

# Phage Integrases for Genome Editing

Michele P. Calos

**Abstract** Phage integrases are prokaryotic site-specific recombinases that perform precise cut-and-paste recombination between their short *attB* and *attP* recognition sequences. These enzymes work in cellular environments ranging from bacteria to mammalian cells and have become useful genome engineering tools. PhiC31 was the first phage integrase to be developed for use in mammalian cells. This integrase has the useful property of being able to recombine its own *attB* and *attP* sites. In addition, phiC31 integrase performs recombination at related native sequences called pseudo *att* sites present in large genomes, which has allowed integration into unmodified genomes. PhiC31 integrase can also be used in conjunction with another phage integrase, Bxb1, which has different recognition sequences and does not recombine at pseudo *att* sites. The properties of these phage integrases have led to a range of applications, summarized here, from creation of transgenic organisms and in vivo gene therapy, to cellular reprogramming and precise genome editing by cassette exchange. The latest system, dual integrase cassette exchange (DICE), uses target phiC31 and Bxb1 *attP* sequences precisely placed in genomes by homologous recombination and is especially useful for iterative genome engineering in pluripotent stem cells.

**Keywords** *attB* site • *attP* site • Bxb1 integrase • Cassette exchange • Embryonic stem cell • Homologous recombination • Induced pluripotent stem cell • phiC31 integrase • Reprogramming • TALEN

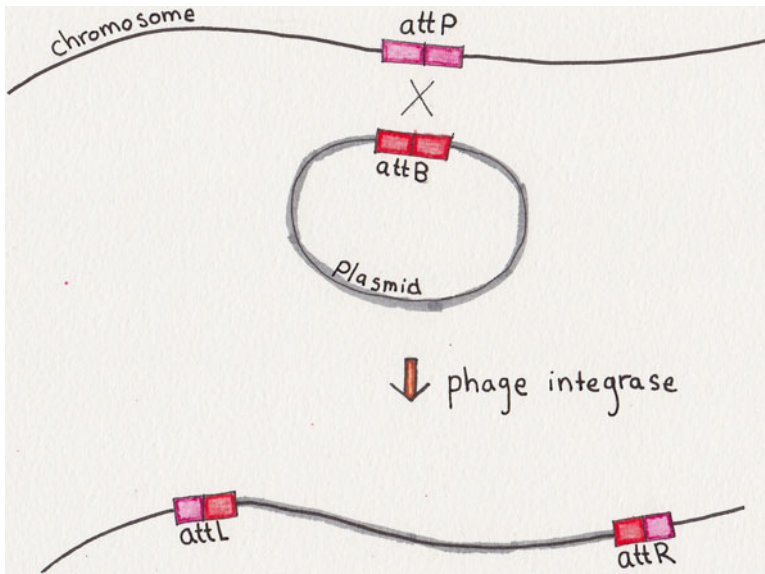
## Utility of Phage Integrases

It was reported in 1998 that phage phiC31 integrase could recombine its *attB* and *attP* recognition sites in *E. coli* and in buffer, independent of its native *Streptomyces* cellular environment [1]. These data suggested that the enzyme was self-contained

---

M.P. Calos, Ph.D. (✉)

Department of Genetics, Stanford University School of Medicine,  
300 Pasteur Drive, Alway Bldg., Room M334, Stanford, CA 94305-5120, USA  
e-mail: [calos@stanford.edu](mailto:calos@stanford.edu)



