ml BSA. Reactions were incubated at 37° C for 45 min, stopped by addition of SDS to a final concentration of 0.1%, and the products identified by electrophoresis through a 0.8% agarose gel.

II.I. Radiolabeling DNA fragments

Labelled DNA fragments used for native gel electrophoresis and DNase I nuclease protection assay were generated by a Klenow polymerase fill-in of the restriction digestion overhang with α-P32 dATP. Fragments containing *attP* were obtained by a BamH I digestion of pMH94 or a BamH I / EcoR I digestion of pCPΔR13. The *attL* fragments were obtained by digestion of pCP24 with BamH I / BsiW I or pJL39, pJL46, pJL51, or pJL52 with EcoR I. Fragments containing *attR* were generated by cutting pCP24 with BamHI / BglII, or pJL39, pJL46, pJL53 or pJL54 with BamHI. To create an *attR* fragment labelled on one end, pJL53 or pJL54 were digested with EcoR I, labelled, and then cut with Pvu II.

II.J. Native gel electrophoresis

The reaction for complex formation was typically carried out in a 10 μl reaction with 20 mM Tris pH 7.5, 25 mM NaCl, 10 mM spermidine, 1 mM DTT, 1 mM EDTA, 0.1 mg/ml BSA and 100 μg/ml salmon sperm DNA. The protein concentration was 170 nM mIHF, 96 nM Int, and 32 nM – 3.2 μM Xis unless otherwise indicated, and 500 cpm of

radiolabelled DNA was used. The reactions were incubated on ice for 30 minutes and then immediately loaded on a 1X TBE, 5% acrylamide gel at 4°C; a 50% sucrose loading buffer was used. In order to look at product formation, SDS was added directly to the reaction or used to soak a gel slice for a second dimesion gel at a final concentration of 0.5%; the DNA was then run on a gel containing 0.1% SDS.

II.K. DNase I nuclease protection assay

Buffer conditions for the DNase I protection assay were nearly identical to those for native gel electrophoresis with the exception of the presence of 5 mM $MgCl_2$ in place of the EDTA. Furthermore, the typical reaction was carried out in a 200 μl volume with 10 – 20 kcpm of DNA. Following 20 minutes on ice, the reactions were moved to 37°C at least 5 minutes before digestion. DNase I was diluted to 1.4 U/μl and 5 μl was added to each reaction. After 1 minute, the digestion was stopped by adding 0.7 ml of stop solution (645 μl ethanol, 5 μl 10 mg/ml tRNA and 50 μl saturated Ammonium Acetate) and placing it in a dry ice / ethanol bath. The reactions were transferred to – 80°C overnight, pelleted and resuspended in sequencing loading buffer. The DNA was run on either 6% or 8% sequencing gel.

II.L. Plasmid recovery from purified genomic DNA

For plasmid rescue, genomic DNA was prepared using a previously described CTAB miniprep protocol (Ausubel *et al.*, 1996) except that the pellet collected from the 3ml saturated culture was resuspended in 500 μl of TE and then treated with lysozyme at a final concentration of 1mg/ml for 1 hour at 37° C. Approximately 1 μg of DNA was used for an *in vitro* excisive recombination reaction as described above and incubated for 2 hours. Samples were then extracted with phenol:chloroform 1:1, chloroform alone, and ethanol precipitated. After resuspending the DNA in 4 μl of TE, 2 μl of the sample was used to transform electrocompetent *E. coli* XLI-Blue. One ml of tryptic soy broth

(TSB) (Ausubel *et al.*, 1996) was added and the cells allowed to recover for 1 hour at 37°C before plating 40 and 400 µl aliquots on solid medium.

II.M. Vector recovery using phasmids

Mycobacterial cultures for infection were grown overnight without Tween 80 by diluting a saturated stock 1/100 or 1/2 for *M. smegmatis* or BCG, respectively. Phage was added to 1 ml of culture at an MOI of 10 and incubated at 37°C. One hour before harvesting, the cells were spun down and resuspended in miniprep solution I supplemented with 20 µg/ml of lysozyme. After completion of incubation at 37°C, the DNA was prepared using an alkaline lysis miniprep (Ausubel *et al.*, 1996), precipitated with 2 volumes of ethanol, and resuspended in 5 ul of TE buffer (10 mM Tris pH 8.0, 1 mM EDTA). *E. coli* was transformed by electroporation with 1.5 µl of the solution, recovered in 1 ml of TSB, and plated on LB plates with the appropriate antibiotics.

II.N. Computer programs

BLAST-P (Altschul *et al.*, 1990) and PSI-BLAST (Altschul *et al.*, 1997) were run using the web interface on the NCBI web site (http://www.ncbi.nlm.nih.gov). Source code for PROBE (Neuwald *et al.*, 1997) was downloaded from the NCBI ftp site (ftp://ncbi.nlm.nih.gov/neuwald/probe1.0/), compiled using the GNU Compiler and run at the University of Pittsburgh on a Solaris using SunOS 5.7. The Mac executable applications, BLASTALL and FORMATDB, were obtained from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/). CLUSTAL-X version 1.8 (Jeanmougin *et al.*, 1998), a mac executable version of CLUSTAL-W was downloaded from the Bio-Web web site (http://web.tiscalinet.it/biologia/clustalx.hqx). The PHYLIP package version 3.573c was downloaded from the Washington University web site (http://evolution.genetics.washington.edu/phylip.html). All Mac applications were run on an iMac with a 333 MHz PowerPC G3 processor, 196 MB of RAM, and using the MAC OS

9.0.4 operating system. Compute pI/MW (Bjellqvist *et al.*, 1993), (Wilkins *et al.*, 1999) was used on the expasy web server (http://www.expasy.ch). Two programs, translate.pl and HTHpred, were written in PERL for this analysis and were run on the iMAC using MacPerl or on the University of Pittsburgh Unix system.

II.O. Sequence acquisition and analysis

The initial list of sequences was obtained using key words 'excisionase', 'excisase', or 'xis' with ENTREZ on the NCBI web site. Searches were done using BLAST-P and PSI-BLAST on the NCBI web site. Some searches were performed using PROBE running on the University of Pittsburgh Unix system and searching a downloaded copy (April, 2000) of either the GenBank nr database or the GenBank protein annotation database. Searches done using BLASTALL were done on miscellaneous groups of proteins that were formatted into databases using FORMATDB. A list of nucleotide records that contained integrase or that were from phages were examined for the presence of an excisionase by translating the sequence using translate.pl, an algorithm which translates a DNA sequence in all 6 reading frames and outputs it in fasta format. Sequence alignments and tree determination were done using CLUSTAL X. Trees were drawn using DrawGram from the PHYLIP package. The isoelectric focusing point, pI, was calculated using Compute pI/MW. The presence of a Helix-Turn-Helix DNA binding motif was calculated using HTHpred, an algorithm that parses a list of sequences and uses the weight matrix and method from Dodd and Egan (Dodd and Egan, 1990) to predict the probability of the motif.

III. IDENTIFICATION AND CHARACTERIZATION OF L5 XIS

(Note: This chapter was previously published as Lewis, J.A. and G.F. Hatfull (2000). "Identification and characterization of mycobacteri-ophage L5 excisionase." *Mol Microbiol* **35**(2): 350-60.**)**

III.A. Introduction

The lysogenization of mycobacteriophage L5 requires integration of the phage DNA into the host genome. While much work has been done characterizing integration (Lee and Hatfull, 1993; Peña *et al.*, 1999; Pena *et al.*, 2000), the excision reaction has proven to be more elusive. Efficient excision of L5 prophages has been observed upon induction of the inducible lysogens L5c^{ts43} and L5c^{ts63} (Donnelly-Wu *et al.*, 1993) and also must occur during phage propagation, but not all the factors necessary to reproduce the reaction *in vitro* have been identified. Many phage recombination systems mediate directionality by using a recombination directionality factor (RDF), so it is a reasonable hypothesis that L5 will utilize a similar protein, though previous to this work, none had been identified. In general, RDFs are small, basic proteins with little sequence similarity. These characteristics have made it difficult to identify this class of proteins by sequence-based searches alone.

Though the L5 *xis* gene has not been identified, there are several lines of evidence suggest a region which may contain it. Often, the RDF gene is located in close proximity to the integrase gene, and there are several small ORFs adjacent to *attP* that are good candidates. Initially, the closest gene (*34*) was considered to be a likely choice (Hatfull and Sarkis, 1993), although subsequent experimental results do not support this

assumption (C.E.A. Pena and G.F. Hatfull unpublished). After comparing the genomes of L5 and D29, it was seen that gene *34* was not conserved, but a different potential gene, *34.1*, was identified and considered to be the best candidate (Ford *et al.*, 1998).

Additional evidence for the location of the Xis gene comes from analysis of the stability of integration-proficient vectors. Plasmids containing the *attP* and *int* genes are able to efficiently transform *M. smegmatis* by integration into the host *attB* site (Lee *et al.*, 1991). The rates of loss for these plasmids are dependent on the presence of additional genes. Plasmid pMH94 contains only the *attP* site and *int* gene (Figure 14A) and shows only minimal loss after multiple generations of growth without selection (Lee *et al.*, 1991). A different plasmid, pMH5, contains a longer segment of L5 DNA which contains 10 additional ORFs that flank the *attP/int* region (Figure 14A). This plasmid transforms *M. smegmatis* at an equivalent efficiency but is lost at a much higher rate (Lee *et al.*, 1991). While the reason for this instability could result from the presence of a gene that is poorly tolerated by *M. smegmatis*, another possibility is that it is due to the presence of the *xis* gene.

III.B. Identification of the L5 *xis* gene

Based on previous observations, the L5 *xis* gene was speculated to be one of the 10 ORFs present in pMH5 but absent in pMH94 (Lee *et al.*, 1991). To verify the quality and reproducibility of the previously reported data on the stability of integrating vectors, pMH5 and pMH94 derivatives (pJL5-pJL8) were constructed that contain the sucrose sensitivity counter-selectable marker (*sacB*) (Pelicic *et al.*, 1996). Transformants containing these plasmids were grown for ~30 generations without selection and then scored for loss of the sucrose-sensitive marker on the integrated DNA marker. The percent of colonies that had lost the integrated vector was 24% for pMH5 derivates and 1.7% for pMH94 derivates.

Figure 14. Identification of L5 Xis

(A) Plasmid pMH5 contains a 4.9kb segment of L5 DNA that contains gene *32* to *42* as well as the *attP* site. Plasmid pMH94 contains only 2.1 kb of this DNA segment including *attP* and gene *33* (*integrase*). Relevant restriction sites are shown above.

(B) A series of kanamycin-resistant plasmids are shown in which segments of L5 DNA derived from pMH5 are fused to an hsp60 promoter. The genes present in each plasmid are shown with arrows indicating orientation relative to the promoter. Each plasmid (with the exception of pJL15) was transformed into a stain of *M. smegmatis* containing a hygromycin-resistant integrated vector (mc²155pJL11), and kanamycin-resistant transformants were screened for loss of hygromycin-resistance, as indicated.

(C) Four strains derived by transformation of mc²155 pJL11 with pJL22 or pMV261 were analyzed by PCR to identify the presence of different recombination sites in mc²155 pJL11 transformed with pJL22 or its parent vector (pMV261). Different combinations of primers gave different size products that indicated the presence of *attB* (361bp), *attP* (221bp), *attL* (316bp), or *attR* (266bp). All four of the pJL22 transformants contain *attB* indicating loss of pJL11. The attachment junctions, *attL* and *attR*, are present in all four transformants carrying the pMV261 vector.

Molecular weights of standards (M) have been indicated in basepairs.

A
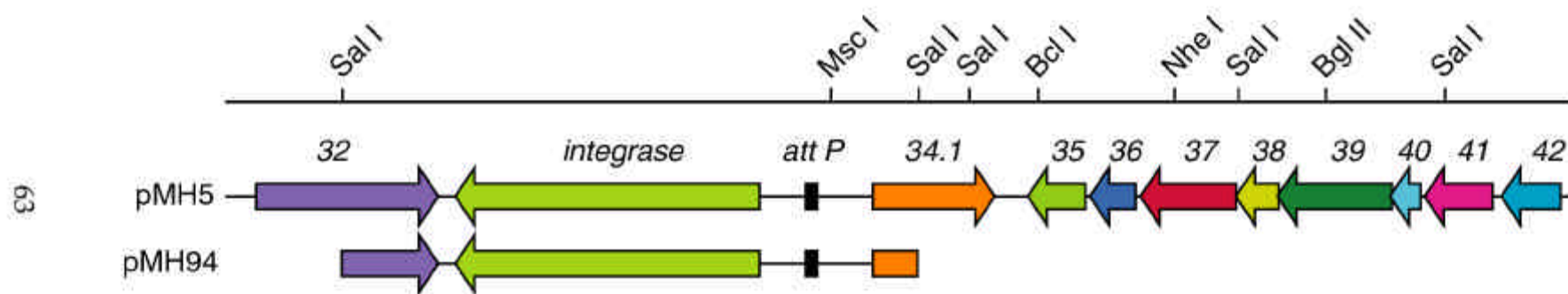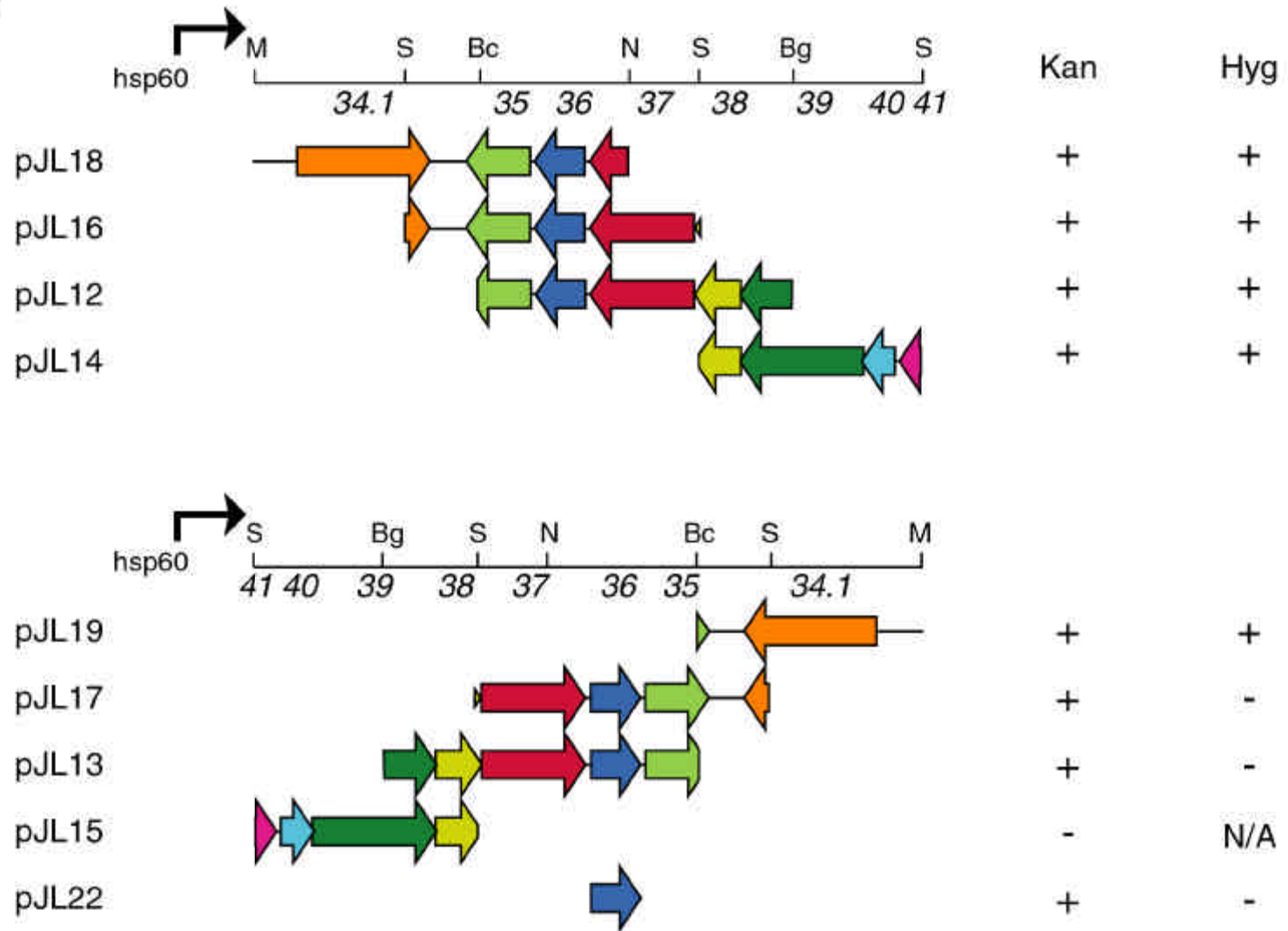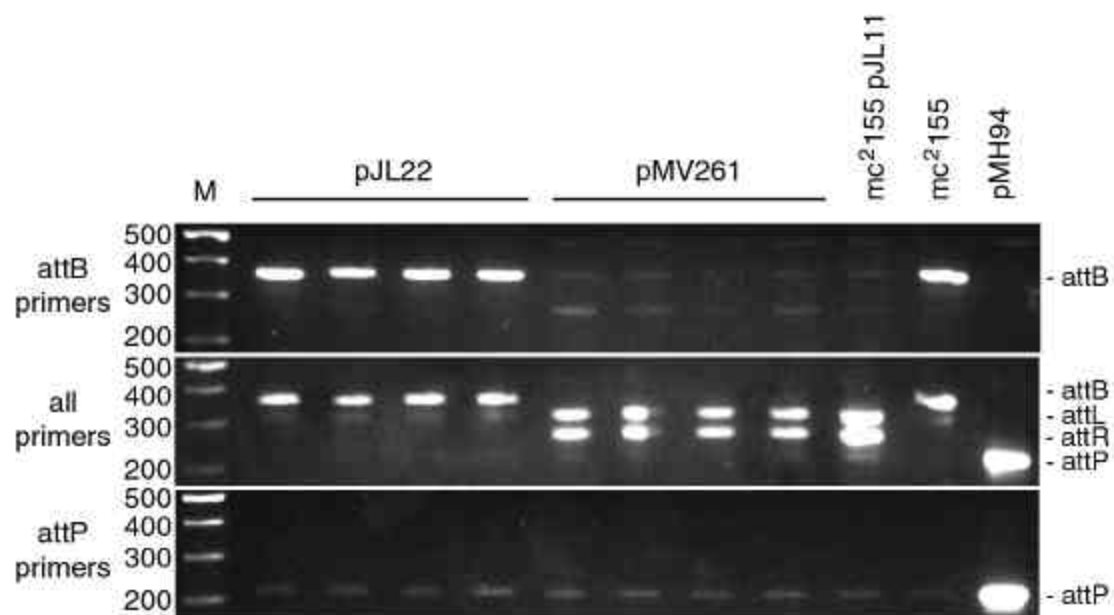
A two-plasmid assay was developed to determine which of the genes in pMH5 causes the instability. The first plasmid, pJL11, is a hygromycin resistant (hygR) derivative of pMH94 that also contains the *sacB* gene. This plasmid will serve as the reporter plasmid to score for excision. The second plasmid contains various fragments of L5 DNA from pMH5 fused downstream of the BCG hsp60 promoter in the kanamycin resistant (kanR) plasmid pMV261 (Stover *et al.*, 1991). The strain mc^{2}155 pJL11 was transformed with each of the plasmids (pJL12 to pJL22; pJL15 did not grow) and grown on plates containing kanamycin. The colonies were scored for loss of hygR, indicating excision of the integrated vector, by patching colonies to plates containing hygromycin. Only two phenotypes were observed in this experiment: either all colonies were hygR or all were hygromycin sensitive (Figure 14B). The smallest region that produces the complete loss of the integrated vector contained gene *36* and *37*. These data taken together with sequence features (discussed below) suggested that *36* is the L5 *xis* gene. A plasmid (pJL22) containing only gene *36* fused to the hsp60 promoter was constructed and used to transform mc^{2}155 pJL11. All colonies obtained from these transformations had lost the hygR marker.

A further verification that loss of hygR was linked to excision of the integrated vector was achieved through a PCR assay. This previously described assay uses four primers that flank the attachment sites; the length of the amplified products indicates the sites present (*attP*, *attR*, *attL* or *attB*) (Peña *et al.*, 1997). Four strains isolated from pMV261 and pJL22 transformation of mc^{2}155 pJL11 were used as templates for the assay. All of the pMV261 transformants contain *attL* and *attR*, while those from pJL22 contain only *attB* (Figure 14C). The ability to transform two of these cured strains with an integrating vector indicates that a functional *attB* site has been regenerated and that the integrating vector has been precisely excised.

III.C. Sequence features of L5 *36*

The identification of recombination directionality factors by sequence-based methods alone has been very difficult, but many that have been identified are small, basic proteins. A total of 63 RDFs have been identified, of which only 10 have pI values that are less than 7 and all but 15 are shorter than 100 amino acids (See Chapter VI). The gp36 protein meets these minimal requirements by being small, only 56 amino acids (a.a.) and highly charged with a predicted pI of 10.16. While gp36 does not strongly match any known Xis proteins, there are several which have weaker similarities that have been identified using BLAST. These are from *Shigella flexneri* phage V (38 % identity in 36 a.a. segment), Salmonella phage P22 (37% identity in a 32 a.a. segment) and Myxococcus phage Mx8 (37% identity in a 43 a.a. segment with a 4 residue insertion). The two most closely related proteins that were identified are gp36 from phage D29 and Rv2657c, from a gene located in a prophage like element of *M. tuberculosis* (Hendrix *et al.*, 1999). While neither of these has been shown to be an RDF, D29 is a recently diverged relative of L5 with an intact recombination system, and Rv2657c is located two ORFs away from an integrase, Rv2659c. An alignment of the proteins shows 8 completely conserved residues and several others which are in most of the proteins (Figure 15). An additional feature that has been observed in all of these proteins is a helix-turn-helix DNA-binding motif (Figure 15). It is likely that this domain provides the specificity for interaction with the substrate DNA.

III.D. Overexpression of L5 gp36 and *in vitro* recombination

To further study the L5 site-specific excision reaction, an *in vitro* excision assay was developed. Previously, an *in vitro* assay had been described for L5 integration (Lee and Hatfull, 1993); a similar system should then be possible for excision. While the integration reaction uses *attP* and *attB* as substrates, its products, *attL* and *attR*, are not

Figure 15. Alignment of putative RDF sequences

BLAST search analysis identified several proteins with similarity to L5 gp36. Two of these, *Shigella flexneri* phage V (SfV Xis) and P22 excisionase (P22 Xis) are known to be functional excisionases, while the two with better BLAST scores, D29 gp36 and *M. tuberculosis* Rv2657c, are predicted from sequence similarity alone. All five protein sequences can be aligned as shown with eight fully conserved residues and 27 others that are present in at least three of the protein sequences. All five of the proteins contain a putative helix-turn-helix DNA binding motif near the N-terminus as indicated.

helix-turn-helix

```
L5 gp36     1 --------MPPRASIQETADYLGVSTKTVRNYIADGRLKAVRLGPR-----LIRVERESVEE  49
D29 gp36    1 --------MPQRASIQQTADFLGVSTKTVRRYIADGRLKAVRLGPR-----LIRVERDSVEA  49
Rv2657c     1 MADAPPLSRRYITISEAAEYLAVTDRTVRQMIADGRLRGYRSGTR-----LVRLRRDEVDG  56
SfV Xis     1 -------MEPYSLTLDEACDFLKISRPTAINWIRTGRLQATRKDPTKSKSPYLTTRQACIAA  55
P22 Xis     1 -------MESHSLTLDEACAFLKISRPTATNWIRTGRLQATRKDPTKPKSPYLTTRQACIAA  55

L5 gp36    50 LMRPIGK-------------------------------------------------------  56
D29 gp36   50 LMRPIGK-------------------------------------------------------  56
Rv2657c    57 AMHPFGGAA-----------------------------------------------------  65
SfV Xis    56 LQSPLHTVQVSAGDGITEERKCHSSAEVKYGTPVSHCRTVKDLNSLLEQRTKGRRQNSMTS  116
P22 Xis    56 LQSPLHTVQVSAGDDITEELKCHYSAEVKPGTPVSHCRTAKDLSSLLGQRTKGRPQSFMTS  116
```

recombined in the integration assay but are the presumed substrates for excisive recombination. Therefore, to test for excision, we used a 7.2 kb plasmid (pCP24) that contains both *attR* and *attL*. The products of excision would be two fragments containing either *attP* or *attB*. Typically we linearized pCP24 so that the products would be a 6.1 kb linear fragment and a 1.1 kb circular piece of DNA.

In order to obtain significant quantities of L5 gp36, the protein was overexpressed in *E. coli* using the pET21 expression systems. A plasmid, pJL20, was constructed in which gene *36* was fused to the T7 promoter in pET21a. Upon induction, a protein with an apparent molecular weight of ~ 6 kDa was observed in strains containing pJL20 but absent in vector controls (Figure 16A). When extracts from induced pET21a cells or uninduced pJL20 cells were used in the *in vitro* recombination reaction, no evidence of excision was observed. However, the extracts containing the induced ~6 kDa protein stimulated the formation of a linear 6.1 kb fragment and a 1.1 kb circular DNA (Figure 16B and data not shown).

The purification of gpXis was assisted by the use of selective denaturation; it was determined that boiling of extracts did not significantly reduce excision activity. A scheme for purification was devised that includes boiling, clarification, ammonium sulphate precipitation at 55% and 80% of saturation and ion exchange chromatography with a weak anion resin (CM, Perseptive Biosystems). Fractions collected off the final column were used in the *in vitro* excision reaction. Only fractions containing the induced ~6 kDa protein showed evidence of excision (Figure 16C), and this activity is independent of visible contaminants when stained with Coomassie Blue. In fact, the level of recombination mimic the amount of protein with peak activity and protein staining coming from column fractions 24 and 25 (Figure 16D) though the experiment was not done quantitatively.

Figure 16. Expression of L5 Xis

(A, B) SDS-PAGE analysis and excision assay of gpXis inductions

E. coli strains containing vector (pET21a) or gp36 expression plasmid (pJL20) were induced with IPTG and samples collected at 0, 30, 60 and 120 minutes after induction. L5 Xis (gp36) is seen as a ~6kD protein in extracts from induced cells containing pJL20. Expression is not seen in uninduced containing pJL20 (0 min) cells or in induced strains containing pMV261. Crude extracts prepared from the pJL20-induced cells were used to test for the ability to mediate excisive site-specific recombination in the *in vitro* excision assay. Extracts containing L5 gp36 (pJL20, 30 - 120 min) are active in promoting excisive recombination as seen in panel B. No recombinant product was formed when using any of the other extracts (from pET21a induced or pJL20 uninduced cells).

(C, D) SDS-PAGE analysis and excision assay of purified gpXis

L5 gp36 obtained from overexpression in *E. coli* was purified to near homogeneity by boiling, clarification, and ion exchange chromatography. Protein fractions eluting from the final carboxymethyl column contain the induced ~6kD protein (gpXis) and few other contaminants as visible by coomassie blue staining. Aliquots of the fraction shown in (C) were used in the excision assay (panel D) and stimulated the generation of products in proportion to the amount of gp36 present. Maximal activity is seen with fractions 24 and 25 which contain the highest concentration of excisionase.

Molecular weights of standards (M) have been indicated in kilodaltons (A,C) or kilobases (B,D).

