1
2
3
4
5
6
7
8
9
10
11 Brujita Integrase: A simple, arm-less, directionless, and promiscuous tyrosine integrase system
12
13
14
15
16
17
18
19
20

21 Running Title: Brujita Integrase
22
23
24
25
26
27
28
29

30 Bryce L. Lunt and Graham F. Hatfull*
31
32
33
34
35
36
37
38
39
40
41
42
43

44 *Corresponding author
45 Department of Biological Sciences
46 University of Pittsburgh
47 gfh@pitt.edu
48
49
50

ABSTRACT

Mycobacteriophage Brujita is an unusual temperate phage in which establishment of superinfection immunity is dependent on chromosomal integration. Integration is mediated by a non-canonical tyrosine integrase lacking an N-terminal domain typically associated with binding to arm-type sites within the phage attachment site (*attP*). This raises the question as to how these integrases (Ints) bind their DNA substrates, if they form higher order protein DNA complexes, and how site selection and recombinational directionality is determined. Here we show that Brujita integrase is a simple recombinase, whose properties more closely resemble those of FLP and Cre than it does the canonical phage integrases. Brujita Int uses relatively small DNA substrates, fails to discriminate between *attP* and *attB*, cleaves attachment site DNA to form a 6-base overlap region, and lacks directional control. Brujita Int also has an unusual pattern of binding to its DNA substrates. It binds to two half sites (B and B') at *attB*, although binding to the B half site is strongly dependent on occupancy of B'. In contrast, binding to the P half site is not observed, even when Int is bound at P'. However, an additional Int binding site (P1) is displaced to the left of the crossover site at *attP*, is required for recombination and is predicted to facilitate binding of Int to the P half site during synapsis. These simple phage integrase systems may reflect ancestral states of phage evolution with the complexities of higher-order complex formation and directional control representing subsequent adaptations.

INTRODUCTION

Temperate bacteriophages typically establish lysogeny through regulation of lytic gene expression, and chromosomal integration to form prophages [1, 2]. Integration is commonly site-specific and mediated by a phage-encoded integrase of either the tyrosine-recombinase or serine-recombinase class of site-specific recombinases [3]. These act by promoting strand exchange between a phage-encoded attachment site (*attP*) and a bacterial attachment site (*attB*) to generate two product sites, *attL* and *attR* (Fig. 1A). The integrases catalyze excisive recombination using *attL* and *attR* as substrates in addition to integration, and are under tight directional control with a Recombinational Directional Factor (RDF) acting as an accessory protein to stimulate excisive recombination and inhibit integration [4].

Many tyrosine integrases have been described and several have been well-characterized biochemically, including those encoded by enterobacteriophage λ and mycobacteriophage L5 [3, 5, 6]. These integrases contain three domains, a catalytic domain that includes the active site tyrosine, a core-binding domain that promotes DNA binding at the sites of strand exchange, and an N-terminal arm-type binding domain that recognizes arm-type integrase binding sites in *attP* that flank the core binding region [6, 7]. As such, the requirements for *attP* and *attB* are different, with *attB* function located within a 25 bp segment, and *attP* function spanning 250 bp [8-11]. *attP* also includes binding sites for an integration host factor that facilitates DNA bending to enable formation of higher-order protein-DNA structures containing Int-DNA bridges in which arm- and core-type binding sites are simultaneously occupied by Int protomers [1, 6, 9, 12, 13], as well as binding sites for the RDF that impose DNA distortions constraining the types of Int-DNA bridges that can be formed [14, 15].

Recently, a new subclass of tyrosine integrases have been described that lack an N-terminal arm-type binding domain and thus must use alternative mechanisms to control directionality [16, 17]. These are components of systems for regulating phage life cycles, where integration is a requirement for lysogenic establishment. A notable characteristic of these is that the *attP* site is located within the gene encoding the phage repressor, such that integration results in loss of the 3' end of the gene and removal of a C-terminal degradation tag that would otherwise inactivate the virally-encoded form of the repressor [17]. Several examples encoded by mycobacteriophages (e.g. phages BPs and Brujita) have been described, although the systems are present in phages infecting other hosts [17]. Although the integrases in these systems appear to lack an N-terminal domain, they have a C-terminal extension that promotes Int degradation [17]. Furthermore, a stabilized mutant derivative of Brujita Int not only mediates higher levels of integrative recombination but also promotes excisive recombination, even in the absence of any additional phage-encoded gene products [17]. Directionality of these systems may thus be facilitated by Int stability and total integrase concentration, without requirement for an RDF.

Here, we explore the requirements at *attP*, *attB*, *attL*, and *attR*, for integrase-mediated recombination *in vivo*, and show that not only is there an apparent lack of directional control, but that site selection is promiscuous, such that *attP* x *attP* or *attB* x *attB* recombine similarly to *attP* x *attB*. However, the site requirements for *attP* and *attB* are different, with *attB* function situated within a ~25 bp segment located around the putative crossover sites, whereas *attP* function requires ~50bp, with the putative crossover sites asymmetrically positioned. DNA binding studies show that Int binds differently to *attP* and *attB*. When binding to *attB*, Int can independently recognize and bind to the B' half site, but occupancy of the B half site is entirely dependent on occupancy of B'. In contrast, binding of Int to the P' half site of *attP* does not promote occupancy of the putative P half site at the left-most scissile bond. However, Int binds

independently to a separate site, P1, located immediately to the left of the putative P half site. Because the Int at P1 is not positioned to play a role in catalysis, we propose that it facilitates recruitment of Int to the P half site during synaptic complex formation, representing a notable departure from other site-specific recombination systems.

RESULTS

Organization of Brujita attachment sites.

The Brujita *attP* and *attB* site locations are defined by their 39 bp common core sequence that is shared by a host tRNA-Thr (Msmeg_6152) gene, extending from the anticodon loop to the 3' end of the gene; there is little or no evident sequence similarity to the left of the common core (Fig. 1A). In other tyrosine integrase systems where *attB* overlaps a host tRNA gene and where the crossover site has been determined, the sites of strand exchange are typically located near the 5' side of the common core and spaced 7 bp apart such that Int cleavage generates 5' extensions [18-21]. In the Brujita attachment sites there is no obvious inverted symmetry around the 5' side of the common core in either *attP* or *attB* that might reflect Int recognition sequences. For discussion purposes we have numbered the leftmost position of the common core as coordinate +1; the base to its immediate left is -1 (Fig. 1A).

An intra-molecular in vivo assay for Brujita integrase-mediated recombination

To establish an *in vivo* assay for Brujita integrase-mediated recombination, first we constructed a chimeric plasmid (pBL50) in which directly oriented Brujita *attP* and *attB* sites separate a hygromycin resistant (Hyg^R) cassette from other plasmid components, origins of replication for both *E. coli* and mycobacteria (oriE and oriM respectively) and a kanamycin resistance (Kan^R)

cassette (see Fig. 1B). Brujita integrase is expected to recombine *attP* and *attB*, leading to loss of hygromycin resistance. Because the specific requirements at *attP* and *attB* are unknown, pBL50 was constructed with relatively large DNA segments, 442 bp and 436 bp for *attP* and *attB* respectively. These large segments are expected to contain all the required sequence elements for recombination.

We evaluated the ability of Brujita integrase to recombine pBL50 by co-transformation with pBL03 into *M. smegmatis*. Plasmid pBL03 contains a copy of the Brujita integrase gene with a single amino acid substitution (A296E) near its C-terminus that promotes protein stability and higher rates of recombination [17], and is expressed from its native promoter. Plasmid pBL03 lacks a mycobacterial origin of replication, an *attP* or integration system, and does not propagate in mycobacteria. Following co-electroporation of pBL50 and pBL03, kanamycin resistant transformants were selected, and screened for their hygromycin resistant phenotype. We observed that 41% had lost hygromycin resistance; in control transformations without the Int containing plasmid, 100% of transformants are hygromycin resistant. Sequencing of hygromycin sensitive pBL50 transformants confirmed recombination between *attP* and *attB*. Brujita Int expressed from pBL03 is thus capable of mediating recombination between these *attP* and *attB* substrates.