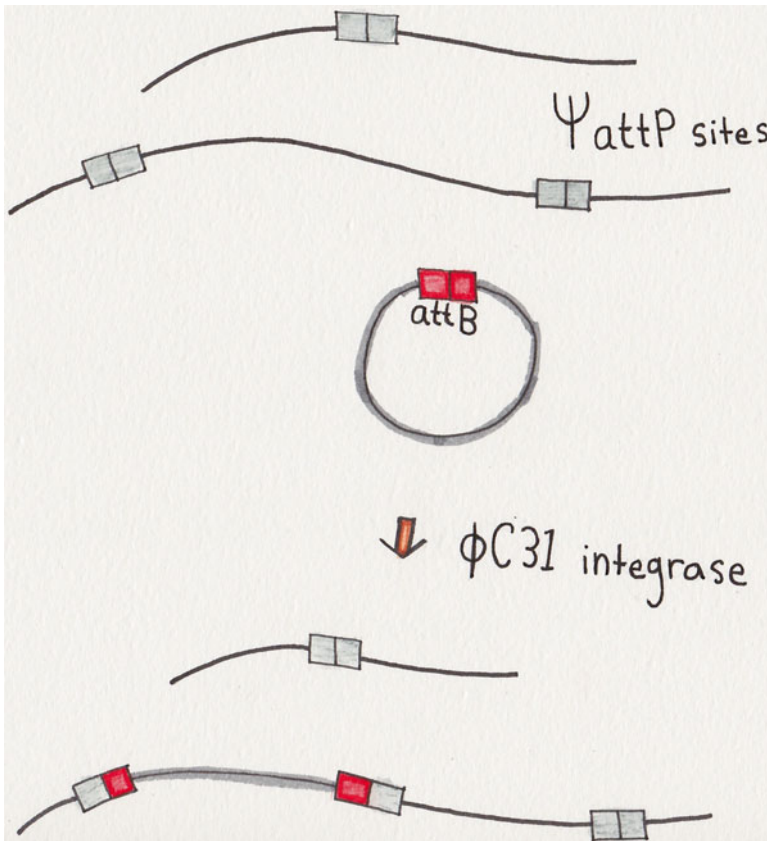
**Fig. 1** Basic phage integrase reaction. Phage integrases in the serine family possess *attB* and *attP* sites originating in the host bacterial and the phage genomes, respectively, that are approximately 34-bp long partially palindromic sequences. Dimers of integrase bind to each *att* site, and upon synapsis of the two sites (upper part of drawing), recombination takes place by a concerted cut-and-paste reaction at the center of the *att* sites, resulting in covalent strand exchange. By this means, a plasmid carrying an *attB* site can be integrated into a chromosome carrying the cognate *attP* site (lower). Hybrid *att* sites consisting of half of an *attB* and half of an *attP* site flank the integration. These sites are not substrates for integrase, so the reaction is unidirectional

and did not require host-specific cofactors for activity. On this basis, we hypothesized that phiC31 integrase might work in foreign cellular environments, such as mammalian cells, where site-specific integration systems were needed. Furthermore, the recombinational behavior reported for various combinations of *att* sites indicated that the recombination reaction was unidirectional [1], which was a desirable feature for high efficiency integration into mammalian genomes.

These predictions were realized in our 2000 study demonstrating activity of phiC31 integrase at its wild-type *attB* and *attP* sites in mammalian cells [2]. The length of the *att* sites was also defined in that study for the first time, at approximately 34-bp. As illustrated in Fig. 1, integrase mediates recombination across the center of the *attP* and *attB* sites, producing recombinant *attL* and *attR* sites that are no longer substrates for integrase, rendering a unidirectional reaction. It is now understood that each *att* site is bound by a dimer of integrase molecules, which mediate a concerted cut-and-paste recombination reaction [3].

## Integration into Pseudo Sites Versus into attP Sites

The *att* site length of ~30-bp suggested the possibility that rare native sequences with partial sequence identity might be adequate to catalyze recombination by phiC31 integrase. While a perfect match to a 30-bp sequence would not be statistically expected, even in large genomes, matches of 16-bp would be expected, and we predicted that this level of identity might be adequate for reaction. This prediction was fulfilled in studies in human and mouse cell lines, which revealed phiC31 integrase-mediated recombination at native sequences, named pseudo *att* sites [4] (Fig. 2).