A

pET21a   pJL20

M   0 30 60 120   0 30 60 120

43-
29-
18.4-
14.3-

6.2-   ◄gpXis

3-

B

pET21a   pJL20   DNA only

M   0 30 60 120   0 30 60 120

10-
8-
                                        ◄Prod.
6-
5-

72

C

Column Load

Column Fractions
20   21   22   23   24   25   26   27

43-
29-
18.4-
14.3-

6.2-   ◄ gpXis

3-

D

Column Fractions
M   20   21   22   23   24   25   26   27   No Xis   Extract

10-
8-
6-   ◄ Prod.
5-
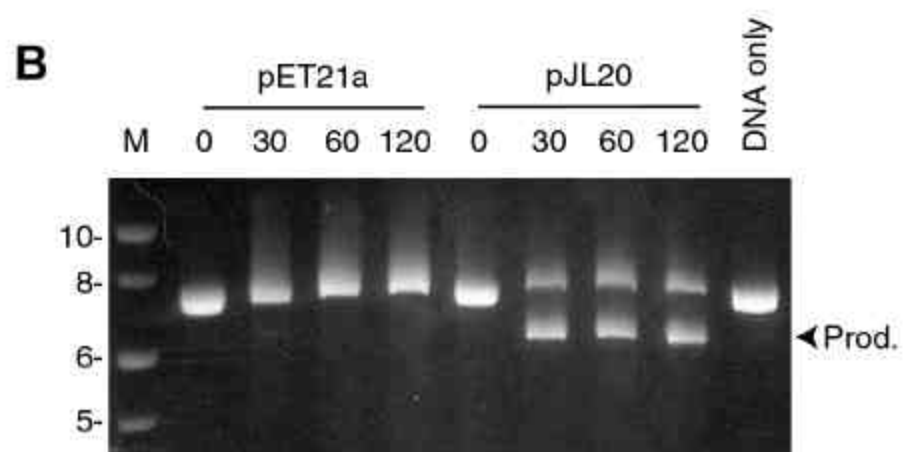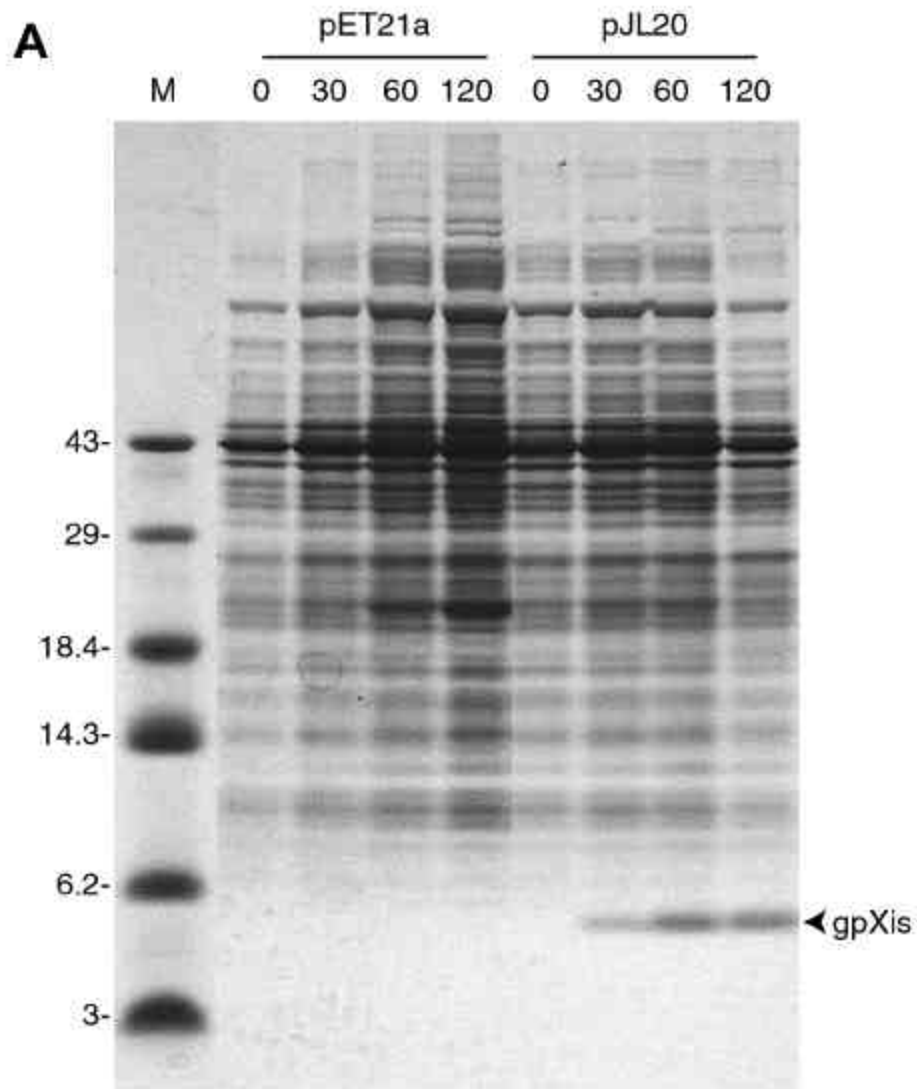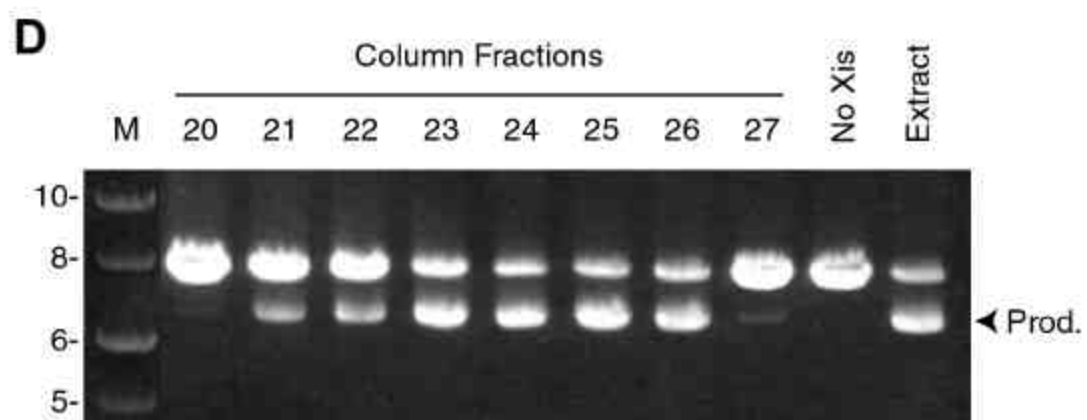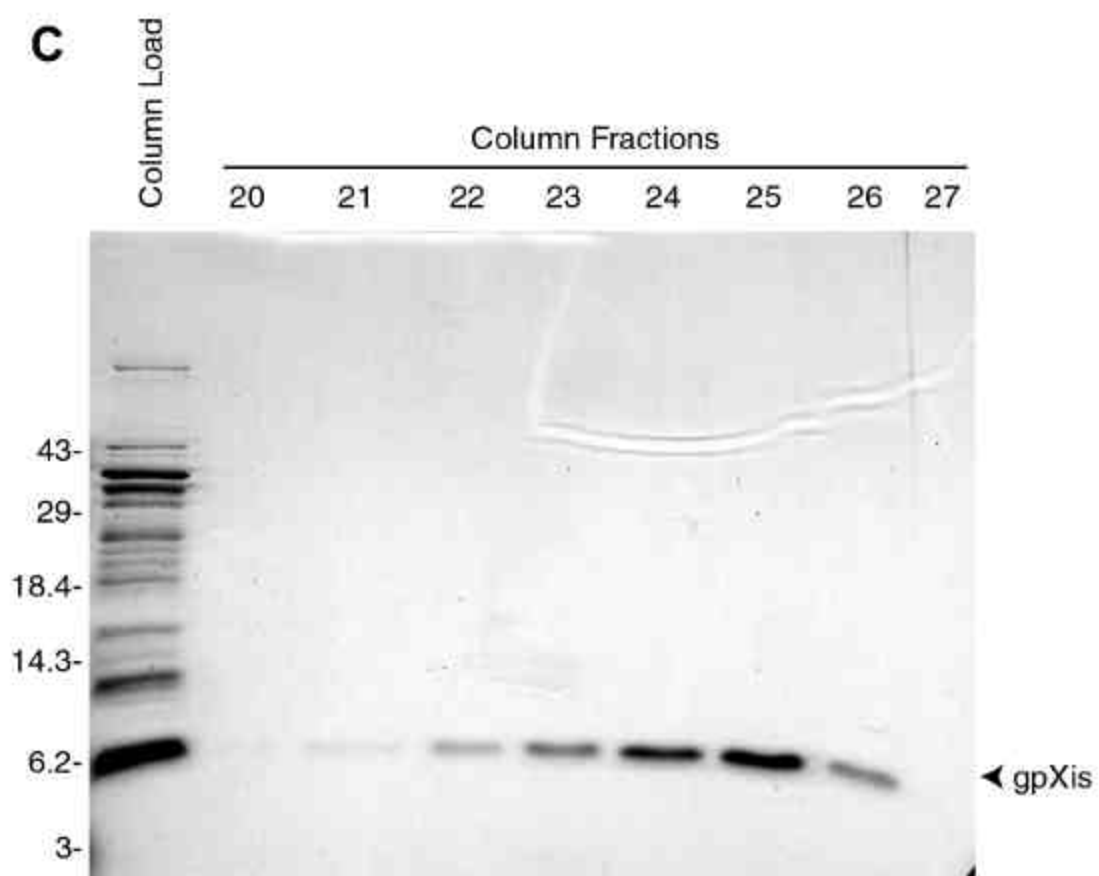
III.E. DNA substrate requirements for *in vitro* excision

It has been shown previously that the L5 *in vitro* recombination reaction is strongly stimulated by supercoiling of either one of the substrate molecules, *attP* or *attB* (Peña, C.E.A. *et al.*, 1998). To test the substrate requirement for excisive recombination, the excision reaction was performed with a variety of substrate combinations: *attL* and *attR* on a single supercoiled plasmid, both sites on a linear DNA segment, and the sites on different linear pieces of DNA (Figure 17A). The efficiency of linear substrates is not affected by the spacing between the sites, but the intramolecular reaction is more efficient.   A modest enhancement of product formation is seen when using a supercoiled substrate and viewed as a function of time (Figure 17B and C). While *attL* and *attR* are efficiently recombined in the presence of gpXis, *attB* and *attP* are not good substrates for recombination under these conditions. Evidence for this is the yield of greater than 50% of DNA recombined in the excision assay, and the presence of Xis in the integration assay results in near complete inhibition of product formation (see Chapter IV).

III.F. Host factor requirement

The integrative recombination reaction has a strict requirement for the mycobacterial integration host factor (mIHF). Previous attempts to substitute *E. coli* extracts or purified HU or IHF proteins for mIHF have not yielded any integration products (Lee and Hatfull, 1993). It was a logical assumption that mIHF would also be necessary for excision, so it was included when establishing the *in vitro* excision assay. This was verified when the excision assay was performed in the absence of mIHF, and no excisive products were formed (Figure 18). It therefore can be concluded that mIHF is an important component in the excision reaction. To determine if the necessity for mIHF in excision was as stringent as in integration, the excision reaction was carried out

Figure 17. Substrate specificity of *in vitro* excision assay

(A) Agarose gel electrophoresis of excision reactions.

Supercoiled substrates containing both *attL* and *attR* (pCP24), are somewhat better substrates than linearized forms of the plasmid in the excision reaction. Plasmid pCP24 was linearized by cutting with either of two restriction enzymes, *Xho* I or *Sac* II, which resulted in the attachment junctions being either 6.1kb or 1.1kb apart, respectively. Both substrates yield equivalent amounts of products indicating that inter-site spacing is not critical. By cutting with both enzymes, *attL* and *attR* are placed on separate linear fragments, and while intermolecular recombination is observed, it is less efficient. After the excision reactions, all DNA molecules were cleaved to completion with both enzymes and the products analyzed by agarose gel electrophoresis.

(B,C) Time course of excisive recombination.

Supercoiled and linear forms of pCP24 substrate DNA were compared in an excisive recombination time course. The linear DNA was generated by cutting pCP24 with *Xho* I, and the supercoiled DNA reactions were digested with *Xho* I after completion of the reaction. The substrates exhibit only small differences in reaction rates over the 90 minute time course. The percentage of substrate DNA that recombined is shown for the linear (■) and supercoiled (●) reactions.

Molecular weights of standards (M) have been indicated in kilobases.

A
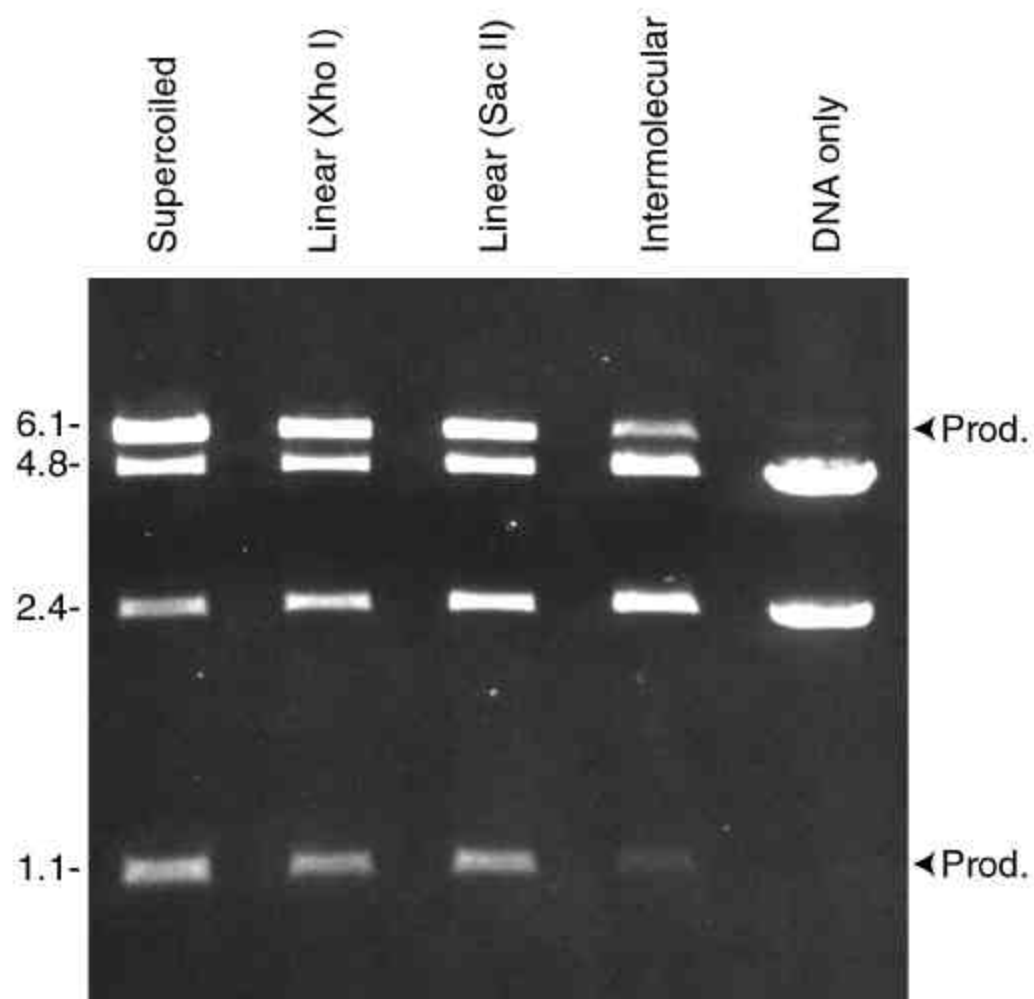
B
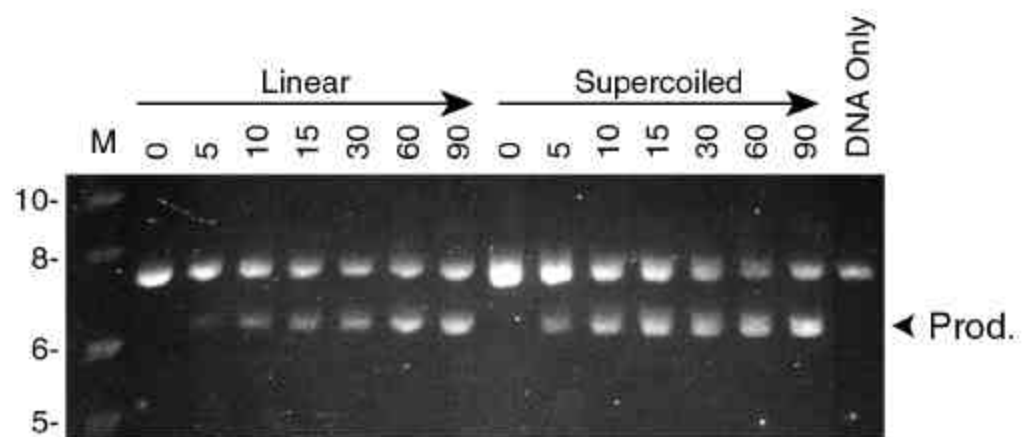
C
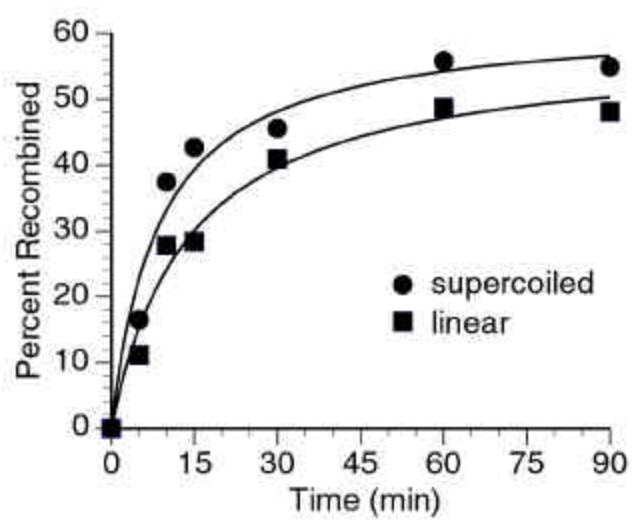
Figure 18. Host factor requirements for excisive recombination.

The requirement for mIHF in the *in vitro* excision and integration reactions was tested by substituting for mIHF with unpurified extracts of *E. coli*. The excision reactions used pCP24 (containing *attL* and *attR*) and integration reactions used the *attP*-containing plasmid pMH94 and a 45bp *attB* DNA. Excisive recombination reactions contained either DNA without added proteins (DNA only), the complete reaction with gpXis, gpInt, and mIHF, or with gpInt and gpXis plus 0, 1/10μl, 1/3μl or 1 μl of an *E. coli* extract, as indicated. The integration reactions were performed similarly except that gpXis was not included. While an *E. coli* extract cannot substitute for mIHF in integration, it can at least partially replace mIHF in the excision reaction. Substrates and products for both excision and integration are indicated. Molecular weights of standards (M) are given in kilobases.
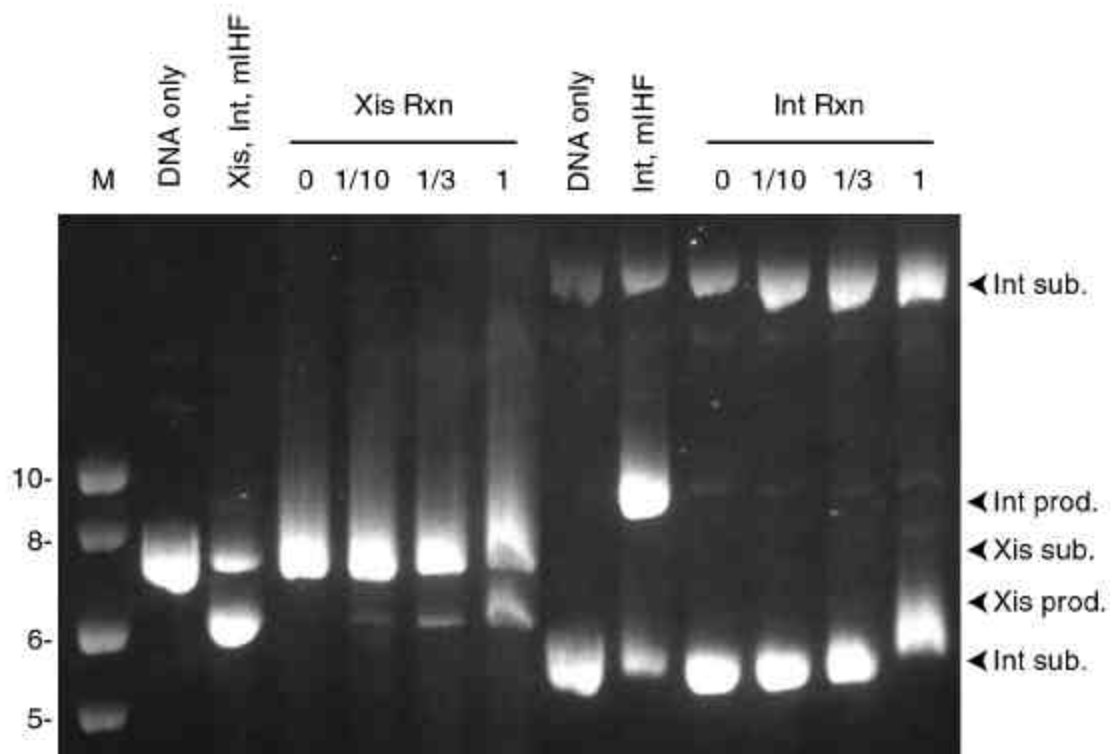
in the presence of *E. coli* extract. With relatively high levels of extract, there is a moderate stimulation of excision, but the same amount of extract has no effect on integration. While an *E. coli* extract is capable of substituting for mIHF in excision, it is not clear what component is responsible.

III.G. Requirements for excision *in vivo*

While the *in vitro* excision reaction requires Int and Xis, the *in vivo* reaction might have different needs. Mycobacterial cells could harbor some factor that takes their place, though none had previously been identified for L5 integration (Peña *et al.*, 1997). The requirement for bacteriophage L5 derived Int and Xis was demonstrated by testing for loss of an integrated vector with different combinations of the proteins. A new vector (pJL35) was derived from pJL11 in which the *int* gene was nearly completely deleted. This vector can be introduced into mycobacteria by using a previously described method in which an *attP* containing plasmid is co-electroporated with a suicide plasmid that contains the *int* gene (Peña, C.E. *et al.*, 1998). Strains contain either pJL11 or pJL35 were then transformed with an Xis-expressing plasmid (pJL22) or its parent vector (pMV261). These transformants were plated on medium which selected for the extrachromosomal plasmid (pJL22 and pMV261) alone or double-selected for the integrated vector and the extrachromosomal vector. The presence of *int* (pJL11) and *xis* (pJL22) prevented any growth on the double-selected medium despite the ability of pJL22 to efficiently transform mc²155pJL11 when not selecting for the integrated DNA. It has already been shown (see above) that these conditions correlate with excision of the integrating vector. When the other three combinations of plasmids were tested, no excision was detected, indicating that both Xis and Int are required for efficient excision.

The stability of these transformants was further examined by growing them for ~30 generations and selecting only for the extrachromosomal vector. While the pJL11 transformants showed the same low-level loss as previously described for pMH94

derivatives (above), strains containing pJL35 that lacked *int* showed no detectable loss. These results indicate that the minimal instability found with the integrating vectors is the result of an Int mediated reaction, and greater stability can be achieved by removal of the *int* gene.

To test the requirement for the P6 and P7 arm-type sites, which are not required for integration (Peña *et al.*, 1997), the ability to excise an integrating vector lacking them was tested. *M. smegmatis* mc²155 was transformed with an integrating plasmid (pCPΔR13) containing a shortened *attP* or one with full-length *attP* (pMH94). Strains containing each of these integrating vectors were transformed with plasmids containing the *xis* gene (pJL33 and pJL34) or its parental plasmid lacking *xis* (pJL32). When these transformants were grown on medium selecting for both the extrachromosomal and integrated vectors, the strains containing Xis expressing plasmids show the same loss of viability as described above. All three extrachromosomal plasmids efficiently transform either strain containing an integrated plasmid when not selecting for both vectors. These data suggest that efficient excision *in vivo* does not require the P6 and P7 arm-type sites.

III.H. Discussion

Gene *36* has been shown to encode the functional RDF of mycobacteriophage L5. The gene product is a 56-residue protein that has a predicted pI of 10.16. The protein does not have sequence similarity to Lambda Xis, but this is not surprising in light of more recent data on RDF classification (see Chapter VI). Using simple sequence comparison methods, the P22 phage Xis was identified as the most similar RDF. The L5 and P22 proteins contain a putative helix-turn-helix DNA binding motif that is likely to be responsible for interaction with the attachment site. L5 Xis has a functional similarity to the Lambda Xis, in that both inhibit integration and stimulate excision.

The protein that is most closely related to L5 gp36 is the homologous protein from mycobacteriophage D29. Since D29 is a lytic phage, at first glance this might seem to be contradictory. However, it is believed that D29 recently suffered a deletion of the phage repressor gene and previously was a temperate phage. This is further supported by the ability of D29 to form a lysogen in the presence of the L5 repressor (Ford *et al.*, 1998). The resulting D29 prophages are integrated in the same *attB* that L5 uses and are spontaneously induced to form phage particles, which presumably requires D29 gp36 for excision. The relatedness of the D29 and L5 integration systems is exemplified by the interchangeability of the Int and *attP* from the two systems (Peña, C.E. *et al.*, 1998). The region near the *xis* gene shows some variation between the two phages. For example, L5 *35* is absent in D29 and while both phages have a leftward terminator, the sequence of the two is unrelated. Despite the variation in this area of the genome, gene *36* has been conserved, suggesting that it is an important gene, such as a phage RDF would be.

Three genes coding for proteins that are similar to L5 gp36 have been identified in *M. tuberculosis*: Rv2657c, Rv2310 and Rv1584c (Cole and Barrell, 1998). Rv2309 is adjacent to the gene Rv2310, which appears to be a partially deleted integrase gene. The strongest match is with Rv2657c, which is located near an integrase gene, Rv2659c, within the φRv2 prophage-like element (Hendrix *et al.*, 1999). Rv1584c is in a similar prophage-like element (φRv1) with the same linear arrangement of *int* and *xis* genes (Hendrix *et al.*, 1999), but the associated recombinase is a member of the large serine-integrase family. This protein family shares a catalytic domain with the invertases and resolvases, but it has a different carboxy-terminal domain (Hatfull and Grindley, 1988; Thorpe and Smith, 1998). In several systems that use a large serine recombinase, RDFs have been identified (Ramaswamy *et al.*, 1997; Breuner *et al.*, 1999), but none of the proteins shows similarity to Rv1584c or L5 Xis (see Chapter VI). However, there is some experimental evidence that Rv1584c acts as an RDF in the φRv1 recombinase

system (Bibb and Hatfull unpublished). Despite the need for an accessory protein, the mechanism of directional control in serine integrases is still unclear.

The ability to transform *M. smegmatis* with plasmids containing the *attP* and *int* genes such as pMH94 and pMH5 has been well-documented (Lee *et al.*, 1991). While it also has been noted that pMH5, which contains more L5 DNA than pMH94, is less stable, the reason for the difference was unknown. The cause for the alternate behavior can easily be explained by the presence of the L5 *xis* gene. The discrepancy in behavior is exaggerated in BCG where pMH5 transforms at 10^{-4} the efficiency of pMH94 (Lee *et al.*, 1991). Furthermore, all pMH5 transformants that have been examined contain a deletion (Lee *et al.*, 1991). This suggests that some gene(s) on the plasmid is not tolerated in BCG. The regulation of expression in BCG is likely to vary from *M. smegmatis* and thus the amount of Xis could be increased to a level which prohibits integration. Toxic effects on the cell cannot be ruled out, as there are several lines of evidence showing that high levels of Xis are harmful. It has been observed that the presence of pJL22, which expresses gp36 fused to the hsp60 promoter, slows the growth of *M. smegmatis*. Additionally, pJL22 is not able to transform BCG, and only plasmids containing a reduced expression of the *xis* gene are tolerated (Chapter V and T. Parish personal communication). When genes *39* and *40* are expressed using the hsp60 promoter (pJL15), the plasmid is not tolerated. While there are likely to be many genetic and metabolic differences between the fast-growing *M. smegmatis* and the slow-growing BCG, the reason for increased expression or sensitivity to certain gene products is unknown.

The timely control of recombination is critical for the longevity of the phage. This requires an inhibition of integration during lytic growth, integration of the prophage upon formation of a lysogen, inhibition of excision during lysogeny and excision upon induction into lytic growth. The location of the *xis* gene offers a means by which the state (integrated or excised) of the phage DNA can be controlled. L5 gene

36 is located at the end of the operon located on the right arm, which has two promoters (P_{left} and P_{71}) to drive expression. Regulation of this transcript is mediated by a novel system in which gp71 binds to a series of stoperator sites spread throughout the genome (Brown *et al.*, 1997). The expression of *36* will be down-regulated by gp71 binding to multiple stoperators during lysogeny and expressed during lytic growth, just like the whole right arm. In the absence of Xis production, the phage DNA will integrate and not be able to excise. Upon induction of lytic growth, the Xis protein will be made and thus the integrated prophage will be excised and further integration will be inhibited.

While many of the requirements for excision have been examined, the functional role of gp36 has not been determined. Based on several observations, speculations can be made about possible mechanisms. In the absence of Xis, *attL* DNA remains in an intasome like complex and *attR* is only observed as free DNA. This would suggest that Xis is required to form an *attR* complex. The difference in mIHF requirements for integration and excision adds to this hypothesis. In gel shift experiments, the intasome complex can be formed in the presence of *E. coli* extracts and the absence of mIHF (A. Lepisto and G.F. Hatfull, unpublished). While the intasome and likely the *attL* complex do not have a strict mIHF requirement, integrative recombination does. If the *attL* complex is an excisive intermediate, synapsis with an *attR* complex that requires Xis but has a reduced specificity for mIHF could be used for excision. A more complete analysis of the functional role of the Xis protein is required to understand the control of directionality and excision in phage L5.

IV. MECHANISM OF L5 EXCISION

IV.A Introduction

The assembly of intricate nucleoprotein complexes is required for integrase-mediated site-specific recombination. In the Lambda site-specific recombination system, unique complexes have been identified for integration and excision. While the assembly of L5 integration complexes has been well studied (Pena *et al.*, 2000), no work has been done examining the mechanism of L5 excision.

There are several lines of evidence that support speculation on how L5 Xis might function. Every RDF that has been studied functions by binding DNA to affect the architecture of the nucleoprotein complexes that form. The presence of a putative helix-turn-helix DNA binding motif suggests that L5 Xis will act similarly by binding DNA. The location of binding can be hypothesized by examining the products from the *in vitro* integration reaction; DNA containing *attL* was found in complex with Int and mIHF, while *attR* was found free. In order for recombination to occur, the *attR* DNA must be able to form a complex that can synapse with *attL*.

Initial predictions suggest that L5 Xis is likely to contribute to the stabilization of an *attR* complex that is part of a synaptic complex in conjunction with an *attL* complex., and the role of Xis is likely to be mediated by direct binding to the DNA. The two main methods that will be used to examine the assembly of excisive complexes will be the native gel electrophoresis assay and the DNase I nuclease protection assay.

IV.B. Binding specificity of L5 Xis

To determine the role of L5 Xis as a DNA binding protein, native gel electrophoresis was used to study protein-DNA interactions. Three attachment sites or junctions (*attP*, *attR* or *attL*) were incubated without protein or with a range of L5 Xis concentrations (0.096 – 3.2 μM). Upon electrophoresis of these samples, specific bands were seen only with *attP* and *attR* DNA and only at the highest concentration of Xis (3.2 μM) (Figure 19A). Some smearing was observed on the gel with all three substrates, but it was observed only at the highest Xis concentration with *attL* or with slightly less Xis (0.96 μM) in the cases of *attP* and *attR* DNA. To verify that the difference in binding to *attL* and *attR* is not the result of variation between individual reactions, a similar series of reactions were carried out, but both DNA molecules were included in the same samples (Figure 19B). Once again, the only specific band is seen at the highest Xis concentration. Furthermore, while nearly all of the *attR* is bound to Xis, there is no visible change in the amount of free *attL*. Another interesting observation is that while the majority of the *attP* and *attR* DNA is in complex at the high Xis level, the same complex is completely absent at lower protein concentrations. This pattern of complex formation is consistent with Xis binding in a cooperative manner.

If L5 Xis binds specifically within *attR* DNA, the location of DNA-protein interactions can be determined by the DNase I nuclease protection assay. The DNase I digestion was performed using *attR* DNA without proteins or with a range of Xis concentrations (32 nM – 3.2 μM). When either strand was looked at, the only region that showed any difference in the presence of Xis was between the P2 and P3 arm-type sites (Figure 20). While the overall region of protection is ~45 bp, this includes several bases in which there is no protection or even an enhancement. Interestingly, these are located 10-11 bases apart, and are likely to fall on the same side of the DNA helix. This type of digestion pattern is often associated with a tight bend in a DNA molecule,

Figure 19. Formation of Xis complexes

One of three recombination substrates, *attP*, *attR* or *attL* was combined with a variety of Xis concentrations (96 nM – 3.2 µM) and run on a native gel (A). The only interactions observed are at high Xis concentration (960 nM – 3.2 µM). Furthermore, the only specific bands are at the highest concentration, and only with *attR* and *attP*. Some smearing is seen with *attL*. To verify that the difference between *attL* and *attR* binding was not due to differences in the reaction conditions, both DNAs were included in the same reaction (B). Under these conditions, nearly all of the *attP* is in complex at the highest Xis concentration, but no change in free *attL* is detectable. In both gels, the formation of complex increase over a very short concentration difference, so Xis binding is likely to be cooperative.

