

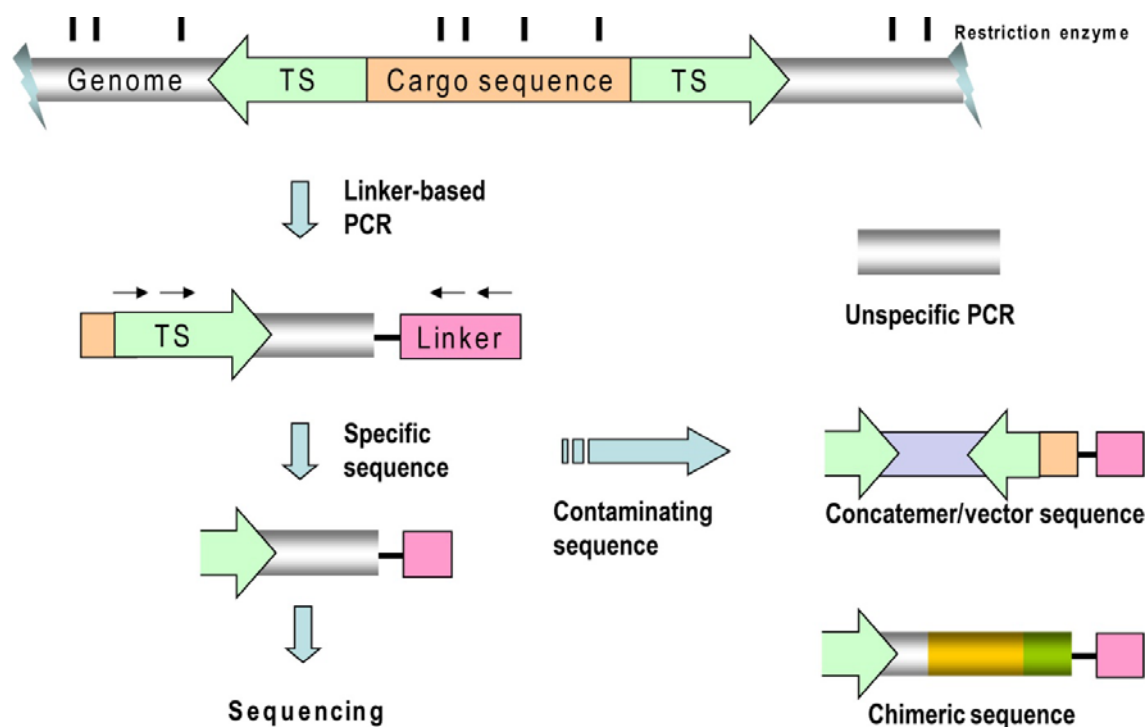
## **Chapter 2.      *iMapper*: A web server for the automated analysis and mapping of insertional mutagenesis sequence data against Ensembl genomes**

### **2.1 Introduction**

Large-scale insertional mutagenesis screening in mice or cultured cells is a rapid and efficient approach for gene discovery. Unlike other gene discovery approaches, the fact that insertional mutagens modify the genome directly, as opposed to overexpressing cDNAs from plasmids or viruses or knocking down genes by shRNA or siRNA transfection, means that they are capable of simultaneously informing us about the function of protein coding genes, regulatory regions, non-coding RNAs, miRNA or indeed any other element of the genome.

Retroviral-based insertional mutagenesis screens have been a valuable tool for the discovery of oncogenes and tumour suppressors in mice (106) and also for gene discovery in cultured cells (107). More recently transposon-based approaches such as the use of the *TcI*-family transposon *Sleeping Beauty* (88,89,108), the *Trichopulsia*-derived transposon *piggyBac* (109), and the *Tribolium castaneum*-derived *TcBuster* (110) have been developed increasing the repertoire of insertional mutagens available as gene discovery tools. To determine where in

the genome an insertional mutagen has inserted, the usual approach is to use a linker-based PCR method, such as vectorette or splinkerette (100). For any insertional mutagenesis screen to cover a significant proportion of the genome it is desirable to perform a screen using hundreds of mice and hundreds if not thousands of cell clones. Thus insertional mutagenesis screens may involve the generation and analysis of tens of thousands of DNA sequence reads from insertion sites. Although linker-based PCR methods are generally specific, non-specific PCR products, chimeric sequences and sequences derived from transposon concatemeric arrays can all represent contaminating sequences within pools of insertion site PCR products (**Figure 2-1**), thus without processing of sequence data the direct mapping of insertion site sequences to the genome may result in the identification of false-positive insertion sites. Therefore, insertion site sequences need to be processed prior to downstream mapping and analysis.



**Figure 2-1. Schematic diagram of linker-based PCR procedure and contaminating sequences**

To perform linker-based PCR, genomic DNA is first digested randomly with a frequent restriction cutter (average size of 300 – 500 bp). The genomic DNA fragments are first ligated with a linker sequence of various design and then subject to PCR amplification. The amplified DNA sequences containing insertional mutagens are subjected for sequencing and genome mapping. The contaminating sequences from linker-based PCR could be derived from non-specific PCR reactions, concatemer sequences from adjacent transposon arrays or chimeric sequences from inter-connection of genomic DNA during ligation step.