**Fig. 2** The pseudo site integration reaction. A subset of serine phage integrases, notably phiC31 integrase, can interact with native chromosomal sequences having only partial identity with their true *attP* site. These sequences are known as pseudo *att* sites. Typically, a number of potential pseudo sites exist in the genome, defined by both DNA sequence and genomic context (three pseudo *att* sites are schematically illustrated in the upper part of the drawing). In the presence of phiC31 integrase and a plasmid bearing a phiC31 *attB* site, integration of the plasmid can occur at a pseudo *att* site, usually in single copy (lower). Integration at pseudo *att* sites is less precise than at genuine *attP* sites, and small deletions in the vicinity of the integration site often occur

The pseudo *att* site recombination reaction mediated by phiC31 integrase was the first instance of this type of “semi-specific” integration behavior into a mammalian genome. The level of specificity attained compared favorably with the integration specificity of the systems available at the time, including random integration of DNA and retrovirus- or transposon-mediated integration, also largely random. PhiC31 integrase-mediated integration at pseudo *att*-sites had an immediate appeal for situations involving integration into unmodified genomes, including in vivo gene therapy and construction of transgenic organisms.

Genomes typically harbor numerous potential pseudo *att* sites, and these sites appear to be utilized in a manner that depends on both the extent of DNA sequence identity and the genomic context; those pseudo *att* sites present in open, transcriptionally active chromatin are apparently more available for recombination [5]. A drawback of phiC31-mediated integration at pseudo *att* sites is the lack of predictability of where the integration reaction will occur, since there are generally multiple possibilities. In addition, integration at pseudo *att* sites is usually somewhat imprecise, often involving loss of several base pairs at the integration sites and sometimes more extensive chromosome rearrangements [5]. Nevertheless, pseudo site-mediated integration allowed integration into native, unmodified genomes in a manner that was orders of magnitude more site-specific than other systems available at the time.

## Use of phiC31 Integrase for Constructing Transgenic Organisms

One application of phage integrases that has been popular is the use of phiC31 integrase to place transgenes into the genome to construct transgenic organisms. Both pseudo *att* sites and authentic *att* sites have been used in this regard. For example, the pseudo *att* site reaction of phiC31 integrase was used to place genes into the genome of the amphibian *Xenopus laevis* [6]. In *Drosophila melanogaster*, the phiC31 *attP* site was first placed into the genome with a randomly integrating P element. The *attP* site was then targeted at high efficiency and specificity by an incoming plasmid bearing the *attB* site, utilizing co-injected mRNA encoding phiC31 integrase [7]. Variations on these themes have now been used to create transgenic organisms in a wide range of species, including fish, birds, amphibians, mammals, insects, and plants [8].

## Gene Therapy Studies Utilizing Integration into Pseudo Sites

The ability of phiC31 integrase to integrate DNA into unmodified mammalian genomes at a relatively low number of positions opened up new possibilities for in vivo gene therapy. The first studies utilizing phiC31 integrase for gene therapy took

advantage of a relatively simple and effective *in vivo* delivery method in mice, hydrodynamic injection, to place plasmids encoding integrase and a human factor IX gene into the liver [9]. Therapeutic levels of factor IX were produced after just one injection of small amounts of plasmid DNA. Site-specific integration of the therapeutic plasmid in hepatocytes was verified, and the integration specificity observed was impressive, since most integration occurred at one hotspot in the genome. Furthermore, the integration was stable, and factor IX production persisted long-term [10]. These features suggested that use of phiC31 integrase for correction of hemophilia might be clinically translatable. Studies in disease model mice for hemophilia A and B were carried out, with long-term expression of human factor VIII and IX observed [11, 12]. Unfortunately, efforts to translate the hydrodynamic DNA delivery method to the livers of larger animals have not been sufficiently effective to date to achieve therapeutic factor levels and enable clinical translation.