A

attP   attR   attL

B

Lane labels (top): DNA Only | Xis (μM): 0.032 | 0.096 | 0.32 | 0.96 | 3.2

← attR / Xis complex

Free attR →

Free attL →

90
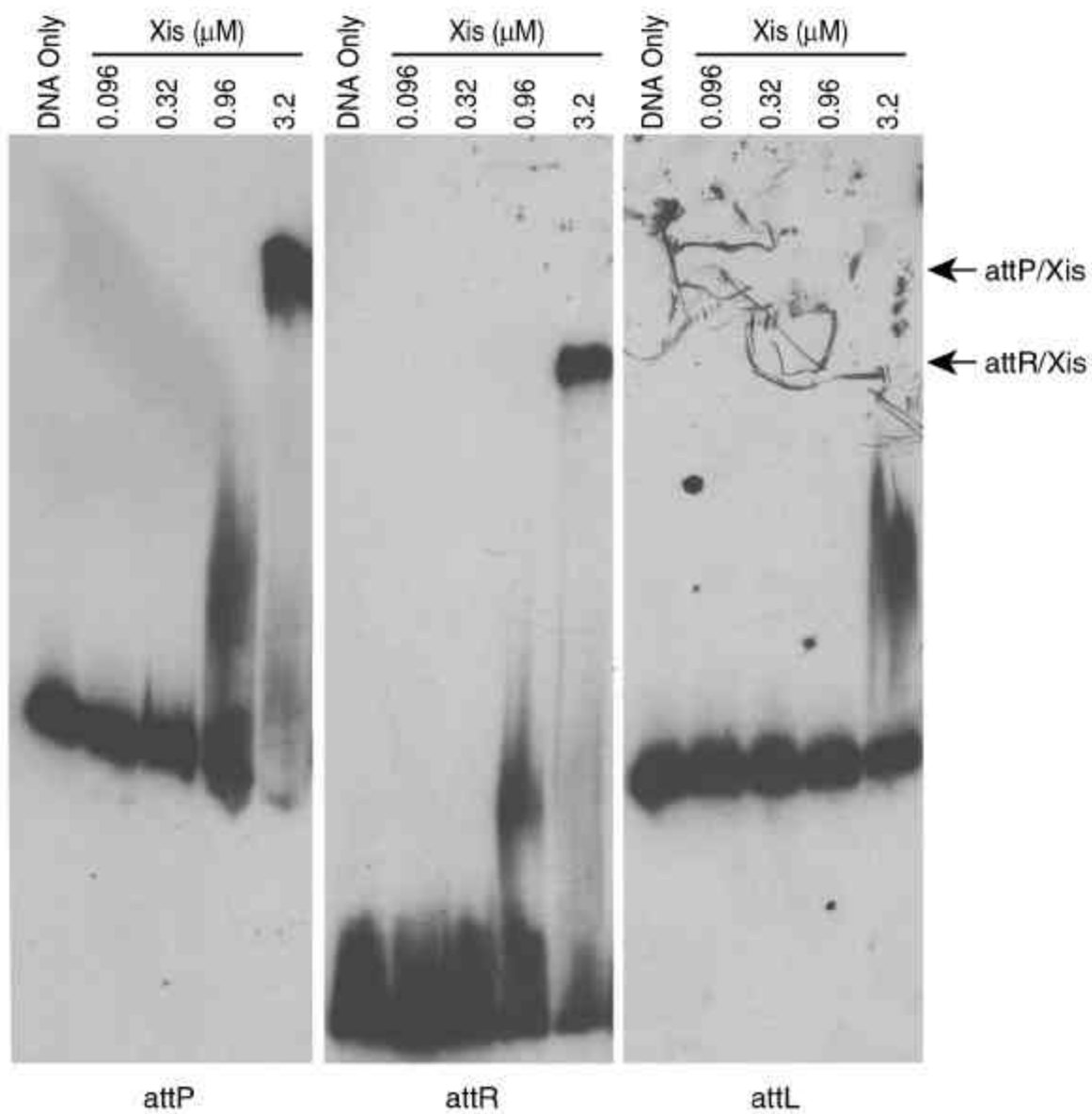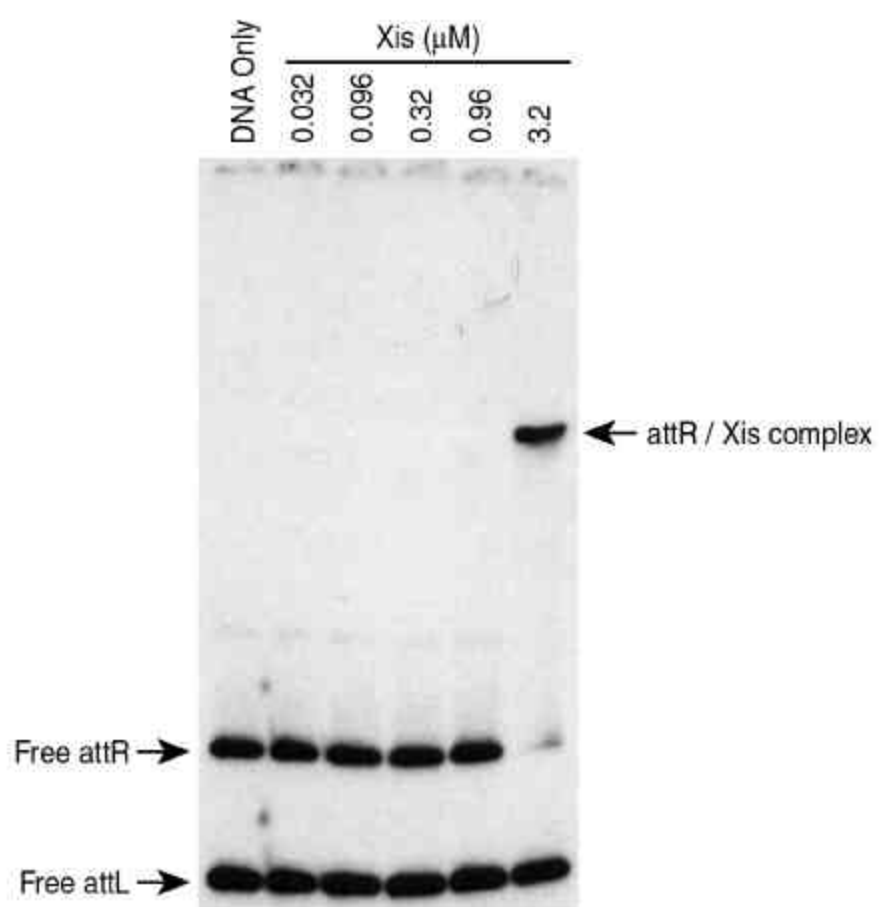
Figure 20. Dnase I protection assay of *attR* with Xis

(A)DNase I nuclease protection assays were performed with *attR* DNA in the presence of a variety of concentrations of Xis (32 nM – 3.2 µM). The only region of either strand that showed protection from digestion was between the P2 and P3 arm-type sites. There is a series of unprotected bases or enhancements spaced evenly (10-11 bases apart) throughout the region. (B) A Planar representation of the footprinted DNA is shown with triangles indicating protected phosphates and pluses indicating enhancements. This pattern shows that L5 Xis binds to this region, but one face of the helix is accessible to the nuclease.

A

Bottom Strand

Top Strand

B

P1

P2

P3
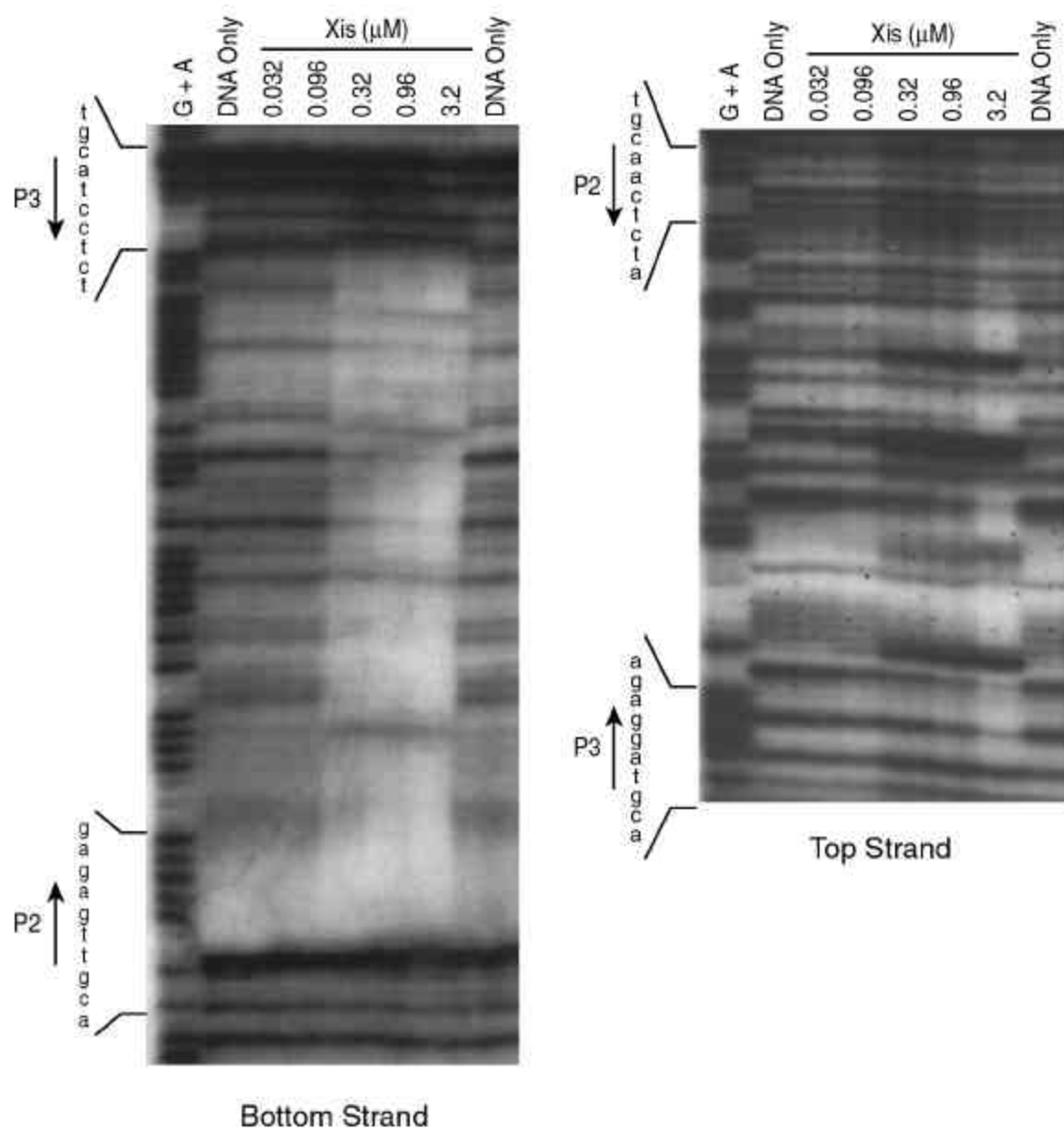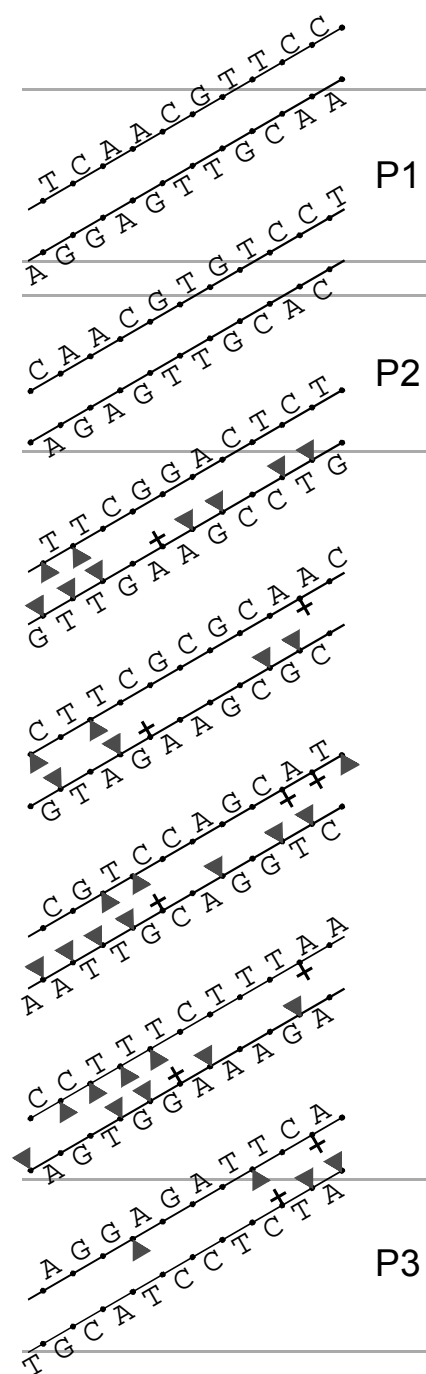
which allows easier access for the nuclease on the outer edge. While the overall area of protection is similar on each strand, a notable difference between the two is the degree of enhancement. On one strand, there is either no protection or moderate enhancement, while on the other strand the enhancements are quite pronounced. This shows that L5 Xis binds to a specific region between the P2 and P3 arm-type sites and in the absence of other proteins, does not bind tightly with the rest of the *attP* site.

An apparent discrepancy between the footprinting and native gel data is the concentration at which protein binding is observed.   In the native gel, an electrophoretically stable complex is only observed at the highest Xis concentration (3.2 μM). However, the footprinting analysis shows protection by Xis at a concentration 10-fold lower (320 nM). One possibility for this difference is in the conditions of each assay; while the components of each reaction are identical, the nuclease protection analyzes complexes in solution and is responsive to more dynamic interactions. A second option is that the interactions required to form the complex at the highest concentration are some how different and more stable than those that produce the protection in the DNase I assay.

IV.C. Higher order *attR* complexes

While it is interesting to see that L5 Xis acts as a specific DNA binding protein on its own, the *attR* complex required for integration will almost definitely contain Int and may also require mIHF. In order to look at complex formation in the presence of these proteins, EMSA was performed in which *attR* DNA was combined with different combinations of Int, mIHF and Xis. Three different complexes were observed: one that contains only Xis, another that is dependent on Xis and Int and a third, which requires all three proteins (Figure 21). The presence of Int and mIHF in the absence of Xis is not sufficient to create an electrophoretically stable complex. The Xis-only complex mimics what was described above and is present in all samples that contain the highest Xis

94

Figure 21. Formation of multi-protein *attR* complexes

By using native gel electrophoresis with *attR* and a variety of protein combinations, three unique complexes can be observed in the presence of Int, mIHF and Xis. The two faster moving complexes have nearly identical mobility, but the protein concentrations required for their formation differ greatly. One is only seen at the highest Xis concentrations regardless of which other proteins are present, while the other is seen in all concentrations of Xis used (96nM – 3.2 μM), but only in the presence of Int and mIHF. A third more diffuse band that has a slower mobility than the other two is only seen in moderate to high concentrations of Xis (320nM – 3.2 μM), and it requires Int. The importance of this third band is questionable, because it is dependent on an *attR* fragment containing flanking plasmid DNA is used and is absent when a shorter but still recombinagenic fragment is used.
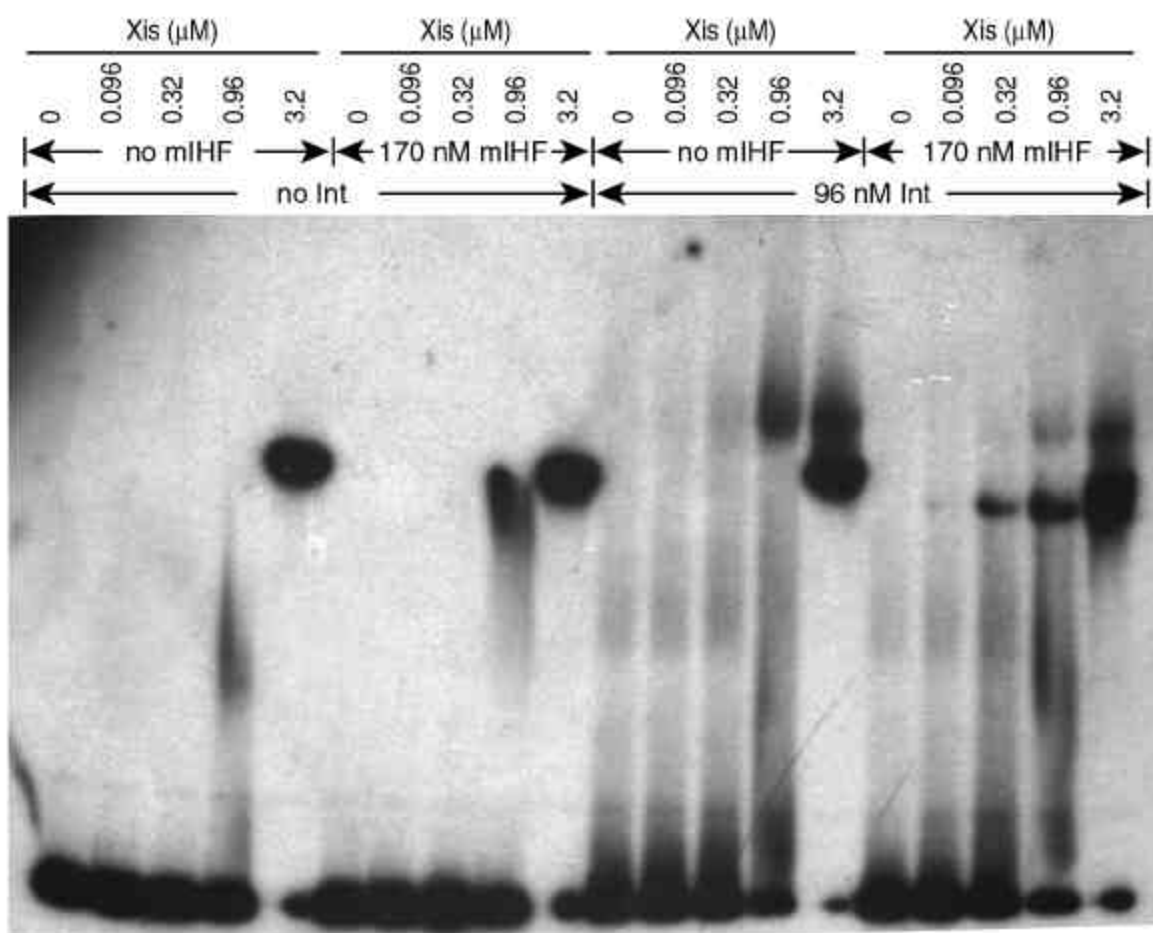
concentration despite the other proteins present. The second complex, which requires Int and Xis, is seen faintly at 0.32 μM Xis and increases with increased concentrations of Xis. However, despite the presence of this complex in this experiment, it is not observed in a reproducible or consistent manner, so its role in the excision reaction is dubious. Furthermore, footprinting analysis shows no difference between Xis-only digestion and the corresponding Xis and Int patterns (data not shown). The third complex that requires all three proteins is present at the lowest Xis concentrations and increases with more Xis. Since this complex is present over an Xis concentration range similar to those required for excision, this is the best candidate for an excisive *attR* complex.

While complexes have been observed using native gel electrophoresis, the makeup of these complexes is unclear, and furthermore, additional protein-DNA interactions could be present that are not stable enough to be seen in the assay. To examine the framework of the identified complexes and to seek out others, DNase I footprinting was performed using *attR* DNA with Int, mIHF and either no Xis or with a range of Xis concentrations (32 nM – 3.2 μM). The concentrations of Int and mIHF (96nM and 170nM respectively) were in a range that yields efficient complex formation and excisive recombination, but neither protein shows protection by itself (data not shown). However, the mIHF alone footprint shows some moderate enhancements, the nature of which is not clear, as these had not been observed in previous assays with mIHF (Pedulla *et al.*, 1996). The combination of Int and mIHF (Figure 22) produces large regions of protection within *attR*. The most significant protection is in the core and immediately adjacent regions and is interspersed with evenly spaced enhancements. The left end of *attR*, from the P3 site through the P1 site, shows little difference from DNA only, and notably, the P1 and P2 arm-type sites are not protected. The protection pattern seen here is quite similar to that previously seen in this region of the *attP* intasomes DNA (Peña *et al.*, 1999). To achieve the amount of protection

Figure 22. Interaction of Int, mIHF and Xis with *attR*

(A) A DNase I footprinting analysis was performed on *attR* DNA in the presence of Int, mIHF and increasing concentrations of Xis. The concentrations of Int and mIHF used in these reactions show little or no interaction with *attR* on there own, but they exhibit cooperative binding and show a pronounced effect on the pattern of *attR* digestion. The protection demonstrated by Int and mIHF is at the core region and to its left through to the P3 arm-type site. Within this region, there are also periodic enhancements or unprotected bases. With the addition of Xis, the most pronounced change is protection in between the P2 and P3 sites that is quite similar to Xis only binding. Other notable changes are protection of the P1 and P2 arm-type sites and increased protection at the core. (B) Specific locations of protection (triangles) and enhancements (pluses) have been indicated.

A

B

TCGTGAACAAGGCTACCGGGTTGCAACTCCTGTGCAACTCTCAGGCTTCAACGCGCTTCTACGAC

AGCACTTGTTCCGATGGCCCAACGTTGAGGACACGTTGAGAGTCCGAAGTTGCGCGAAGATGCTG

P_{int}

P1   P
            2

CTGCAATTTCTTTCCACTTAGAGGATGCAGCCGAGAGGGGGTAAAAACCTATCTTGACCGGCCCA

GACGTTAAAGAAAGGTGAATCTCCTACGTCGGCTCTCCCCCATTTTTGGATAGAACTGGCCGGGT

P3

TATGTGGTCGGCAGACACCCATTCTTCCAAACTAGCTACGCGGGTT

ATACACCAGCCGTCTGTGGGTAAGAAGGTTTGATCGATGCGCCAA

Core

observed in this footprint, the majority of the *attR* DNA must be bound, but under identical conditions no complex was observed using native gel electrophoresis (Figure 21). With the addition of Xis a new pattern develops in which protection and enhancements in the region between the P2 and P3 sites are observed and the P1 and P2 arm-type sites are protected. The digestion pattern between P2 and P3 is nearly identical to Xis alone footprints but with a slight change in position of the enhancements. Protection of the P1, P2 and core sites, taken together with the likely bending indicated by the enhancements, is consistent with a model in which Int molecules form a bridge between the arm-type sites and core with a sharp bend stabilized by Xis and mIHF proteins.

IV.D. The *attL* complex

Excisive recombination requires both *attL* and *attR* to be together in complex. While the formation of *attR* complexes requires Xis, the *attL* DNA has been observed in complex as a product of *in vitro* integration (Pena *et al.*, 2000). It is still possible that Xis interaction with *attL* is required to form an excisive complex. A native gel with Xis as the only protein (Figure 19A and B), showed that Xis binds *attR* and *attP* better than *attL*, which is consistent with Xis being required only to the left of core. To ensure that no additional *attL* complexes form that require any combination of Int, mIHF and Xis, mobility of *attL* DNA was observed in the presence of mIHF, a variety of Int concentrations (9.6 nM – 96 nM) and either without Xis or with a range of Xis (96nM – 3.2μM) (Figure 23). Only one unique complex is observed with all combinations of proteins. While the amount of complex is dependent on the quantity of Int, there is no detectable effect by Xis at any Int concentration. At the highest Int levels, which are the same used for *in vitro* excision, nearly all of the *attL* DNA is in complex, so it is likely that this complex is the predominant *attL* species present in the excision reaction.

Figure 23. An *attL* complexes independent of Xis

The formation of an *attL* complex can be detected by native gel electrophoresis. This complex is dependent on Int and mIHF, but Xis is not required. To determine if Xis had any affect on this complex or contributed to another, reactions containing mIHF and a variety of Int (9.6 nM – 96 nM) and Xis (96 nM – 3.2 μM) concentrations were run on a native gel. The presence of Xis had no detectable effect on *attL*.
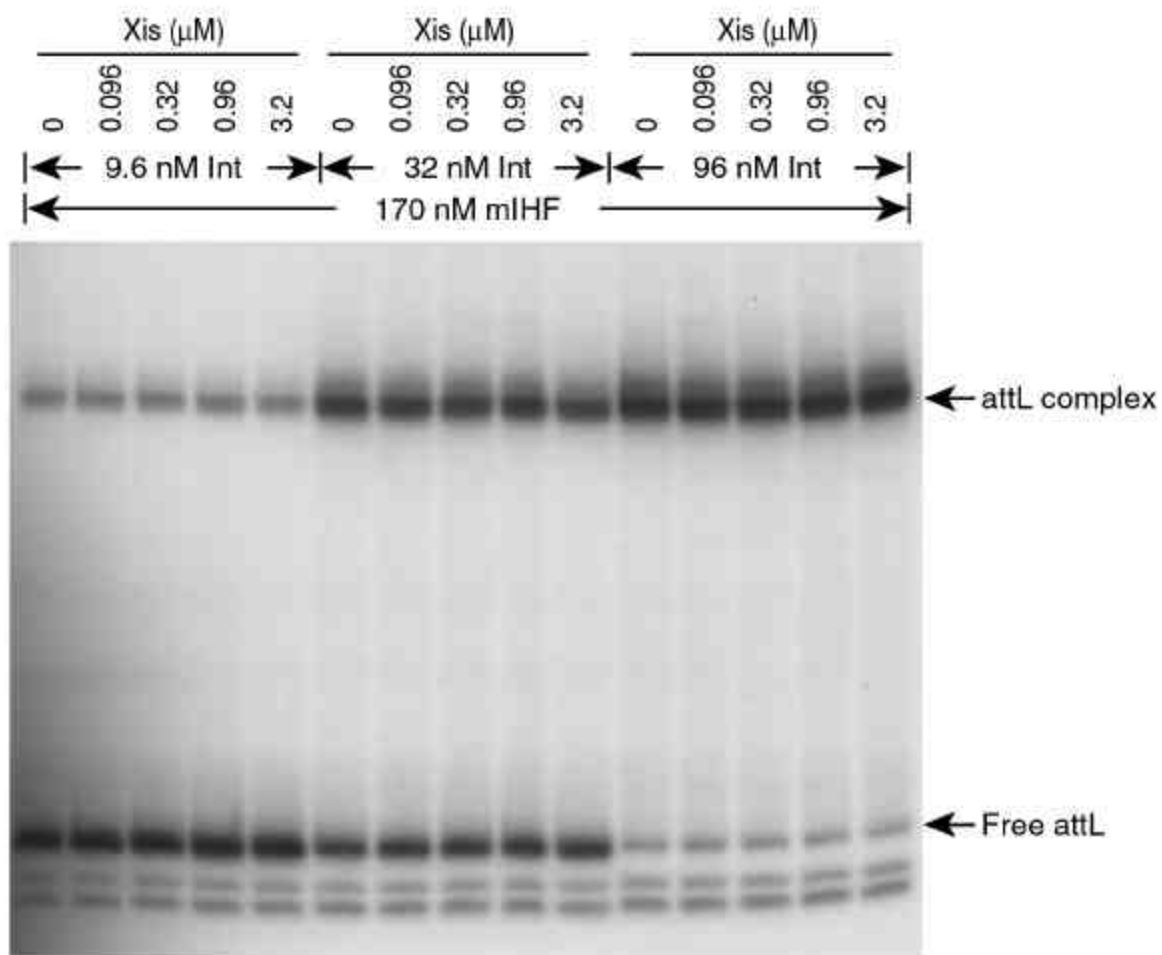
IV.E. Synapsis and excision

Putative excisive recombination complexes have been identified for both the *attL* and *attR* attachment junctions. However, a complex containing both DNA substrates is required for excision to occur. In an attempt to identify a synaptic complex, native gel electrophoresis was performed with both substrates present in the same reaction with Int, mIHF and either no Xis or a variety of Xis concentrations (96 nM – 3.2 μM). Furthermore, the reactions were incubated on ice or at 37°C, and in addition to using wild type integrase, a subset of the reactions used a catalytic defective mutant Y349F (C.J. Wadsworth and G.F. Hatfull unpublished observations). When either integrase was incubated on ice, the primary complexes are the same as those seen with *attR* alone (Figure 24A). A second complex was seen, but its formation does not require Xis and in fact, the quantity of the complex is reduced by Xis. When the reactions containing wild type integrase are incubated at 37°C, additional complexes are formed in the presence of Xis, but these are not seen with the catalytic mutant of Int. One possibility is that the mutant Int is defective in complex formation, but the mutant Int is effective at forming all the complexes seen on ice and also is able to generate integrative complexes (C.J. Wadsworth and G.F. Hatfull unpublished). An alternate explanation is that the complexes are formed from the products of excision (*attP* and *attB*). To look at the DNA content of the complexes, a reaction containing wt Int, mIHF and 3.2 μM Xis was first separated by native gel electrophoresis, and then the lane was sliced out of the gel, soaked in an SDS solution and run on a denaturing gel (Figure 24B). The results of this experiment show the DNA content of the four dominant complexes: two of them are the *attR* complexes previously characterized, one contains only *attP* and the fourth has both *attB* and *attP*. While new complexes are observed when *attR* and *attL* are combined, they do not contain these substrates, but instead are composed of the

Figure 24. Search for an excisive synaptic complex

In an attempt to try to capture an excisive synaptic complex, native gel electrophoresis (A) was used in conjunction with two temperatures of incubation (0°C or 37°C) and either the wildtype (wt) integrase or a Y349F mutant; *attR* and *attL* were included, but only *attR* was labeled. On ice, little difference is seen between the wt and mutant. Both show the two *attR* complexes previously described, and in the absence of Xis a synaptic complex is detected. Since it is lost with the addition of Xis, it is unlikely that it is involved in excision. A variety of complexes are formed in the presence of Xis and wt Int at 37°C that are not seen with the mutant Int. To determine if these contain substrate or product, a lane from a 3.2 μM Xis reaction was soaked in SDS and run through a second dimension denaturing gel. This clearly shows that *attR* is only present in the previously described complexes and that the new complexes contain *attP* and *attB*.