We used a similar scheme to test for Brujita Int-mediated recombination in *E. coli*. First we modified pBL50 (to create pBL90; Fig. 1B) by introducing the *ccdB* gene adjacent to the Hyg^R cassette to enable direct selection for recombination between *attP* and *attB* [22]. We also constructed a pBAD-derived plasmid (pBL83) that expresses Int A296E from the arabinose promoter. Competent Stellar *E. coli* cells carrying pBL83 were induced with arabinose and transformed with pBL90, selecting for kanamycin resistance. Transformants selected for loss of *ccdB* were readily recovered at a frequency of 7×10^4 colony forming units per microgram DNA

(colony forming units/ microgram DNA; cfu/μg) compared to control transformations into Int-lacking cells, which yielded <10 colonies per 50 ng DNA per transformation (Fig. 2A). Transformation of pBL90 into a non-*ccdB* selection host suggests that integration occurs at a frequency of 5-10%, although differences of cell competencies confounds a more precise measurement of efficiency. Because Brujita integrase-mediated recombination between these *attP* and *attB* substrates is reasonably efficient in *E. coli*, recombination does not require a mycobacterial specific host factor, although we cannot rule out the role of a common host factor shared by *M. smegmatis* and *E. coli*.

Requirements for recombination at *attP*

The ability to select for recombinants in *E. coli* using the *ccdB* system provides a simple assay for determining the site requirements for recombination. A series of plasmid derivatives were constructed that contain progressively larger deletions from either the right- or left-hand side of *attP* and tested for the generation of recombinants by transformation into Stellar *E. coli* expressing integrase. Removal of most of the right arm sequences leaving just 14 bp to the right of the left end of the common core sequences (coordinate +14) has only a minor impact on recombination (Fig. 2A), but further resection (e.g. pBL146) reduces recombination to near-background levels. Removal of sequences from the left end to leave 35 bp to the left of the common core (i.e. pBL105) has little impact on recombination, but removal of an additional 5 bp or 10 bp strongly reduces recombination (Fig. 2A). A 49 bp substrate extending from -35 to +14 (pBL173) recombines with only about a 2-fold reduction relative to the wild-type substrate. These data indicate a notable difference between Brujita integrase-mediated recombination and other well-studied tyrosine integrase systems such as λ [8] and L5 [9]. In these systems, *attP* is considerably larger (~250 bp) because of inclusion of the arm-type integrase binding sites, which appear to be absent from Brujita *attP*, consistent with the absence of an N-terminal arm-

type binding domain in the integrase protein.

Requirements for recombination at *attB*

A similar process was used to examine *attB* site requirements. Removal of sequences from the right hand side of the site to +39 (i.e. pBL116, Fig. 2B) has no effect on recombination, but further deletion to +14 (pBL147) reduces recombination about 3-fold. Because this endpoint is within the common core and the right-side deletion to +14 in *attP* shows only a small impact on recombination frequency, it is likely that the vector sequences adjacent to the +14 position contribute to Int recognition, and that the +14 position is close to the boundary for Int binding. Deletion from the left side of *attB* to the -10 or the -5 positions causes about 2-fold reduction in recombination, but removal to coordinate +1 eliminates recombination (Fig. 2B). A substrate containing the 14 bp from coordinates -5 to +9 does support recombination, but the efficiency is about 20-fold lower than the wild-type substrate. As noted above, adjacent vector sequences may contribute to Int recognition in some substrates. Nonetheless, it is clear that Brujita *attP* and *attB* differ in their length requirements, even though neither is long enough to include arm-type Int binding sites and IHF binding sites. We also note that a substrate (pBL176) combining the smallest active *attP* and *attB* sites (those represented in pBL173 and pBL204, Fig. 2) recombines poorly (~1% of pBL90), presumably in part by combining the inefficiencies of the individual sites (data not shown).

Brujita integrase-mediated excisive recombination

The apparent lack of *attP* arm-type integrase binding sites and an integrase arm-type binding domain suggest that the directional control of integration and excision differs for Brujita and related systems from the canonical λ system. We therefore constructed similar test plasmids in

which the attachment sites are *attL* and *attR* (see Fig. 1C), placed in direct orientation such that excisive recombination will generate products in which the *ccdB* gene is lost and can be selected in Int-expressing Stellar cells. Plasmid pBL144 (Fig. 1C) carrying *attL* and *attR* (440 bp and 438 bp, respectively) sites recombines in the presence of integrase, and with similar efficiency to *attB* x *attP* integrative recombination, indicating that Brujita integrase lacks directional control and does not require any other phage-encoded RDF. Resection from the P' side of *attL* to +14 (pBL178) causes about a 3-fold reduction in recombination, consistent with the *attP* observations (Fig. 2C). Resection of the B arm of *attL* to -10 has little effect on recombination, but removal of an additional five base pairs to -5 further reduces recombination, again defining the requirements as similar to those for the B arm of *attB* (Fig. 2B). An *attL* site containing the 24 bp from -5 to +19 is functional, but about 10-fold less efficient than the pBL144 substrate (Fig. 2C).

Similar analysis of *attR* shows that deletion of the B' arm to +14 has little impact on recombination, removal of a further 5 bp reduces recombination more than 10-fold, and resection to +4 eliminates activity (Fig. 2D). Deletion of the P arm of *attR* shows that loss of sequences beyond -35 substantially impacts recombination. A 49 bp substrate from -35 to +14 (pBL223) has about a 2-fold reduction in recombination relative to the pBL144 parent substrate (Fig. 2D). These data show that there are seemingly no site requirements for *attL* x *attR* excisive recombination other than those encompassed within *attP* and *attB* that are required for integration. The precise endpoints for recombination at the P, P', B and B' halves of each site are unclear because of the likely contributions of vector sequences substituting for the Int recognition requirements. Nonetheless, the requirements at the P' and B' arms are similar – as expected because they share the common core, whereas the length requirements for the B and P arms are clearly different (see Fig. 1A).

Brujita *attP* x *attP* and *attB* x *attB* recombination

These observations indicate several unusual aspects of the Brujita integrase-mediated recombination. Perhaps most importantly, the lack of arm-type sites confounds site identities and the question of how recombinationally functional pairs of sites are selected. For example, although *attP* and *attB* are clearly distinct, *attL* and *attR* are effectively identical to *attB* and *attP* respectively (see Fig. 1A). The lack of RDF requirements further illustrates the apparent lack of strong site selection or recombination control. However, because *attB* and *attP* differ in their size needs, we asked whether Int discriminates between these for recombination.

We constructed similar plasmids to those described above, but which contain either two *attB* sites (pBL224) or two *attP* sites (pBL225; Fig. 1D, E). Using large site substrates, both of these recombine, albeit with somewhat reduced efficiency relative to pBL90 (*attP* x *attB*; Fig. 2E). Smaller *attB* sites containing the segment from -35 to +39 also support recombination, but further reduction of the sites (e.g. pBL228) renders them inactive (Fig. 2E). A substrate in which one of the *attB* sites contains from coordinates -35 to +39 and the other from -5 to +14 (pBL227) supports some recombination, but it is about 100-fold reduced from pBL90. We confirmed for pBL224 and pBL225 that recombination is integrase-dependent.

Mapping the Brujita crossover points

Tyrosine recombinases use a sequential strand cleavage mechanism in which each strand is cut 6-8 bp apart generating 5' extensions [3]; most tyrosine integrases appear to cleave 7 bp apart. The intervening 6-8 bp segment (the overlap region) is typically identical in both recombinational partner sites and introduction of heterology in the overlap typically interferes with recombination [23]. Restoration of homology with sequence departures from the wild-type

site can restore recombination activity, unless those specific base pairs are also required for integrase recognition [23, 24]. Although the sites of Int cleavage are not known for Brujita integrase, we predict that they are close to the 5' side of the common core where they are located in attachment sites of other integrase systems [18] (Fig. 1A).

We constructed three series of mutant substrates derived from pBL90 (Fig. 1B) in which we individually mutated each of the first nine base pairs at the 5' end of the common core within *attB*, the corresponding mutation in *attP*, or with identical substitutions in both *attP* and *attB* (Fig. 3). Each of these was then tested for recombination. For the single site *attB* mutants, changes at +6, +7 and +8 have minor reductions in recombination, changes +3, +4, and +5 reduce recombination 5-20 fold, and changes at +1, +2 and +9 reduce recombination about 100-fold (Fig. 3). The behaviors of each of the mutants with corresponding changes in *attP* are similar (Fig. 3), most notably with little or no impact on substitutions at positions +6, +7, and +8, modest reductions (5-20 fold) with mutations at +3, +4 and +5, and substantial reductions (100-fold) at positions +1, +2, and +9 (Fig. 3).

For the double mutants, substitutions at positions +2 and +3 show increased recombination relative to the single substitutions and are likely to be within the overlap region (Fig. 3). The +1 position is also likely within the overlap region, and the single substitutions are strongly defective in recombination, with the double mutant showing no further reduction and a small increase relative to the single change mutants. Substitutions at both +5 and +9 are more reduced in the double mutants than in either of the respective single substitutions, and are candidates for positions important for Int recognition. Single substitutions at +6, +7, and +8 have little impact on recombination when in either *attP* and *attB* and not unexpectedly the double mutants behave similarly (Fig. 3).