Many other gene therapy studies in animals have been carried out utilizing phiC31 integrase for genomic integration of plasmids carrying therapeutic genes in a variety of tissues and species (reviewed in [8, 13, 14]). For example, DNA was delivered by electroporation to rat retina, and long-term expression of a marker gene was observed [15]. Electroporation was also utilized to deliver plasmid DNA carrying the *DYSTROPHIN* gene to mouse muscle in a model of Duchenne muscular dystrophy [16]. While proof of principle for successful long-term delivery of plasmid DNA and site-specific integration into the chromosomes was demonstrated in these rodent studies, their clinical translation awaits effective delivery methods for these plasmid-based strategies that can be translated to large animals.

## Utilizing phiC31 Integrase for Reprogramming Mammalian Cells

One approach to circumvent the difficulty of effective delivery of plasmid DNA to the body is to deliver DNA first to cells *in vitro*, since effective transfection methods exist for most cell types, then deliver the transfected cells carrying integrated transgenes to the body. If the cells are immortal, then single cells with defined integration sites can be cloned, permitting the integration site to be defined, an important safety feature.

Cells that can be cloned include pluripotent stem cells, such as embryonic stem cells (ESC). In 2006, induced pluripotent stem cells (iPSC) were described, in which ESC-like cells could be derived from somatic cells by addition of four transcription factor genes [17, 18]. These iPSC hold extensive potential for regenerative medicine, because they are immunologically matched to the patient, can be grown to large numbers, are susceptible to genetic engineering methods, and are free of political or ethical issues. For genetic diseases, the iPSC from a patient can be corrected *in vitro*, for example by integration of the relevant therapeutic gene, then used in a therapeutic strategy involving *in vitro* differentiation, followed by transplantation to the appropriate tissue or organ.

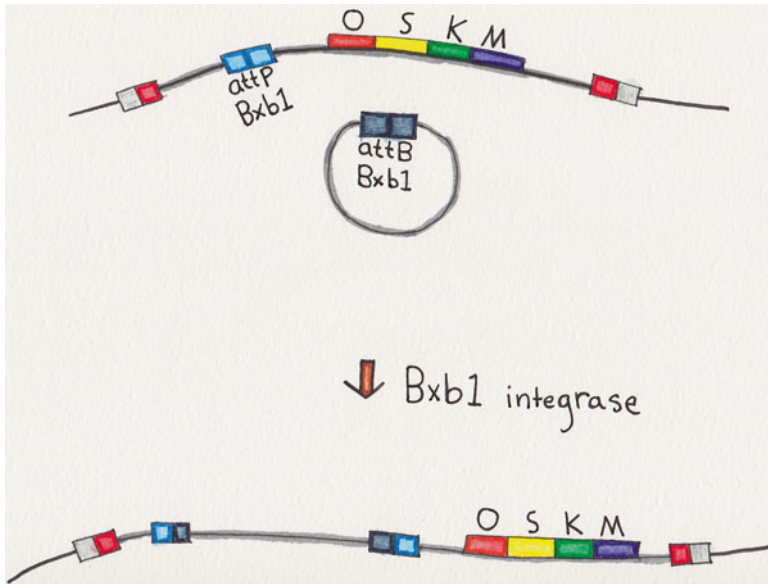
Reprogramming therefore opened up a vast number of potential therapeutic strategies. However, the initial methods to generate iPSC utilized retroviruses to integrate the four transcription factors into the genome. This methodology generally resulted in numerous, random integration events scattered about the genome, increasing the risk of insertional mutagenesis and resulting tumorigenesis [17, 18]. To overcome this problem, we devised a strategy utilizing phiC31 integrase to introduce one copy of a reprogramming cassette carrying all four transcription factors at a single, safe site in the mouse genome [19].

In this method, a reprogramming plasmid (p4FLR, 11.9-kb) was constructed that carried all four transcription factors identified by Yamanaka for reprogramming, along with strategically located recombinase recognition sites. The cDNA sequences for the murine *cMyc*, *Klf4*, *Oct4*, and *Sox2* genes were linked by 2A sequences, facilitating their polycistronic transcription from the CAG promoter. The plasmid also carried a phiC31 integrase *attB* site to mediate integration into the genome at pseudo *att* sites (Fig. 2). Two *loxP* sites flanked the reprogramming cassette, so that it could be deleted after reprogramming by transient transfection with a plasmid expressing Cre resolvase [19]. This step was important, to render the iPSC less tumorigenic and more amenable to differentiation.