A

B

Electophoretic Mobility Shift Assay

SDS Denaturing Gel

← attP

← attR
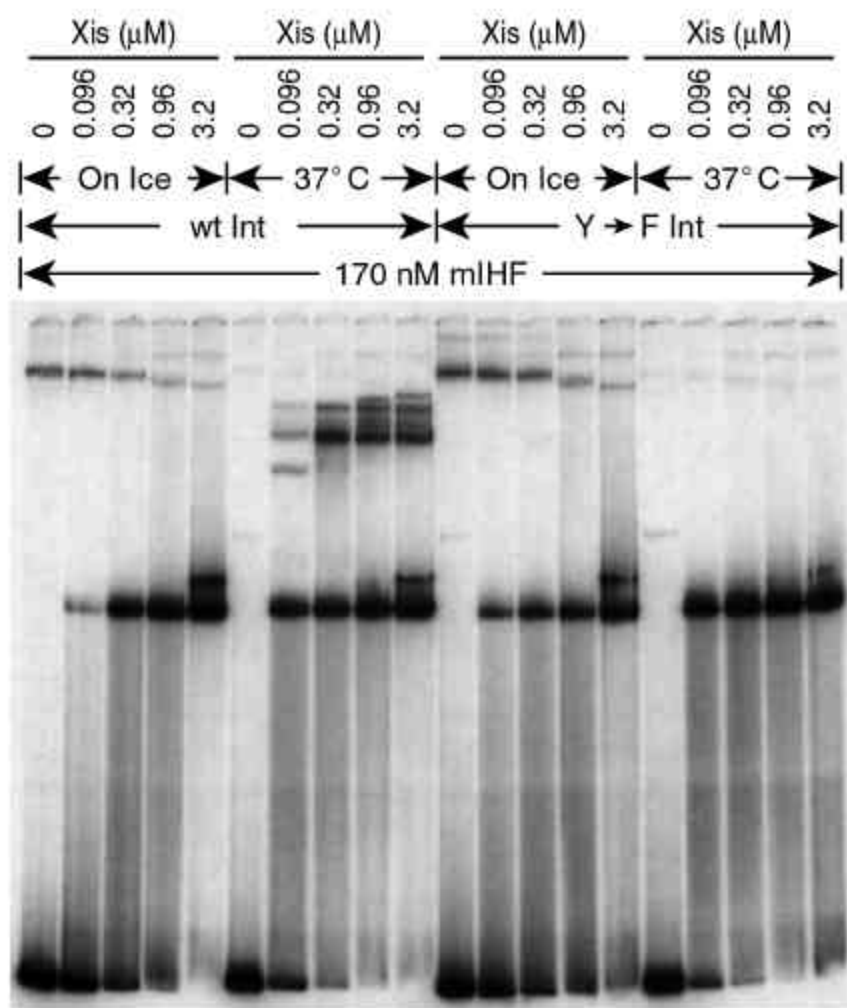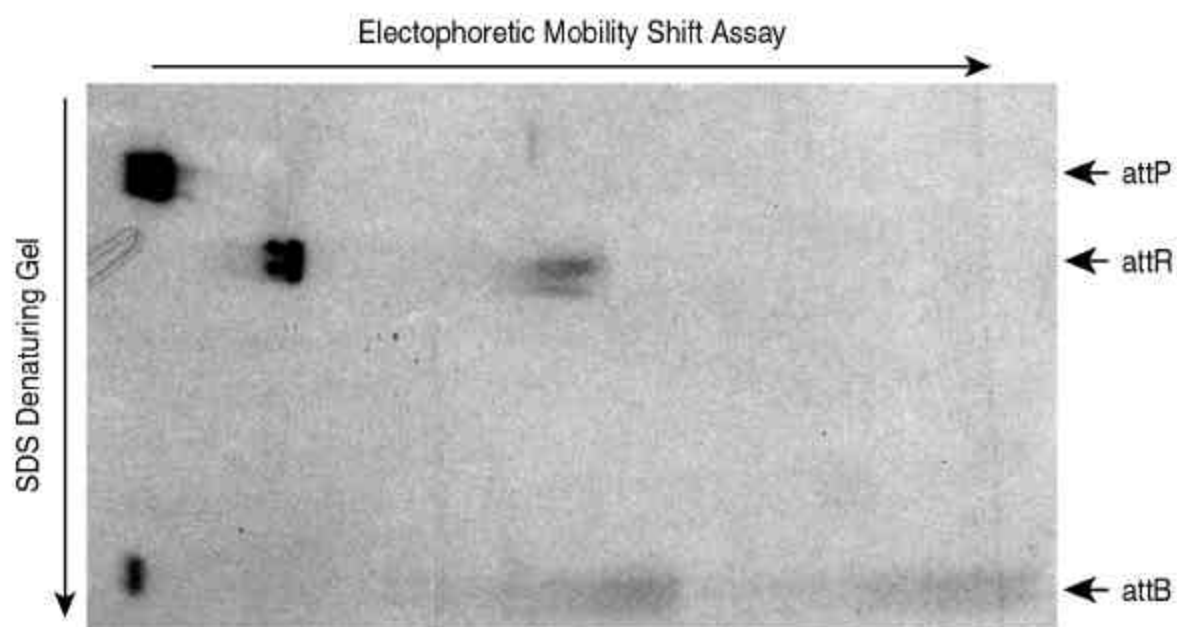
← attB

products, *attP* and *attB*. A synaptic complex must occur, but it is either short-lived, or it is not stable in native gel electrophoresis.

The relevance of the observed complex and protein-DNA interactions is contingent on the ability of the conditions used to support excision. To determine the amount of Xis required for excision, *in vitro* excision reactions were performed with a variety of Xis concentrations (9.6 nM – 3.2 µM) and monitored over time (0 – 40 min.). The reactions were run on a denaturing gel, and amounts of product were quantified. Since all conditions are identical to previous native gel and footprinting analysis, these reactions will be useful in determining their validity. With 9.6 nM Xis, there is little if any recombination, but slightly more Xis (32 nM) stimulates excision (Figure 25). All the higher levels of Xis have increased but similar rates, so the maximal rate is achieved at 96 nM but is not inhibited by concentrations as high as 3.2 µM.. These results indicate that all the protein concentrations and other conditions used in the above analysis are suitable for excision, so complexes seen throughout the Xis concentration range are potential candidates for excisive intermediates.

IV.F. Inhibition of integration

The function of an RDF is not only to stimulate excision, but in many cases, it also inhibits integration. To determine what role L5 Xis plays in inhibition of integration, potential interactions with *attP* were examined. It has already been shown that Xis and *attP* are able to form a complex (Figure 19). To look for higher order complexes, a native gel analysis was performed with *attP*, Int, mIHF and either without Xis, or with a variety of Xis concentrations (96 nM – 3.2 µM) (Figure 26). With no Xis, the only complex is the previously described intasome. At the lowest Xis concentration an additional complex is formed, and at the highest two more are present. The formation of the Xis dependent complexes with *attP* appears to mimic the three seen with *attR*: one formed at low concentrations requiring all three proteins, an inconsistent

Figure 25. Rates of excision

The rates of excision were tested using buffer conditions and Int and mIHF concentration that are identical to those in the native gel electrophoresis analysis. The amount of product formation was measured over a 40 minute time course for Xis concentration ranging from 9.6 nM to 3.2 µM. Results are given as a percentage of the total DNA that has been converted to product. The rate of excision reaches its maximum at a concentration of 96 nM, but is not inhibited by concentrations as high as 3.2 µM. At the two concentration lower than 96 nM, the rates are significantly reduced.

Rates of Excision

Legend
- ■ 3.2 μM Xis
- ▲ 960 nM Xis
- ▼ 320 nM Xis
- ◆ 96 nM Xis
- ● 32 nM Xis
- □ 9.6 nM Xis

Time (min)
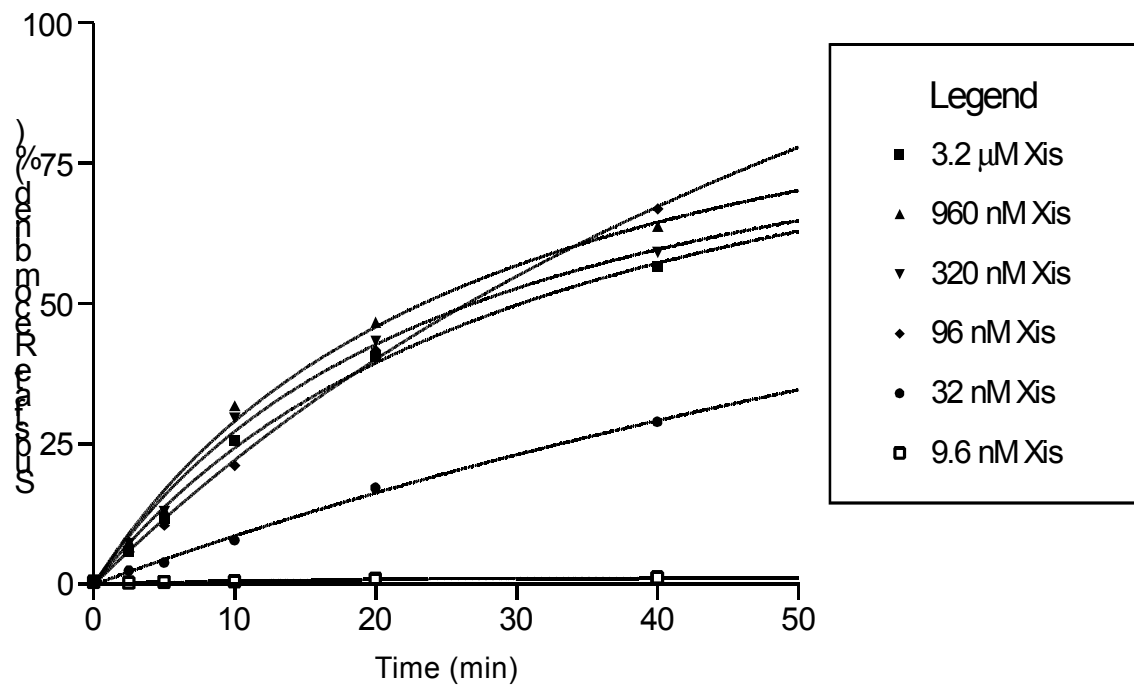
Figure 26. Formation of Xis dependent *attP* complexes

Native gel electrophoresis was used to look at *attP* complexes in the presence of Int, mIHF and a range of Xis concentration (96 nM – 3.2 µM). In the absence of Xis, the previously described intasome is present. With the addition of Xis, three complex can be observed. The concentration of Xis required to form each of them mimics the three complexes seen with *attR*. All three can be seen at the highest Xis concentration, but only two can be seen at the next lower level of Xis. One Xis dependent complex can be observed across the whole range of concentrations.
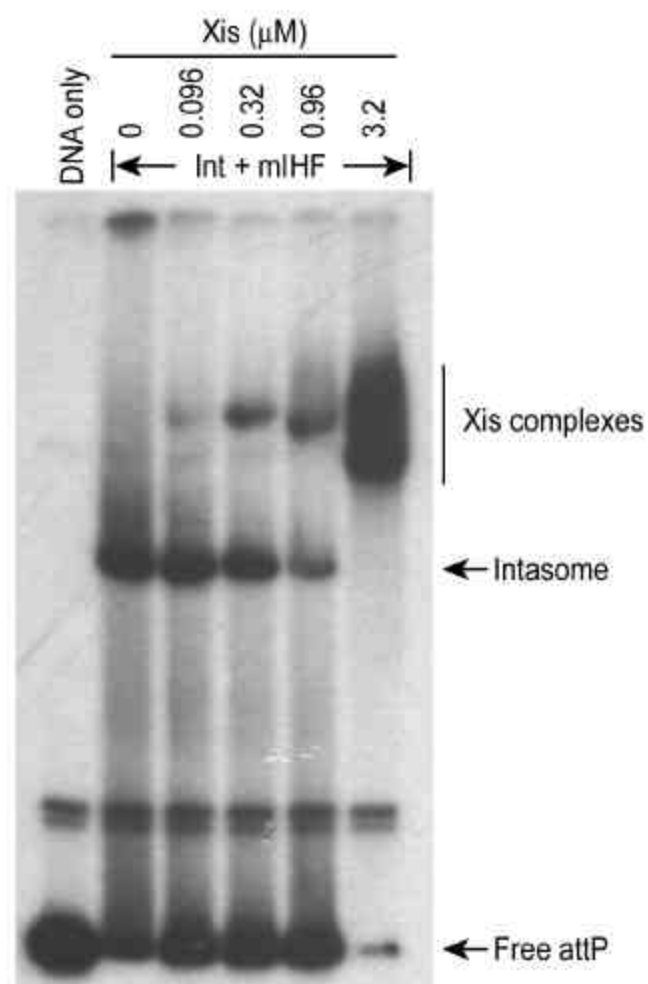
one forming at higher concentrations with Int and Xis, and an Xis only complex at the highest Xis level, which has a mobility only slightly slower than the complex containing all three proteins. The Int / Xis complex, like its *attR* counterpart, is not reproducibly seen, so its role in inhibition is uncertain. The other two are potential complexes involved in inhibition of integration.

To examine the composition of the *attP* complexes, the location of protein-DNA interactions was determined by DNase I footprinting. Digestion of DNA in the presence of Xis alone produced patterns nearly identical to those observed with *attR*: protection between P2 and P3 with periodic enhancements (data not shown). While it is interesting to see where Xis binds, the interactions in the presence of mIHF and Int were examined to gain an understanding of the inhibition of integration. The pattern in the presence of Int and mIHF are comparable to published results of the intasome (Pena *et al.*, 2000) where there is protection from the P5 to P3 site through the core with some periodic enhancements; the P1 and P2 sites are vacant (Figure 27). As Xis is added to the reactions, protection mimicking the Xis only reactions is seen between the P2 and P3 sites. Furthermore, there is minimal protection at the P1 and P2 arm-type sites. However, even at the highest Xis concentrations, there is no change in the pattern of digestion over the rest of *attP* from the P3 to P5 sites. The footprint shows that the P4 and P5 sites remain occupied in the inhibited complex, and that in at least a subset of the complexes the P1 and P2 sites are also bound.

The presence of Int at P1/P2 and P4/P5 pairs of arm-type sites in the inhibited complexes brings into question the mechanism of inhibition. If four Int molecules are involved, then not only will the *attP* core be bound, but there is the potential to recruit *attB*. To look for synaptic complexes (with *attP* and *attB*) in the presence of Xis, native gel electrophoresis was used with *attP*, Int, mIHF and either without or with a variety of Xis and *attB* concentrations. In the absence of Xis, the integrative complexes that have been previously described, intasome, synaptic complex 1 and synaptic complex 2 (Pena

Figure 27. Binding of Xis to integrative complexes

(A) The DNase I nuclease protection assay was used to look at binding of Xis to *attP* in the presence of Int and mIHF. These are conditions in which Xis can be inhibitory to integration. In the absence of Xis, the Int and mIHF bind to *attP* to produce the intasome, and the protection pattern is similar to those previously described. There is protection starting at the rightmost binding site, P5, through the core and to the P3 region. Periodic enhancements are observed throughout the protected area. The most notable change observed with the addition of Xis is protection in the P2 to P3 region with enhancements similar to those observed with *attR*. The P1 and P2 sites show some modest protection at higher Xis concentrations, but the amount does not indicate total occupancy of the sites. In the regions of *attP* to the right of P3, through the core and to P5, Xis has little or no effect on the protection pattern. (B) The sequence has been marked to indicate the locations of protection (triangles) and enhancements (pluses).

A

115

B

```
GGCTACCGGGTTGCAACTCCTGTGCAACTCTCAGGCTTCAACGCGCTTCTACGACCTGCAATTTC
CCGATGGCCCAACGTTGAGGACACGTTGAGAGTCCGAAGTTGCGCGAAGATGCTGGACGTTAAAG
              P1              P2

TTTCCACTTAGAGGATGCAGCCGAGAGGGGGTAAAAACCTATCTTGACCGGCCCATATGTGGTCG
AAAGGTGAATCTCCTACGTCGGCTCTCCCCCATTTTTGGATAGAACTGGCCGGGTATACACCAGC
         P3

GCAGACACCCATTCTTCCAAACTAGCTACGCGGGTTCGATTCCCGTCGCCCGCTCCGCTGGTCAG
CGTCTGTGGGTAAGAAGGTTTGATCGATGCGCCCAAGCTAAGGGCAGCGGGCGAGGCGACCAGTC
                        Common Core

AGGGTGTTTTCGCCCTCTGGCCATTTTTCTTTCCAGGGGTCTGCAACTCTTGTGCGACTCTTCTG
TCCCACAAAAGCGGGAGACCGGTAAAAAGAAAGGTCCCCAGACGTTGAGAACACGCTGAGAAGAC
                                        P        P
                                        4        5
```

116

et al., 2000) are observed (Figure 28). At the lower concentrations of Xis (96 and 320 nM), the *attP* / Xis complex (above) is present, but is reduced with the addition of *attB*; at 320 nM Xis, the amount of integrative complexes are also slightly reduced in the presence of *attB*. When the amount of Xis is increased to 960 nM, both synaptic complexes are missing, and the amount of intasome is greatly reduced. At the highest Xis level, the two most prominent *attP* / Xis complexes are present. The only integrative complex remaining is the intasome, but the amount is greatly reduced and only detectable without *attB*. In addition to previously observed complexes, one additional complex is observed containing *attP*, *attB* and Xis, but only at the highest level of Xis. Despite the observation of a synaptic complex containing Xis, it is only observed at very high Xis concentrations. The reduction of integrative complexes at lower concentrations and their replacement by an *attP* / Xis complex is more likely to represent the means of inhibition.

For the assumptions about the complexes' involvement in inhibition to be valid, the protein concentrations and conditions must be suitable for integration in the absence of Xis, but not in its presence. To test the concentrations of Xis required for inhibition, *in vitro* integration reactions were performed either without Xis or with a variety of Xis concentrations (32 nM – 3.2 μM). Samples were taken over time (0 – 40 min), denatured, run on a gel containing SDS and the amount of product was quantified. Though a significant amount of integration was seen without Xis, even the lowest level of Xis showed some inhibition (Figure 29). When 96 nM or more of Xis was present, maximal inhibition was achieved, but a minimal amount of recombination was still observed. The conditions used are suitable for integration, but even low levels of Xis are effective at inhibiting integrative product formation. However, part of the low accumulation of products could be a result of excisive recombination and further decreased by the lack of supercoiling for integrative substrates. For these reasons, rates of integration in inhibited reactions must be viewed with caution.

Figure 28. Analysis of Xis complexes inhibiting integration

The inhibition of integration can be achieved at multiple points in the assembly pathway. To differentiate between disruption of the intasome to prevent *attB* capture versus the prevention of a synaptic complex from becoming active, native gel electrophoresis was used to determine if Xis dependent complexes are formed with *attB*. Labeled *attP* was included with Int, mIHF, a variety of Xis concentrations (96 nM – 3.2 μM) and increasing amounts of *attB*. In the absence of Xis, all previously described integrative complexes are present. Without *attB*, the previously described Xis dependent complexes and the reduction of intasome are observed. When increasing amounts of Xis are added to reactions containing *attB*, a reduction in synaptic complexes 1 and 2 is seen. At the highest levels of Xis, the only Xis independent complex remaining is a small amount of intasome. Under all of the conditions, there is only one complex that forms in response to Xis and *attB*, and it is only seen at the highest Xis concentration (3.2 μM).
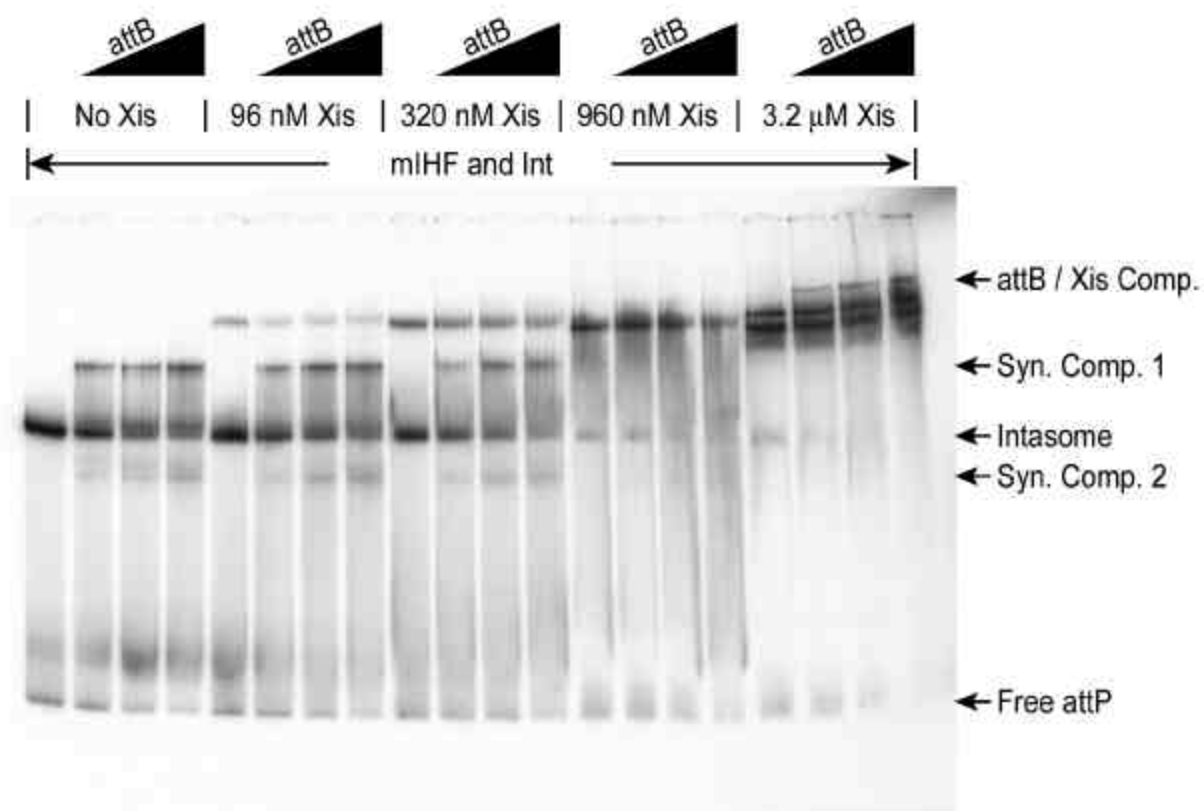
Figure 29. Inhibition of integration

To determine if the condition used in the analysis of binding of Xis to *attP* are meaningful to inhibition of integration, the amount of integration and inhibition by addition of Xis was quantified. The *in vitro* recombination assays were carried out with conditions nearly identical to the native gel electrophoresis. After incubation at 37ºC for 2.5 to 40 min., the reactions were stopped by addition of SDS, and the DNA was run on a denaturing gel and then quantified. The units used to show amount of recombination are a ratio of the maximal amount of product obtained in the no Xis reaction. The overall levels of integration were relatively low, and the total DNA could not be quantified because it was not within a linear range with the products. While only modest amounts of recombination were observed in the absence of Xis, addition of 32 nM Xis shows significant reduction in integration. When higher levels of Xis were added, recombination was nearly completely stopped, though some low-level product formation is seen at even the highest Xis concentration.

Inhibition of Integration
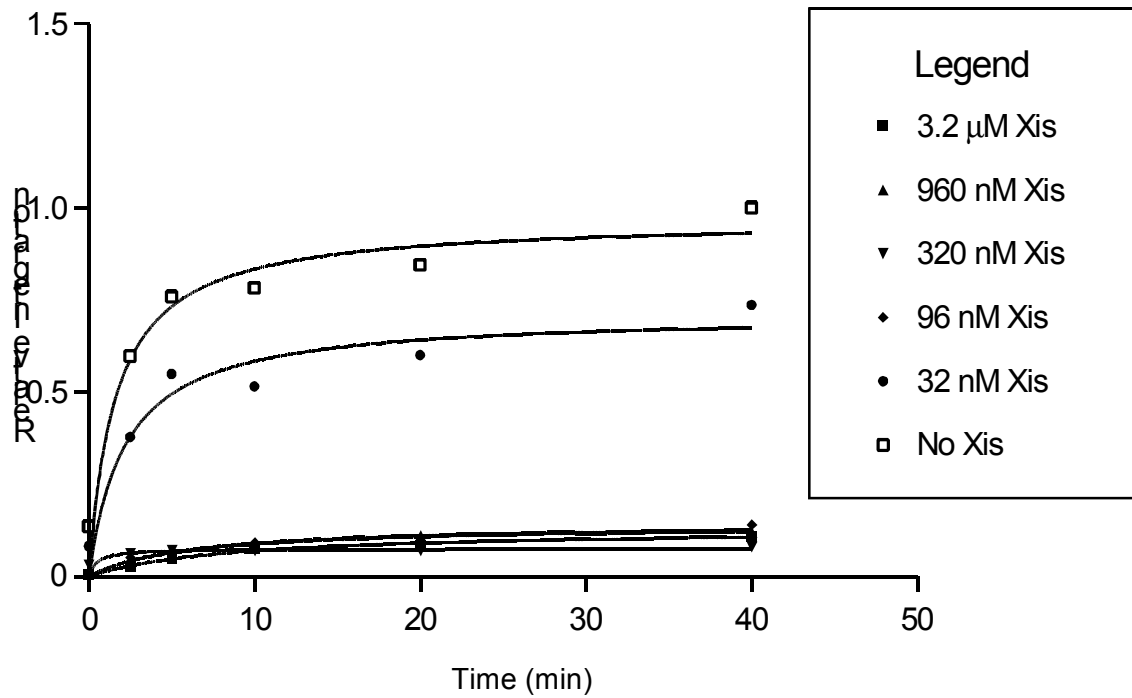
IV.G. Discussion

IV.G.i. Binding of Xis and its functional role

The function of L5 Xis appears to be mediated by site-specific binding to sequences between the P2 and P3 arm-type sites. This interaction has been shown in both *attP* and *attR*, but is absent with *attL*. The role of the DNA binding is likely to be in bending of the DNA (Figure 30A). This is supported by the spacing of the enhancements in the Xis footprint which are approximately one helical turn apart and are suggestive of a bend. This sort of mechanism is consistent with other recombination accessory proteins such as IHF and mIHF, which have both been shown to bend DNA (Robertson and Nash, 1988; Pedulla and Hatfull, 1998).

Although speculation on the sequence of binding sites must be done with caution, there are four imperfect repeats (consensus CTTCNAC) in the region where Xis binds (Figure 30B and C). In the 7 bp sequence, 3 locations are completely conserved and another 3 bases are in 3 of the 4 repeats. Furthermore, the sites are located at 10 bp intervals, which would be the required phasing to introduce a bend in the DNA. While data does not yet exist to show that these are the recognition sites, they are within the region of Xis binding and do represent a sequence feature that could be involved

Another interesting aspect of Xis binding is that it appears to be highly cooperative. This is highlighted by the native gel electrophoresis data with Xis-only complexes, in which there is no complex visible at lower concentrations, but a small increase in protein to the highest concentrations results in nearly all of the DNA being in complex. Another example of this is the footprint analysis with Xis alone. While the amount of protection increases slightly from 320 nM to 3.2 μM of Xis, there is no detectable protection below 320 nM, so there is a sharp drop-off of binding below this concentration. This behavior is consistent with cooperative binding

Figure 30. Model for Xis binding

L5 Xis binds to a region between the P2 and P3 arm-type sites. The most likely role for Xis is to bend the DNA (A) to allow the formation of alternate architectures. While the specific sequences involved in site recognition have not been characterized, there are four imperfect tandem repeats spaced 10 bases apart in the region where Xis binds (B). Within the 7 bp repeat, three bases are present in all four and another three location are the same in three of the repeats (C). The consensus sequence for the repeat is CTTCNAC.

A

B

```
GTGCAACTCTCAGGCTTCAACGCGCTTCTACGACCTGCAATTTCTTTCCACTTAGAGGATGCAG
CACGTTGAGAGTCCGAAGTTGCGCGAAGATGCTGGACGTTAAAGAAAGGTGAATCTCCTACGTC
```

P2 →   ← P3

C

attR repeats
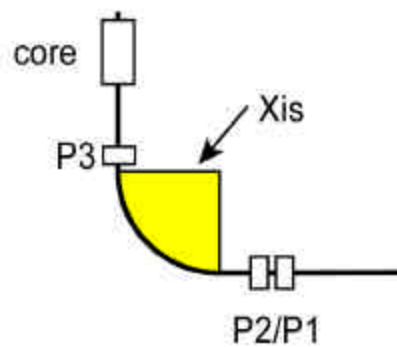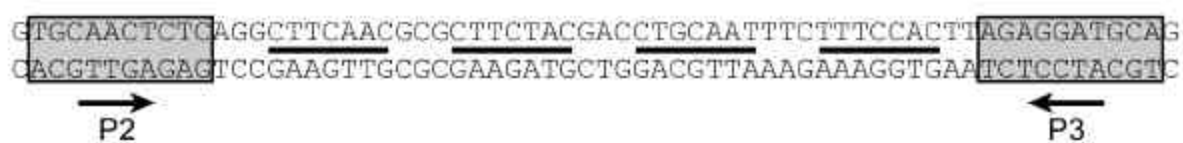
consensus   **CTTCNAC**
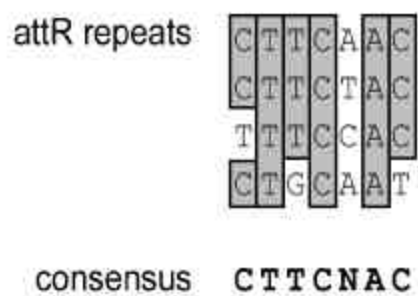
IV.G.ii. Excisive complexes

Complexes containing either *attR* or *attL* have been observed in the presence of Xis. The complex containing Int, Xis and mIHF is the most likely candidate for excisasome-R, the *attR* excisive complex. This is the only *attR* complex present at all the concentrations of Xis that were competent for recombination, and it also contains all the proteins necessary for excision. The complex can be represented by a simple model, in which the P1 and P2 arm-type sites are bound with Int molecules that bridge to the core-type binding sites (Figure 31A). Two different accessory proteins, Xis and mIHF, facilitate the tight bend required to bring the core and arm sites into proximity. Excisasome-L is the only *attL* complex that was detected and it is also present in recombinagenic conditions. This complex has already been characterized (Pena *et al.*, 2000), and it contains Int molecules making bridges between the arm-type sites and the core. Unlike the *attR* complex, Xis is not needed for this structure, and only mIHF is required to mediate the tight bend (Figure 31A). Both of these complexes must come together before excision can occur.

A synaptic complex containing both excisosomes has not yet been identified. A putative model for such a complex (Figure 31A) would not require any rearrangement of the two individual complexes and could be held together by Int-Int protein interactions. The rapidity of excision with preformed complexes supports the idea that no significant rearrangement is needed. If weak protein-protein binding holds them together, this could explain the inability to trap it before recombination has occurred. Also, another possible reason for why only product or individual substrate complexes are detected is that catalysis occurs rapidly after synapsis or even within the polyacrylamide gel used in native gel electrophoresis. Despite the inability to observe the synaptic complex, the conditions used support extensive recombination, so the predominant complexes, excisasomes-R and L, are likely involved.

Figure 31. Models for excision and inhibition of integration

(A) The role of Xis in excision is in bending *attR* to mediate the formation of excisosome-R; a complex in which Int forms intramolecular bridges between the arm-type and core sites. This complex also requires mIHF to assist in bending the DNA. Excisosome-L is the corresponding *attL* comples, but Xis is not required. While no synaptic complex has been observed, the *attL* and *attR* must come together. A simple model for a synaptic complex is bringing together of the two excisosomes, which could be mediated by protein-proteins interactions. (B) The presence of Xis prevents the formation of the integrative synaptic complex. This synaptic complex contains Int, mIHF, *attP* and *attB* in a condensed nucleoprotein complex. The Xis protein prevents the formation of the synaptic complex by preventing capture of *attB*, but the P4/P5 to core bridges are not disrupted. It is unclear if the P1 and P2 sites are occupied in the inhibited complex because minimal protection was seen in DNase I analysis. The Xis dependent *attP* / *attB* synaptic complex that was observed in native gels are not likely to be involved in inhibition because of the high Xis concentration required. It is likely to be a modified form of synaptic complex 2 that has Xis bound where it does not interfere with Int binding.

A

excisosome-R

excisosome-L

synatpic
complex ?