A major time-limiting step for insertional mutagenesis studies is the processing of tens of thousands of sequence reads in an accurate and efficient way, which includes: (1) identifying and eliminating the contaminating sequences (mainly concatemeric or chimeric sequences); (2) identifying the mutagen tag sequence in the sequencing reads to avoid any unspecific PCR products; (3) mapping the processed sequencing reads to the host genome to identify the insertion sites; and (4) obtaining a list of candidate genes overlapping with these insertion sites. Insertional mutagenesis screens are usually performed by experimental biologists who may not have the computing knowledge required to process large-scale sequencing reads, and therefore have to rely on collaborations with computational biologists who have the software and the computational skills to do large-scale genomic mapping. This has greatly limited the efficiency of insertional mutagenesis screens that have been carried out in laboratories lacking bioinformatic support. Although there are online genome browsers such as Ensembl or UCSC genome browser which enable the mapping individual genomic sequences, so far there has been no software or online web tools that could specifically facilitate the analysis and mapping of large numbers of sequencing reads generated from insertional mutagenesis screens. Therefore, developing a bioinformatics tool specifically for insertional mutagenesis sequence analysis could not only facilitate the analysis of insertional mutagenesis data for my PhD projects, but could also provide a solution for the community in better analyzing and processing their experiment results.

## **2.2 Aim and summary of the project**

The aim of this project is to develop a powerful but simple to use bioinformatics tool for insertional mutagenesis dataset analysis. The main features of this tool should include:

1. Efficient and accurate processing of insertion site sequence data and analysis against genomes of many model organisms, including human, mouse, rat, zebrafish, *Drosophila*, and *Saccharomyces cerevisiae* genomes.
2. Output of annotated sequence reads with links to genome browsers so that insertion sites can be viewed in the context of gene structures and other genomic features.
3. Output of processed sequence data in FASTA and GFF file format to allow insertion site sequence data to be analyzed in any sequence analysis package and could be displayed as a DAS track against the Ensembl genome.

4. Output a graphical chromosome *KaryoView* showing insertion sites against an ideogram of each chromosome.
5. Since this tool should be open to public access, ideally this tool is to be developed as an online web-based tool. The interface should be friendly and the operation should be easy to use by scientists with limited computer knowledge.

Based on these expectations, a web-based server called *iMapper* (Insertional Mutagenesis Mapping and Analysis Tool) was developed for the efficient analysis of insertion site sequence reads against vertebrate and invertebrate Ensembl genomes. Taking linker-based PCR sequence reads as input, *iMapper* first scans the sequence to identify a tag sequence (TS) derived from the end of the insertional mutagen using a local sequence alignment (LSA) algorithm. *iMapper* then scans the downstream sequence for user defined contaminating sequences, then processes the sequences to identify the restriction site sequence used for linker ligation during the insertion site PCR, clips out the genomic sequence between the tag sequence and first restriction enzyme cutting site and presents this sequence to a rapid mapping algorithm called SSAHA (102). Insertion sites can then be navigated in Ensembl in the context of other genomic features such as gene structures. *iMapper* also generates FASTA and GFF files of the clipped sequence reads and provides a graphical overview of the mapped insertion sites against a *karyotype*. *iMapper* is designed for high-throughput applications and can efficiently process tens of thousands of DNA sequence reads in a short time.