The phiC31-mediated method was successful for reprogramming mouse embryonic fibroblasts and adult mesenchymal stem cells at efficiencies comparable to retroviral methods, without the handling hazards and random integration risks associated with viruses. Individual iPSC clones were analyzed by ligation-mediated PCR to determine the integration site, with the result that approximately one-third of the iPSC carried a single integration of p4FLR. Many different mouse pseudo *attP* sites were utilized in the collection of iPSC analyzed. Six of 14 sites were intergenic, and of those, two were considered to be safe sites, in terms of being distant from known cancer genes and other hazards. A plasmid expressing Cre resolvase was transfected into two representative iPSC clones, and precise deletion of the reprogramming cassette was demonstrated. Pluripotency of the iPSC before and after Cre excision was demonstrated, including ability of the iPSC to generate teratomas, as well as chimeric mice [19].

## Site-Specific Integration of a Therapeutic Gene at a Pre-integrated *attP* Site

We built further on the concepts in the initial reprogramming study in order to create a stronger reprogramming cassette that would supply a higher percentage of single-copy iPSC clones. We also included sequences on the reprogramming plasmid to permit us to integrate a therapeutic gene site-specifically into the iPSC [20]. The new reprogramming plasmid, pCOBLW, carried a more favorable order of the reprogramming genes, placing the Oct gene first and the Myc gene last (OSKM; Fig. 3). We also added the WPRE element to enhance transcription of the reprogramming cassette. pCOBLW was able to reprogram cells more effectively with



**Fig. 3** Integration into a Bxb1 *attP* site placed by phiC31 integrase. We developed a reprogramming plasmid carrying the Oct-Sox-Klf-Myc (OSKM) reprogramming cassette and also a Bxb1 *attP* site. PhiC31 integrase was used to integrate the reprogramming plasmid at a pseudo *attP* site, producing iPSC (upper part of diagram). After determining that the integration location was safe by DNA sequencing and bioinformatics analysis, a therapeutic gene borne on a plasmid carrying a Bxb1 *attB* site was integrated precisely at the Bxb1 *attP* site resident in the integrated plasmid in the iPSC, resulting in site-specific integration of the therapeutic gene at a safe location. Cre resolvase was then applied to remove unwanted sequences, including the reprogramming genes and plasmid backbone sequences, by recombining between strategically located *loxP* sites in the plasmids

only a single copy, reflected by 93 % of iPSC generated with this plasmid being single-copy [20].

Along with the *loxP* sites we previously included for use in excision of reprogramming genes and plasmid sequences, we included the *attP* site of Bxb1 integrase, as a target for addition of a therapeutic gene to the integrated reprogramming plasmid (Fig. 3). Bxb1 is a serine integrase related to phiC31, but its *att* sites are completely distinct and do not cross-react with those of phiC31. Bxb1 is active on its own *att* sites in mammalian cells, but does not recognize native pseudo *att* sites at a measurable frequency [21]. Insulator sequences were included, such that they would flank the therapeutic gene after integration, to reduce position effects on the therapeutic gene and on neighboring genomic sequences.

To carry out this reprogramming strategy, we nucleofected pCOBLW into fibroblasts derived from the *mdx* mouse, along with a plasmid encoding phiC31 integrase. Integration of the vector through its phiC31 *attB* site occurred at pseudo *att* sites. An iPSC line with integration at a safe site was chosen for addition of a therapeutic gene. In this case, we used the full-length cDNA for mouse dystrophin, which

is the gene affected in Duchenne muscular dystrophy. We carried out gene addition by nucleofection of the iPSC with a plasmid carrying the genes for dystrophin, a promoterless puromycin resistance gene, and the Bxb1 *attB* site, along with a plasmid encoding Bxb1 integrase. Correct integrants were identified by puromycin selection, since a promoter was located adjacent to the target *attP* site on the pCO-BLW plasmid. Correct site-specific integration was verified by PCR, and correct integrants were subjected to transient exposure to Cre recombinase to remove the reprogramming cassette and unwanted plasmid sequences [20].