Legend

Int

Xis

mIHF
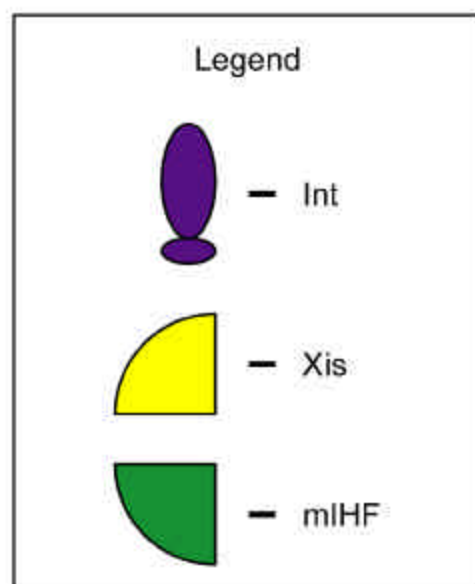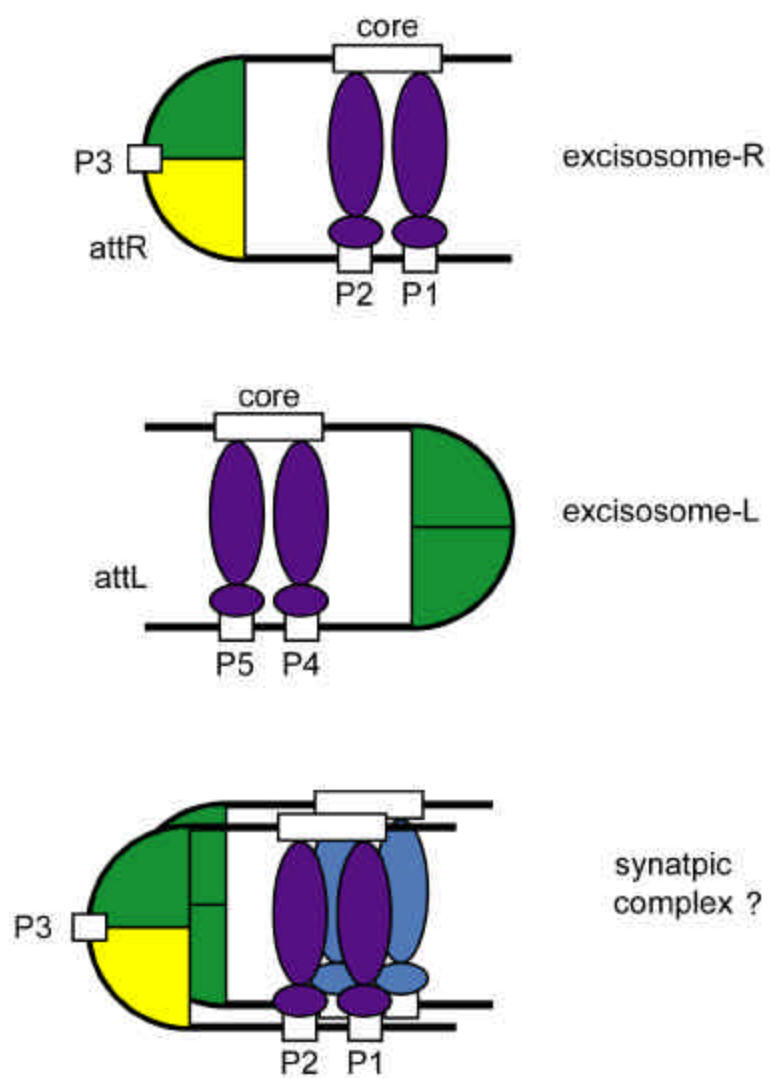
B
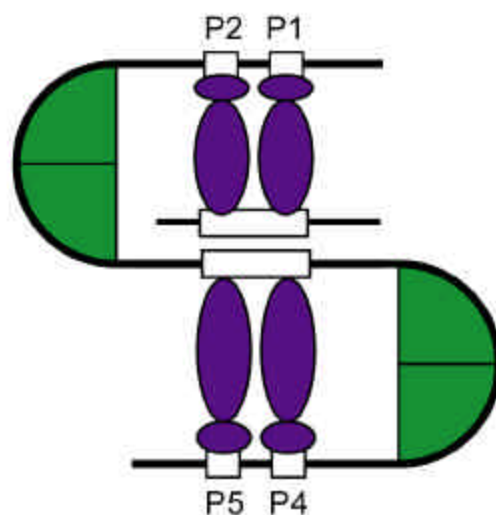
Integrative Complex

Inhibitive Complexes

or

attB / Xis complex

Legend

Int

Xis

mIHF

P2 P1

P5 P4

P3

core

128
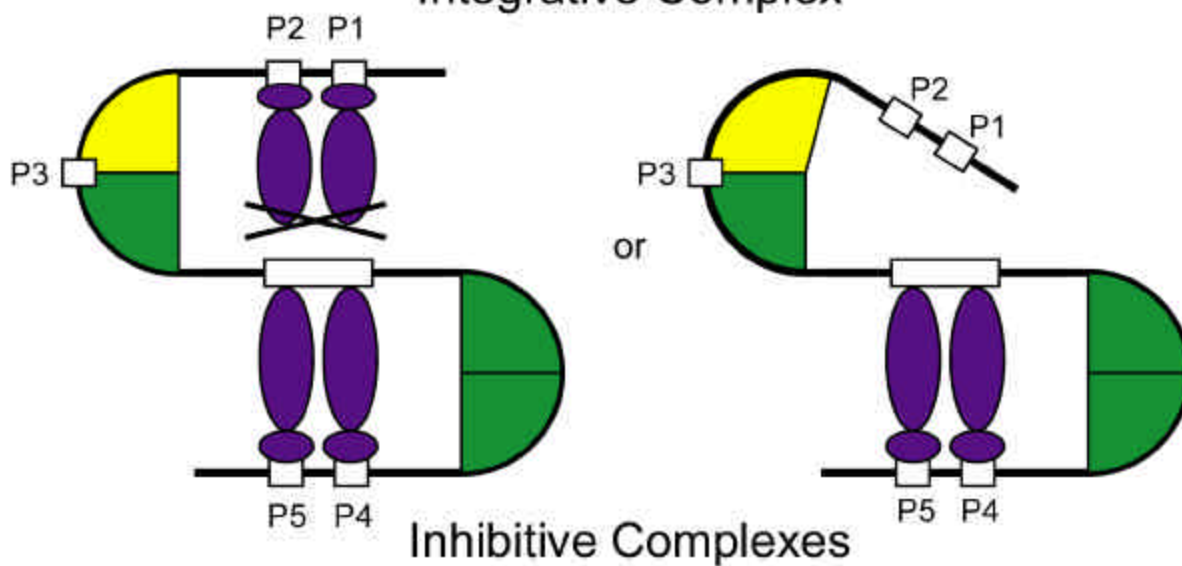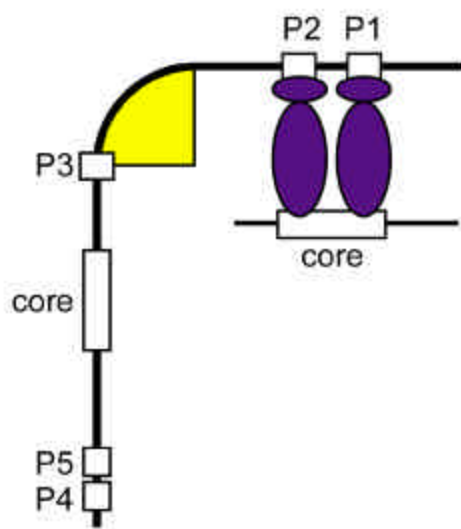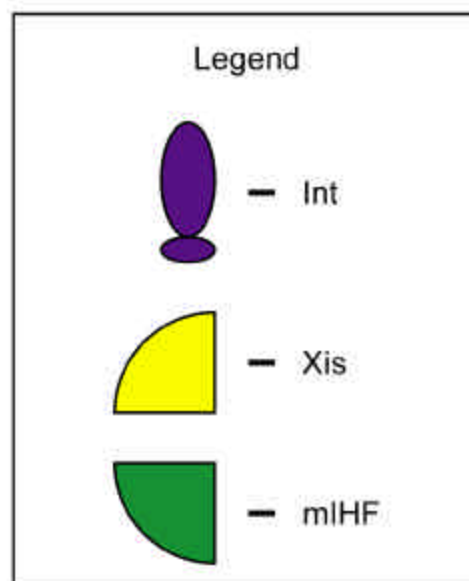
IV.G.iii. Inhibition of integration

The role L5 Xis plays in inhibition of integration appears to be mediated by DNA bending in the region between the P2 and P3 arm-type sites. There is no apparent change in the footprinting pattern on the right side of *attP*. The protection of the P4 and P5 arm-type sites and the core by Int and the intermediate regions by mIHF are unaffected by Xis concentration. The only differences observed are seen in the region Xis binds (P2 – P3) and a slight amount of protection at the P1 / P2 arm-type sites. While the presumed presence of Ints at these sites suggest that inhibition might occur after recruitment of *attB*, synaptic complexes containing Xis are only seen at the highest Xis concentration, a level far above where inhibition was seen. Furthermore, the P1 / P2 arm-type sites protection is much less than that seen within the same region in excisosome-R. This indicates that the more likely role of Xis is in creating a conformation that prevents binding of Int molecules to the arm-type sites, but a model in which some occupancy occurs cannot be completely ruled out (Figure 31B).

A potential mechanism for directional control that cannot be totally ruled out is the excision of integrative products at a rate that prevents their accumulation. Some of the inhibition observed in these assays can be a result of the substrate conformation used in the assay. It has been observed previously that integration has a relatively strict DNA supercoiling requirement (Peña, C.E.A. *et al.*, 1998) that excision does not share. The DNA used in many of these assays is linear and thus somewhat inhibitory to integration. Furthermore, even if a supercoiled substrate were used, the products of excision in the *in vitro* reaction would be relaxed. While the rates of excision in the conditions used could to greater than the integration, there are quite obvious interactions with *attP* that prevent formation of integrative complexes and are presumably the method of inhibition of integration.

IV.G.iv. mIHF requirement

The L5 integration reaction shows a specific requirement for the mIHF accessory protein, and attempts to substitute this factor with *E. coli* extracts or other purified proteins were unsuccessful (Lee and Hatfull, 1993). However, intasome complexes can be formed in the presence of just Int and *E. coli* extracts (A. Lepisto and G.F. Hatfull unpublished). Furthermore, it has been shown that the excision reaction can occur in the absence of mIHF, if supplemented with *E. coli* extracts (Chapter III). Since the intasome and excisosome-L do not have an absolute requirement for mIHF, this specificity must arise from sequences to the left of core. Also, in excisosome-R, the region between P3 and core utilizes mIHF and presumably *E. coli* extracts in its absence. This would suggest that the region where Xis binds is also the location where mIHF is stringently required. Obviously, the necessity for mIHF in integration might vary in the formation of an active complex, but based on formation of intermediate complexes, it can be localized to the left end of *attP*.

V. UTILIZATION OF L5 XIS AS A GENETIC TOOL

(Note: This chapter was previously published as Lewis, J.A. and G.F. Hatfull (2001). "Control of directionality in integrase-mediated recombination: examination of recombination directionality factors (RDFs) including Xis and Cox proteins." *Nucleic Acids Res* **29**(11): 2205-16.**)**

V.A. Introduction

The integration system from mycobacteriophage L5 has been used to develop an extremely powerful tool for introducing DNA into mycobacterial hosts. By including *attB* and *int* on a plasmid, mycobacteria can be efficiently transformed by integrating a single copy of the vector into the host's genome at the *attB* site (Lee *et al.*, 1991). While these integrated plasmids are quite stable, the loss rates can be reduced to undetectable levels by removing the *int* gene from the integrating vector and providing it in trans on a non-replicating vector during the transformation step (Peña, C.E. *et al.*, 1998).

There are several advantages that make integrating plasmids preferable to extrachromosomal plasmids.   Integrated vectors are more stable than currently available extra-chromosomal plasmids (Chapter III and unpublished observations). Since only a single copy of the DNA is integrated, there are no multicopy affects that can create undesired phenotypes (Banerjee *et al.*, 1994; Barsom and Hatfull, 1996). Furthermore, the system allows large pieces of DNA to be integrated and transform cells at a high efficiency (Lee and Hatfull, 1993). The major disadvantageous with this system has been the inability to remove integrated DNA. There are two reasons for which removal of DNA is desirable: to cure strains and to recovery DNA into *E. coli* for

further analysis (Pascopella *et al.*, 1994). With the identification of the L5 Xis gene, methods can be developed that will facilitate the removal of integrated vectors to expand the utility of this system.

V.B. Curing of mycobacterial strains containing integrated vectors

The ability to cure strains of integrated vectors is useful for a multitude of tasks including examination of second-site mutations and verification of phenotypes obtained by transformation. Even though low loss rates are detectable with integrating vectors containing the *int* gene, the negligible levels of loss makes it extremely difficult to recover these strains without a counter-selectable marker. A simple plasmid based method for curing *M. smegmatis* strains was observed during the *in vivo* identification of the L5 *xis* gene (Chapter III). In these experiments, a plasmid (pJL22) containing the hsp60 promoter expressing the *xis* gene was shown to be incompatible with an integrated vector. In other words, transformation with pJL22 results in 100% loss of the integrated vector. This method requires a subsequent removal of the extrachromosomal plasmid after the excision of the integrated DNA, but its loss is at a high enough frequency that cells lacking the plasmid can be readily recovered.

A reasonable assumption can be made that a similar method for curing could be used in slow-growing mycobacteria. Observations made in collaboration with Tanya Parish revealed that pJL22 is unable to efficiently transform *M. tuberculosis* (T. Parish, personal communication)*.* It is possible that the BCG hsp60 promoter is more active in slow-growers or *M. tuberculosis* is more sensitive to L5 Xis. To overcome this problem, two different plasmids (pJL28 and pJL29) were constructed in which a temperature-sensitive mutant allele of the L5 repressor gene *71* with its own promoter was placed upstream of 10 tandem copies of the gp71 binding sites followed by the *xis* gene. Presumably with this construct, Xis should be expressed at a much lower level than from pJL22, though specific expression levels have not been measured. When pJL28

and pJL29 were used to transform *M. tuberculosis*, transformants were grown on media selecting for just the extrachromosomal plasmids or both the integrating and extrachromosomal DNA. Nearly 1000-fold more colonies were observed when selecting for pJL28 or pJL29 alone, and the integrating vector had been lost (T. Parish, personal communication). The use of plasmids expressing L5 Xis provides an efficient method for curing mycobacteria of integrated vector, though expression levels must be moderated for slow-growing species.

V.C. Recovery of integrated plasmids by *in vitro* recombination

While having a single copy of a stably integrated vector is extremely useful for genetic studies, these traits are contradictory to obtaining significant quantities of purified DNA. The most reliable way to do this is to shuffle the vector into *E. coli*, where it can be replicated as a high copy plasmid and easily purified. To be able to transform *E. coli*, the vector DNA must first be excised from the bacterial genome. An efficient method for excising the DNA is the *in vitro* excision reaction; in assays using plasmid DNA, greater than 50% of the substrate was recombined (Chapter III). To demonstrate that this was a plausible method to recover integrated DNA, genomic DNA was isolated for strains of *M. smegmatis* containing the integrated vector pJL11 or the extrachromosomal plasmid pCP10. This DNA was used in the *in vitro* recombination reaction with varying protein combinations and then used to transform *E. coli* (Table 1). All samples isolated from strains containing pCP10, regardless of proteins used, yielded 10^3-10^4 *E. coli* transformants. While a small number of background colonies were obtained, a significant number of transformants from pJL11 samples was only achieved when Int, mIHF and Xis were used in the excision reaction. The purified plasmid DNA from 18 of the colonies was digested and shown to have the same pattern as pJL11; analysis of background colonies revealed that none appeared to contain plasmid DNA (data not shown). To verify that a functional *attP* had been

Table 1. Purification of integrated plasmids

Proteins used *in vitro*	pJL11[a,b]	pCP10[a,c]
No Protein	0	3850
Int, mIHF	1	6010
Xis Only	0	18400
Int, mIHF, Xis	226	8750

a. number of colonies obtained by transformation of *M. smegmatis* with ~0.5 μg of DNA from *in vitro* reaction with listed proteins.
b. DNA isolated from mc²155 pJL11 (integrated)
c. DNA isolated from mc²155 pCP24 (extrachromosomal)

regenerated, two samples of DNA recovered from pJL11 strains were used to retransform *M. smegmatis* and both did so efficiently. The *in vitro* excision reaction is a simple and quick method to recover an integrated vector from mycobacteria to *E. coli*.

V.D. Development of a shuttle phasmid system to express L5 Xis

The use of *in vitro* excision has been shown as a plausible method to recover DNA into *E. coli*, but it requires the use of three different purified proteins. This is an acceptable method in a laboratory that already has these proteins or that intends to recovery a significant number of samples so that the effort required for purifying the proteins is worthwhile. However, a system that uses fewer components and is easier to implement would be desirable for most users. The challenge in developing an *in vivo* method is that the *xis* gene must be expressed at a high level, but only within the short time window when recovery will be attempted. It has already been shown that the presence of Xis in a cell results in the loss of the integrated vector, so constitutive expression would not be acceptable.

For these reasons, a phage-derived system was developed in which high levels of Xis expression can be achieved immediately following infection of the cell. Three phasmids were constructed from a temperature sensitive, non-replicating shuttle phasmid derivative of bacteriophage TM4 (phAE87) (Bardarov *et al.*, 1997). Phasmid phJL1 contains the *xis* gene fused to the hsp60 promoter and two other phasmids (phJL2 and phJL3) contain both the *int* and *xis* genes fused to hsp60 promoters (Figure 32). Since these phage are non-replicating at 37°C, extended incubations can be used after infection to enhance excision without lysis of the cell. To determine if these phages were capable of stimulating excision, strains containing integrated DNA were infected and samples were taken over a period of time. The amount of excision was verified by looking at the quantity of attachment sites using a previously described PCR assay (Peña *et al.*, 1997). There is a small increase in the amount of *attB* over the time of

Figure 32. Phasmid delivery of L5 Xis

Three shuttle phasmids were constructed that express L5 Xis. The phage are derivatives of a temperature sensitive, non-replicating mutant of TM4 that also has an *E. coli* origin of replication and Lambda cos sites. In one of the constructs, the L5 *xis* gene was put downstream of the hsp60 promoter. The other two had identical inserts but in opposite orientations. In these phage, both *xis* and *int* are downstream of hsp60 promoters.

phJL1

phJL2 and phJL3
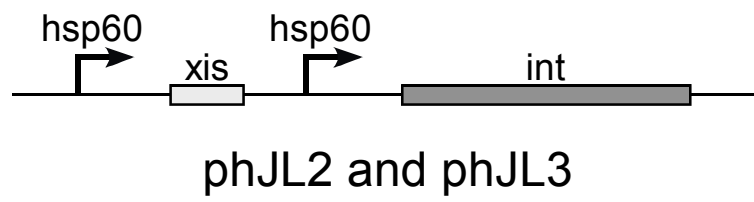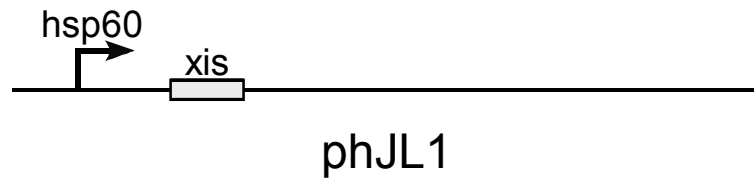
infection indicating that excision is occurring (Figure 33). However, little reduction is seen in *attL* and *attR* levels, so the percent of excised DNA is small. The shuttle phasmid system provides a mechanism to excise integrated DNA in a defined time frame.

V.E. Phasmid-based recovery of integrated vectors

While integrated DNA is excised using the shuttle phasmids, the vector must still be recovered in *E. coli*. Electroduction of plasmids from *M smegmatis* directly into *E. coli* has been successful for extrachromosomal plasmids (Baulard *et al.*, 1992), but these methods have not been efficient enough to recover integrated vectors (data not shown). A modified alkaline-lysis mini-prep was used to purify vector DNA that had been excised from *M. smegmatis* using one of the three phages that express Xis (phJL1 – phJL3), and then the DNA was used to transform *E. coli* by electroporation. To assure the versatility of this system, three different integrated constructs were tested: a plasmid (pJL11) containing *attP* and *int*, a plasmid (pJL35) with *attP* but lacking *int*, and a cosmid (1A9) possessing both genes. *E. coli* transformants were recovered with each vector, but optimal time, quantity of colonies and phage requirements varied (Table 2). The plasmid recovery yielded ~10^3 colonies when phJL2 or phJL3 was used, with the maximum achieved after 14 hours of infections. After infection with phJL1 (Xis only), the pJL11 samples produced only ~100 colonies, while the pJL35 samples quantity of colonies and phage requirements varied (Table 2). The plasmid recovery yielded ~10^3 colonies when phJL2 or phJL3 was used, with the maximum achieved after 14 hours of infections. After infection with phJL1 (Xis only), the pJL11 samples produced only ~100 colonies, while the pJL35 samples which lack the *int* gene yielded a number of transformants equal to background levels. Digestion of DNA preparations showed that recovered plasmids had identical restriction patterns to the appropriate construct, pJL11

Figure 33. Excision of integrated vectors

To determine whether the phasmid is stimulating excision, a PCR analysis detecting attachment sites was performed. Four primers that flank *attP* and *attB* amplify unique length fragments for each of the four sites (*attP, attB, attR* and *attL*). In this assay, infection of BCG by phJL2 was stopped at specific time points from 1 to 14 hours. When all of the primers were used, there is little change in the amount of *attL* and *attR*, but also a little *attP* is detectable. When the *attB* primers are used by themselves, that site is also detected.
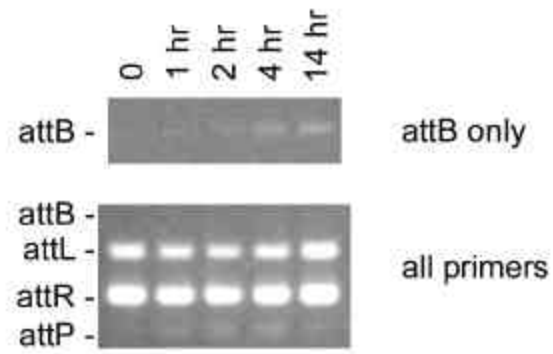
Table 2. Recovery of integrated vectors from *M. smegmatis*

DNA	Type	Phage	0 hrs	3 hrs	6 hrs	14 hrs
--	--	phJL2	4	0	0	0
pMH94	plasmid	phAE87	0	3	16	5
pJL11	plasmid	phJL1	27	24	83	115
pJL11	plasmid	phJL2	--	740	1680	1200
pJL11	plasmid	phJL3	--	600	1000	2100
pJL35	no int plasmid	phJL1	6	2	0	1
pJL35	no int plasmid	phJL2	--	660	1320	6820
pJL35	no int plasmid	phJL3	--	660	480	2760
1A9	cosmid	phJL1	0	13	23	30
1A9	cosmid	phJL2	--	140	298	198
1A9	cosmid	phJL3	--	42	175	156

or pJL35. This system provides an efficient method to recover integrated DNA from *M. smegmatis* for replication in *E. coli*.

To verify that excision was a result of Xis expression and not merely phage infection, the parent phage, phAE87, was used to infect a strain carrying an integrated vector (pMH94); a different vector was used due to antibiotic conflicts. Though a small number of background colonies were seen, the numbers are significantly lower than those for recovered vectors. To determine the cause of the background seen under all conditions, many of the transformants were used to prepare DNA and then digested with restriction enzymes (data not shown). While almost all grew, the majority of those obtained from the pMH94 strains failed to grow; one that did contained DNA that matched pMH94. DNA from other colonies had several different restriction patterns: some matched pJL11 while others had novel patterns. Some of the transformants recovered in negative controls might be the result of cross-contamination, but there is a minimal but consistent recovery of the pJL11 vector in the absence phage infection. This is consistent with the observation of an Int mediated excision of integrated DNA that is responsible for the low-level loss observed with integrating vectors; none of the colonies recovered in the absence of *int* contained pJL35.

Cosmid recovery offers an additional challenge because the larger size results in a lower transformation efficiency and an increased potential for isolation of DNA with partial deletions. This inefficiency is reflected in the lower number of colonies obtained in the recovery (Table 2). As with plasmid recovery, the number of transformants from samples infected by phJL2 and phJL3 was greater than with phJL1; the optimal yields were ~200 or 30 respectively. Furthermore, the yield seems to be somewhat reduced at the 14-hour time point as compared to 6 hours. DNA preparations were obtained for transformants obtained from all three phage infections. Every colony that was analyzed contained DNA that matched the restriction pattern of 1A9 (Figure 34). The use of

Figure 34. Recovery of full-length cosmid DNA

The phage delivery system for recovering DNA has yielded many transformants, but the quality of the DNA needs to be verified. Three transformants obtained from independent isolations using each of the three phasmids (phJL1 –phJL3) were grown and the DNA was purified. Digestion of the DNA with Pvu II shows an identical restriction pattern to the cosmid, 1A9, which was used to transform the original *M. smegmatis* strain.

144
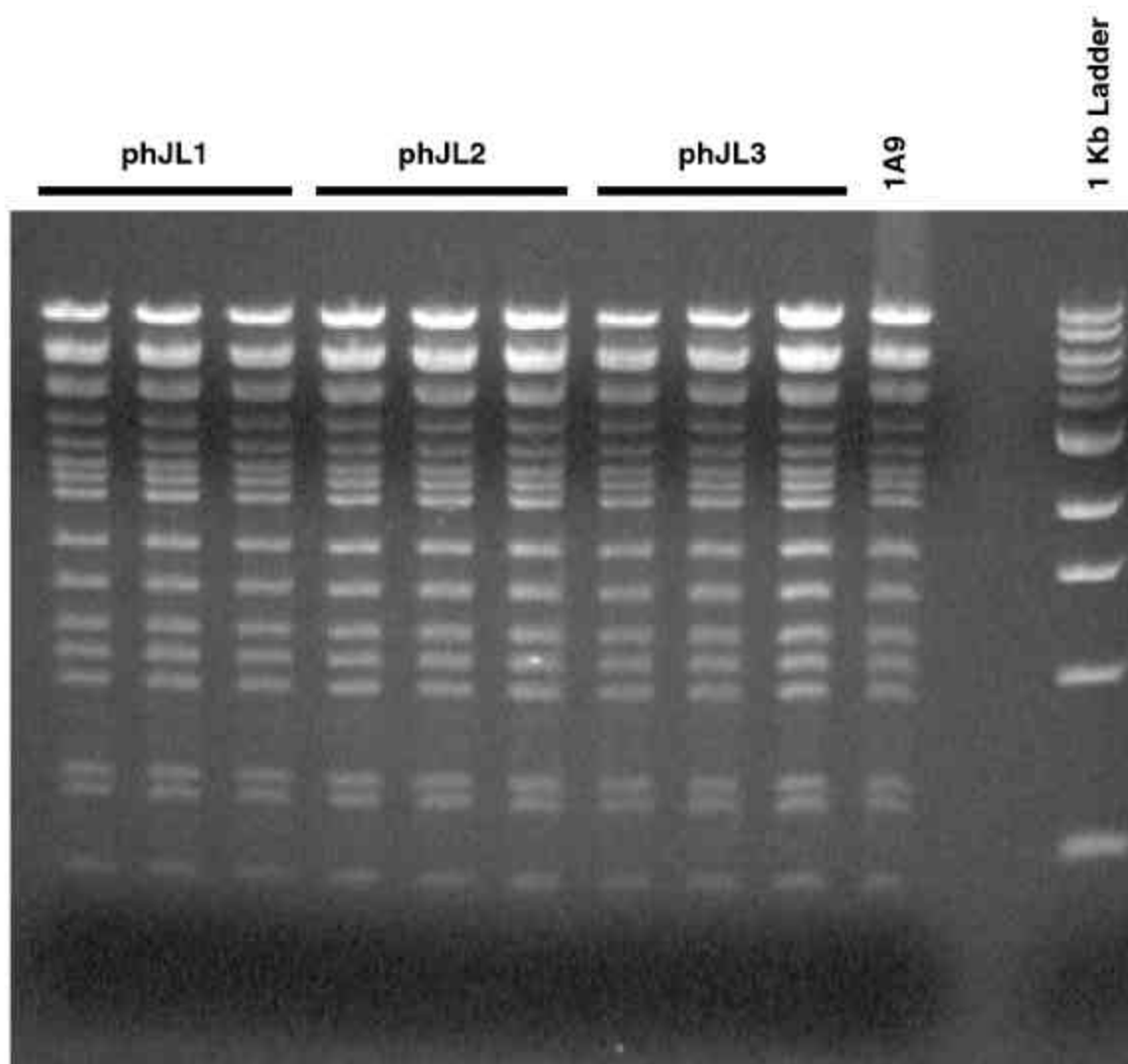
shuttle phasmids provides an efficient method to successfully recover intact integrating vectors from *M. smegmatis* that are the size of cosmid DNA.

Due to metabolic and genetic differences between slow- and fast-growing strains of mycobacteria, the success in recovering DNA from *M. smegmatis* using this phage method cannot be directly translated into the same ability in slow-growers such as *M. bovis* BCG. A method similar to the smegmatis experiments was used to test for recovery from BCG. A strain containing an integrated vector, pJL11, was infected with one of the three phages (phJL1 – phJL3), the cells were lysed using a modified alkaline lysis protocol, and the purified DNA was used to transform *E. coli*. Approximately 10^4 colonies were obtained for all three phages (Table 3). with the optimal yield at 14 hours. Digestions of the DNA isolated from these colonies were identical to pJL11. The recovery of a cosmid size insert, 1A9, has proven more difficult. Initial experiments yielded a reduced number of colonies versus the plasmid recovery, consistent with smegmatis work, but digests of the DNA indicated that all transformants contained only a portion of the original cosmid (data not shown). Two possibilities exist: either the DNA was deleted before the recovery, or it is a shortcoming of the recovery process. Attempts to transform BCG with 1A9, yielded only a few colonies with an extremely low efficiency of 10-20 cfu / µg. Five individual 1A9 transformants of *M. bovis* BCG were used in the recovery assay and in all but one case numerous *E. coli* transformants were obtained. DNA from individual *E. coli* colonies was purified and digested. Interestingly, all DNA preparations obtained from an individual BCG transformant had identical restriction patterns, but each BCG colony yielded a different vector, which was a subset of 1A9 (data not shown). The largest fragment recovered was nearly 40 kb in length, so the ability to recover plasmid size vectors into *E. coli* has been successfully achieved. Difficulties in cosmid size recovery appear to be more related to the ability of the cosmid used (1A9) to transform BCG then in the recovery method. Despite the failure in isolating full-length constructs, large pieces (~40 kb) have still been obtained.

Table 3. Recovery of DNA from *M. bovis* BCG

DNA	Type	Phage	0 hrs	5 hrs	14 hrs
pJL11	plasmid	phJL1	30	3.9×10^3	2.6×10^4
pJL11	plasmid	phJL2	--	1.2×10^4	2.6×10^4
pJL11	plasmid	phJL3	--	3.5×10^3	1.1×10^4
1A9	cosmid		?	?	?