The behaviors of the single and double mutants at the +7, and +8 positions strongly suggest that the right-most Int-mediated DNA cleavage is to the left of position +7. Not only do the single and double mutants at +7, and +8 have little or no effect on recombination frequency, sequencing of the *attL* product site shows that in all progeny analyzed the mutation segregates with the B' or P' half site, and thus are to the right of the cleavage point (Fig. 3). Interestingly, substitutions at the +6 position also have little impact on recombination frequency, but sequencing of sites shows a distribution of mutant and wild type alleles into *attL*, whether the mutation originates in either *attB* or *attP*. Position +6 is thus presumably within the overlap region, notwithstanding that a mismatch at that position is well-tolerated for strand exchange (Fig. 3). Likewise, for changes at the +1 position – at least in *attB* – the wild-type and mutant alleles are distributed in *attL*, consistent with +1 being within the overlap region (Fig. 3). Brujita Int thus cleaves *att* DNA six bases apart and we assume that, like other tyrosine recombinases, cleavage generates 5' extensions.

Binding of Brujita Int to *attP* and *attB* sites

To explore the requirements for Brujita Int to bind to its attachment sites, we first constructed two Int expression plasmids containing different translation initiation sites. One of these uses the annotated translation start site at coordinate 28,055 and expresses a 297-residue protein that we designate as Int_{short}. The second uses an upstream start site at coordinate 29,011 to express a 318-residue protein designated Int_{long}. Prior mutational analysis showed that alteration of the annotated start codon at 28,055 has no effect on *in vivo* recombination integration frequency, whereas alteration of the start codon at 29,011 reduces integration more than 10-fold, suggesting that the Int_{long} is the best candidate for the protein expressed from the phage genome [17]. Initial binding studies showed that Int_{short} and Int_{long} proteins bind similarly to *attB* and *attP* DNA substrates, although Int_{short} binds with about a 5-fold reduced affinity (Fig. S1).

Int_{long} was used for all subsequent experiments, and is simply referred to as Int.

Int binds to *attP* DNA to form two complexes (complex 1 and complex 2; Fig. 4A). Complex 1 forms at lower protein concentrations and behaves as an intermediate in the formation of complex 2 (Fig. 4A). Int forms two similar complexes with *attB* DNA, and an additional, weakly forming complex at high protein concentrations (complex 3) is also observed (Fig. 4B).

Stoichiometries of Int-DNA binding

Brujita Int elutes from a size-exclusion column at a volume consistent with it being predominantly monomeric in solution (data not shown), and there are two simple explanations of the binding patterns. First, complex 1 and complex 2 could each consist of a single protomer of Int bound to two different sites, with complex 3 (with *attB*) containing two Int protomers occupying both sites. Second, complex 1 and complex 2 could contain one and two protomers of Int, respectively (with *attB* complex 3 perhaps containing additional protomers). We constructed an Int expression vector in which Brujita Int is expressed as a fusion with maltose binding protein (MBP), and which binds to both *attB* and *attP* forming complexes with slower mobilities (Fig. 4C, D); we then used mixtures of Int and MBP-Int proteins to determine if any of the complexes contain more than one Int protomer (Fig. 4C, D). With *attP* DNA we observed one hybrid complex migrating with mobility intermediate to the complex 2's formed by each protein, indicating that complex 2 contains two Int protomers (Fig. 4C). Complex 1 presumably contains only a single protomer. A similar pattern was observed with *attB* DNA, except an additional hybrid complex was observed with intermediate mobility to the two complex 3 bands, suggesting that this also contains two Int protomers.

Because of the apparent promiscuity of site selection for recombination *in vivo*, it is plausible

that in reactions containing only a single input DNA (either *attP* or *attB*), Int could mediate formation of recombinant products and synaptic intermediates containing two DNA fragments; however, recombinant products would be indistinguishable from input DNA substrates. We therefore examined complex formation using two differently-sized DNA fragments containing either *attP* or *attB* (Fig. 4E, F). Using *attP* substrates of 292 bp and 89 bp all of the complexes can be accounted for by binding to the individual DNA fragments and no synaptic complexes are seen (Fig. 4E). We also do not see any recombinant product DNAs or product DNA complexes. Similar observations were made with *attB* DNA, although an additional slower migrating complex was observed when the two DNAs were used (Fig. 4F). However it seems unlikely that this is either a synaptic complex or a product complex, which would be expected to migrate with intermediate mobilities compared to those seen with the individual substrates.

Requirements for Brujita Int binding to *attB*

These observations suggest that there are at least two Int binding sites in *attB* DNA. It is expected that Int would bind to both the B and B' binding sites to facilitate cleavage of each of the two strands during recombination. We predict that complex 2 contains two Int promoters bound to the B and B' half site, but complex 1 could either be a mixture of a single Int protomer bound at either B or B', or a single complex with Int bound to just one of the two sites. To dissect this binding pattern we performed a detailed mutagenesis study using a large set of deletion and base substitution substrates. Examples of key binding patterns are shown in Figure 5A, and summaries of all mutants tested are shown in Figures 6 and 7; Int binding data are shown in Figure S1 and 2.

First, we demonstrated that all of the binding requirements lie within a region of 37-44 bp encompassing putative B and B' half sites and correspond closely to the *attB* segment required

for recombination *in vivo* (Fig. 5A). Int binds to attB44 and attB37 (containing from coordinate -14 to +23) substrates with similar affinities to the larger attB345 DNA and form similar patterns of protein-DNA complexes (Figs. 5, 6, S1); reduction to shorter fragments (i.e. attB33, attB29) greatly reduces the affinity of Int binding (Fig. 6). When *attB* is truncated from its B side (e.g. attB28), complex 2 is lost, but complexes 1 and 3 are still observed (Fig. 5A). However, when *attB* is truncated from its B' side, DNA binding is essentially eliminated (attBL1, Fig. 5A). A simple explanation for these binding patterns is that complex 1 contains a single protomer of Int bound at the B' half site, and that occupancy of the B half site to form complex 2 is dependent on Int binding to the B' site, presumably through cooperative interactions between Int protomers. The behaviors of additional substrates (Fig. 6) suggests that the boundaries of the B' half site are between coordinates +4 and +21; the boundaries for B half site binding are less clear because of the dependency on B' binding, but the requirements do not extend to the left of -14 (Fig. 6). We also observed that Int binding varies as the sequence information between -5 and -14 is replaced by either non-specific DNA or by the plasmid vector sequences present in the substrates used for recombination (Fig. 2), consistent with the prediction that this region is important for Int recognition. The identity and functional relevance of complex 3 is unclear, and although it is observed with some substrates in which the B half is lost (e.g. attBL26), it is not seen in others (e.g. attBR6; Fig. 5A).

Using a series of single base transversion substitutions we examined the requirements for Int binding to the B and B' half sites (Figs. 5, 7). First, B' binding was tested using derivatives of a parent substrate attBR6 (-7 to +23; Figs. 5A, 7). Most substitutions have either no effect on binding (+4, +6, +12, +13, +14, +15, +16, +19, +20, +21, +22, +23), or only modest reduction in binding affinity (+5, +7, +8, +11, +17, +18). However, substitutions at both +9 and +10 effectively eliminate binding (attBR6SDM8, attBR6SDM9; Fig. 7) suggesting that the bases at the 5'-AG dinucleotide play a particularly important role in Int recognition. Next, we examined

single base substitutions in the context of a larger parent substrate (attB37) containing coordinates -14 to +23. Substitutions within the B' region mirror those seen with the smaller substrate, with changes at +9 and +10 (attB37SDM20 and attB37SDM27; Fig. 5A) strongly interfering with all complex formation. Substitutions in the B half of *attB* have little or no effect on complex 1 formation as expected, but some alter complex 2. The most notable of these are substitutions at -3 and -4 (attB37SDM6 and attB37SDM26, Figs. 5A and 7), which strongly interfere with complex 2 formation. Interestingly, the 5'-AG at positions -3 and -4 correspond to the symmetrically related – with regard to the proposed crossover points – 5'-AG at +9 and +10. Thus although the bases at this dinucleotide are critical for binding at both the B and B' half site, it is clearly not sufficient for Int recognition as there is no binding to the B half site in the absence of B'. Interestingly, substitutions at -1 and +1 (attB37SDM3 and attB37SDM16) also strongly reduce complex 2 formation, although the symmetrically related substitutions at +6 and +7 have only modest impacts on binding either in the *attB* context (attB37SDM9 and attB37SDM10) or in the B' half site context (i.e. attBR6SDM5 and attBR6SDM6) (Fig. 7). We note in particular that the same substitutions at the 5'-T position at +1 and +6 (5'- T+1A and 5'- T+6A, attB37SDM16 and attB37SDM9, respectively) have notably different impacts on Int binding.