We then carried out *in vitro* differentiation of the iPSC into muscle precursor cells, which were subsequently engrafted in a hind limb muscle of the *mdx* mouse model of Duchenne muscular dystrophy. This study provided a model for generation of iPSC without random integration, site-specific integration of the full-length dystrophin gene, and precise excision of unwanted reprogramming and plasmid sequences. In addition, proof-of-principle for the differentiation and engraftment of the cells was provided, suggesting a potential therapeutic approach for muscular dystrophy.

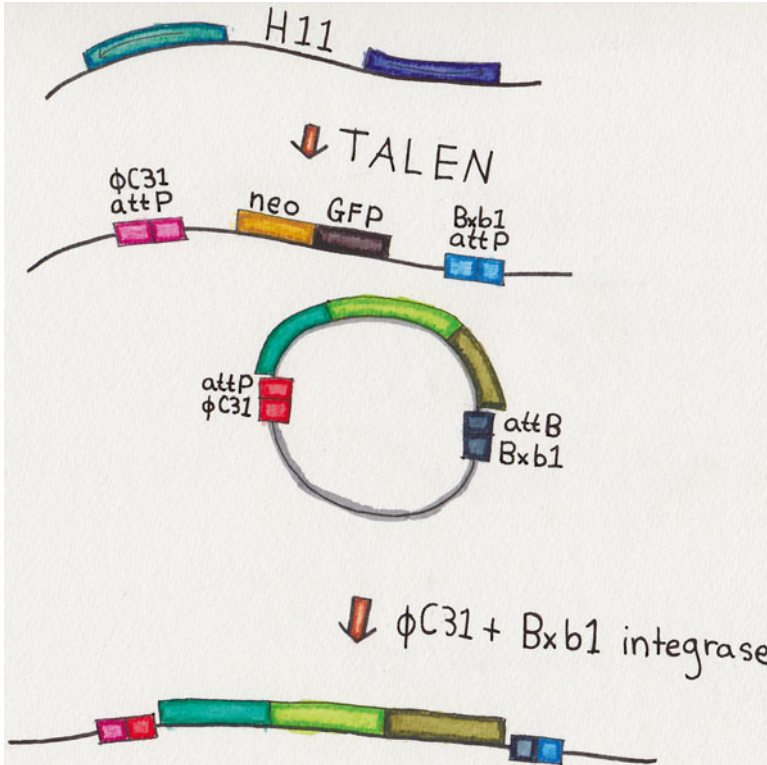
## The DICE System: Combining Homologous Recombination with phiC31 and Bxb1 Integrases for Cassette Exchange

While phiC31-mediated integration into pseudo *att* sites provides a convenient method for genomic integration into unmodified genomes, it is laborious to analyze a set of clones to find an integration site that is safe and desirable. In pluripotent stem cells, such as ESC and iPSC, as well as in immortalized cell lines, another strategy is available that allows the user to control the site of integration precisely, via homologous recombination. If homologous recombination is used to position *attP* sites for integrases, these sites can then be targeted precisely for integration and /or cassette exchange of incoming genes. We developed a strategy called Dual Integrase Cassette Exchange, or DICE that utilizes precise placement of *attP* sites for cassette exchange [22].

In the DICE method, a “landing pad” carrying *attP* sites for phiC31 integrase and Bxb1 integrase, is positioned in the genome by homologous recombination. In our study, we used an intergenic, safe, transcriptionally active site called H11 on human chromosome 22 as the destination for the landing pad. Neomycin resistance and GFP genes placed between the *attP* sites served as markers for selection and screening of clones carrying the landing pad (Fig. 4). We utilized TALENs targeted to H11 to stimulate the frequency of homologous recombination. This strategy was particularly valuable in iPSC having significant disease pathology that depressed the rate of homologous recombination.