V.F. Discussion

The curing of strains containing integrated DNA has been successfully achieved by introducing plasmids that express L5 Xis. Plasmids with varying levels of expression are required for slow- and fast-growing strains, but the system is extremely efficient in both cases. Two methods have been developed to recover integrated DNA. A simple method using the *in vitro* excision reaction provides a straight forward method to excises plasmids from genomic DNA. A second system uses a phage derived shuttle phasmid to express L5 Xis in mycobacterial cells in a timely fashion which stimulates the excision of integrated vectors. In both systems, the excised DNA can then be used to transform *E. coli*. The later method has been used in *M. smegmatis* to successfully recover large and small vectors and DNA that lacks the *int* gene. Use of this system in slow-growing strains has met with limited success, but the efficiency of plasmid recovery and difficulty in cosmid transformation suggest that cosmid recovery should be established easily.

While introducing plasmids to cure strains of integrated DNA is extremely proficient (nearly complete loss of insert), the removal of the plasmid creates another step needed to create the desired strain. If the phages used to recover DNA could be used to cure strains, the non-replicating phenotype of the phage would reduce the time need to obtain the final genotype. Multiple attempts were made trying to cure strains using the phage, but all were unsuccessful. The reasons for the failure are not entirely clear, but the low level of excision, as seen by the PCR assay (Figure 33), might be a factor. Since the DNA in only a fraction of the cells is being excised, it could be difficult to isolate this small portion. It should be noted that in one of the initial recovery assays, PCR analysis showed a much higher percent DNA being excised (data not shown), but it has not been possible to replicate these results. The longevity of the excised DNA compared to the phage DNA might be another factor making curing difficult. For

curing, the DNA must be excised long enough either to be degraded and thus not reintegrated or to be excised during host genome replication. If the phage DNA is degraded more quickly than the vector and the host genome has not replicated, the amount of Xis will drop off and allow the vector to be reintegrated, resulting in no curing. Multiple factors are likely to contribute to the inability to cure strains using phage methods, however it is still possible that further optimization could make this into a feasible method for curing cells.

The majority of the integrating vectors used today contain both *attP* and *int*. The main advantage to this method is the high efficiency of transformation. However, the presence of the *int* gene has been shown to contribute to a low level excision of the integrated DNA. The alternate system in which the *int* gene is provided on a non-replicating plasmid offers stability similar to any other region of the genome. With the proposal to genetically modify BCG vaccine strains using an integrating vector (Stover *et al.*, 1991), an increased stability is highly desirable. It is important to develop tools that will allow the manipulation of integrated DNA with or without the *int* gene. The recovery of integrated vectors lacking *int* has been successfully accomplished using the phage recovery system. The same has not been demonstrated for plasmid mediated curing, but it seems quite likely that the inclusion of *int* on the plasmid should produce the desired result. Alternate source of Int have been used for integration with the two-plasmid system and in recovery using the phage system. There is no reason to assume that the same could not be applied to the curing of cells.

Throughout the course of these experiments, numerous differences have been observed between the fast-grower *M. smegmatis* and the slow-grower BCG. The recovery of DNA is 10 fold more efficient from BCG than from *M. smegmatis*. There are several reasons why this difference could have occurred. It could be an experimental difference, such as more cells used in BCG or a statistical variation, but there are some known disparities between the two species that could contribute to it. It has been

reported that isolation of high quality genomic DNA for pulse field electrophoresis has been more difficult from *M. smegmatis* than BCG. It has been speculated that one of the reasons is that BCG is easier to lyse, and others suggest that there is a nuclease problem. Either of these could affect the quality or yield of DNA used to transform *E. coli*. The hsp60 promoter used in the Xis expression constructs originates from BCG, so it is possible that it is more active in its native species, and thus more Xis is made. A further consideration is the amount of infection. While L5 is known to infect both species, there could be a difference in rate of infection or titer under the conditions used in the experiment. It has been noted previously that pMH5 is unable to transform BCG, and the most likely reason is a difference in amount of expression. An alternate explanation for the observed differences could be the presence of some factor within BCG which increases excision activity, and that could explain both the higher recover rates and the alternate behavior of pMH5 in the two species.

With the high efficiency of plasmid recovery from BCG, the recovery of cosmid size DNA should be relatively simple. However, this has not been the case, and even though large subsets of the transformed cosmid (1A9) have been recovered, a full-length copy has not. While the initial thought is that there is a deficiency in the phage recovery system, additional work supports the notion that the real problem is the inability to get full-length transformants of 1A9 in BCG. The DNA from each individual BCG transformants has the same restriction pattern even when isolated in independent recovery assays, but the DNA from each BCG transformant is different. The isolation of only partially deleted constructs taken together with the low transformation efficiency (10-20 cfu/μg) suggests that something contained on 1A9 is not tolerated in BCG. If the deletions in 1A9 are present within the identified transformants, then the phage recover system is already working efficiently. Furthermore, if a cosmid that is able to transform BCG is used, complete recovery should be observed using the current method.

The inability to transform 1A9 is not the only instance when differences in lethality have been observed between BCG and *M. smegmatis*. Another example is pMH5, in which the only BCG transformants isolated, all contain DNA with deletions (Lee *et al.*, 1991). A third example of a difference between slow- and fast-growers is the inability to transform *M. tuberculosis* with pJL22 (T. Parish personal communication). While all these examples show an increased sensitivity in slow-growers, it is reasonable to assume that examples in the other direction exist, but it is interesting to speculate that slow-growing mycobacteria might be extremely sensitive to gene dosage.

While the tools described here are useful on there own, their utility can be expanded to a variety of other systems. One such example is the removal of any DNA fragments that is flanked by attachment junctions, whether it is a plasmid or any genomic location. This could include the loss of a selectable marker introduced by homologous recombination or by an independent recombination system. It can be used to simplify genomic screens by offering an efficient means to recover the library inserts, and it offers an alternate to extremely inefficient homologous recombination for verifying the essentiality of a gene. The ability to effectively manipulate a DNA fragment is a powerful addition to the genetic toolbox used to study mycobacteria.

VI. IDENTIFICATION AND CLASSIFICATION OF RDFs

VI.A. Introduction

The control of directionality in many integrase-mediated site-specific recombination reactions is facilitated by a small DNA-binding protein known as a recombination directionality factors (RDF). This diverse group of proteins has been shown to play architectural roles in determining the substrate specificity for their cognate recombinases. Most of these proteins have been identified as part of phage recombination systems using tyrosine integrases, but this is only a subset of the variety represented. They have been found in systems as wide-ranging as transposons (Tn5276, (Rauch and de Vos, 1994)) and plasmids (pSAM2, (Sezonov *et al.*, 1998)) and also in conjunction with large serine recombinases (Ramaswamy *et al.*, 1997; Breuner *et al.*, 1999). A further divergence is the Cox family of proteins, which also have a transcriptional regulation activity in contrast to other RDFs that only have a directionality activity.

Previous to this work, no systematic attempt had been made to identify or classify the RDF class of proteins. This is due in part to the difficulty in identifying these proteins. While many have been identified by experimental means, sequence based searches have met with limited success because of the small size (most are < 100 a.a) and minimal similarity between the proteins. The diversity of this class is further exemplified by the charge of the proteins; while many are highly basic, others are acidic.

In this work, an attempt has been made to exhaustively search the sequence databases to identify as many candidates RDFs as possible. Two means were used to complete this task: the first was to find all proteins that are annotated as RDFs and then use them as query sequences in database searches to identify new members. A variety of computational techniques were used to analyze sequence similarities and amino acid composition. The information gathered from the study has allowed a careful consideration of evolutionary relationships among the RDF class of proteins.

VI.B. Determination of annotated RDFs

The first step in compiling the list of RDFs was to identify all proteins contained in GenBank that are annotated as an excisionase, excisase or xis. The combination of these three searches produced a total of 204 matches, of which 55 are duplicate records with identical accession numbers, and another 62 are redundant records that contain identical proteins. The list of the 87 unique sequences that remained was further reduced because 39 of the proteins were only included because of cross-references to RDFs (24 integrases and 15 other proteins).

A close look at the 48 remaining protein sequences reveals that some of them are extremely unlikely to function as RDFs. For example, in mycobacteriophage L5 and D29, gene 34 was originally annotated as the phage excisase due to its proximity to the *int* gene (Hatfull and Sarkis, 1993; Ribeiro *et al.*, 1997). At a later point, a different gene, *34.1*, was determined to be a better candidate because of its conservation between the two phages (Ford *et al.*, 1998). However, there is now strong experimental evidence that shows that *36* encodes the phage RDF (Chapter III). Unfortunately, since the annotation of genes *34* and *34.1* has not been updated, they were included in the initial list, and furthermore, similarity to *34.1* was used to assign an inaccurate function to orf53 from pREAT701 (Takai *et al.*, 2000). All of these genes were removed from the RDF list. Three putative *xis* genes from plasmid pME2200 and phages phiAR29 and A2 were

152

initially characterized based exclusively on their proximity to the *int*, and further analysis has not revealed any characteristics shared with other RDFs. While the possible excision function of these genes cannot be ruled out, the evidence is not strong enough to include them. Gene SCE39.01c has also been removed because it is a partial gene located at the end of a sequencing cosmid; gene SCE29.20c encodes the full-length gene. The 38 proteins in the remaining list all either have experimental evidence indicating excision activity or reasonable sequence similarity to those that have.

Text-based searches are only effective in identifying records that use specific nomenclature, so some proteins were likely to be missed. An example is the Cox family of proteins, which in addition to an RDF activity, these proteins also play a role in transcriptional regulation. There are four known members (encoded by phages P2, HP1, K139 and WPhi) of this group, but only P2 was annotated as an "excisionase". The other three were missed in the initial search, but were identified in subsequent sequence searches. Finally, there have been RDFs identified in association with the serine integrases encoded by phage TP901 and *Anabaena* XisF that were missed in the text queries. Experimental data indicates that directionality in the TP901 system requires a single RDF (Breuner *et al.*, 1999), and in Xis-F-mediated recombination, two proteins, XisI and XisF, have been shown to be necessary (Ramaswamy *et al.*, 1997). With the inclusion of these three proteins, the total number of proteins identified in text-based searches is 41.

VI.C. Searches for unidentified RDFs

The list of RDFs was expanded by searching for proteins in the GeneBank non-redundant (nr) database that have sequence similarities to known proteins. A variety of search algorithms were used including PROBE (Neuwald *et al.*, 1997), BLAST (Altschul *et al.*, 1990) and PSI-BLAST (Altschul *et al.*, 1997), and each RDF was used as a query sequence. Early in the project, it was realized that the small size and sequence diversity

made it extremely difficult to identify related proteins based on sequence identity alone. This was highlighted when searching with PROBE, an algorithm in which iterative BLAST searches are performed in conjunction with model building. In many of these searches, once weak identities from non-RDF sequences were included in model building, future iterations included few if any true RDFs. While this program has been successful for closely related families (Neuwald *et al.*, 1997), it is ineffective with this varied group of small proteins. The BLAST and PSI-BLAST algorithms, which allow interactive control, were found to be more valuable and were the primary methods used to identify new candidates.

The low information content of RDFs, resulting from their diversity and size, prevents statistically high scores for matching proteins even if they are homologous. For this reason, a secondary criteria was used for identification, in which any putative RDF (with an E <10) was investigated further in an attempt to locate nearby integrase genes. Since RDFs only play an architectural role, a recombinase is absolutely required. For genes from bacterial genomes, a region up to 50 kb was searched, as this is the size of a typical bacteriophage, and if the *int* and RDF genes flank the *attP*, they would be at the ends of the integrated prophage. Using this dual criterion method, a total of 16 new RDFs were identified.

Three additional putative RDFs were identified that are not associated with recombinase genes, but which had substantial similarity (E < 10^{-4}) to other RDFs; these are encoded by phage DR1455, phi-R67 and TM1. Two closely related genes (B2168_C1_172 from *M. leprae* and SCE68.26c from *Steptomyces coelicolor*) have moderate BLAST scores (E =10^{-3}), but no intact, cognate recombinases have been identified. In *M. leprae*, a pseudogene with a good match (E = 10^{-22}) to the integrase family was found ~41 kb from B2168_C1_172, but in *S. coelicolor* a thorough search has been hindered by the incomplete sequencing of the genome. Despite the failure to completely match the dual

criteria used for other RDFs, these five proteins have been considered as good candidates and have been added to the list.

Since many RDFs are short ORFs and some even overlap the RDF (e.g Lambda (Davies, 1980)), it would be relatively easy to miss them while annotating a sequence. To search for these potential genes, GeneBank nucleotide records containing the keyword 'bacteriophage' or 'integrase' were translated in 6 frames (using translate.pl, unpublished) and formatted into a searchable database (using FORMATDB from NCBI). The sequences were searched using each of the previously identified RDFs as a query for BLAST. Of the two significant hits, one was a previously identified psuedogene DLP12 (Lindsey *et al.*, 1989) but was not included because it was incomplete. The other (Pspu, Table 4), found in *Pseudomonas putida*, has a strong similarity to the Lambda family of RDFs (see below), and like Lambda *xis*, it partially overlaps the *int* gene. This was the last protein added to the list, making a total of 63 known or putative RDFs. The complete list of RDFs with additional information can be viewed in Table 4.

An attempt has been made to include as many true RDFs as possible, while avoiding false positives. While the rigidity of the screening should limit this possibility, there is a small chance that non-RDFs have been included. Furthermore, there were numerous sequences identified in the course of compiling the list that were reasonable candidates but were excluded because they did not meet all of the criteria. To view these borderline sequences, visit the web site http://www.pitt.edu/~gfh/rdf.html.

VI.D. Recombination directionality factor classification

The greater diversity seen in the RDF class of proteins gave an initial indication that the proteins might not be a single homologous family. When attempts to align all 63 of the proteins were problematic, and no conserved residues were observed, it was realized that the proteins needed to be divided into smaller, more related groupings.

Figure 35. Phenogram of RDF proteins

A tree based on degrees of similarities between RDF proteins was calculated with CLUSTALX (using the default parameters from http://web.tiscalinet.it/biologia/) and drawn with the DrawGram program. The vertical bars indicate groups of RDFs that stay together during multiple cycles of tree generation. The groups are named (as shown on the right) according to a member for which there is experimental evidence of RDF activity.

P22
- gi|2465478| SfV
- gi|75990| P22
- gi|4099029| SfX
- gi|6118035| APSE-1
- gi|2105132| Mx8
- gi|9635627| D3

L5
- gi|2358239| D29
- gi|15892| L5
- gi|1550698| Rv2657c
- gi|467073| B2168_C1_172
- gi|5123673| SCE68.26c

pSAM2
- gi|6459215| DR1455
- gi|1076058| pSE101
- gi|152673| pSE211
- gi|3043524| pSAM2
- gi|4490998| SCE29.20c
- gi|8517283| TM1
- gi|3261643| Rv2310
- gi|2960174| Rv3750c
- gi|3378297| pNL1

SLP1
- gi|140147| P4
- gi|93825| phi-R73
- gi|4106643| yp43
- gi|312936| SLP1
- gi|9944850| SgiI
- gi|5453491| Tn4555

Tn916
- gi|75987| Tn1545
- gi|532534| Tn916
- gi|3243184| Tn5382
- gi|8100683|Tn1549

P2
- gi|5824357| WPhi
- gi|76820| P2

HP1
- gi|459180| HP1
- gi|1679810| S2
- gi|4530499| K139
- gi|141342| phi-R67

Lambda
- gi|15761| HK022
- gi|6901614| HK97
- gi|801887| 434
- gi|4490351| H19J
- gi|215134| Lambda
- gi|4520377| Pspu
- gi|7466710| e14
- gi|215449| P21
- gi|75989| phi-80
- gi|3294479| Gifsy-1
- gi|2924238| TP901-1
- gi|3337276| 186
- gi|5824336| 16-3
- gi|5881594| VT2-Sa
- gi|1787608| ydaQ

Tn5276
- gi|6782410| ICESt1
- gi|1361379| Tn5252
- gi|1075727| Tn5276
- gi|1613875| xisH
- gi|684926| mv4
- gi|1926325| phig1e
- gi|7476829| Rv1584c

L54a
- gi|4894317| pX01
- gi|76011| L54a
- gi|1877428| T12
- gi|455128| 11
- gi|1613876| xisI

157

Table 4. The 63 RDFs analyzed in this study

Family[1]	Name[2]	Location[3]	Host	Evid[4]	Xis gi #	Rec type[5]	Rec gi #	References
L5	B2168_C1_172	Prophage	*Mycobacteria leprae*	P	467073	ND	ND	(Robison, 1994)
L5	DR1455	Prophage	*Deinococcus radiodurans*	P	6459215	Ser	6459214[6]	(White *et al.*, 1999)
L5	L5	Phage	mycobacteria	E	15892	Tyr	465416	(Lewis and Hatfull, 2000)
L5	Rv2657c	Prophage	*Mycobacteria tuberculosis*	P	1550698	Tyr	1550700	(Cole *et al.*, 1998; Hendrix *et al.*, 1999)
L5	SCE68.26c	Prophage	*Streptomyces coelicolor*	P	5123673	ND	ND	(Murphy *et al.*, 1999)
L5	D29	Phage	mycobacteria	P	2358239	Tyr	3172283	(Ford *et al.*, 1998; Peña, C.E. *et al.*, 1998)
P22	APSE-1	Phage	*Acyrthosiphon pisum*	P	6118035	Tyr	6118033	(van der Wilk *et al.*, 1999)
P22	P22	Phage	*Escherichia coli*	E	75990	Tyr	76009	(Leong *et al.*, 1986)
P22	SfV	Phage	*Shigella flexneri*	P	2465478	Tyr	2465477	(Huan *et al.*, 1997)
P22	SfX	Phage	*S. flexneri*	P	4099029	Tyr	4099030	(Guan, 1996)
PSAM2	orf277(pNL1)	Plasmid	*Sphingomonas aromaticivorans*	P	3378297	Tyr	3378303	(Romine *et al.*, 1999)
PSAM2	TM1	Prophage	*Arthrobacter sp. TM1*	P	8517283	ND	ND	(Gartemann *et al.*, 1999)
PSAM2	pSAM2	Plasmid	*Streptomyces ambofaciens*	E	3043524	Tyr	3043525	(Boccard *et al.*, 1989; Raynal *et al.*, 1998)
PSAM2	pSE101	Plasmid	*Saccharopolyspora erythraea*	P	1076058	Tyr	541467	(Brown *et al.*, 1994)
PSAM2	pSE211	Plasmid	*S. erythraea*	E	152673	Tyr	152674	(Katz *et al.*, 1991)
PSAM2	Rv2310	Prophage	*M. tuberculosis*	P	3261643	Tyr	1524291	(Cole *et al.*, 1998)
PSAM2	Rv3750c	Prophage	*M. tuberculosis*	P	2960174	Tyr	2960175	(Cole *et al.*, 1998),
PSAM2	SCE29.20c	Prophage	*S. coelicolor*	P	4490998	Tyr	4490997	(Murphy *et al.*, 1999)
SLP1	P4	Phage	*E. coli*	P	140147	Tyr	15176	(Halling *et al.*, 1990)
SLP1	phi-R73	Phage	*E. coli*	P	93825	Tyr	93827	(Sun *et al.*, 1991)
SLP1	SLP1	Plasmid	*S. coelicolor*	E	312936	Tyr	312937	(Brasch and Cohen, 1993)
SLP1	yp43	Prophage	*Yersinia pestis*	P	4106643	Tyr	4106629	(Buchrieser *et al.*, 1998)
HP1	HP1	Phage	*Haemophilus influenzae*	E	459180	Tyr	459175	(Esposito and Scocca, 1994),
HP1	K139	Phage	*Vibrio cholerae*	P	4530499	Tyr	4530503	(Nesper *et al.*, 1999)
HP1	phi-R67	Prophage	*E. coli*	P	141342	ND	ND	(Hsu *et al.*, 1990; Dodd and Egan, 1996)
HP1	S2	Phage	*Haemophilus influenzae*	P	1679810	Tyr	1679808	(Skowronek and Baranowski, 1997)
P2	P2	Phage	*E. coli*	E	76820	Tyr	6136261	(Yu and Haggard-Ljungquist, 1993)

Table 4. (continued)

Family[1]	Name[2]	Location[3]	Host	Evid[4]	Xis gi #	Rec type[5]	Rec gi #	References
P2	WPhi	Phage	*E. coli*	P	5824357	Tyr	5824355	(Liu and Haggard-Ljungquist, 1999)
L54a	L54a	Phage	staphylococcal	E	76013	Tyr	76011	(Ye and Lee, 1989)
L54a	pXO1	Plasmid	*Bacillus anthracis*	P	4894317	Tyr	4894320	(Okinaka *et al.*, 1999)
L54a	T12	Phage	*Streptococcus pyogenes*	P	1877428	Tyr	1877429	(McShan *et al.*, 1997)
Tn916	Tn1545	Transposon	*Streptococcus pneumoniae*	E	75987	Tyr	76007	(Poyart-Salmeron *et al.*, 1989)
Tn916	Tn1549	Transposon	*Enterococcus faecalis*	P	8100683	Tyr	8100684	(Garnier *et al.*, 2000)
Tn916	Tn5382	Transposon	*Enterococcus faecium*	P	3243184	Tyr	3243185	(Carias *et al.*, 1998)
Tn916	Tn916	Transposon	*E. faecalis*	E	532534	Tyr	532535	(Flannagan *et al.*, 1994; Marra and Scott, 1999)
Lambda	434	Phage	*E. coli*	P	801887	Tyr	215353	(Baker *et al.*, 1991)
Lambda	e14	Prophage	*E. coli*	P	7466710	Tyr	3024035	(Blattner *et al.*, 1997)
Lambda	H19J	Phage	*E. coli*	P	4490351	Tyr	4490352	(Schmidt, 2000)
Lambda	HK022	Phage	*E. coli*	E	15761	Tyr	15760	(Yagil *et al.*, 1989)
Lambda	HK97	Phage	*E. coli*	P	6901614	Tyr	6901614	(Juhala *et al.*, 2000)
Lambda	Lambda	Phage	*E. coli*	E	215134	Tyr	215133	(Nash, 1981; Ball and Johnson, 1991a)
Lambda	Pspu	Prophage	*Pseudomonas putida*	P	4520377[7]	Tyr	4520380	(Nashimoto, 1999)
Lambda	P21	Phage	*E. coli*	E	215449	Tyr	215450	(Schneider, 1992; Wang *et al.*, 1997)
Tn5276	ICESt1	Transposon	*Streptococcus thermophilus*	P	6782410	Tyr	6782411	(Burrus *et al.*, 2000)
Tn5276	Tn5252	Transposon	*Streptococcus pneumoniae*	P	1361379	Tyr	1361380	(Kilic *et al.*, 1994)
Tn5276	Tn5276	Transposon	*L. lactis*	E	1075727	Tyr	1075733	(Rauch and de Vos, 1994)
misc	11	Phage	staphylococcus	E	455128	Tyr	166159	(Ye *et al.*, 1990)
misc	16-3	Phage	*Rhizobium meliloti*	E	5824336	Tyr	5824335	(Semsey *et al.*, 1999)
misc	186	Phage	*E. coli*	E	3337276	Tyr	3337277	(Dodd *et al.*, 1993)
misc	D3	Phage	*Pseudomonas aeruginosa*	P	9635627	Tyr	9635596	(Kropinski, 2000)
misc	Gifsy-1	Phage	*Salmonella typhimurium*	P	3294479	Tyr	3294478	(Figueroa-Bossi and Bossi, 1999)
misc	mv4	Phage	*Lactobacillus plantarum*	P	684926	Tyr	684925	(Dupont *et al.*, 1995)
misc	Mx8	Phage	*Myxococcus xanthus*	P	2105132	Tyr	2149006	(Salmi *et al.*, 1998)
misc	phi-80	Phage	*E. coli*	E	75989	Tyr	75992	(Leong *et al.*, 1986)
misc	phig1e	Phage	Lactobacillus	P	1926325	Tyr	1926326	(Kodaira *et al.*, 1997)

Table 4. (continued)

Family[1]	Name[2]	Location[3]	Host	Evid[4]	Xis gi #	Rec type[5]	Rec gi #	References
misc	SgiI	Prophage	*Salmonella enterica DT104*	P	9944850	Tyr	9944851	(Boyd *et al.*, 2000)
misc	Tn4555	Transposon	*Bacteroides fragilis*	P	5453491	Tyr	5453489	(Tribble *et al.*, 1999)
misc	VT2-Sa (933W)	Phage, converting	*E. coli*	P	5881594	Tyr	5881593	(Miyamoto *et al.*, 1999; Plunkett *et al.*, 1999)
misc	ydaQ	Prophage	*E. coli*	P	1787608	Tyr	1787607	(Blattner *et al.*, 1997)
misc	Rv1584c	Prophage	*M. tuberculosis*	P	7476829	Ser	7476830	(Cole *et al.*, 1998)
misc	TP901-1	Phage	*L. lactis*	E	2924238	Ser	6808404	(Breuner *et al.*, 1999)
misc	xisH	Chromosomal	anabaena	E	1613875	Ser	1075645	(Ramaswamy *et al.*, 1997)
misc	xisI	Chromosomal	anabaena	E	1613876	Ser	1075645	(Ramaswamy *et al.*, 1997)

1. The family groupings based on sequence similarity.
2. The name used for the protein in this chapter.
3. Location of the gene as found in database searches.
4. Type of evidence indicating functionality as recombination directionality factor (Experimental or Putative)
5. The type of recombinase as classified by its catalytic residue (Tyrosine or Serine)
6. The recombinase gene has a frame shift and is therefore not annotated as a protein. It is contained within this nucleotide record (88-1096).
7. The xis gene has not been annotated as a protein. It is in this nucleotide record (4773-4531).

To analyze potential groups, phenograms representations of guide trees created using CLUSTAL-X (available from BioWeb at http://www.web.tiscalinet.it/biologia) were analyzed. After multiple generations of the guide trees were examined (representative phenogram in Figure 35), some of the proteins were repeatedly grouped in the same clusters. These sets of proteins give a good indication of their relatedness. This is supported by PSI-BLAST analysis which identified matching groups.

A total of 10 distinct groups have been identified that together include 46 of the proteins. Each of the groups contains at least one member with experimental data supporting RDF activity, so the families were named according to these proteins (i.e. P22, L5, pSAM2, SLP1, TN916, L54a, P2, HP1, Lambda and Tn5276 families). Another 17 of the proteins did not have significant similarity to more than one other protein and were generically classified as miscellaneous. Database scans with PSI-BLAST provide support for the 10 families, and in 4 of them (Tn5276, Lambda, Tn916 and L54a), the searches converged on a list of proteins that contained all the members of the family but did not include any other putative RDFs. Since members of the L5, pSAM2, SLP1 and P22 groups showed a variety of crossover in search results, there is evidence for homology between these proteins. Joining three of these families has created the L5-SAM-SLP1 family. However, the P22 family was not included because of several insertions that are present in all of its members, but lacking in the larger L5-SAM-SLP1 family. The P2 and HP1 subgroups where originally classified independently based on phenogram analysis, but the proteins show a great deal of similarity and also function as *cox* transcriptional regulatory genes. For these reasons, these two groups have been combined to form the P2-HP1 family.

The relatedness of each family is illustrated by their sequence alignments (Figure 36). The L5-SAM-SLP1 family is the largest and contains 17 members. The alignment of the core ~50 amino acids does not show any completely conserved residues, but there are many locations in which the similar residues are found in all the proteins (Figure

Figure 36. Sequence alignments of RDF proteins.

RDF proteins within individual families are shown using alignments derived from CLUSTALX analysis (using the default parameters as in Fig. 1). Amino acid residues that are identical in 65% of the sequences are highlighted in red and residues that are similar among at least 75% of sequences are shown in blue. Similarity groupings were based on positive scoring substitutions as determined by the BLOSUM 85 substitution matrix (Henikoff and Henikoff, 1992). The location of a putative helix-turn-helix DNA binding motif is shown above the L5-SAM-SLP1 and P22 families. In cases where RDFs from different sources are of identical sequence, only one was used in the alignment; the view of each alignment is limited to a 150 residue segment containing the most similar region of the sequences.