Requirements for Brujita Int binding to attP

The recombination requirement for an *attP* substrate longer than *attB* suggests that Int binding differs between the two sites. However, the B' and P' half sites are wholly contained within the common core and thus any differences must lie on the P arm side of *attP*. A similar dissection was performed as for *attB* and examples of key binding patterns are shown in Fig. 5B, and summaries of all mutants tested are shown in Figure 8; Int binding data are shown in Figures S4 and S5.

434

435 All of the *attP* requirements for binding appear to be contained within coordinates -42 and +21
436 (*attP*63) and further reduction to 49 bp (coordinates -35 to +14, *attP*49) eliminates Int binding
437 (Fig. 8). Several substrates illustrate the difference between recognition of *attP* and *attB*. Firstly,
438 Int binding to *attP*37 forms only complex 1, whereas the *attB* substrate containing the same
439 coordinates (*attB*37, -14 to +23) forms three complexes (Fig. 5B, 8); moreover, the *attP*37
440 complex has a similar relative mobility to *attB*37 complex 1, suggesting it contains only a single
441 protomer of Int. A smaller substrate, *attP*28, lacking a full P half site, shows no difference in
442 binding as compared to *attP*37. Because the *attP*37 substrate spans the P and P' regions
443 equivalent to B and B', the simple interpretation is that Int does not bind at a position
444 corresponding to a P half site (Fig. 8), even when the P' half site is occupied, in contrast to Int
445 binding to *attB*. Secondly, Int binds to substrates completely lacking the P' half site (e.g. *attP*30),
446 although the requirements do not extend to the right of coordinate -13 (Fig. 5B, 8). Int binding to
447 these substrates forms a single complex, and we propose that complex 1 formed with the whole
448 *attP* site contains a mixture of two different complexes, one in which one protomer is bound at
449 the P' half site, and a second in which a protomer is bound at this second site, which we
450 propose to call P1 (see Fig. 1A); *attP* complex 2 presumably contains two protomers, one bound
451 at P1 and the other bound at P', and dual occupancy does not involve cooperative interactions.

452

453 Single base substitutions at this P1 site based on a parent DNA substrate containing
454 coordinates -50 to -13 (*attP*12) show that positions -32 and -33 are particularly important for Int
455 binding and substitutions there strongly interfere with complex formation (Fig. 8, 9). These
456 correspond to a 5'-AG dinucleotide suggesting that the P1 site is in direct orientation to the P'
457 site and that these correspond to the P' +9 and +10 positions. Substitutions at other positions
458 (e.g. -31, -34, -35) also have substantial negative impacts on Int binding, and substitutions to
459 the left of -37 do not alter Int binding affinity (Fig. 8). Importantly, derivatives of the whole *attP*

site (attP67) containing either substitutions at the 5'-AG in P' (+9, +10) or in P1 (-32, -33) both eliminate complex 2 formation and form a single complex 1 that share the same mobility (Fig. 8, 9). Furthermore, substitutions at these positions together in both the P1 and P' sites completely eliminates binding to *attP*, and substitutions in the P site does not influence binding (Fig. 8, 9). This not only confirms the model proposed for *attP* recognition, but illustrates that Int does not appear to bind a P half site at an equivalent position to the B half site, even when Int occupies the P' and P1 sites. This is unexpected, because binding at the P half site position is presumably needed for nucleophilic attack of the DNA at *attP*.

Int binding to the P half site is required for recombination

The recombination assays described above indicate that the P1 site is needed for recombination. A plausible role is that it promotes binding to the P half site, but only in the process of synapsis. Thus although the P site is not occupied in *in vitro* DNA binding assays, binding to P is presumably required for recombination *in vivo*. We therefore predict that a mutation substituting the -3 and -4 positions would impact recombination without affecting the *in vitro* DNA binding patterns. To test this we constructed a series of pBL90 derivatives in which the critical 5'-AG dinucleotides are substituted in the B or B' half sites in *attB*, or in the P', P1 and putative P site in *attP* (Fig. 9A). As expected, the substitutions in the B' and P' half sites strongly reduce recombination; not only are these positions important for Int binding, but substitutions at the +9 position were shown to interfere with recombination (Fig. 3). The substitutions in the B and P1 sites also strongly reduce recombination confirming the requirement for Int binding at those sites (Fig. 9A). Interestingly, substitutions at the -3 and -4 positions in *attP* also strongly reduces recombination, though not as severely as the other positions, consistent with the need for Int to bind at the P half site, even though no binding is observed at the P site in *attP* DNA.

DISCUSSION

Integration-dependent immunity systems represent an unusual variation on the regulation of lysogeny in temperate phages. The integrases of these systems are characterized by having a C-terminal tag for proteolysis, but also appear to lack an N-terminal arm-type binding domain, conferring a domain organization that more closely resembles simple tyrosine recombinases such as Cre, FLP, and XerCD [25, 26]. Although thousands of tyrosine integrase sequences have been deposited in the public databases, there are at least a hundred whose alignments suggest that, like Brujita, they lack an N-terminal domain. This organization may be relatively common, even though it represents a significant departure from the well-studied λ integrase prototype [6].

Deletion analyses of *attP* and *attB* demonstrate that the sites are relatively short, sharing a common end-point within the common core, and with *attP* having an extended left arm relative to *attB*. Because the P' and B' sites are identical, *attR* is effectively identical to *attP*, and *attL* is the same as *attB* (Fig. 1A). In this regard, it is perhaps not surprising that Brujita Int mediates *attL* x *attR* excisive recombination in the apparent absence of a phage-encoded RDF (Fig. 2). The distinction between the requirements at *attP* and *attB* could have conferred site-selectivity such that *attP* and *attB* would not recombine with themselves. However, this does not appear to be the case, and *attP* x *attP* and *attB* x *attB* recombination is also observed. Site-selection thus appears to be completely promiscuous (at least among *att* sites; there is as yet no evidence for use of secondary *att* sites). We note that small differences in recombination efficiencies may be magnified in vivo with an Int that does not carry a stabilizing mutation.

The positions of integrase binding sites agree well with the site requirements for recombination.

At the P' and B' sites, resection to the +14 position has variable impacts depending on whether it is in the *attP*, *attB*, *attL* or *attR* contexts (Figs. 2 and 3) consistent with +14 being close to the extremity of the B'/P' binding requirement, and adjacent vector sequences contributing to site function. Int binding studies are consistent with this and suggest that there are no sequence requirements to the right of the +19 coordinate (Fig. 7). On the left side of *attB*, resection to -10 has minimal impact on recombination, and binding studies suggest that there are no sequence-specific requirements to the left of the -11 position (which is in the symmetrically related position to +17). Likewise, at the left end of *attP*, resection to -35 has little impact on recombination, and there appears to be no sequence-specific requirements to the left of position -37.

Mutational analysis identifies the bases at a 5'-AG dinucleotide that appear to be particularly important for Int recognition. Mutations at these positions in both the B and B' half sites essentially eliminates binding, and the behaviors of substitutions at the 5'-AG at positions -32 and -33 supports identification of a similar Int binding site (P1) oriented in the direct orientation as P'. Why the P1 site is required for recombination is unclear, and it is not positioned such that an Int protomer could mediate nucleophilic attack at the crossover sites. It is plausible that it facilitates the binding of Int at the P site during synapsis, as the putative P site is not occupied even when both P' and P1 are occupied on a linear substrate *attP* DNA. The lack of Int binding at the P site is a peculiar observation, especially as Int binding to the B' site provides strong cooperative interactions that facilitate binding at the B site. Even though the critical 5'-AG dinucleotide is present in the P region (at -3, -4), Int binding at P' does not offer cooperative binding to P regardless of the occupancy of P1. Cooperative interactions from the Int bound at P1 could promote binding to the P site, although these would require different protein-protein interactions than those facilitating B site binding, given the relative orientations of the sites (see Fig. 1A). Preliminary data show that small insertions (2-5 bp) between -18 and -19 substantially reduced recombination efficiencies, consistent with this interpretation (data not shown).