Once a line has been constructed that carries a landing pad, it can be used to position incoming genes at the desired site by an efficient and site-specific cassette exchange reaction mediated by phiC31 and Bxb1 integrases. To obtain cassette exchange, the landing pad line is nucleofected with a donor plasmid carrying the





**Fig. 4** DICE, dual integrase cassette exchange. To carry out integration by DICE, a landing pad bearing phiC31 and Bxb1 integrase *attP* sites is placed into the genome at a desired location by homologous recombination. We inserted a landing pad in human ESC and iPSC at the H11 locus on chromosome 22, a safe site where transcription is ubiquitous. The frequency of homologous recombination can be stimulated by the use of TALENS to make double-strand breaks at the target site. Selectable and screenable markers such as neomycin resistance and green fluorescent protein can be used to facilitate identification of correct integrants (*upper part* of drawing). Once the landing pad is inserted in the genome, it can be targeted readily by introducing plasmids carrying the genes one wants to integrate, such as a therapeutic gene, marker genes, or transcription factors, flanked by phiC31 and Bxb1 *attB* sites, along with plasmids encoding the phiC31 and Bxb1 integrases. Site-specific cassette exchange occurs, with the genetic information between the *attB* sites now present in the chromosome (*lower*)

genes of interest flanked by *attB* sites for phiC31 and Bxb1 integrases, along with plasmids encoding the two integrases. During cassette exchange, the neomycin and GFP genes will be removed from the landing pad and the donor genes will be inserted in their place (Fig. 4). Donor genes can include therapeutic genes and/or genes for selection, screening, tracking engraftment, and so on.

The DICE system is particularly valuable when there is a need to construct a series of parallel cell lines with different genes inserted into the same precise location. During cassette exchange, the content, direction, and position of the incoming

genes are completely controlled, so the outcome is predictable. For example, the method allowed us to construct rapidly a series of human ESC and iPSC lines carrying all combinations of three neural transcription factors, to evaluate their roles in differentiation of dopaminergic neurons [22].

## Reversal of the Integration Reaction with $\phi$ C31 Excisionase

When utilizing the site-specific integration reaction mediated by  $\phi$ C31 integrase at its own *attB* and *attP* sites, the products of this reaction, *attL* and *attR* (Fig. 1) are different in sequence from the starting *attB* and *attP* sites and do not act as substrates for the integrase [1, 23]. Phage integrase systems generally also encode a small protein called an excisionase or Recombination Directionality Factor (RDF) that can bind to the integrase and alter its specificity so that the *attL* and *attR* sequences are now substrates for the integrase, resulting in reversal of the integration reaction.

The RDF for  $\phi$ C31 integrase was recently identified [24], leading to the possibility that this protein could be used to reverse the integration reaction when the enzyme was used in mammalian cells. We created assay plasmids to test this hypothesis and found that the  $\phi$ C31 RDF could, in combination with  $\phi$ C31 integrase, efficiently reverse the integration reaction [25]. Therefore, the availability of the  $\phi$ C31 RDF adds an additional tool that may be useful in future strategies employing phage integrases.

**Acknowledgements** M.P.C. thanks Victoria Ellis for creating the figures and the California Institute for Regenerative Medicine for financial support.

