BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 3 2012, pages 306–310
doi:10.1093/bioinformatics/btr672

Genome analysis

Advance Access publication December 6, 2011

InFiRe — a novel computational method for the identification of insertion sites in transposon mutagenized bacterial genomes

Olga Shevchuk †, Louisa Roselius †, Gabriele Günther, Johannes Klein, Dieter Jahn, Michael Steinert * and Richard Münch

Institute of Microbiology, Technische Universität Braunschweig, Spielmannstrasse 7, 38106 Braunschweig, Germany

Associate Editor: John Quackenbush

ABSTRACT

Motivation: InFiRe, Insertion Finder via Restriction digest, is a novel software tool that allows for the computational identification of transposon insertion sites in known bacterial genome sequences after transposon mutagenesis experiments. The approach is based on the fact that restriction endonuclease digestions of bacterial DNA yield a unique pattern of DNA fragments with defined sizes. Transposon insertion changes the size of the hosting DNA fragment by a known number of base pairs. The exact size of this fragment can be determined by Southern blot hybridization. Subsequently, the position of insertion can be identified with computational analysis. The outlined method provides a solid basis for the establishment of a new high-throughput technology.

Availability and implementation: The software is freely available on our web server at www.infire.tu-bs.de. The algorithm was implemented in the statistical programming language R. For the most flexible use, InFiRe is provided in two different versions. A web interface offers the convenient use in a web browser. In addition, the software and source code is freely available for download as R-packages on our website.

Contact: m.steinert@tu-bs.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 19, 2011; revised on November 10, 2011; accepted on November 23, 2011

1 INTRODUCTION

Transposons are mobile genetic elements capable of random insertions into host genomes. This fundamental feature has been employed for decades to inactivate genes, leaving a detectable signature at the locus of integration. This well-developed genetic technique is applicable to the genomes of viruses, prokaryotes and eukaryotes (Hayes, 2003; Largaespada, 2009; Vilen *et al*., 2003; Yergeau and Mead, 2009). Various modifications of traditional transposon mutagenesis facilitated the development of new techniques, such as signature target mutagenesis (Hensel *et al*., 1995), transposon-mediated differential hybridization (Chaudhuri *et al*., 2009), genetic footprinting (Smith *et al*., 1995) and gene expression analysis (Judson and Mekalanos, 2000; Opperman *et al*., 2003). The crucial step of all these methods is the localization of the transposon insertion sites. In the course of the last years, several methods for the localization of transposons have been developed. Among them are the Vectorette PCR, inverse PCR and the cloning of restricted chromosomal DNA fragment into vectors with subsequent amplification (Arnold and Hodgson, 1991; Ochman *et al*., 1988). Single-primer PCR allows a rapid identification of insertion sites, but often requires optimization of the amplification conditions (Karlyshev *et al*., 2000). With the improvement of DNA sequencing techniques, it became possible to determine the transposon insertion site from the bacterial chromosomal DNA (Hoffman *et al*., 2000; Qimron *et al*., 2003).

Here, we present a novel computational method for the identification of insertion sites in transposon-mutagenized bacterial genomes. The functionality of the algorithm was successfully demonstrated using a mini-Tn*10* transposon *Legionella pneumophila* Corby library.

2 METHODS

2.1 InFiRe algorithm

To improve the efficiency of transposon mutagenesis experiments, we developed an algorithm for the identification of insertion sites in transposon-mutagenized bacterial genomes. The strategy of this procedure is outlined in Figure 1. It is based on the restriction digestions of genomic DNA in combination with Southern blot hybridization. In the first step, the genomic DNA is cleaved with different restriction endonucleases. Each restriction enzyme digest produces a unique pattern of DNA fragments with defined sizes. In the second step, the sizes of the fragments with an inserted transposon are determined by Southern blot hybridization with a transposon-specific probe. The size of the fragment in the original genome equals the obtained fragment size minus the length of the transposon. Finally, using the derived fragment size pattern the most probable genomic position of the transposon can be calculated by the InFiRe software. Afterwards, confirmation of the accurate determination of insertion sites can be performed by PCR.