538

539 Several features of the Brujita integration system suggest that it shares more in common with
540 simple site-specific tyrosine recombinase systems such as FLP and Cre than the prototypical λ
541 integrase-like systems. For example, there are clearly no arm-type sites and little or no apparent
542 regulation of directionality. There is also no evidence of the role of a host factor, and the sizes of
543 the attachment sites suggest there is little space available for host factor binding. The presence
544 of a third recombinase binding site flanking those at the crossover site is reminiscent of the third
545 FLP binding site at FRT [27, 28]. Although FLP binding to this third FRT site has been shown, it
546 is different from Brujita P1 in that it is dispensable for recombination [26, 29]. The crossover
547 points for Brujita recombination are positioned 6 bp apart as deduced from mutations in the
548 overlap region, and are symmetrically positioned relative to the binding sites identified at B and
549 B'. XerC/XerD and Cre also cleave with a 6 bp overlap [30, 31] whereas FRT has an 8 bp
550 overlap [26, 31, 32]. However, phage integrase systems are typically restricted to a 7 bp overlap
551 region [6, 18-21]; alternative overlaps have been proposed but can be difficult to confidently
552 assign [33, 34].

553

554 The promiscuity in site selection and lack of directionality in Brujita integrase-mediated
555 recombination has several notable consequences for regulation of the phage life cycles. First, it
556 is plausible that discrimination between integrative and excisive recombination is mediated in
557 part by total levels of the integrase protein resulting from the balance of gene expression and
558 protein turnover, although mutationally stabilized Int stimulates both integrative and excisive
559 recombination. We note that Brujita lysogens release relatively high levels of phage particles,
560 which is consistent with relatively high levels of prophage excision. Second, concatemers of
561 phage DNA are potentially at risk of recombinational reduction by Int-mediated *attP* x *attP*
562 recombination, which could interfere with DNA packaging. Preliminary observations in related
563 system suggest, however, that Int is not expressed during lytic growth thus avoiding this

564 potential issue.

565

566 Finally, it has been proposed that the relative simplicity of the integrase-dependent immunity

567 systems could reflect an ancestral state in evolution of phage life cycle regulation, given its

568 relative simplicity and smaller number of components than deployed in more complex systems.

569 Likewise, the relative simplicity of the Brujita integration system suggests that these represent

570 an ancestral state upon which the addition of N-terminal Int domains, arm-type sites, and

571 recombination directionality factors represent subsequent adaptations. It is not clear whether this

572 simple integrase system is also ancestral to other simple recombinase systems such as XerC/D,

573 Cre and FLP, although this remains an attractive hypothesis.

574

575

EXPERIMENTAL PROCEDURES

Bacterial growth

M. smegmatis cultures were grown in 7H9 broth or 7H10 agar (Difco) with carbenicillin (50 µg/ml), cyclohexamide (10 µg/ml), and appropriate antibiotics. *E. coli* cultures were grown in LB broth or agar (Difco) with appropriate antibiotics.

Plasmid construction

Plasmid pBL61 expressing Int_{long} was constructed as follows: plasmid pET21a was digested with Nde I and Hind III and a 957 bp Int_{Long} fragment corresponding to Brujita coordinates 29011 - 28057 was PCR amplified and joined to pET21 using Gibson Assembly (Invitrogen). Plasmid pBL60 expressing Int_{short} was constructed similarly using a 894 bp PCR fragment containing Brujita coordinates 28948 – 28057. Plasmid pBL51 expressing an MBP-Int fusion was constructed by linearizing plasmid pLC3 [35] with Nde I and Hind III and ligating it with the 957 bp Int_{Long} PCR product. Plasmid pBL83 was made by digestion of pBAD28, which contains an arabinose inducible promoter, [36] with Kpn I and Xba I and ligating it with a PCR amplified fragment consisting of Brujita Int_{Long} cleaved with Kpn I and Xba I. All Int expressing plasmids contain a stabilizing mutation (A296E).

Plasmids containing *attP* and *attB* were created from a pJL37 backbone [37]. The hsp60 promoter was removed by digestion with Xba I and Nde I and replaced with a 999 bp hygromycin resistance marker, PCR amplified from pYUB854 [38]. *attP* constructs of differing lengths, shown in Figures 2, 3, and 9 were PCR amplified and inserted into the Bam HI and Nde

I sites, and *attB* constructs were similarly inserted into the Not I and Kpn I sites using either Infusion (Clontech) or Gibson Assembly (Invitrogen). All plasmid constructs were verified by Sanger sequencing. Counter-selection plasmids (pBL90 derivatives) were modified from pBL50 derived plasmids by digestion with Dra I and cloning of a PCR amplified *ccdB* gene and upstream sequence (including the promoter region, 669 bp total) from pDONR/Zeo (Thermo Fisher Scientific). All *ccdB* plasmids were grown in One Shot *ccdB* survival competent cells (Thermo Fisher Scientific).

Attachment site mutants were generated using Quick Change Site-Directed Mutagenesis (Stratagene) to generate mutants shown in Figure 3. Mutants were PCR amplified and cloned into the pBL90 backbone as described above. Mutants were recombined using integrase expressing pBL83 cells as before. Following recombination, plasmid was extracted from transformed colonies and sequenced to determine the presence of the mutation.

Expression and purification of Brujita Int

Plasmid pBL61 was transformed into BL21-CodonPlus (DE3)-RIPL cells (Agilent). An isolated colony was used to inoculate 50 ml of LB and antibiotics and incubated overnight at 37°C. After incubation, the cells were pelleted and used to inoculate 2 liters of autoinduction media [39]. The cells were incubated while shaking for 18 hours at room temperature, pelleted, and frozen at -80° C. Thawed pellets were resuspended in 10 ml/gram lysis buffer (10 mM phosphate pH 7.0, 400 mM NaCl, 5 mM BME, 5 mM imidazole, 5% glycerol), protease inhibitors added (PMSF, leupeptin, aprotinin, and pepstatin) and lysed by three passes through a homogenizer. The lysed cells were spun at 30,000 x g in a Sorvall Lynx 6000 centrifuge for 40 minutes at 4° C. Untagged Integrase protein was isolated using nickel affinity chromatography as native

integrase has been found to bind well to nickel, followed by heparin affinity chromatography. Integrase containing fractions were further purified by size exclusion chromatography, dialyzed into freezer storage buffer, (10 mM Phosphate pH 7.0, 400 mM NaCl, 50% glycerol, 5 mM BME) and stored at -20° C.

In vivo recombination assays

To assay for integrase-mediated recombination in *M. smegmatis*, 50 ng of an *int*-expressing plasmid pBL03 [17] and 50 ng plasmid pLB50 were co-electroporated into *M. smegmatis* mc²155 and transformants were recovered on solid media containing kanamycin. Colonies were then picked onto plates containing either kanamycin or kanamycin and hygromycin. Recombination frequency was recorded as the percentage of transformants isolated on kanamycin related to those selected with both antibiotics.

For *E. coli* recombination assays, plasmid pBL83 was transformed into Stellar competent cells (Clontech), an isolated colony selected, and grown in LB broth with antibiotics overnight to saturation. Three mls of overnight culture were inoculated into 100 mls fresh broth with antibiotics and grown at 37° C with vigorous shaking to O.D.₆₀₀ = 0.5. Forty-five minutes after inoculation, arabinose was added to final concentration of 40 mM to induce Int expression. Cells were pelleted and washed 4 times with 10% ice-cold glycerol, frozen on dry ice, and stored at -80° C until use. Un-induced control cells were made similarly without the addition of arabinose. Recombination was measured by electroporating 50 ng of substrate plasmid DNA containing directly oriented attachment sites into Stellar *E. coli* cells carrying pBL83 and selecting kanamycin resistant transformants. Recombination frequencies were determined as cfu/ g DNA. Error values represent the average efficiency ± one standard error of the mean value for three separate transformations (six for the experiment shown in Fig. 9).

Electrophoretic Mobility-Shift Assays

EMSAs were performed by established protocols [40, 41]. DNA oligonucleotides were 5' radiolabelled with γ -³²P ATP and T4 polynucleotide kinase (Roche). Non-oligo DNAs (*attP292* and *attB345*) were PCR amplified and gel purified prior to radiolabelling as above. Binding reactions were prepared by adding 5 ng radiolabelled probe, one μ g calf-thymus DNA, binding buffer (20 mM Tris pH 8.0, 10 mM EDTA, 25 mM NaCl, 10 mM Spermidine, 1 mM DTT) and one μ l of varying concentrations of Int in a final volume of 10 μ l. After 30 minutes incubation at room temperature, the binding reactions were run on a 5% native TBE polyacrylamide gel at 250 V for 2-3 hours. Detection was performed using autoradiography and phosphorimaging.

ACKNOWLEDGMENTS

We thank Carlos Guerrero for excellent technical support. We also thank Aubrey Lowen, Swarna Mohan, Adam Wier, and Chris Amrich for technical assistance, and Andy VanDemark for the RIPL cell strains. This work was supported by a National Science Foundation pre-doctoral fellowship 1247842 (for fellow 2013141947) to BL.