L5-SAM-SLP1

HTH

```
gi|2358239|D29        --------------------------------------------------MPQRA-----SIQTMDFLGVSTKTVRRYIADIRLK-AVRLLPRLIIVERDSVEALMRP--IGK----------------------------------
gi|15892|L5           --------------------------------------------------MPPRA-----SIQTIDYLGVSTKTVRNYIADIRLK-AVRLLPRLIIVERESVEELMRP--IGK----------------------------------
gi|1550698|Rv2657c    ------------------------------MCAFPSPSLGWTVSHETERPGMADAPPLSRRYITISEAIEYLAVTDRTSVRQMIADIRLR-GYHSITRLVGLKRDEVDGAMHP--FGGAA--------------------------------
gi|467073|B2168_C1_172 ------------------------------MTSTNGPSARDSAGKSVRDTGSFEGQQPRTQFITVAEVEILMRVSKMIVYRAVHNGELP-AVRVG-RSFIVHAKAVHDMLETSYFDVG--------------------------------
gi|5123673|SCE68.26c  ------------------------------MAAAG---------------ERPLNEVQFITVAEVEISVHRVSKMIVYRAVHSIHLP-AIRVG-RSFIVPEQAVHEYLRESYVGVETI-------------------------------
gi|1076058|pSE101     --------------MST-----------------------PAPASLPRPRESGDATRLLATVEQAARRLSVGRTTMFKIKSIDVD-SVR1G-HARIVPVEALTAVIDR-LARGAA--------------------------------
gi|152673|pSE211      --------------MTTNVTQLAATLASLAALLAEQQPAPEPEPEPAARRLPNRVLLATVPEVMARLKVGHGKTKTYAIVASIETE-SVR1G-RLRIIPRTALDDVAARLIAQQSAN-------------------------------
gi|3043524|pSAM2      --------------------------------MTTVTP----ELLATVPEVMARLKVGHSKVYDIIRTRRLA-SIKID-GARIVPTDAVRDFIQDQLGEAI------------------------------------------
gi|4490998|SCE29.20c  ----------------------------MTAATAVATAELLATVPEVMARLKVGHSKVYDIIRSRRLV-SIKID-GARIIPAAAVRDFVLGQIEEAA------------------------------------------
gi|8517283|TM1        MRPKNGSQPSPQFFLSGADAGDQVEIPHEIYEVLVKTVEAMRKGLAITTITPSSMTLATQQAIELLGVTRPTVVRLLDAIKIP-FEKPM-THRIIKLEDVLAFKEVRKAEQYEIALDS--LRVLAINSFAAINRNKKHIVIAAMRRHQHKE-
gi|3261643|Rv2310     --------------------------------MVAALHAGKAVTIAPQSMTLATQQAIELLGVTRPTYVRLLDAIKIP-FEKPM-THRIIKLEDVLAFKEVRRAEQYEIALDS--LRVLAINSFAAINRNKKHIVIAAMRRHQHKE-
gi|2960174|Rv3750c    ---------------------------NAGQPMTLPEPVRDALYNVVLALSQGKGISLVPRHLKLITQRAIDLLNISRPTLVRLLEDIRIP-FEKPG-RHRIVSLDALLEYQQETRSNRRALGE----------------------------
gi|140147|P4          ----------------------------MQAVFSSPSPAPVTFMLPLDITQERFLRVPEVMHLGLSHSTIYLIRKEIEFPPQVSLIGGKNVAMLHSEVTANHKAGRKRGYDA----------------------------------
gi|4106643|yp43       -------------------------------------------------MTSYQLARLRQVEQKTGLKRSQTYLYMKEIAFPRSIKIIPASVAWLESEIDEWINKKLSDR-----------------------------------
gi|3378297|pNL1       MAQKAGHAKLRVPDPKTRKPVEIALTPAMSDLFLRLLRHIGSGDAVTLVPIQQMITTQQAIDLLNISRPTYIIKIEKIDIP-HSMVI-RHRILKAEDVFAYKAKKDETRAKINDDLLSIIDAILY------------------------
gi|6459215|DR1455     -------------------MWWPHVRFACLYMLVHEGVAPSARSEERRSVMTLASQFDNLPKLLKVSEVIDFTGTHERTVRRWIKDIRLG-AVEHF-SGLIVPRRSLWRFLGLEFALSA--------------------------------
gi|312936|SLP1        --------------------------------MARPKMLKLSEVLEEITMSHAAFYRTRARGKAPKLIKLPNGHIICRRSDLDAWATHEIAA----------------------------------------------
ruler                 ........40........50........60........70........80........90.......100.......110.......120.......130.......140.......150.......160.......170.......180.
```

P22

HTH

```
gi|2465478|SfV        MGYDSRLDRLAATSWYPFFNNVTARGEIHEPYSSLLDKACDPIKIIRPIAIHWIITGRLQATVRQDAISKSPYEITRQACIAALGSPLITVQVSAGCGITIERKCHRGAAEYAFVSYVSHCFTVIDLNSLICRTKGHRQNSMES------
gi|75990|P22          -----------------------------MEEEIFPREEIEASYLAVVDKGITVWIEMDDLQAAKIMKDKPSSPVRICKSDIIAALKAVRNSNAVI-VQVVKALCFNYAIARRYTAIESRQMDQELDALKHPTGKRRKNTIV-----
gi|4099029|SfX        -----------------------------MEEEIFPREEIEASYLAVVDKGITVWIEMDDLQAAKIMKDKPSSPVRICKSDIIAALKAVRNSNAVI-VQVVKALCFNYAIARRYTAIESRQMDQELDALKHPTGKRRKNTIV-----
gi|6118035|APSE-1     --------------------MRMVNKATIYRQRIAKIIGIKEDILSIMCKLRLIISYQSRNGYYKKSPYLFIKEACIEIVKNSINILSEIRID--VIRSKKCQQ1HNVIWFDAATQTHILAEKRLRNHIAQRTMNKLRSYTINEELNYGE
ruler                 .....10........20........30........40........50........60........70........80........90.......100.......110.......120.......130.......140.......150....
```

P2-HP1

```
gi|4530499|K139       -MLTYEVAP-------------------------------------ESSYFTCEKYSVISILPMGTIRQYIAESRIIIK--PKTK----TKEKPLANHVIMH----------EIAARHEAMQVLG------
gi|141342|phi-R67     -MLTYLFCDPSISFVPIWVLVTITTGSPQANRRCCIMTPNISTYLNTHVIVIERYSEITLIIDTINDILADVRFIRHRERKDK---KREKVMTNGAALT--------VDALTCNVVFN------
gi|459180|HP1         -MSKQNAICIN------------------------------IHMEQVYMFREEFYAKKLDVITRTTIDALRQQCVEKCIKHKNDEGEETERLVILEDIVIA----------VRNAKNAFQI------
gi|1679810|S2         -MNNGQNAICIN-------------------------------VQLHAEVYILKKYIEDITLIILDKVRKRAAKELIADKK-----AERRSVILNNIIIA--------KQAAKQE------
gi|5824357|WPhL       HEVNDYVIQYP------------------------------IDAIHTVKFREIIKTGAVRRIDKRKCIVIELRDPSKPNARVRKWFPEFNRAVREEPYNIPVEQRDAWLLWMI------
gi|76820|P2           ISK---QVTLM---------------------TDAIPYQFFKLIIKITGAVRRIDKRKCIVIDMTDPQSASGRAIEYMYTPRWNNGLKLLAVESIPKEIRDGWLMWLIGEPR
ruler                 1........10........20........30........40........50........60........70........80........90.......100.......110.......120.......130
```

L54a

```
gi|4894317|pXO1       MDYLSLEEVCDRVGLAKNQLG-TLIKYKLIKPINLATWKADGGYRFEQEDVRLEELAKDSLTIKEAAAFLNKSKNVHNAAKDGILPFKEITKGKSTERLYLKKDLEIFKERIENRSKEESKDKKQHLSLYLDEKVLEAIKKKAEKKGYN
gi|76011|L54a         -------MD--IKFIDAEIA-RIKKS-------VEMKSSE--FDIYDDEQLLKLKIQ--EQIMLAIRKSIYKD----IYL----------FLIN----------------------------------------------------
gi|1877428|T12        -------PHGWNYNSYKANNNNILIIRTLED--IYKIN----TLMRDQYKIGSNEI-NAFQQDIEQDKNQIALLFN--------EHQILENIDN----------------------------------------------------
ruler                 1........10........20........30........40........50........60........70........80........90.......100.......110.......120.......130.......140.......150.
```

Tn5276

```
gi|6782410|ICESt1     MLILAEVPIENCKAEWGGLALMKRPEGKLATLANLAPENSMKPQDKDFVILSGIRTIWIHFSGPLNFLKYRENSIPIEVG
gi|1361379|Tn5252     -NNIKNIGLRKFHCRMAYIIKKSILAFWGGINKYTLARWIKRSIKEKRTGSMFVINPINLAFIILEGIFESFLRGKLIKTI---
gi|1075727|Tn5276     -MEENVK---------DLRRISELFDG-VPIKTLANLDAFEMSH-RTENVFIILRISHIRSTINVQYKSFSIETKQKIIESTI-
ruler                 1........10........20........30........40........50........60........70........80.
```

Lambda

```
gi|7466710|e14        HLQMLFLEIWRAAEKYRSNPFPSVSSTLAHYAKKQNEFCIHAMSQERLWRVREDNELAIGEHVTIVLKKNDSLLAQRLLSNGSQTA--
gi|215449|P21         HSRLITLGDVAKESEGDLAPIERVLKKYAGGKMMAIAITVIGWYWMIDRNSRFIGTLABIQLFINANPKIQALIADGC-----
gi|215134|Lambda      --MYLTLQENNARQ--RRRFRLSEIVRRVRESCRIIFGPVFIDGREYLFHESIVKIQD---LNKIFVTGG-LIKKAFRNGKK--IKS-
gi|4520377|Pspu       -MCKVIIDINGHDHR-RTPFSNMEERRWSREGLITIAPVRIGNSYYVEADVCYREPEPLIKVIVGSP-IVNLYASARYGIQAA
ruler                 1........10........20........30........40........50........60........70........80.
```

Tn916

```
gi|8100683|Tn1549     NNN--EMPIWEKYTLTIEEASKYFRIGENKLRKLAEENPASGWVIMNGNRIQIKRKQFEKIIDTLGTI
gi|3243184|Tn5382     NNN--IVPIWEKYTLTIEEASKYFRIGENKLRKLAEENPASGWVIMNGNRIQITT------SSSRK-
gi|532534|Tn916       IKGTHIPIWSRYTLTLEEASKYFRIGENKLRLAEENKNANSLIIMNGNRIQISKNQFEKTIDELGTI
ruler                 1........10........20........30.........40........50........60...
```

163

36). A notable difference between the members is the distance between the beginning of the sequence and the relatively well-conserved core region. While this could suggest evolutionary differences, it may also be explained by erroneous assignment of translation start sites. The 6 members of the P2-HP1 family do not have any conserved residues but do have several locations with similar chemical characteristics (Figure 36).

Each of the 5 remaining groups contains fewer members, and as might be expected, they tend to have more similarities than the above families. A 110 residue segment of the P22 family contains 21 conserved residues. Only three members of the Tn916 family have been shown, but they are extremely similar with 42 identical residues in a ~70 a.a. stretch. (Tn1545 Xis was excluded from the alignment because it is identical to Tn916, but since the cognate integrases are slightly different it has been included as a separate RDF.) Alignment of the full-length proteins from the Lambda, Tn5276 or Tn916 families show identity at 10, 18 or 42 locations, respectively. (HK97, HK022, 434 and H19J Xis have been removed from the Lambda family alignment because they are identical to Lambda Xis). The L54a family shows the most considerable variation, but the members are still more similar to each other than any other RDFs. Furthermore, the charge properties of these proteins (see below) are unique to this group.

VI.E. Protein charge

While it has been previously reported that many RDFs are basic, with predicted pIs in the range of 9-10 (Ye *et al.*, 1990; Lewis and Hatfull, 2000), no systematic verification of this has been performed. Analysis of the pIs of all the identified RDFs indicates that only about half or them are in this narrow range, but as predicted, the majority is basic (Figure 37). Only 10 of the RDFs have pI values of < 7 (listed in Table 5), and a closer examination reveals that five of these, three from the L54a family and two (Mv4 and phig1e) from the miscellaneous group, might have a shared ancestry.

Figure 37. Distribution of pI values among RDF proteins

The predicted isoelectric focusing point (pI) was calculated for each RDF using Compute pI/MW (http://www.expasy.ch). The number of RDFs with pIs within 0.5 pH intervals was determined and plotted. The majority of the proteins are basic with only 10 of the 63 RDFs having a pI < 7. Five of the seven proteins (Lambda, HK97, HK022, 434, H19J) in the 11-11.5 range have identical sequences.
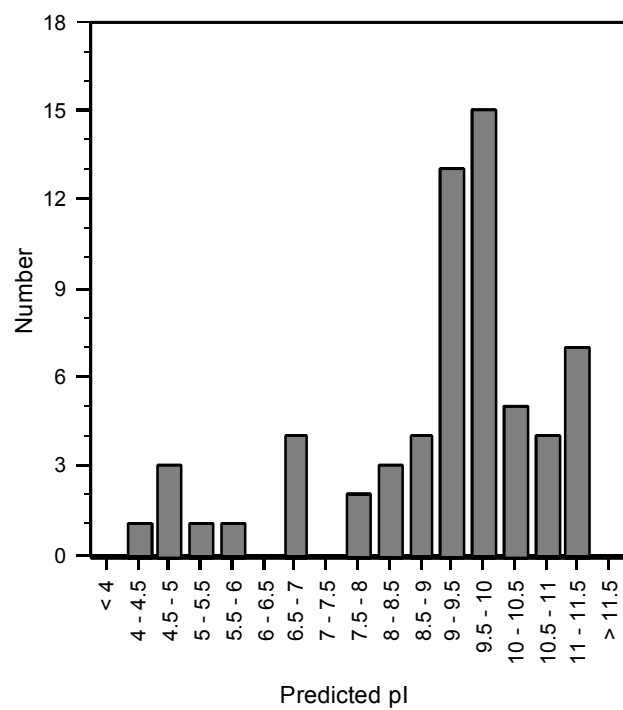
Table 5. Acidic recombination directionality factors

Name	pI	RDF Family	Int Family
Mv4	6.51	misc	LC3
phig1e	4.57	misc	LC3
L54a	4.53	L54a	LC3
T12	5.15	L54a	LC3
pXO1	6.87	L54a	misc
11	4.13	misc	phi11
WPhi	6.58	P2	P2
orfA	6.93	pSAM2	N/A
xisH	4.78	misc	Serine
xisI	5.88	misc	Serine

While the Mv4 and phig1e Xis are not similar enough to be grouped in the L54a family, the Int for these and two of the L54a family are members of the LC3 family of integrases. Other RDFs associated with LC3 integrases are basic and in the Tn5276 family (Esposito and Scocca, 1997) (Table 6). This distinction amongst protein charges suggests that a new RDF was acquired more recently than the LC3 recombination systems diverged from other integrases.

VI.F. Other RDF properties

While all RDFs are assumed to be DNA-binding proteins, the particular domain involved has not been determined. Some work has suggested that the helix-turn-helix DNA binding (HTH) motif has this responsibility in some RDFs (Salmi *et al.*, 1998; Lewis and Hatfull, 2000). Using the methods of Dodd and Egan (Dodd and Egan, 1990), the program HTHpred was written to predict the presence of this motif. Twenty-three of the putative RDFs have a probability greater than 25% of contain a HTH motif, and three others scored 2.2 which is just below the 25% threshold with a score of 2.5. All 26 proteins are reasonable candidates for containing the domain. Nineteen of these RDFs are in the L5, P22, pSAM2 or SLP1 families, so only 3 (pSE101, pSE211 and SLP1) of the 22 members in these families were not predicted to contain the HTH motif. None of the Lambda group of RDFs contains this motif.

VI.G. Evolutionary considerations

The examination of these known and putative RDFs has provided insight into their evolution. The first thing that can be considered is how closely the historical paths of the integrating element coincide with that of their hosts. The members of the Lambda family are all encoded by similar phages that infect related hosts, but this is not the case for other families. The L5 group has members that are encoded by mycobacteriophages and a putative prophage in *Deinococus radiodurans*. Another

Table 6. RDF versus integrase classifications

Name[1]	RDF Family	Int Family[2]	Name[1]	RDF Family	Int Family[2]
L5	L5	FRAT1	HP1	HP1	P2
Rv2657c	L5	FRAT1	K139	HP1	P2
D29	L5	FRAT1	S2	HP1	P2
APSE-1	P22	P22	P2	P2	P2
P22	P22	P22	WPhi	P2	P2
SfV	P22	P22	434	Lambda	Lambda
SfX	P22	P22	e14	Lambda	Lambda
pNL1	pSAM2	misc	H19J	Lambda	Lambda
pSAM2	pSAM2	misc	HK022	Lambda	Lambda
pSE101	pSAM2	pSE	HK97	Lambda	Lambda
pSE211	pSAM2	pSE	Lambda	Lambda	Lambda
SCE29.20c	pSAM2	pSE	Pspu	Lambda	Lambda
P4	SLP1	P4	P21	Lambda	Lambda
phi-R73	SLP1	P4	11	misc	phi11
SLP1	SLP1	P4	16-3	misc	misc
yp43	SLP1	P4	186	misc	P2
Tn1545	Tn916	Tn916	D3	misc	P22
Tn1549	Tn916	Tn916	Gifsy-1	misc	Phi-80
Tn5382	Tn916	Tn916	mv4	misc	LC3
Tn916	Tn916	Tn916	Mx8	misc	misc
ICESt1	Tn5276	LC3	phi-80	misc	Phi-80
Tn5252	Tn5276	LC3	phig1e	misc	LC3
Tn5276	Tn5276	LC3	SgiI	misc	misc
L54a	L54a	LC3	Tn4555	misc	misc
pXO1	L54a	misc	VT2-Sa (933W)	misc	phiCTX
T12	L54a	LC3	ydaQ	misc	p4

1. The name of the RDF gene as described in Table 1. Only proteins for which a tyrosine integrase was found are included.
2. Int family was based on Esposito's classifications (Esposito and Scocca, 1997).

subfamily, pSAM2, has members which are found in hosts as varied as *Sphingomonas, Arthobacter sp., Streptomyces ambofaciens, Saccharopolyspora erythraea, Mycobacterial tuberculosis*, and *S. coelicolor* and in both plasmids and prophages. If these groups are truly homologous, then the RDFs have spread throughout the phage, plasmid and bacterial population by widespread lateral transfer.

Secondly, the evolutionary relationships between RDFs and their cognate integrases can be investigated. Independent classifications of the proteins from the integrase family (Esposito and Scocca, 1997) can be compared to family groupings for RDFs. For 52 of the RDFs, full-length tyrosine integrases have been identified, and 40 of these were already classified (Esposito and Scocca, 1997), while the other 14 were only recently completed (D. Esposito, personal communications). When looking at RDFs that have been grouped, the associated integrase proteins form corresponding families (Table 6). An example of this is the Tn916 families of Ints and RDFs, in which all four RDFs form the same group as their cognate recombinase. While RDFs participate in extensive lateral exchange throughout the microbial world, these observations suggest that the integrase and RDF are relocated as a genetic pair.

The one notable exception to this pattern is with the Tn5276 and L54a families of RDFs. While these two groups are relatively distinct, 5 of the 6 total proteins (ICESt1, Tn5252 and Tn5276 from the Tn5276 family and L54a and T12 from the L54a family) have cognate integrases that are members of the same family (LC3). The variation could be an artifact from the methods used for grouping, but in addition to the sequence alignment data, there is a distinct difference in overall protein charge between the two groups of RDFs (above, Figure 37, Table 5). Despite the general trend of co-evolution of Int and RDF pairs, this is an example where the two genes have different evolutionary paths.

The final evolutionary consideration is whether all the RDFs have evolved from one common ancestor or if there are multiple origins. Based on the examination of the

identified RDFs, it seems probable that there are at least four independent origins for this class of protein. The first unique group, L54a, has little or no sequence similarity with other RDFs, and unlike the majority of RDFs, the protein charge is quite acidic. Members of the second distinct family, P2-HP1, are *cox* genes, which in addition to the RDF activity, are also transcriptional regulators. The third distinguishing characteristic is the presence of a helix-turn-helix DNA binding motif in proteins from the L5, SAM, SLP1 and P22 families. This domain is not observed in the Tn916, Tn5276 or Lambda families. The presence of a specific DNA binding domain and the difference in primary structure suggests that these proteins could have an independent origin. The three remaining families (Tn916, Tn5276 and Lambda) have a significant amount of sequence diversity, but are all small, basic proteins with no characterized domain structures; this prevents a clear interpretation of evolutionary relationships. While any speculation on origins of proteins must be done with caution, the small size and function as a DNA-binding, architectural protein with no catalytic activity makes it more plausible to acquire a new gene that can assume the role of a RDF.

VI.H. Final considerations

While extensive database searches have been performed and a great diversity of RDFs with multiple origins has been identified, there is still the question of the completeness of the list. There have been more than 120 tyrosine integrases that have been described (Esposito and Scocca, 1997), and it is a reasonable assumption that the majority of these should have an accessory protein that mediates directional control. The list of RDFs identified in this work is only about half of that number, so there are still a large number of RDFs that have yet to be recognized despite the presence of their sequence in existing databases. The methods of sequence identification used here have not been sufficient to identify these proteins, so a different approach will be necessary. Perhaps more advanced models developed from the RDF families can be useful, but

171

complete detection will require experimental dissection of novel recombination systems.

The architectural role of Xis proteins in recombination has been shown in Lambda and L5, and it is probable that other RDFs associated with tyrosine integrase (Int-Y) will have a similar function. However, in systems that utilize a large serine recombinase, the mechanism of recombination, including directional control, is not been well studied. Only two such systems have been investigated biochemically (encoded by ϕC31 and TP901), and in both, the minimal sites required for recombination are less than 60 bp (Thorpe *et al.*, 2000; Breuner *et al.*, 2001), which is much shorter than the *attP* in tyrosine integrase (Int-Y) systems. This DNA requirement indicates that the nucleoprotein complex must be very different in Int-S recombination. An RDF has only been identified in one of these studied systems (TP901) (Breuner *et al.*, 1999), but no data exists on how it functions. Another Int-S associated putative RDF, Rv1584c, which has minimal experimental data supporting RDF activity (L. Bibb and G.F. Hatfull unpublished), has some weak similarity to the L5 family. An RDF in the L5 family (Rv2657c) is in a prophage-like element (ϕRv2) in *M. tuberculosis* that has an Int-Y. Interestingly, Rv1584c is in a similar element (ϕRv1) with a colinear arrangement of the RDF and *int* genes, adding support to the RDF function. The only other accessory proteins identified in directional control for Int-S mediated recombination are in *Anabaena*, but in this case, two proteins XisH and XisI are required. Even though there are several Int-S systems with experimental data indicating a requirement for RDFs, the role they play in directional control has not been determined.

The information contained here is only a beginning and will grow as new RDFs are identified and new family groupings are characterized. This collection and classification of RDFs and putative RDFs should be useful in annotation of future phage, plasmid and prophage sequencing projects. To aid in this process of discovery, a web site describing known and putative RDFs and borderline proteins has been

established at http://www.pitt.edu/~gfh/rdf.html. An equivalent site characterizing tyrosine recombinases can be found at http://members.home.net/domespo /trhome.html.

VII. CONCLUSION

VII.A. Control of directionality in L5 site-specific recombination

The control of directionality in L5 integrase-mediated site-specific recombination was shown to be controlled by a small, basic accessory protein, gp36, termed Xis. The presence of the gene *in vivo* promotes efficient excision of an integrated vector; the *int* gene must also be expressed. To assay conditions required for excision, the protein was overexpressed, purified and used to establish an *in vitro* excision assay. The requirement for supercoiled substrates is not as strict as in the integration reaction, and there is also a difference in mIHF specificity. While mIHF is absolutely required for integration, *E. coli* extracts are able to substitute for this role, but the *E. coli* protein participating has not been identified

The role of L5 Xis in excision was determined to be mediated by binding to a region within *attR* between the P2 and P3 arm-type sites. Binding to this region creates a bend in the DNA that is required for complex formation. In excisosome-R, an intramolecular bridge is formed between the arm- and core-type sites by Int molecules. While Xis is necessary for part of the bend, a second accessory protein, mIHF, is also required to support complex formation. Excisosome-L is analogous to the intasome complex previously characterized for the integration reaction; both contain integrase-mediated intramolecular bridges between the core and the P4 and P5 arm-types with mIHF facilitating the bend. For excision to occur, the two excisosomes must synapse, but this complex has not been identified. However, Int-Int protein interaction could

174

bring the two together and would require no significant rearrangement of either complex.

The inhibition of integration is also mediated by Xis binding to the region between the P2 and P3 arm-type sites. The presence of Xis prevents the formation of the intasome and both synaptic complexes, but notably in the inhibited complex, the P4/P5 to core Int bridging found in the intasome and synaptic complex 1 does not appear to be disrupted in any way. The mechanism of inhibition is by preventing the capture of *attB*. While there is minimal binding to the P1 and P2 arm-type sites, presumably by Int, it is more likely that the presence of Xis in this complex prevents Int from binding to these sites.

VII.B. Mycobacterial toolbox

The identification of L5 *xis* has been critical in expanding the versatility of tools that utilize the L5 integration system. By placing the L5 *int* and *attP* genes on a plasmid, the DNA can be used to efficiently and stably transform mycobacteria by integration into the bacterial *attB* site (Lee *et al.*, 1991). This system has been used to introduce large pieces of DNA (Stover *et al.*, 1991; Pascopella *et al.*, 1994) without the multi-copy effects associated with extrachromosomal plasmids (Banerjee *et al.*, 1994; Barsom and Hatfull, 1996). A specific issue that has been lacking with these vectors is the ability to remove the plasmid in order to cure the strain or to transform *E. coli* for amplification of the DNA. Methods have been developed to accomplish both of these goals. The curing of strains has been achieved by introducing plasmids that express Xis, which causes a close to total loss of integrated DNA. It was necessary to use plasmids with different levels of expression in fast and slow growing strains of mycobacteria because of sensitivity to Xis in *M. tuberculosis*. Two different procedures were created to recover integrated vectors into *E. coli*. The first takes advantage of the *in vitro* excision reaction to remove the vector from purified genomic DNA. The second technique makes use of

non-replicating shuttle phasmids that express Xis and stimulate the excision of the integrate. For both systems, the excised plasmid can be used to transform *E. coli*. While the use of these methods just in combination with the integrating vector system should greatly enhance mycobacterial genetics, it is easy to envision new tools that utilize these methods that would allow a broad range of genetic manipulation.

VII.C. Recombination directionality factors

Many systems have been identified that use small, accessory proteins to mediate directional control. These recombination directionality factors (RDFs) are an extremely diverse group of proteins that have been difficult to identify due to their small size and sequence diversity. A systematic search of sequence databases has identified 63 putative or known RDFs that are associated with both serine and tyrosine recombinases. Comparison of sequence similarities and amino acid composition has revealed that the proteins can be divided into at least 10 distinct family groupings. Interestingly, the cognate recombinases can be placed into congruent groups, suggesting a co-evolution of the two proteins. However, the integrases are likely to have a common ancestor, but the RDFs appear to be derived from at least four different ancestral origins. While this classification will provide insight into directional control of site-specific recombination, it will also be helpful in future identification of novel RDFs.

VII.E. Specificity in Mycobacteriophage L5 recombination

Much of the specificity in recombination is achieved through cooperative interactions between multiple proteins and binding sites. Binding by any individual protein involved in recombination is achieved at a higher concentration than when different proteins are present. This is seen quite readily when looking at the mIHF and Int alone footprints (data not shown), which show little if any protection. However, when these proteins are combined in the assay at the same concentrations, there is

significant protection of much of the attachment site. Furthermore, Xis binding to *attR* occurs at a much lower concentration when Int and mIHF are present. A further example of cooperativity is demonstrated by the Xis alone complex formation, which changes from no complex to involvement of nearly all of the DNA in an extremely short range of protein concentration.

The specificity and control of recombination is essential for the continued viability of the bacteriophage. The system of recombination that has evolved uses multiple proteins and binding sites that must act together to control recombination. An additional protein, the RDF, is only expressed at times when excision is desired. To add to the specificity, all of these proteins require cooperative interactions to allow binding. Only when all the proteins are present and intricate nucleoprotein complexes have been constructed is efficient recombination possible.

VII.D. Abnormal recombination

In general, the L5 recombination reaction shows a great deal of specificity in determining directionality, but this control is not absolute. The low level of Int-mediated loss of integrated vectors occurs in the absence of Xis. The inhibition of integration in the *in vitro* recombination reaction is not complete; even at the highest concentrations of Xis, there is detectable formation of integrative products. The difference between specific and unregulated recombination can be estimated from the experimental data. There is about 10-fold difference in product formation in inhibition of excision and about a 1 % loss of integrated vectors. While these numbers are achieved through different techniques, it suggests that excision is more tightly controlled than inhibition of integration.

VII.E. Future Work

VII.E.i. Arm-type site alignment

While the study of the L5 excision system has generated a better understanding of L5 recombination, there are still a number of questions for examination. The occupancy of the Int mediated bridge between arm-type and core sites has been characterized in both integrative and excisive complexes, but it is possible that rearrangements could occur during different phases of recombination. The asymmetrical arrangement of the core and arm-type sites creates specific restrains on how the DNA can interact with the integrase tetramer. These limits are a result of trying to maintain consistent chemical environments for the binding domains of each Int molecule. The inverted repeats of the core sites require a binding site that is radial to the proteins (Figure 38A), which is consistent with the structural data of integrases complexes (Guo *et al.*, 1997). The tandem repeats of the arm-type sites require more of a tangent wrapping of the DNA around the tetramer (Figure 38A). Furthermore, the orientation of the asymmetric crossover in the core and the tandem alignment of the P1/P2 and P4/P5 sites across the core imparts further restraints.

Two possible arrangements of arm-type and core interactions exist. One is cis conformation in which either the P1/P2 or the P4/P5 pair of arm-type sites is involved in Int bridges to only one core, but the other possibility is a trans conformation is in which an arm-type pair is bound to Int molecules that are each bound to a different core molecule (Figure 38B). Each of these conformations requires a different wrapping of the DNA molecules to satisfy the arm-type binding requirements. While one model for complex formation is a simple folding of the DNA, in actuality a wrapping of at least one-of the arms around the complex is often required (Figure 38B).

Another consideration is that upon completion of recombination, a cis conformation of substrates would produce a trans arrangement of products, and visa

**Figure 38.** Theoretical restrains on complex formation

Due to site orientation and the requirement of identical chemical environments for each molecule in the Int tetramer, there are certain limitations to the conformation of bound DNA (A). Binding of inverted repeats requires a radial arrangement of binding sites, while the direct repeats found in the arm-type sites need a tangential alignment. The radial alignment of the core sites has been observed in the Cre / loxP co-cyrstals. A second consideration in complex formation is the conformation of the arm-type sites to core bridges. The alignment that has been predict from most observed complexes is a cis conformation, which indicates that the pair of Int molecules that bind a pair of arm-type sites each binds the same core. The alternate conformation, trans, has each Int molecule bound to a pair of arm-type sites binding different cores. Both conformations are likely to exist at some point during recombination because recombination changes the conformation. The tandem alignment of the P1/P2 and P4/P5 arm-type sites across the molecule imparts further restraints on how the DNA must wrap around the complex.
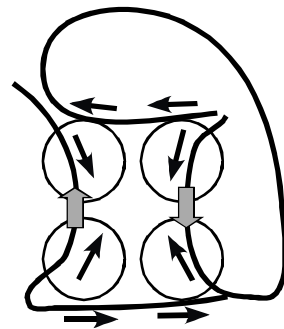
A
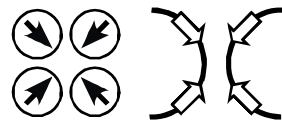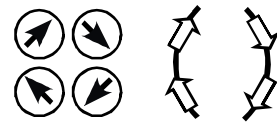
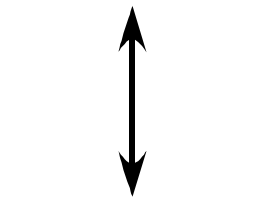core-type sites   arm-type sites
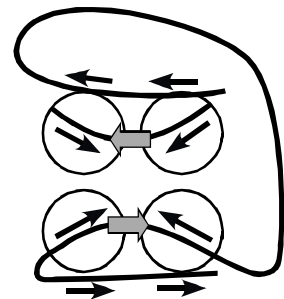
B
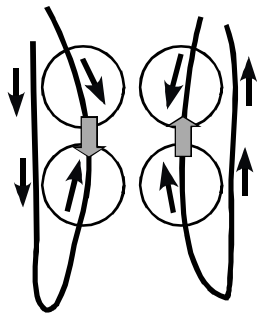
attR / attL cis   attR / attL trans

attP / attB trans   attP / attB cis

versa (Figure 38B). While all substrates that have been identified are presumed to be cis, this might not be the case in active synaptic complexes. It is clear that intasome formation involves the P4 / P5 pair of arm-type sites, and *attB* capture requires P1 and P2 (Pena *et al.*, 2000). Also, excisosomes only contain one core and arm-type pair, so the most likely arrangement is in cis. After recombination *attL* remains in a presumably cis complex with Int and mIHF, so there must be a rearrangement of the Int molecules either during integration or after release; the reverse must also occur for excision. The theory of the cis arrangement in the recombination intermediates is partially based on the assumption that only two Int molecules are present in some of the intermediate complexes, but if four are present, then the excisosomes could have alternate arrangements. However, theoretically, only two Int molecules are required to mediate simple bridges, but if four are present, formation of higher order complexes would be complicated by a significant rearrangement and loss of proteins. The usage of arm-type sites is critical to the specificity of recombination, so a thorough understanding of how they are utilized will ultimately be required to determine the details of L5 recombination.