2.2 Statistics

In order to determine the number of restriction enzymes required for InFiRe analyses, we developed two statistical models. The explicit model calculates the number of enzymes based on the genome sequence information, whereas the theoretical model estimates this number based on a random sequence of given length. Both models produce similar results.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

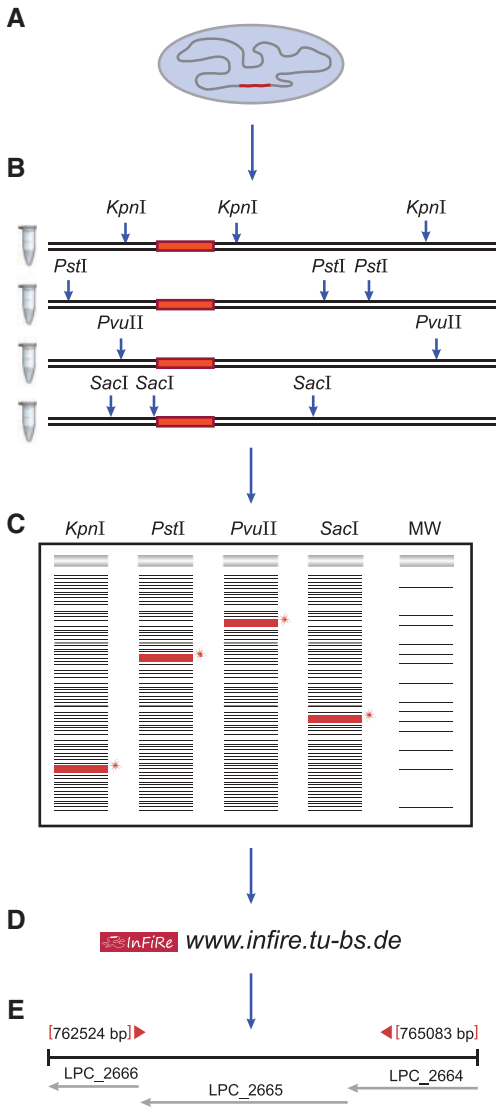

Fig. 1. Procedure for the identification of the transposon insertion in bacterial genomes. Chromosomal DNA from a transposon insertion mutant (**A**) is digested with a set of restriction enzymes (**B**). Each restriction digestion results in a unique pattern of DNA fragments. Separation of the DNA fragments by agarose gel electrophoresis followed by Southern blot hybridization with a transposon-specific probe allows determination of the approximate size of chromosomal fragments containing the transposon (**C**). The pattern of the derived fragment sizes allows the calculation of the most probable genomic position of the transposon via the InFiRe software (**D** and **E**).

2.3 Explicit model

The outlined explicit statistics approach calculates the number of digestions and shows the most suitable combination of restriction enzymes with an error probability. In the first step, the fragment size distribution for each digestion i is calculated for a given genome sequence. Fragment sizes x are distributed almost exponentially. Assuming a random and unbiased DNA sequence, the approximated distribution $P_1(x)$ of the fragment sizes is calculated by Equation (1) (Fig. 2).

$$P_1(X > x) \leftarrow -e^{-\frac{x}{m}}, \quad \forall x \geq 0 \tag{1}$$

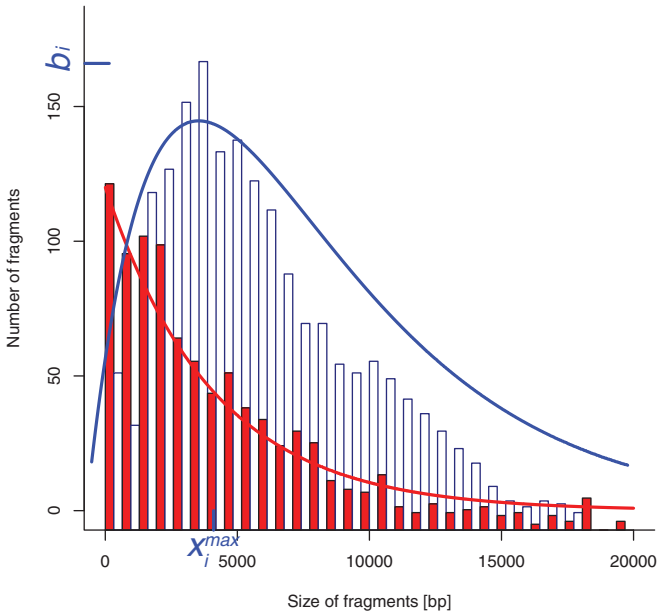

Fig. 2. Example for the DNA fragment sizes distribution using the *L.pneumophila* Corby chromosome and the restriction enzyme *Eco*RI (red bars). An observational accuracy of 20% results in an left-peaked distribution (blue bars). The maximum number of derived fragments is marked with b_i with a fragment size of $x_i^{\max}$. The course of the red line indicates the theoretical fragment size distribution [Equation (1)]. The blue line indicates the theoretical fragment size distribution with an 20% observational accuracy of the fragment sizes [Equation (2)].