REFERENCES

- [1] Landy A. Dynamic, structural, and regulatory aspects of lambda site-specific recombination. *Annu Rev Biochem.* 1989;58:913-49.
- [2] Herskowitz I, Hagen D. The lysis-lysogeny decision of phage lambda: explicit programming and responsiveness. *Annu Rev Genet.* 1980;14:399-445.
- [3] Grindley ND, Whiteson KL, Rice PA. Mechanisms of site-specific recombination. *Annu Rev Biochem.* 2006;75:567-605.
- [4] Lewis JA, Hatfull GF. Control of directionality in integrase-mediated recombination: examination of recombination directionality factors (RDFs) including Xis and Cox proteins. *Nucleic Acids Res.* 2001;29:2205-16.
- [5] Pena CE, Kahlenberg JM, Hatfull GF. Assembly and activation of site-specific recombination complexes. *Proc Natl Acad Sci U S A.* 2000;97:7760-5.
- [6] Landy A. The lambda Integrase Site-specific Recombination Pathway. *Microbiol Spectr.* 2015;3:MDNA3-0051-2014.
- [7] Ross W, Landy A. Bacteriophage lambda int protein recognizes two classes of sequence in the phage att site: characterization of arm-type sites. *Proc Natl Acad Sci U S A.* 1982;79:7724-8.
- [8] Landy A, Ross W. Viral integration and excision: structure of the lambda att sites. *Science.* 1977;197:1147-60.
- [9] Peña CE, Lee MH, Pedulla ML, Hatfull GF. Characterization of the mycobacteriophage L5 attachment site, attP. *J Mol Biol.* 1997;266:76-92.
- [10] Radman-Livaja M, Biswas T, Ellenberger T, Landy A, Aihara H. DNA arms do the legwork to ensure the directionality of lambda site-specific recombination. *Curr Opin Struct Biol.* 2006;16:42-50.

708 [11] Mizuuchi M, Mizuuchi K. The extent of DNA sequence required for a functional bacterial
709 attachment site of phage lambda. *Nucleic Acids Res.* 1985;13:1193-208.

710 [12] Pedulla ML, Lee MH, Lever DC, Hatfull GF. A novel host factor for integration of
711 mycobacteriophage L5. *Proc Natl Acad Sci U S A.* 1996;93:15411-6.

712 [13] Miller HI, Kikuchi A, Nash HA, Weisberg RA, Friedman DI. Site-specific recombination of
713 phage lambda: the role of host gene products *Cold Spring Harb Symp Quant Biol.*
714 1979;43:1121-6.

715 [14] Lewis JA, Hatfull GF. Control of Directionality in L5 Integrase-mediated Site-specific
716 Recombination. *J Mol Biol.* 2003;326:805-21.

717 [15] Abremski K, Gottesman S. Purification of the bacteriophage lambda xis gene product
718 required for lambda excisive recombination. *J Biol Chem.* 1982;257:9658-62.

719 [16] Broussard GW, Hatfull GF. Evolution of genetic switch complexity. *Bacteriophage.*
720 2013;3:e24186.

721 [17] Broussard GW, Oldfield LM, Villanueva VM, Lunt BL, Shine EE, Hatfull GF. Integration-
722 dependent bacteriophage immunity provides insights into the evolution of genetic switches. *Mol*
723 *Cell.* 2013;49:237-48.

724 [18] Peña CE, Stoner JE, Hatfull GF. Positions of strand exchange in mycobacteriophage L5
725 integration and characterization of the attB site. *J Bacteriol.* 1996;178:5533-6.

726 [19] Hauser MA, Scocca JJ. Site-specific integration of the *Haemophilus influenzae*
727 bacteriophage HP1. Identification of the points of recombinational strand exchange and the
728 limits of the host attachment site. *J Biol Chem.* 1992;267:6859-64.

729 [20] Smith-Mungo L, Chan IT, Landy A. Structure of the P22 att site. Conservation and
730 divergence in the lambda motif of recombinogenic complexes. *J Biol Chem.* 1994;269:20798-
731 805.

732 [21] Kolot M, Malchin N, Elias A, Gritsenko N, Yagil E. Site promiscuity of coliphage HK022
733 integrase as a tool for gene therapy. *Gene Ther.* 2015;22:521-7.

734 [22] Hiraga S, Jaffe A, Ogura T, Mori H, Takahashi H. F plasmid ccd mechanism in *Escherichia*
735 *coli*. *J Bacteriol.* 1986;166:100-4.

736 [23] Weisberg RA, Enquist LW, Foeller C, Landy A. Role for DNA homology in site-specific
737 recombination. The isolation and characterization of a site affinity mutant of coliphage lambda. *J*
738 *Mol Biol.* 1983;170:319-42.

739 [24] Weisberg RA, Gotesman M. The stability of *Int* and *Xis* functions. In: Hershey AD, editor.
740 *Bacteriophage Lambda*. New York: Cold Spring Harbor Lab. Press; 1971. p. 489-500.

741 [25] Van Duyne GD. A structural view of cre-loxp site-specific recombination. *Annu Rev Biophys*
742 *Biomol Struct.* 2001;30:87-104.

743 [26] Sadowski P. Site-specific recombinases: changing partners and doing the twist. *J Bacteriol.*
744 1986;165:341-7.

745 [27] McLeod M, Craft S, Broach JR. Identification of the crossover site during FLP-mediated
746 recombination in the *Saccharomyces cerevisiae* plasmid 2 microns circle. *Mol Cell Biol.*
747 1986;6:3357-67.

748 [28] Senecoff JF, Bruckner RC, Cox MM. The FLP recombinase of the yeast 2-micron plasmid:
749 characterization of its recombination site. *Proc Natl Acad Sci U S A.* 1985;82:7270-4.

750 [29] Gronostajski RM, Sadowski PD. Determination of DNA sequences essential for FLP-
751 mediated recombination by a novel method. *J Biol Chem.* 1985;260:12320-7.

752 [30] Blakely GW, Davidson AO, Sherratt DJ. Binding and cleavage of nicked substrates by site-
753 specific recombinases XerC and XerD. *J Mol Biol.* 1997;265:30-9.

754 [31] Ghosh K, Van Duyne GD. Cre-loxP biochemistry. *Methods.* 2002;28:374-83.

755 [32] Van Duyne GD. Cre Recombinase. *Microbiol Spectr.* 2015;3:MDNA3-0014-2014.

756 [33] Sylwan L, Frumerie C, Haggard-Ljungquist E. Identification of bases required for P2
757 integrase core binding and recombination. *Virology.* 2010;404:240-5.

758 [34] Coddeville M, Ritzenthaler P. Control of directionality in bacteriophage mv4 site-specific
759 recombination: functional analysis of the *Xis* factor. *J Bacteriol.* 2010;192:624-35.

[35] Owens RM, Hsu FF, VanderVen BC, Purdy GE, Hestende E, Giannakas P, et al. M. tuberculosis Rv2252 encodes a diacylglycerol kinase involved in the biosynthesis of phosphatidylinositol mannosides (PIMs). *Mol Microbiol.* 2006;60:1152-63.

[36] Guzman LM, Belin D, Carson MJ, Beckwith J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol.* 1995;177:4121-30.

[37] Bibb LA, Hatfull GF. Integration and excision of the *Mycobacterium tuberculosis* prophage-like element, ϕ Rv1. *Mol Microbiol.* 2002;45:1515-26.

[38] Bardarov S, Bardarov Jr S, Jr., Pavelka Jr MS, Jr., Sambandamurthy V, Larsen M, Tufariello J, et al. Specialized transduction: an efficient method for generating marked and unmarked targeted gene disruptions in *Mycobacterium tuberculosis*, *M. bovis* BCG and *M. smegmatis*. *Microbiology.* 2002;148:3007-17.

[39] Studier FW. Protein production by auto-induction in high density shaking cultures. *Protein expression and purification.* 2005;41:207-34.

[40] Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, et al. *Current Protocols in Molecular Biology*. New York: Wiley Intersciences; 1996.

[41] Villanueva VM, Oldfield LM, Hatfull GF. An Unusual Phage Repressor Encoded by *Mycobacteriophage* BPs. *PLoS One.* 2015;10:e0137187.