VII.E.ii. Mechanism of directional control with serine integrase

The serine integrases are a relatively new system of site-specific recombination. While work has started on understanding how these proteins function, much still needs to be done. One particular aspect of these systems that is poorly understood is the mechanism of directional control. While RDFs have been identified in several of these systems, the means by which they work is unclear. The small size of the attachment site suggests that specificity must be achieved in a different way than with tyrosine integrases.

VII.E.iii. Regulation of Mycobacteriophage L5 recombination

This work has been aimed at identifying the mechanism of directional control in L5 site-specific recombination, but it has not addressed the means by which it is coordinated with the phage life cycle. During lytic growth, the P_{left} promoter is on, so Xis is being made, and integration is inhibited. Expression of the *xis* gene must be repressed during lysogeny to prevent ill-timed excision, and presumably the phage repressor, gp71, maintains this state. However, the mechanism by which the phage is induced from lysogenic to lytic growth has not been studied. While temperature sensitive mutants of gp71 have been identified that allow induction, there is no reason to believe that this induction mechanism is related to normal function. The choice between lysogeny and lytic growth is an important switch in the phage life cycle, and understanding it would offer new insight into Mycobacteriophage L5.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abremski, K. and S. Gottesman (1979). "The form of the DNA substrate required for excisive recombination of bacteriophage lambda." *J Mol Biol* **131**(3): 637-49.

Abremski, K. and S. Gottesman (1982). "Purification of the bacteriophage lambda xis gene product required for lambda excisive recombination." *J Biol Chem* **257**(16): 9658-62.

Altschul, S.F., W. Gish, et al. (1990). "Basic local alignment search tool." *J Mol Biol* **215**(3): 403-10.

Altschul, S.F., T.L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res* **25**(17): 3389-402.

Arciszewska, L.K. and D.J. Sherratt (1995). "Xer site-specific recombination in vitro." *Embo J* **14**(9): 2112-20.

Ausubel, F.M., R. Brent, et al. (1996). Current Protocols in Molecular Biology. New York, Wiley Intersciences.

Baker, J., R. Limberger, et al. (1991). "Recombination and modular exchange in the genesis of new lambdoid phages." *New Biol* **3**(3): 297-308.

Ball, C.A. and R.C. Johnson (1991a). "Efficient excision of phage lambda from the Escherichia coli chromosome requires the Fis protein." *J Bacteriol* **173**(13): 4027-31.

Ball, C.A. and R.C. Johnson (1991b). "Multiple effects of Fis on integration and the control of lysogeny in phage lambda." *J Bacteriol* **173**(13): 4032-8.

Banerjee, A., E. Dubnau, et al. (1994). "inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis." *Science* **263**(5144): 227-30.

Bardarov, S., J. Kriakov, et al. (1997). "Conditionally replicating mycobacteriophages: a system for transposon delivery to Mycobacterium tuberculosis." *Proc Natl Acad Sci U S A* **94**(20): 10961-6.

Barsom, E.K. and G.F. Hatfull (1996). "Characterization of Mycobacterium smegmatis gene that confers resistance to phages L5 and D29 when overexpressed." *Mol Microbiol* **21**(1): 159-70.

Bauer, C.E., S.D. Hesse, et al. (1986). "Mutational analysis of integrase arm-type binding sites of bacteriophage lambda. Integration and excision involve distinct interactions of integrase with arm-type sites." *J Mol Biol* **192**(3): 513-27.

Baulard, A., C. Jourdan, et al. (1992). "Rapid mycobacterial plasmid analysis by electroduction between Mycobacterium spp. and Escherichia coli." *Nucleic Acids Res* **20**(15): 4105.

Better, M., S. Wickner, et al. (1983). "Role of the Xis protein of bacteriophage lambda in a specific reactive complex at the attR prophage attachment site." *Cell* **32**(1): 161-8.

Bishop, A.J. and R.H. Schiestl (2000). "Homologous recombination as a mechanism for genome rearrangements: environmental and genetic effects." *Hum Mol Genet* **9**(16): 2427-334.

Bjellqvist, B., G.J. Hughes, et al. (1993). "The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences." *Electrophoresis* **14**(10): 1023-31.

Blattner, F.R., G. Plunkett, 3rd, et al. (1997). "The complete genome sequence of Escherichia coli K-12 [comment] [see comments]." *Science* **277**(5331): 1453-74.

Boccard, F., T. Smokvina, et al. (1989). "The integrated conjugative plasmid pSAM2 of Streptomyces ambofaciens is related to temperate bacteriophages." *Embo J* **8**(3): 973-80.

Boyd, D.A., G.A. Peters, et al. (2000). "Partial characterization of a genomic island associated with the multidrug resistance region of Salmonella enterica Typhymurium DT104." *FEMS Microbiol Lett* **189**(2): 285-91.

Brasch, M.A. and S.N. Cohen (1993). "Excisive recombination of the SLP1 element in Streptomyces lividans is mediated by Int and enhanced by Xis." *J Bacteriol* **175**(10): 3075-82.

Breuner, A., L. Brondsted, et al. (1999). "Novel organization of genes involved in prophage excision identified in the temperate lactococcal bacteriophage TP901-1." *J Bacteriol* **181**(23): 7291-7.

Breuner, A., L. Brondsted, et al. (2001). "Resolvase-like recombination performed by the TP901-1 integrase." *Microbiology* **147**(Pt 8): 2051-63.

Brown, D.P., K.B. Idler, et al. (1994). "Characterization of the genes and attachment sites for site-specific integration of plasmid pSE101 in Saccharopolyspora erythraea and Streptomyces lividans." *Mol Gen Genet* **242**(2): 185-93.

Brown, K.L., G.J. Sarkis, et al. (1997). "Transcriptional silencing by the mycobacteriophage L5 repressor." *Embo J* **16**(19): 5914-21.

Buchrieser, C., C. Rusniok, et al. (1998). DNA sequence of the 102 kbases unstable region of Yersinia pestis, GenBank. **2000**.

Burrus, V., Y. Roussel, et al. (2000). "Characterization of a novel integrative element, ICESt1, in the lactic acid bacterium Streptococcus thermophilus." *Appl Environ Microbiol* **66**(4): 1749-53.

Bushman, W., J.F. Thompson, et al. (1985). "Control of directionality in lambda site specific recombination." *Science* **230**(4728): 906-11.

Bushman, W., S. Yin, et al. (1984). "Determinants of directionality in lambda site-specific recombination." *Cell* **39**(3 Pt 2): 699-706.

Calmette, A. and C. Guerin (1908). "Sur quelques proprietes du bacille tuberxulwux cultive sur la bil." *C. R. Acad, Sci.* **147**: 1456-1459.

Carias, L.L., S.D. Rudin, et al. (1998). "Genetic linkage and cotransfer of a novel, vanB-containing transposon (Tn5382) and a low-affinity penicillin-binding protein 5 gene in a clinical vancomycin-resistant Enterococcus faecium isolate." *J Bacteriol* **180**(17): 4426-34.

Carrasco, C.D., K.S. Ramaswamy, et al. (1994). "Anabaena xisF gene encodes a developmentally regulated site-specific recombinase." *Genes Dev* **8**(1): 74-83.

Christiansen, B., M.G. Johnsen, et al. (1994). "Characterization of the lactococcal temperate phage TP901-1 and its site-specific integration." *J Bacteriol* **176**(4): 1069-76.

Cole, S.T. and B.G. Barrell (1998). "Analysis of the genome of Mycobacterium tuberculosis H37Rv." *Novartis Found Symp* **217**: 160-72.

Cole, S.T., R. Brosch, et al. (1998). "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence." *Nature* **393**(6685): 537-44.

Colloms, S.D., J. Bath, et al. (1997). "Topological selectivity in Xer site-specific recombination." *Cell* **88**(6): 855-64.

Craig, N.L. and H.A. Nash (1983). "The mechanism of phage lambda site-specific recombination: site- specific breakage of DNA by Int topoisomerase." *Cell* **35**(3 Pt 2): 795-803.

Davies, R.W. (1980). "DNA sequence of the int-xis-Pi region of the bacteriophage lambda; overlap of the int and xis genes." *Nucleic Acids Res* **8**(8): 1765-82.

Dodd, I.B. and J.B. Egan (1990). "Improved detection of helix-turn-helix DNA-binding motifs in protein sequences." *Nucleic Acids Res* **18**(17): 5019-26.

Dodd, I.B. and J.B. Egan (1996). "The Escherichia coli retrons Ec67 and Ec86 replace DNA between the cos site and a transcription terminator of a 186-related prophage." *Virology* **219**(1): 115-24.

Dodd, I.B., M.R. Reed, et al. (1993). "The Cro-like Apl repressor of coliphage 186 is required for prophage excision and binds near the phage attachment site." *Mol Microbiol* **10**(5): 1139-50.

Donnelly-Wu, M.K., W.R. Jacobs, Jr., et al. (1993). "Superinfection immunity of mycobacteriophage L5: applications for genetic transformation of mycobacteria." *Mol Microbiol* **7**(3): 407-17.

Dupont, L., B. Boizet-Bonhoure, et al. (1995). "Characterization of genetic elements required for site-specific integration of Lactobacillus delbrueckii subsp. bulgaricus bacteriophage mv4 and construction of an integration-proficient vector for Lactobacillus plantarum." *J Bacteriol* **177**(3): 586-95.

Esposito, D. and J.J. Scocca (1994). "Identification of an HP1 phage protein required for site-specific excision." *Mol Microbiol* **13**(4): 685-95.

Esposito, D. and J.J. Scocca (1997). "The integrase family of tyrosine recombinases: evolution of a conserved active site domain." *Nucleic Acids Res* **25**(18): 3605-14.

Figueroa-Bossi, N. and L. Bossi (1999). "Inducible prophages contribute to Salmonella virulence in mice." *Mol Microbiol* **33**(1): 167-76.

Flannagan, S.E., L.A. Zitzow, et al. (1994). "Nucleotide sequence of the 18-kb conjugative transposon Tn916 from Enterococcus faecalis." *Plasmid* **32**(3): 350-4.

Ford, M.E., G.J. Sarkis, et al. (1998). "Genome structure of mycobacteriophage D29: implications for phage evolution." *J Mol Biol* **279**(1): 143-64.

Franz, B. and A. Landy (1995). "The Holliday junction intermediates of lambda integrative and excisive recombination respond differently to the bending proteins integration host factor and excisionase." *Embo J* **14**(2): 397-406.

Fullner, K.J. and G.F. Hatfull (1997). "Mycobacteriophage L5 infection of Mycobacterium bovis BCG: implications for phage genetics in the slow-growing mycobacteria." *Mol Microbiol* **26**(4): 755-66.

Gally, D.L., J. Leathart, et al. (1996). "Interaction of FimB and FimE with the fim switch that controls the phase variation of type 1 fimbriae in Escherichia coli K-12." *Mol Microbiol* **21**(4): 725-38.

Garnier, F., S. Taourit, et al. (2000). "Characterization of transposon Tn1549, conferring VanB-type resistance in Enterococcus spp." *Microbiology* **146**(Pt 6): 1481-9.

Gartemann, K.-H., J. Fiedler, et al. (1999). Direct Submission, GenBank. **2000**.

Goodman, S.D., S.C. Nicholson, et al. (1992). "Deformation of DNA during site-specific recombination of bacteriophage lambda: replacement of IHF protein by HU protein or sequence-directed bends." *Proc Natl Acad Sci U S A* **89**(24): 11910-4.

Guan, S.a.V., N.K. (1996). Direct Submission, GenBank. **2000**.

Guo, F., D.N. Gopaul, et al. (1997). "Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse." *Nature* **389**(6646): 40-6.

Hallet, B. and D.J. Sherratt (1997). "Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements." *FEMS Microbiol Rev* **21**(2): 157-78.

Halling, C., R. Calendar, et al. (1990). "DNA sequence of satellite bacteriophage P4." *Nucleic Acids Res* **18**(6): 1649.

Han, Y.W., R.I. Gumport, et al. (1993). "Complementation of bacteriophage lambda integrase mutants: evidence for an intersubunit active site." *Embo J* **12**(12): 4577-84.

Han, Y.W., R.I. Gumport, et al. (1994). "Mapping the functional domains of bacteriophage lambda integrase protein." *J Mol Biol* **235**(3): 908-25.

Hatfull, G.F. (1994). "Mycobacteriophage L5: A toolbox for tuberculosis." *ASM News.* **60**: 255-260.

Hatfull, G.F., L. Barsom, et al. (1994). "Bacteriophages as tools for vaccine development." *Dev Biol Stand* **82**: 43-7.

Hatfull, G.F. and N.D.F. Grindley (1988). Resolvases and DNA-Invertases: a Family of Enzymes Active in SIte-Specific Recombination. Washington, D.C., American Society of Microbiology.

Hatfull, G.F. and G.J. Sarkis (1993). "DNA sequence, structure and gene expression of mycobacteriophage L5: a phage system for mycobacterial genetics." *Mol Microbiol* **7**(3): 395-405.

Hendrix, R.W., M.C. Smith, et al. (1999). "Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage." *Proc Natl Acad Sci U S A* **96**(5): 2192-7.

Henikoff, S. and J.G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." *Proc Natl Acad Sci U S A* **89**(22): 10915-9.

Hsu, M.Y., M. Inouye, et al. (1990). "Retron for the 67-base multicopy single-stranded DNA from Escherichia coli: a potential transposable element encoding both reverse transcriptase and Dam methylase functions." *Proc Natl Acad Sci U S A* **87**(23): 9454-8.

Huan, P.T., D.A. Bastin, et al. (1997). "Molecular characterization of the genes involved in O-antigen modification, attachment, integration and excision in Shigella flexneri bacteriophage SfV." *Gene* **195**(2): 217-27.

Jeanmougin, F., J.D. Thompson, et al. (1998). "Multiple sequence alignment with Clustal X." *Trends Biochem Sci* **23**(10): 403-5.

Juhala, R.J., M.E. Ford, et al. (2000). "Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages." *J Mol Biol* **299**(1): 27-51.

Katz, L., D.P. Brown, et al. (1991). "Site-specific recombination in Escherichia coli between the att sites of plasmid pSE211 from Saccharopolyspora erythraea." *Mol Gen Genet* **227**(1): 155-9.

Kikuchi, A., E. Flamm, et al. (1985). "An Escherichia coli mutant unable to support site-specific recombination of bacteriophage lambda." *J Mol Biol* **183**(2): 129-40.

Kilic, A.O., M.N. Vijayakumar, et al. (1994). "Identification and nucleotide sequence analysis of a transfer-related region in the streptococcal conjugative transposon Tn5252." *J Bacteriol* **176**(16): 5145-50.

Kim, S., L. Moitoso de Vargas, et al. (1990). "Mapping of a higher order protein-DNA complex: two kinds of long-range interactions in lambda attL." *Cell* **63**(4): 773-81.

Kochi, A. (1991). "The global tuberculosis situation and the new control strategy of the World Health Organization." *Tubercle* **72**: 1-6.

Kodaira, K.I., M. Oki, et al. (1997). "Genome structure of the Lactobacillus temperate phage phi g1e: the whole genome sequence and the putative promoter/repressor system." *Gene* **187**(1): 45-53.

Kropinski, A.M. (2000). "Sequence of the genome of the temperate, serotype-converting, pseudomonas aeruginosa bacteriophage D3 [In Process Citation]." *J Bacteriol* **182**(21): 6066-74.

Kwon, H.J., R. Tirumalai, et al. (1997). "Flexibility in DNA recombination: structure of the lambda integrase catalytic core." *Science* **276**(5309): 126-31.

Lee, M.H. and G.F. Hatfull (1993). "Mycobacteriophage L5 integrase-mediated site-specific integration in vitro." *J Bacteriol* **175**(21): 6836-41.

Lee, M.H., L. Pascopella, et al. (1991). "Site-specific integration of mycobacteriophage L5: integration- proficient vectors for Mycobacterium smegmatis, Mycobacterium tuberculosis, and bacille Calmette-Guerin." *Proc Natl Acad Sci U S A* **88**(8): 3111-5.

Leong, J.M., S.E. Nunes-Duby, et al. (1986). "Structural and regulatory divergence among site-specific recombination genes of lambdoid phage." *J Mol Biol* **189**(4): 603-16.

Lewis, J.A. and G.F. Hatfull (2000). "Identification and characterization of mycobacteriophage L5 excisionase." *Mol Microbiol* **35**(2): 350-60.

Lindsey, D.F., D.A. Mullin, et al. (1989). "Characterization of the cryptic lambdoid prophage DLP12 of Escherichia coli and overlap of the DLP12 integrase gene with the tRNA gene argU." *J Bacteriol* **171**(11): 6197-205.

Liu, T. and E. Haggard-Ljungquist (1999). "The transcriptional switch of bacteriophage WPhi, a P2-related but heteroimmune coliphage." *J Virol* **73**(12): 9816-26.

Mahairas, G.G., P.J. Sabo, et al. (1996). "Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent M. bovis." *J Bacteriol* **178**(5): 1274-82.

Marra, D. and J.R. Scott (1999). "Regulation of excision of the conjugative transposon Tn916." *Mol Microbiol* **31**(2): 609-21.

McShan, W.M., Y.F. Tang, et al. (1997). "Bacteriophage T12 of Streptococcus pyogenes integrates into the gene encoding a serine tRNA." *Mol Microbiol* **23**(4): 719-28.

Miyamoto, H., W. Nakai, et al. (1999). "Sequence analysis of Stx2-converting phage VT2-Sa shows a great divergence in early regulation and replication regions." *DNA Res* **6**(4): 235-40.

Mizuuchi, K., M. Gellert, et al. (1978). "Involement of supertwisted DNA in integrative recombination of bacteriophage lambda." *J Mol Biol* **121**(3): 375-92.

Mizuuchi, K., R. Weisberg, et al. (1981). "Structure and function of the phage lambda att site: size, int-binding sites, and location of the crossover point." *Cold Spring Harb Symp Quant Biol* **45 Pt 1**: 429-37.

Mizuuchi, M. and K. Mizuuchi (1980). "Integrative recombination of bacteriophage lambda: extent of the DNA sequence involved in attachment site function." *Proc Natl Acad Sci U S A* **77**(6): 3220-4.

Mizuuchi, M. and K. Mizuuchi (1985). "The extent of DNA sequence required for a functional bacterial attachment site of phage lambda." *Nucleic Acids Res* **13**(4): 1193-208.

Moitoso de Vargas, L., S. Kim, et al. (1989). "DNA looping generated by DNA bending protein IHF and the two domains of lambda integrase." *Science* **244**(4911): 1457-61.

Moitoso de Vargas, L. and A. Landy (1991). "A switch in the formation of alternative DNA loops modulates lambda site-specific recombination." *Proc Natl Acad Sci U S A* **88**(2): 588-92.

Moitoso de Vargas, L., C.A. Pargellis, et al. (1988). "Autonomous DNA binding domains of lambda integrase recognize two different sequence families." *Cell* **54**(7): 923-9.

Murphy, L., D. Harris, et al. (1999). Direct Submission, EMBL. **2000**.

Nash, H.A. (1981). "Integration and excision of bacteriophage lambda: the mechanism of conservation site specific recombination." *Annu Rev Genet* **15**: 143-67.

Nash, H.A. and C.A. Robertson (1981). "Purification and properties of the Escherichia coli protein factor required for lambda integrative recombination." *J Biol Chem* **256**(17): 9246-53.

Nashimoto, H. (1999). Direct Submission, NCBI GenBank. **2000**.

Nesbit, C.E., M.E. Levin, et al. (1995). "Transcriptional regulation of repressor synthesis in mycobacteriophage L5." *Mol Microbiol* **17**(6): 1045-56.

Nesper, J., J. Blass, et al. (1999). "Characterization of the major control region of Vibrio cholerae bacteriophage K139: immunity, exclusion, and integration." *J Bacteriol* **181**(9): 2902-13.

Neuwald, A.F., J.S. Liu, et al. (1997). "Extracting protein alignment models from the sequence database." *Nucleic Acids Res* **25**(9): 1665-77.

Numrych, T.E., R.I. Gumport, et al. (1990). "A comparison of the effects of single-base and triple-base changes in the integrase arm-type binding sites on the site-specific recombination of bacteriophage lambda." *Nucleic Acids Res* **18**(13): 3953-9.

Numrych, T.E., R.I. Gumport, et al. (1991). "A genetic analysis of Xis and FIS interactions with their binding sites in bacteriophage lambda." *J Bacteriol* **173**(19): 5954-63.

Nunes-Duby, S.E., R.S. Tirumalai, et al. (1994). "Lambda integrase cleaves DNA in cis." *Embo J* **13**(18): 4421-30.

Okinaka, R.T., K. Cloud, et al. (1999). "Sequence and organization of pXO1, the large Bacillus anthracis plasmid harboring the anthrax toxin genes." *J Bacteriol* **181**(20): 6509-15.

Pabo, C.O., R.T. Sauer, et al. (1979). "The lambda repressor contains two domains." *Proc Natl Acad Sci U S A* **76**(4): 1608-12.

Parsons, R.L., P.V. Prasad, et al. (1988). "Step-arrest mutants of FLP recombinase: implications for the catalytic mechanism of DNA recombination." *Mol Cell Biol* **8**(8): 3303-10.

Pascopella, L., F.M. Collins, et al. (1994). "Use of in vivo complementation in Mycobacterium tuberculosis to identify a genomic fragment associated with virulence." *Infect Immun* **62**(4): 1313-9.

Pedulla, M.L. and G.F. Hatfull (1998). "Characterization of the mIHF gene of Mycobacterium smegmatis." *J Bacteriol* **180**(20): 5473-7.

Pedulla, M.L., M.H. Lee, et al. (1996). "A novel host factor for integration of mycobacteriophage L5." *Proc Natl Acad Sci U S A* **93**(26): 15411-6.

Pelicic, V., J.M. Reyrat, et al. (1996). "Expression of the Bacillus subtilis sacB gene confers sucrose sensitivity on mycobacteria." *J Bacteriol* **178**(4): 1197-9.

Pena, C.E., J.M. Kahlenberg, et al. (2000). "Assembly and activation of site-specific recombination complexes." *Proc Natl Acad Sci U S A* **97**(14): 7760-5.

Peña, C.E., M.H. Lee, et al. (1997). "Characterization of the mycobacteriophage L5 attachment site, attP." *J Mol Biol* **266**(1): 76-92.

Peña, C.E., J. Stoner, et al. (1998). "Mycobacteriophage D29 integrase-mediated recombination: specificity of mycobacteriophage integration." *Gene* **225**(1-2): 143-51.

Peña, C.E., J.E. Stoner, et al. (1996). "Positions of strand exchange in mycobacteriophage L5 integration and characterization of the attB site." *J Bacteriol* **178**(18): 5533-6.

Peña, C.E.A., J.M. Kahlenberg, et al. (1998). "The role of supercoiling in mycobacteriophage L5 integrative recombination." *Nucleic Acids Res* **26**(17): 4012-4018.

Peña, C.E.A., J.M. Kahlenberg, et al. (1999). "Protein-DNA complexes in mycobacteriophage L5 integrative recombination." *J Bacteriol* **181**(2): 454-61.

Plunkett, G., 3rd, D.J. Rose, et al. (1999). "Sequence of Shiga toxin 2 phage 933W from Escherichia coli O157:H7: Shiga toxin as a phage late-gene product." *J Bacteriol* **181**(6): 1767-78.

Poyart-Salmeron, C., P. Trieu-Cuot, et al. (1989). "Molecular characterization of two proteins involved in the excision of the conjugative transposon Tn1545: homologies with other site-specific recombinases." *Embo J* **8**(8): 2425-33.

Ramaswamy, K.S., C.D. Carrasco, et al. (1997). "Cell-type specificity of the Anabaena fdxN-element rearrangement requires xisH and xisI." *Mol Microbiol* **23**(6): 1241-9.

Rauch, P.J. and W.M. de Vos (1994). "Identification and characterization of genes involved in excision of the Lactococcus lactis conjugative transposon Tn5276." *J Bacteriol* **176**(8): 2165-71.

Raynal, A., K. Tuphile, et al. (1998). "Structure of the chromosomal insertion site for pSAM2: functional analysis in Escherichia coli." *Mol Microbiol* **28**(2): 333-42.

Ribeiro, G., M. Viveiros, et al. (1997). "Mycobacteriophage D29 contains an integration system similar to that of the temperate mycobacteriophage L5." *Microbiology* **143**(Pt 8): 2701-8.

Rice, P.A., S. Yang, et al. (1996). "Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn." *Cell* **87**(7): 1295-306.

Richet, E., P. Abcarian, et al. (1986). "The interaction of recombination proteins with supercoiled DNA: defining the role of supercoiling in lambda integrative recombination." *Cell* **46**(7): 1011-21.

Richet, E., P. Abcarian, et al. (1988). "Synapsis of attachment sites during lambda integrative recombination involves capture of a naked DNA by a protein-DNA complex." *Cell* **52**(1): 9-17.

Robertson, C.A. and H.A. Nash (1988). "Bending of the bacteriophage lambda attachment site by Escherichia coli integration host factor." *J Biol Chem* **263**(8): 3554-7.

Robison, K. (1994). Direct Submission, GenBank. **2000**.

Romine, M.F., L.C. Stillwell, et al. (1999). "Complete sequence of a 184-kilobase catabolic plasmid from Sphingomonas aromaticivorans F199." *J Bacteriol* **181**(5): 1585-602.

Salmi, D., V. Magrini, et al. (1998). "Genetic determinants of immunity and integration of temperate Myxococcus xanthus phage Mx8." *J Bacteriol* **180**(3): 614-21.

Schmidt, H.a.K., H. (2000). Chromosomal insertion sites of Shiga toxin 1-converting bacteriophage H19J, GenBank. **2000**.

Schneider, S.J. (1992). Site-specific Recombination of lambdoid phage 21 into the *icd* gene of *Escherichia coli*. Stanford, CA, Stanford University.

Semsey, S., I. Papp, et al. (1999). "Identification of site-specific recombination genes int and xis of the Rhizobium temperate phage 16-3." *J Bacteriol* **181**(14): 4185-92.

Sezonov, G., A.M. Duchene, et al. (1998). "Replicase, excisionase, and integrase genes of the Streptomyces element pSAM2 constitute an operon positively regulated by the pra gene." *J Bacteriol* **180**(12): 3056-61.

Sherratt, D.J., L.K. Arciszewska, et al. (1995). "Site-specific recombination and circular chromosome segregation." *Philos Trans R Soc Lond B Biol Sci* **347**(1319): 37-42.

Skowronek, K. and S. Baranowski (1997). "The relationship between HP1 and S2 bacteriophages of Haemophilus influenzae." *Gene* **196**(1-2): 139-44.

Snapper, S.B., L. Lugosi, et al. (1988). "Lysogeny and transformation in mycobacteria: stable expression of foreign genes." *Proc Natl Acad Sci U S A* **85**(18): 6987-91.

Sonoda, E., M. Takata, et al. (2001). "Homologous DNA recombination in vertebrate cells." *Proc Natl Acad Sci U S A* **98**(15): 8388-94.

Stark, W.M., D.J. Sherratt, et al. (1989). "Site-specific recombination by Tn3 resolvase: topological changes in the forward and reverse reactions." *Cell* **58**(4): 779-90.

Stover, C.K., V.F. de la Cruz, et al. (1991). "New use of BCG for recombinant vaccines." *Nature* **351**(6326): 456-60.

Sun, J., M. Inouye, et al. (1991). "Association of a retroelement with a P4-like cryptic prophage (retronphage phi R73) integrated into the selenocystyl tRNA gene of Escherichia coli." *J Bacteriol* **173**(13): 4171-81.

Takai, S., S.A. Hines, et al. (2000). "DNA sequence and comparison of virulence plasmids from rhodococcus equi ATCC 33701 and 103 [In Process Citation]." *Infect Immun* **68**(12): 6840-7.

Thompson, J.F., L.M. de Vargas, et al. (1987a). "Protein-protein interactions in a higher-order structure direct lambda site-specific recombination." *J Mol Biol* **195**(3): 481-93.

Thompson, J.F., L. Moitoso de Vargas, et al. (1987b). "Cellular factors couple recombination with growth phase: characterization of a new component in the lambda site-specific recombination pathway." *Cell* **50**(6): 901-8.

Thorpe, H.M. and M.C. Smith (1998). "In vitro site-specific integration of bacteriophage DNA catalyzed by a recombinase of the resolvase/invertase family." *Proc Natl Acad Sci U S A* **95**(10): 5505-10.

Thorpe, H.M., S.E. Wilson, et al. (2000). "Control of directionality in the site-specific recombination system of the streptomyces phage phiC31 [In Process Citation]." *Mol Microbiol* **38**(2): 232-41.

Tirumalai, R.S., H.J. Kwon, et al. (1998). "Recognition of core-type DNA sites by lambda integrase." *J Mol Biol* **279**(3): 513-27.

Tribble, G.D., A.C. Parker, et al. (1999). "Genetic structure and transcriptional analysis of a mobilizable, antibiotic resistance transposon from Bacteroides." *Plasmid* **42**(1): 1-12.

van der Wilk, F., A.M. Dullemans, et al. (1999). "Isolation and characterization of APSE-1, a bacteriophage infecting the secondary endosymbiont of Acyrthosiphon pisum." *Virology* **262**(1): 104-13.

Wang, H., C.H. Yang, et al. (1997). "Integration specificities of two lambdoid phages (21 and e14) that insert at the same attB site." *J Bacteriol* **179**(18): 5705-11.

White, O., J.A. Eisen, et al. (1999). "Genome sequence of the radioresistant bacterium Deinococcus radiodurans R1." *Science* **286**(5444): 1571-7.

Wilkins, M.R., E. Gasteiger, et al. (1999). "Protein identification and analysis tools in the ExPASy server." *Methods Mol Biol* **112**: 531-52.

Winoto, A., S. Chung, et al. (1986). "Directional control of site-specific recombination by bacteriophage lambda. Evidence that a binding site for Int protein far from the crossover point is required for integrative but not excisive recombination." *J Mol Biol* **192**(3): 677-80.

Yagil, E., S. Dolev, et al. (1989). "Determinants of site-specific recombination in the lambdoid coliphage HK022. An evolutionary change in specificity." *J Mol Biol* **207**(4): 695-717.

Ye, Z.H., S.L. Buranen, et al. (1990). "Sequence analysis and comparison of int and xis genes from staphylococcal bacteriophages L54a and phi 11." *J Bacteriol* **172**(5): 2568-75.

Ye, Z.H. and C.Y. Lee (1989). "Nucleotide sequence and genetic characterization of staphylococcal bacteriophage L54a int and xis genes." *J Bacteriol* **171**(8): 4146-53.

Yin, S., W. Bushman, et al. (1985). "Interaction of the lambda site-specific recombination protein Xis with attachment site DNA." *Proc Natl Acad Sci U S A* **82**(4): 1040-4.

Yu, A. and E. Haggard-Ljungquist (1993). "The Cox protein is a modulator of directionality in bacteriophage P2 site-specific recombination." *J Bacteriol* **175**(24): 7848-55.