The median of the fragment length m is calculated by $m := 4^l$, where l is the length of the occurring restriction site. Under consideration of an observational accuracy a of the fragment sizes derived from the Southern blot, the intervals of the fragment length distribution widens with increasing fragment sizes and results in an left-peaked distribution $P_2(x)$. This distribution is calculated by:

$$P_2(x) \leftarrow e^{-\frac{x-a\cdot x}{m}} - e^{-\frac{x+a\cdot x}{m}}, \quad \forall x \geq 0 \tag{2}$$

In the next step, the maximum achieved number of the fragments b_i at the size $x_i^{\max}$ is calculated for each digestion i.

$$b_i := \max\left(P_2^i(x)\right) = P_2^i\left(x_i^{\max}\right), \quad \forall x \geq 0 \tag{3}$$

In order to find the maximum number of required digestions, solely fragment sizes within the interval Q_i are considered since $x_i^{\max}$ represents the worst case scenario.

$$Q_i = \left[x_i^{\max} - a\cdot x_i^{\max}, x_i^{\max} + a\cdot x_i^{\max}\right] \tag{4}$$

Finally, for every i (ordered by b_i) all overlaps between the fragments in Q_j, with $j \in [1,2,\ldots,i]$ are computed until the number of overlaps is less than one. The value obtained for i is the number of required digestions.

The results of the statistics approach are scored by an error probability e. The error probability is computed in a combinatorics approach by the number of overlaps o, the number of fragments $f \in Q$ and the number of all fragments a of every digestion.

$$e(i) := \prod_{j=1}^{i-1}\left(\frac{a_j - o_j}{a_j}\right)^{f_{(j+1)}} \tag{5}$$

2.4 Theoretical model

Besides this explicit statistics, a theoretical statistics approach that estimates the number of required digestions by use of the genome size g and the

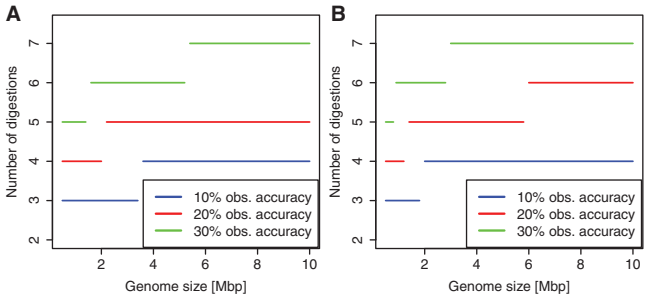

Fig. 3. Theoretical numbers of required restriction enzymes digestions for an InFiRe analysis. Diagram shows the number of required digestions for **(A)** single restriction enzyme reactions and **(B)** double restriction enzyme reactions. Both diagrams show the solution of the statistic approach with a length of restriction site $l=6$ at different observational accuracies (blue 10%, red 20% and green 30%).

observational accuracy was developed [Equations (1) and (2); Fig. 2]. The maximum number of the fragments b with size x and observational accuracy a is calculated by the theoretical distribution $P_2(x)$.

$$b := \max\left(P_2(x)\right), \quad \forall x \geq 0 \tag{6}$$

Taking the maximum number of fragments, it is possible to calculate the number of required digestions. With every digestion, the approximated number of possible genome position of the insertion z is calculated. This number z is inversely proportional to the number of restriction digests and has to be $z \leq 1$. The pseudocode below approximates the number of digestions:

`n(1)` is the approximated number of all used fragments.
`c(1)` is the ratio of fragment size b to the number of all fragments `n(1)`.
`l` is the length of the restriction site.
`k` is the number of restriction enzymes per digestion.

```
n(1) = k*(g/4^l);
c(1) = b/n(1);
i = 1;
while (n(i)*c(i)≥1) {
  i = i+1;
  c(i) = c(i-1)*c(1);
  n(i) = n(1)+n(i-1);
}
```

After the loop, `i` reflects the number of required digestions. It is possible to use more digestions. In Figure 3, the calculated number of required digestions is shown for single and double digestions at various observational accuracies.