FIGURE LEGENDS

Figure 1. Brujita attachment sites and recombination assays. A. Organization of attachment sites used for Brujita integration. *attP* and *attB* share a 39 bp common core sequence (boxed) that overlaps a host tRNA-thr gene. Integrase-mediated strand exchange occurs within the common core to form *attL* and *attR* product sites, and mutations analysis (see below) suggest stand cleavages at the left end of the common core as indicated by the vertical arrows. DNA binding studies show that Int binds to two half sites, B, and B', in *attB* and to P' and a leftmost site designated P1 in *attP*. Int does not bind to the putative P half site *in vitro*, but the site is required for recombination *in vivo*. Because the recombination requirements do not extend beyond the common core to its right, the B' and P' sites are identical, and thus *attP* and *attR* as well as *attB* and *attL* are functionally equivalent. The regions required for *in vivo* recombination are highlighted in red. 5'-AG dinucleotides important for Int binding are shown in bold. **B.** Plasmids used for recombination assays *in vivo*. Plasmids pLB90, pBL144, pBL224, and pBL225, each contain an *E. coli* replication origin (oriE), a mycobacterial origin of replication (oriM), kanamycin and hygromycin resistance genes, and the counter-selectable *ccdB* gene [22]. Plasmid pBL90 contains Brujita *attP* and *attB* sites in direct orientation such that recombination between them yields two products, one containing the origins of replication and kanamycin resistance, the other carrying hygromycin resistance and *ccdB*; plasmids pBL144 (*attL* x *attR*), pBL224 (*attB* x *attB*) and pBL225 (*attP* x *attP*) are constructed similarly. Plasmid pBL50 (not shown) is similar to pBL90 but lacks the *ccdB* gene.

Figure 2. Length requirements of the Brujita attachment sites. *attP* (A), *attB* (B), *attL* (C), and *attR* (D) length requirements. Plasmid pBL90 contains 442 bp *attP* DNA (phage coordinates 28970 to 29411) and a 436 bp *attB* site with the common cores centrally located. When

transformed into *E. coli* cells expressing Brujita Int (pBL90+) and *ccdB* is counter-selected, approximately 8×10^4 colonies per microgram (cfu/ g) DNA are recovered; recombination is Int-dependent and few colonies are recovered in the absence of Int (pBL90-). Plasmid derivatives of pBL90 were constructed in which segments of either *attP* or *attB* (**A** and **B**, respectively) were deleted from the left or right ends, as indicated by the deletion endpoint coordinates (see Fig. 1A), and the numbers of recombinants recovered reported as a percent of control (shown at right). Derivatives of pBL144 (see Fig. 1) were similarly tested for *attL* (**C**) and *attR* (**D**) length requirements. **E.** Int-mediated recombination of pBL224 (*attB* x *attB*) and pBL225 (*attP* x *attP*) and derivatives. Errors values represent the standard errors of three separate transformations.

Figure 3. Mutational analysis of the putative overlap region in *attP* and *attB*. Derivatives of plasmid pBL90 were constructed with single base substitutions between coordinates +1 and +9 (see Fig. 1A) in either *attP*, *attB*, or both *attP* and *attB* as indicated, and tested for recombination. Positions of the base substitutions are shown, together with the numbers of recombinant products in which the substitution is present in the *attL* product.

Figure 4. Binding of Brujita Int to *attP* and *attB* DNA substrates. **A.** Binding of Brujita Int to *attP* DNA. The positions of free DNA (*attP*) and complexes (cmplx1, cmplx2) are indicated. Int concentrations are, from left to right, 0, 0.15, 0.61, 2.4, 9.8, 39, 156, and 625 nM. **B.** Binding of Brujita Int to *attB* DNA. The positions of free DNA (*attB*) and complexes (cmplx1, cmplx2, and cmplx3) are indicated. Int concentrations are as in part A. **C.** Binding of Int and MBP-Int (156 nM and 137 nM, respectively) separately and in combination (as indicated) to *attP* DNA. The positions of complexes and hybrid complexes containing both types of Int are indicated. **D.** Binding of Int and MBP-Int (156 nM and 137 nM, respectively) separately and in combination (as indicated) to *attB* DNA. The positions of complexes and hybrid complexes containing both types of Int are indicated. **E.** Binding of Int to two differently sized *attP* DNA substrates (89 bp,

and 292 bp, attP89 and attP292, respectively; as indicated), separately, and in combination. **F.** Binding of Int to two differently sized *attB* DNA substrates (44 bp and 345 bp; attB44 and attB345, respectively; as indicated). The nature of the lowest migrating complex with both DNAs (marked ?) is unclear.

Figure 5. Int binding to mutant derivatives of *attB* and *attP*. **A.** Binding of Int to representative *attB* DNA substrates. Int concentrations are, from left to right, 0, 0.15, 0.61, 2.4, 9.8, 39, 156, and 625 nM. See Figures 6 and 7 for substrates sequences. **B.** Binding of Int to representative *attP* DNA substrates. Int concentrations are as in part A. See Figure 8 for DNA sequences.

Figure 6. Summary of Brujita Int binding to *attB* substrates. *attB* DNA substrates of varying lengths were tested for ability to bind Int. The intensities of complex 1, 2 and 3 (as shown in Figures 4 and 5) are scored from weakest (-) to strongest (+++), with the +++ designation corresponding to a similar intensity to input DNA (see Figures 4 and 5). Overall binding affinities (Kds) are shown, calculated as the Int concentration at which 50% of the input DNA is bound, regardless of the complexes formed. Base positions in upper case black type correspond to wild-type *attB* sequences, and those in blue lower case type are non-native *attB* sequences corresponding to either plasmid derived sequences (substrates attB37plas) or random sequence. Individual gel assays are shown in Figures 5 and S1.

Figure 7. Summary of Brujita Int binding to *attB* substrates with single base substitutions. *attB* substrates with single base changes from the wild-type sequence were tested for Int binding. Complex intensities and affinities are as described for Figure 6. Individual gel assays are shown in Figures S2 and S3.

Figure 8. Summary of Brujita Int binding to mutant *attP* substrates. *attP* DNA substrates of varying lengths or with mutations were tested for ability to bind Int. The intensities of complex 1 and 2 (as shown in Figures 4 and 5) are scored from weakest (-) to strongest (+++), with the +++ designation corresponding to a similar intensity to input DNA (see Figures 4 and 5). Overall binding affinities (K_ds) are shown, calculated as the Int concentration at which 50% of the input DNA is bound, regardless of the complexes formed. Base positions in upper case black type correspond to wild-type *attP* sequences, and those in blue lower case type are non-native *attP* sequences corresponding to either plasmid derived sequences (substrates attP67plas) or random sequence. Individual gel assays are shown in Figures S4 and S5.

Figure 9. Brujita binding at the P half site is required for recombination. A. *attP* and *attB* substrates were mutated in B', B, P', P1, or P half sites as indicated (X), changing the 5'-AG bases in each half sites important for Int binding to 5'-TC. The recombination frequencies are shown compared to wild-type (pBL90). **B.** Int binding to *attP* substrates with 5'-AG mutations. From left to right, wild-type *attP*, mutations in P', mutations in P1, mutations in both P' and P1, mutations in P. Concentrations from left to right for attP67 through attP67SDM3; 0, 2.4, 9.8, 39, 156, 625 nM. Concentrations for attP67SDM4; 0, 0.15, 0.61, 2.4, 9.8, 39, 156, and 625 nM.

Figure 10. A model for Brujita Int-mediated recombination. Int protomers (yellow orbs) bind independently to B', P, and P' half sites. Binding at B' facilitates recruitment of an Int protomer to the B half site, presumably through subunit-subunit interactions, whereas Int binds independently to P1 and P' half sites, but not to P. We propose that recruitment of Int to the P half site occurs at synapsis, followed by cleavage at the crossover sites (vertical arrows), strand exchange, and release of *attL* and *attR* products.