## References

1. Thorpe HM, Smith MCM. In vitro site-specific integration of bacteriophage DNA catalyzed by a recombinase of the resolvase/invertase family. *Proc Natl Acad Sci U S A*. 1998;95:5505–10.
2. Groth AC, Olivares EC, Thyagarajan B, Calos MP. A phage integrase directs efficient site-specific integration in human cells. *Proc Natl Acad Sci U S A*. 2000;97:5995–6000.
3. Rutherford K, Van Duyn G. The ins and outs of serine integrase site-specific recombination. *Curr Opin Struct Biol*. 2014;24:125–31.
4. Thyagarajan B, Olivares EC, Hollis RP, Ginsburg DS, Calos MP. Site-specific genomic integration in mammalian cells mediated by phage  $\phi$ C31 integrase. *Mol Cell Biol*. 2001;21:3926–34.
5. Chalberg TC, et al. Integration specificity of phage  $\phi$ C31 integrase in the human genome. *J Mol Biol*. 2006;357:28–48.
6. Allen BG, Weeks DL. Transgenic *Xenopus laevis* embryos can be generated using  $\phi$ C31 integrase. *Nat Methods*. 2005;2:975–9.
7. Groth AC, Fish M, Nusse R, Calos MP. Creation of transgenic *Drosophila* by using the site-specific integrase from phage  $\phi$ C31. *Genetics*. 2004;166:1775–82.

8. Geisinger J, Calos MP. Site-specific recombination using phiC31 integrase. In: Renault S, Duchateau P, editors. Site-directed insertion of transgenes. Dordrecht: Springer Science; 2013. p. 211–39.
9. Olivares EC, et al. Site-specific genomic integration produces therapeutic factor IX levels in mice. *Nat Biotechnol.* 2002;20:1124–8.
10. Chavez C, et al. Kinetics and longevity of phiC31 integrase in mouse liver and cultured cells. *Hum Gene Ther.* 2010;21:1287–97.
11. Keravala A, et al. Long-term phenotypic correction in factor IX knockout mice by using phiC31 integrase-mediated gene therapy. *Gene Ther.* 2011;18:842–8.
12. Chavez C, et al. Long-term expression of human coagulation factor VIII in a tolerant mouse model using the phiC31 integrase system. *Hum Gene Ther.* 2012;23:390–8.
13. Chavez C, Calos M. Therapeutic applications of the phiC31 integrase system. *Curr Gene Ther.* 2011;11:375–81.
14. Karow M, Calos M. The therapeutic potential of phiC31 integrase as a gene therapy system. *Expert Opin Biol Ther.* 2011;11:1287–96.
15. Chalberg TC, Genise HL, Vollrath D, Calos MP. PhiC31 integrase confers genomic integration and long-term transgene expression in rat retina. *Invest Ophthalmol Vis Sci.* 2005;46:2140–6.
16. Bertoni C, et al. Enhancement of plasmid-mediated gene therapy for muscular dystrophy by directed plasmid integration. *Proc Natl Acad Sci U S A.* 2006;103:419–24.
17. Takahashi K, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell.* 2007;131:861–72.
18. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell.* 2006;126:663–76.
19. Karow M, et al. Site-specific recombinase strategy to create induced pluripotent stem cells efficiently with plasmid DNA. *Stem Cells.* 2011;29:1696–704.
20. Zhao C, et al. Recombinase-mediated reprogramming and dystrophin gene addition in mdx mouse induced pluripotent stem cells. *PLoS One.* 2014;9(4), e96279.
21. Keravala A, et al. A diversity of serine phage integrases mediate site-specific recombination in mammalian cells. *Mol Genet Genomics.* 2006;276:135–46.
22. Zhu F, et al. DICE, an efficient system for iterative genomic editing in human pluripotent stem cells. *Nucleic Acids Res.* 2014;42(5), e34.
23. Thorpe HM, Wilson SE, Smith MCM. Control of directionality in the site-specific recombination system of the *Streptomyces* phage phiC31. *Mol Microbiol.* 2000;38:232–41.
24. Khaleel T, Younger E, McEwan A, Varghese A, Smith M. A phage protein that binds phiC31 integrase to switch its directionality. *Mol Microbiol.* 2011;80:1450–63.
25. Farruggio A, Chavez C, Mikell C, Calos M. Efficient reversal of phiC31 integrase recombination in mammalian cells. *Biotechnol J.* 2012;7(11):1332–6.