The explicit and theoretic statistics approaches were compared by random simulation. This was performed in 360 runs by 10 randomly selected lists of restriction enzymes that were applied to 12 different prokaryotic genomes with observational accuracies of 10, 20 and 30% in each case. Hereby, in the explicit approach the maximum number of required digestions was in average slightly smaller than in the theoretical approach (mean value of difference 0.58). This was confirmed by a t-test with a P-value of $2.2 \cdot 10^{-16}$. The results demonstrate that in practice the theoretical approach is sufficient, since the number of digestions is only slightly overestimated.

3 RESULTS

3.1 InFiRe analysis workflow

The InFiRe web interface is structured into three steps: (i) selection of genome and replicons; (ii) selection of transposon; and (iii) digestion parameters (Fig. 4). These steps reflect the workflow of an InFiRe analysis and should be used in the given order. In the first step, selection of the genome and replicon, all organisms from sequenced bacterial genomes available at the NCBI can be chosen (Fig. 4A). One genome can consist of several replicons, e.g. chromosomes or plasmids. In the second step, the user selects the transposon (Fig. 4B). The software provides three possibilities to enter a transposon: selection from the list of transposons, entering a transposon sequence into the text field or uploading of a FASTA files. Since it is necessary for the InFiRe analysis that the applied enzymes do not cut inside the transposon sequence, the list of permitted restriction enzymes (Enzymes) is assigned after transposon selection. In the third step, the restriction enzymes applied for the digestions can be chosen (Fig. 4C). It should be taken into consideration that one digestion can be performed by

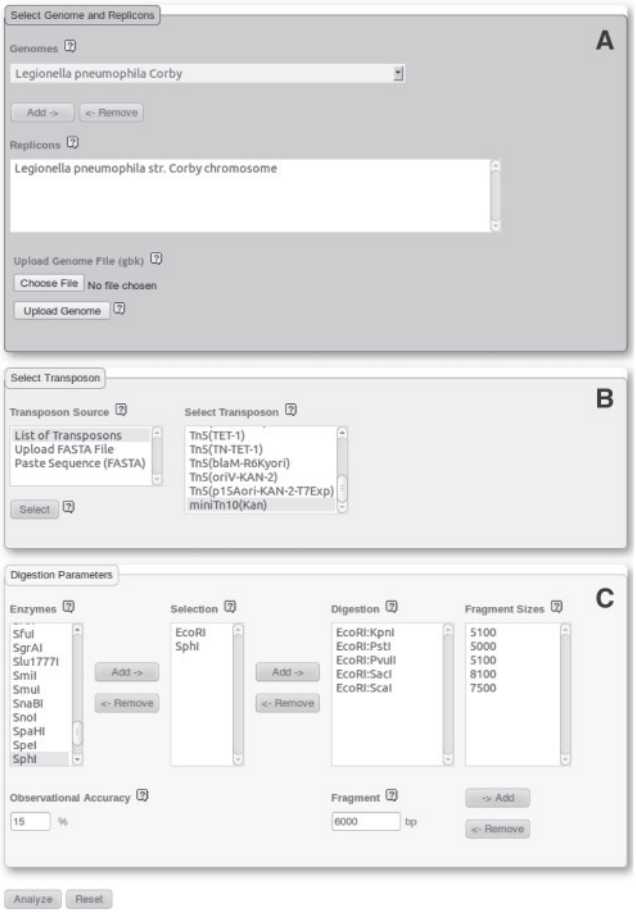

Fig. 4. Screenshot of InFiRe, accessible at www.infire.tu-bs.de/analyze. The use of the InFiRe web interface is divided into three steps: selection of genome and replicons of interest **(A)**, selection of the transposon **(B)** and choice of the digestion parameters **(C)**. Genome and (or) replicons are selected from a list. Transposons can be chosen from a list, by entering a transposon sequence into the text field or by uploading of a FASTA files. The restriction enzymes applied for the digestions can be chosen from the selection list 'Enzymes'. This list shows all restriction enzymes, which do not cut inside the transposon sequence. Finally, the approximated size of the hybridized fragments can be entered in the field 'Fragment Sizes' and analyzed.