Figure 1

Figure 2

Figure 3

Plasmid	Mutation Position	<i>attB</i>	<i>attP</i>	Recomb ⁿ (%)	Mutation in <i>attL</i>
pBL90	wt	+1 TAATGAATA +9	+1 TAATGAATA +9	100 ± 25%	N/A
pBL152	+1	aAATGAATA	TAATGAATA	0.5 ± 0.3%	1/5
pBL153	+2	TtATGAATA	TAATGAATA	0.4 ± 0.3%	2/5
pBL154	+3	TAtTGAATA	TAATGAATA	5.2 ± 2.5%	7/13
pBL155	+4	TAAaGAATA	TAATGAATA	17.7 ± 2.4%	2/13
pBL156	+5	TAATcAATA	TAATGAATA	11.0 ± 4.6%	2/10
pBL157	+6	TAATGtATA	TAATGAATA	43.9 ± 16.8%	2/17
pBL158	+7	TAATGAtTA	TAATGAATA	41.5 ± 11.6%	0/12
pBL159	+8	TAATGAAaA	TAATGAATA	66.0 ± 17.9%	0/11
pBL160	+9	TAATGAATt	TAATGAATA	1.3 ± 0.7%	0/9
pBL161	+1	TAATGAATA	aAATGAATA	0.1 ± 0.1%	0/4
pBL162	+2	TAATGAATA	TtATGAATA	1.2 ± 0.6%	3/7
pBL163	+3	TAATGAATA	TAtTGAATA	4.0 ± 1.0%	7/10
pBL164	+4	TAATGAATA	TAAaGAATA	43.5 ± 12.1%	9/10
pBL165	+5	TAATGAATA	TAATcAATA	26.7 ± 5.6%	9/10
pBL166	+6	TAATGAATA	TAATGtATA	122.2 ± 6.3%	4/10
pBL167	+7	TAATGAATA	TAATGAtTA	192.3 ± 31.9%	9/9
pBL168	+8	TAATGAATA	TAATGAAaA	96.7 ± 18.7%	10/10
pBL169	+9	TAATGAATA	TAATGAATt	3.4 ± 1.1%	8/8
pBL213	+1	aAATGAATA	aAATGAATA	0.9 ± 0.4%	
pBL214	+2	TtATGAATA	TtATGAATA	34.0 ± 12.7%	
pBL215	+3	TAtTGAATA	TAtTGAATA	50.1 ± 8.4%	
pBL216	+4	TAAaGAATA	TAAaGAATA	29.8 ± 10.3%	
pBL217	+5	TAATcAATA	TAATcAATA	0.6 ± 0.3%	
pBL218	+6	TAATGtATA	TAATGtATA	37.4 ± 28.8%	
pBL219	+7	TAATGAtTA	TAATGAtTA	27.3 ± 12.5%	
pBL220	+8	TAATGAAaA	TAATGAAaA	13.1 ± 6.9%	
pBL221	+9	TAATGAATt	TAATGAATt	0.2 ± 0.2%	

Figure 3

Figure 4

Figure 5

A

Figure 6

Figure 7

		Complex			Kd
		1	2	3	
<p> -30 -20 -10 +1 +10 +20 +30 +40 GCCGCCCTTAGCTCAGTCGGTAGAGCGATTCACTCGTAATGAATAGGTCGGGGGTTTCGATTCCCCAGGCGGGCTCTCTTT attB CGGCGGAATCGAGTCAGCCATCTCGCTAAGTGAAGTATTACTTATTCAGCCCCCAAGCTAAGGGGGTCCGCCGAGAGAAA B B' </p>					
attBR6	TCGTAATGAATAGGTCGGGGGTTTCGA	+++	-	-	18
attBR6SDM1	TCGTAATtAATAGGTCGGGGGTTTCGA	+++	-	-	59
attBR6SDM3	TCGTAATaAATAGGTCGGGGGTTTCGA	++	-	-	162
attBR6SDM4	TCGTAAaGAATAGGTCGGGGGTTTCGA	+++	-	-	28
attBR6SDM2	TCGTAATcAATAGGTCGGGGGTTTCGA	++	-	-	144
attBR6SDM5	TCGTAATGtATAGGTCGGGGGTTTCGA	+++	-	-	32
attBR6SDM6	TCGTAATGAtTAGGTCGGGGGTTTCGA	+++	-	-	123
attBR6SDM7	TCGTAATGAAaAGGTCGGGGGTTTCGA	+++	-	-	129
attBR6SDM8	TCGTAATGAATtGGTCGGGGGTTTCGA	+	-	-	-
attBR6SDM9	TCGTAATGAATAcGTCGGGGGTTTCGA	-	-	-	-
attBR6SDM10	TCGTAATGAATAGcTCGGGGGTTTCGA	++	-	-	102
attBR6SDM11	TCGTAATGAATAGGaCGGGGTTTCGA	+++	-	-	37
attBR6SDM12	TCGTAATGAATAGGTgGGGGGTTTCGA	+++	-	-	9
attBR6SDM13	TCGTAATGAATAGGTCcGGGGTTTCGA	+++	-	-	45
attBR6SDM14	TCGTAATGAATAGGTCGcGGGTTTCGA	+++	-	-	60
attBR6SDM15	TCGTAATGAATAGGTCGGcGGTTTCGA	+++	-	-	22
attBR6SDM16	TCGTAATGAATAGGTCGGGcGTTTCGA	++	-	-	95
attBR6SDM17	TCGTAATGAATAGGTCGGGGcTTTCGA	++	-	-	100
attBR6SDM18	TCGTAATGAATAGGTCGGGGGaTCGA	+++	-	-	29
attBR6SDM19	TCGTAATGAATAGGTCGGGGGTaCGA	+++	-	-	37
attBR6SDM20	TCGTAATGAATAGGTCGGGGGTTgGA	+++	-	-	21
attBR6SDM21	TCGTAATGAATAGGTCGGGGGTTCoA	+++	-	-	16
attBR6SDM22	TCGTAATGAATAGGTCGGGGGTTTCGt	+++	-	-	16
attB37	GAGCGATTCACTCGTAATGAATAGGTCGGGGGTTTCGA	++	+++	+	29
attB37SDM7	GAcCGATTCACTCGTAATGAATAGGTCGGGGGTTTCGA	+++	+++	+	26
attB37SDM15	GAGgGATTCACTCGTAATGAATAGGTCGGGGGTTTCGA	++	+++	+	18
attB37SDM4	GAGCcATTCACTCGTAATGAATAGGTCGGGGGTTTCGA	++	+++	+	25
attB37SDM14	GAGCGtTTCCTCGTAATGAATAGGTCGGGGGTTTCGA	+++	+++	+	13
attB37SDM24	GAGCGAaTCACTCGTAATGAATAGGTCGGGGGTTTCGA	++	++	+	35
attB37SDM5	GAGCGATaCACTCGTAATGAATAGGTCGGGGGTTTCGA	+++	++	+	25
attB37SDM25	GAGCGATTgACTCGTAATGAATAGGTCGGGGGTTTCGA	++	+++	+	25
attB37SDM13	GAGCGATTtCTCGTAATGAATAGGTCGGGGGTTTCGA	+++	+	-	15
attB37SDM26	GAGCGATTCAgTCGTAATGAATAGGTCGGGGGTTTCGA	+++	-	-	27
attB37SDM6	GAGCGATTCAcCGTAATGAATAGGTCGGGGGTTTCGA	+++	-	-	26
attB37SDM11	GAGCGATTCACTgGTAATGAATAGGTCGGGGGTTTCGA	+++	+	-	14
attB37SDM3	GAGCGATTCACTcTAATGAATAGGTCGGGGGTTTCGA	+++	-	-	20
attB37SDM16	GAGCGATTCACTCGaAATGAATAGGTCGGGGGTTTCGA	+++	+	-	12
attB37SDM17	GAGCGATTCACTCGTtATGAATAGGTCGGGGGTTTCGA	+++	++	+	17
attB37SDM18	GAGCGATTCACTCGTAtTGAATAGGTCGGGGGTTTCGA	++	+++	+	19
attB37SDM19	GAGCGATTCACTCGTAAaGAATAGGTCGGGGGTTTCGA	++	+++	+	24
attB37SDM1	GAGCGATTCACTCGTAATcAATAGGTCGGGGGTTTCGA	++	+++	++	113
attB37SDM9	GAGCGATTCACTCGTAATGtATAGGTCGGGGGTTTCGA	+++	++	+	17
attB37SDM10	GAGCGATTCACTCGTAATGAtTAGGTCGGGGGTTTCGA	+++	++	+	74
attB37SDM2	GAGCGATTCACTCGTAATGAAaAGGTCGGGGGTTTCGA	++	+++	+	119
attB37SDM20	GAGCGATTCACTCGTAATGAATtGGTCGGGGGTTTCGA	+	+	-	>625
attB37SDM27	GAGCGATTCACTCGTAATGAATAcGTCGGGGGTTTCGA	-	-	-	-
attB37SDM28	GAGCGATTCACTCGTAATGAATAGcTCGGGGGTTTCGA	+++	+++	-	87
attB37SDM21	GAGCGATTCACTCGTAATGAATAGGaCGGGGTTTCGA	+++	+++	-	57
attB37SDM22	GAGCGATTCACTCGTAATGAATAGGTgGGGGTTTCGA	+++	++	+	15
attB37SDM23	GAGCGATTCACTCGTAATGAATAGGTCcGGGGTTTCGA	+++	+++	++	61
attB37SDM8	GAGCGAaTaCTCGTAATGAATAGGTCGGGGGTTTCGA	+++	+	-	14
attB37SDM12	GAGCGAccCAtTCGTAATGAATAGGTCGGGGGTTTCGA	+++	+	-	17

Figure 8

Figure 9

A

Plasmid		Recomb (%)
pBL90		100 ± 32.6
pBL258		2.0 ± 0.7
pBL259		2.6 ± 0.7
pBL260		1.0 ± 0.2
pBL261		0.8 ± 0.2
pBL262		14.6 ± 6.3

B

Figure 10