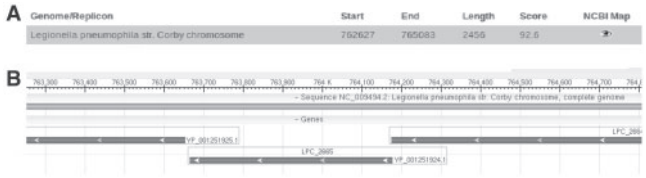

Fig. 5. Screenshot of the derived InFiRe results. **(A)** The list of potential targets was narrowed down to the one fragment. **(B)** Visualization of the obtained results in the NCBI Sequence Viewer.

utilization of multiple enzymes. The number of required restriction digestions and its dependence from observational accuracy of DNA fragment lengths is shown in Figure 3.

For determination of the DNA fragment sizes, Southern blot hybridization needs to be performed (Supplementary Materials). The approximated size of the hybridized fragments is entered in the corresponding field and analyzed.

The results of an InFiRe analysis contains the replicon name of the organism, start, end and the length of the restriction fragments (Fig. 5A). The matches are ordered by a score, which describes the percental accuracy between the estimated and the calculated fragment sizes and reflects the reliability of the prediction. Every location is linked to the NCBI map viewer that provides a visualization of the corresponding genome fragment and further genomic information about the sequence, genes and database links (Fig. 5B). For confirmation of transposon insertion in the predicted DNA fragment, the primer(s) can be designed with Primer-BLAST (Sayers *et al*., 2011). Verification of the transposon insertion can be performed by PCR using primer binding inside the transposon sequence and primer binding outside the predicted DNA fragment or two primers binding outside the predicted DNA fragment.

3.2 Application of the InFiRe software and experimental verification of the prediction

We successfully applied the algorithm in a case study using a *L.pneumophila* Corby mini-Tn*10* transposon library. After the screening for *L.pneumophila* mutants, which are attenuated in intracellular survival within host cells, we received a set of bacterial strains with unknown insertion sites (Shevchuk and Steinert, 2009). We applied the described protocol for the identification of insertion sites in the generated mutants. During the design of the experiment, we chose enzymes that do not cut inside the mini-Tn*10* transposon sequence. The number of restriction digestions was statistically calculated and resulted in five digestions. For every digestion, we used a combination of two restriction enzymes per digest. This adaptation of the method decreased the size of expected DNA fragments and resulted in an increase of the observational accuracy. The exact size of the fragments was determined by Southern blot hybridization (Southern, 2006, 1975). Figure 6 represents the obtained digestion pattern for one of the analyzed *L.pneumophila* Corby::Tn*10* mutants. The approximated size of hybridized fragments were analyzed by InFiRe and the most probable position of insertion was calculated. Accordingly, transposon insertion was found in *L.pneumophila* Corby chromosome between 762 524 bp and 765 083 bp. The PCR amplification and sequencing confirmed the insertion at position 763 194 bp (Fig. 5).

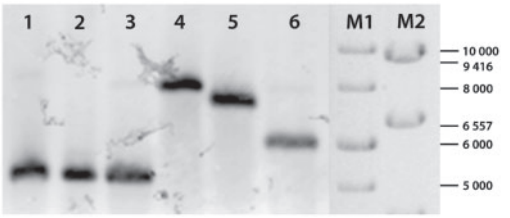

Fig. 6. Digestion pattern of *L.pneumophila* Corby::Tn*10* mutagenized DNA analyzed by Southern blot hybridization. The chromosomal DNA was digested by different combinations of restriction enzymes. Line 1: *Eco*RI/*Kpn*I-HF; line 2: *Eco*RI/*Pst*I; line 3: *Eco*RI/*Pvu*II-HF; line 4: *Eco*RI/*Sac*I-HF; line 5: *Eco*RI/*Sca*I-HF, line 6: *Eco*RI/*Sph*I-HF, M1-1kb DNA Ladder SM0311 (Fermentas), M2- Lambda DNA/*Hind*III Marker SM0101 (Fermentas). The blot was hybridized with a DIG-labeled mini-Tn*10* specific probe and visualized after instruction of the manufacturer.

4 DISCUSSION

In this study, we developed a new approach for the prediction of transposon insertion sites in transposon-mutagenized bacteria. InFiRe is based on a simple algorithm and allows the simultaneous determination of insertion sites by standard restriction digestion combined with Southern blot hybridization. The method is applicable to all sequenced organisms and therefore has a great potential to bring benefits in a wide range of applications.

In comparison to existing methods, InFiRe has several advantages. First, the method does not require intensive experimental optimization, therefore numerous insertion mutants can be analyzed simultaneously. Second, InFiRe overcomes the difficulties associated with the amplification and sequencing of GC-rich genomic fragments. Lastly, the method allows for the determination of exact position of a transposon insertion in any genome. This is especially important for identification of the transposons, which are integrated in repetitive sequences. Although long repeated sequences are carefully annotated in sequenced genomes, less attention was paid to investigate the biological role of these structures (Hahn, 2009; Hill, 1999; Romero *et al*., 1999). Thus, InFiRe may also open new ways to analyze the functions of repeated sequences in prokaryotic genomes.

ACKNOWLEDGEMENTS

We thank Stefan Leupold for his help during development of the software and Bernd Hoppe for financial management. We gratefully acknowledge the support of Barbara Schulz in critical proofreading of the manuscript. We are grateful to Stanislav Rosenblit for assistance in web design.

Funding: Deutsche Forschungs-gemeinschaft (DFG, grant SFBTR51); Bundesministerium für Bildung und Forschung (BMBF) medical infection genomics (grant no. 0315831A, 0315833A).

Conflict of Interest: none declared.

REFERENCES

Arnold,C. and Hodgson,I.J. (1991) Vectorette PCR: a novel approach to genomic walking. *PCR Methods Appl.*, **1**, 39–42.

Chaudhuri,R.R. *et al.* (2009) Comprehensive identification of *Salmonella enterica* serovar typhimurium genes required for infection of BALB/c mice. *PLoS Pathog.*, **5**, e1000529.

Hahn,M.W. (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. *J. Hered.*, **100**, 605–617.

Hayes,F. (2003) Transposon-based strategies for microbial functional genomics and proteomics. *Annu. Rev. Genet.*, **37**, 3–29.

Hensel,M. *et al.* (1995) Simultaneous identification of bacterial virulence genes by negative selection. *Science*, **269**, 400–403.

Hill,C.W. (1999) Large genomic sequence repetitions in bacteria: lessons from rRNA operons and Rhs elements. *Res. Microbiol.*, **150**, 665–674.

Hoffman,L.M. *et al.* (2000) Transposome insertional mutagenesis and direct sequencing of microbial genomes. *Genetica*, **108**, 19–24.

Judson,N. and Mekalanos,J.J. (2000) TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. *Nat. Biotechnol.*, **18**, 740–745.

Karlyshev,A.V. *et al.* (2000) Single-primer PCR procedure for rapid identification of transposon insertion sites. *Biotechniques*, **28**, 1078, 1080, 1082.

Largaespada,D.A. (2009) Transposon mutagenesis in mice. *Methods Mol. Biol.*, **530**, 379–390.

Ochman,H. *et al.* (1988) Genetic applications of an inverse polymerase chain reaction. *Genetics*, **120**, 621–623.

Opperman,T. *et al.* (2003) Microbial pathogen genomes - new strategies for identifying therapeutic and vaccine targets. *Expert. Opin. Ther. Targets*, **7**, 469–473.

Qimron,U. *et al.* (2003) Reliable determination of transposon insertion site in prokaryotes by direct sequencing. *J. Microbiol. Methods*, **54**, 137–140.

Romero,D. *et al.* (1999) Repeated sequences in bacterial chromosomes and plasmids: a glimpse from sequenced genomes. *Res. Microbiol.*, **150**, 735–743.

Sayers,E.W., et al. (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.

Shevchuk,O. and Steinert,M. (2009) Screening of virulence traits in *Legionella pneumophila* and analysis of the host susceptibility to infection by using the *Dictyostelium* host model system. *Methods Mol. Biol.*, **470**, 47–56.

Smith,V. *et al.* (1995) Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proc. Natl Acad. Sci. USA*, **92**, 6479–6483.

Southern,E. (2006) Southern blotting. *Nat. Protoc.*, **1**, 518–525.

Southern,E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.*, **98**, 503–517.

Vilen,H. *et al.* (2003) A direct transposon insertion tool for modification and functional analysis of viral genomes. *J. Virol.*, **77**, 123–134.

Yergeau,D.A. and Mead,P.E. (2009) Transposon-mediated transgenesis in the frog: New tools for biomedical and developmental studies. *Front. Biosci.*, **14**, 225–236.