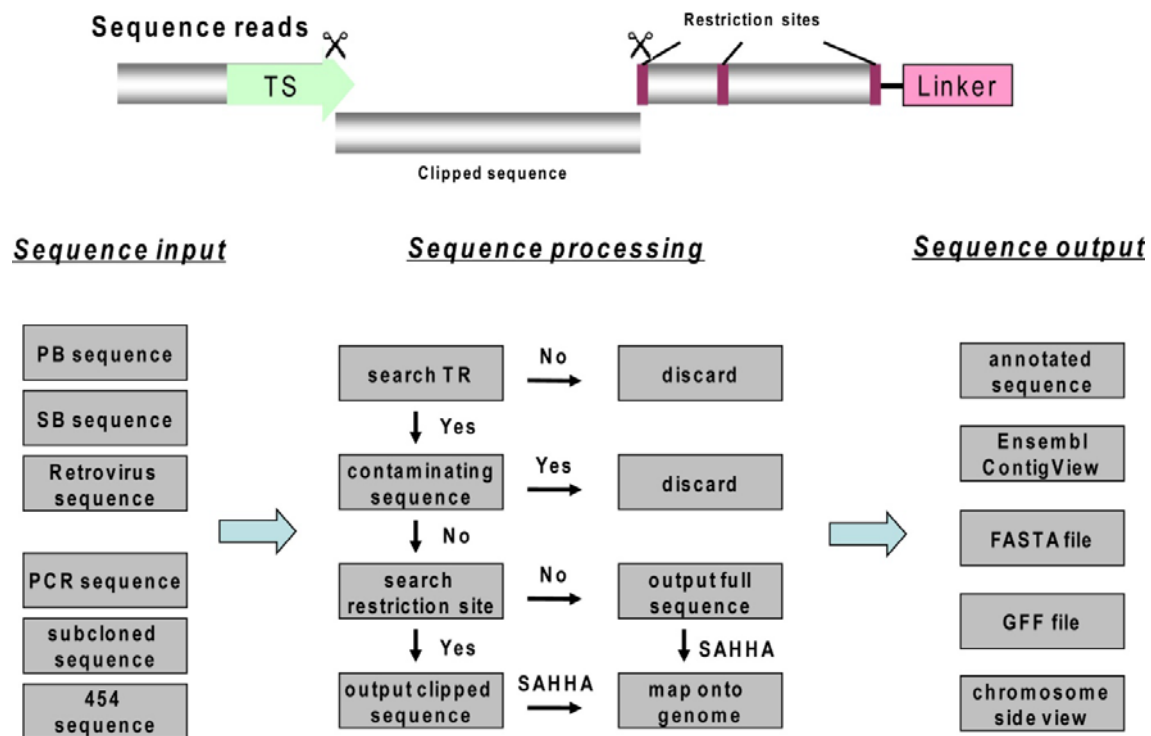
## 2.3 Materials and Methods

### 2.3.1 Architecture of the program

The *iMapper* interface is web-based (**Figure 2-3**) to accept sequence information and other parameters defined by the users. The sequence analysis module within *iMapper* is developed using Perl and CGI script to accept and process information passed from the interface. The processed sequence is then passed to a SAHHA server to map the sequence to an Ensembl genome. Processed sequence information including the tag sequence, genomic sequence and mapping positions are then fed back to the user's computer to generate an output webpage with processed sequence information. The output webpage also generates links to run against the Ensembl applications to generate additional output information such as the Ensembl *ContigView* and generation of a chromosome *KaryoView* graph of the insertion sites.

### 2.3.2 Sequence Processing

The procedure used by the code for sequence processing is shown in **Figure 2-2**. Taking linker-based PCR sequence reads as input, *iMapper* first scans the sequence to identify a tag sequence (TS) derived from the end of the insertional mutagen using a local sequence alignment (LSA) algorithm. *iMapper* then scans the downstream sequence for user defined contaminating sequences such as concatemeric sequences or vector sequences from the end of insertional mutagens, then processes the sequences to identify the restriction site sequence used for linker ligation during the insertion site PCR, clips out the genomic sequence between the tag sequence and first restriction enzyme cutting site and presents this sequence to a rapid mapping algorithm called SSAHA (102). The annotated sequence is then displayed on the output webpage with links to other features of the insertion sequence including: (1) navigation in the Ensembl *ContigView* to view other genomic features such as gene structures; (2) FASTA and GFF files of the clipped sequence reads; (3) a graphical overview of the mapped insertion sites against a *KaryoView* picture generated by Ensembl.



**Figure 2-2. The workflow of *iMapper***

*iMapper* is a sequence mapping and analysis tool designed to process insertion site sequence read data generated using ligation mediated PCR methods such as Vectorsite and Splinkerette. *iMapper* identifies a user-defined tag sequence and restriction enzyme site within a DNA sequence read and maps the intervening sequence to a user-defined Ensembl genome. The sequence input, processing and output formats of the software are shown.

### 2.3.3 Performance Test

To determine the optimal length of the mutagen tag sequence and optimal percentage threshold for sequence tag identification, a published dataset of 1920 *piggyBac* traces was chosen for analysis (109). The optimal length of the mutagen tag sequence was tested by fixing the percentage threshold to 80% and generating a graph of 'Accuracy' vs. 'Coverage' by increasing the tag sequence length for tag sequence identification using local sequence alignment algorithm. The optimal percentage threshold for sequence tag identification was tested by fixing the tag sequence length to 17 bp (PB tag sequence used: TATCTTTCTAGGGTTAA) and generating a graph of 'Accuracy' vs. 'Coverage' by increasing the percentage threshold for tag sequence identification. The value for 'Accuracy' was determined by the presence of *piggyBac* 'TTAA' signature sequence at the end of the tag sequences been identified, the value for 'Coverage' was calculated by using the number of sequences containing tag sequences divided by the number of total sequences.

### 2.3.4 Chromosome *KaryoView* graph

To display the chromosome side view graph of mapped insertion sites a *Sleeping Beauty* dataset containing 4032 sequence reads generated from mouse tumours using shot gun cloning was used for analysis. After *iMapper* processing, users can click on the 'Generate chromosome side view graph' link on the output page to enter the Ensembl *KaryoView* browser. In the next page, users click 'continue' to enter a configure page: here users can define the format and display to generate the side view graph. Users may then click the 'Finish' button and a *KaryoView* picture is generated. To generate a side view graph for multiple datasets, users can choose 'Add more data' in the first setup page and then copy and paste the GFF file generated from another dataset to the text box before following the same procedure to display the multiple datasets on one *KaryoView* picture.


## 2.4 Results

### 2.4.1 Program interface

The front end of *iMapper* is a user-accessible web interface generated using Perl and CGI (**Figure 2-3**), allowing users to submit their sequence data to the *iMapper* server in FASTA format or plain text. There are also different sections and options on the interface page where



users can define *iMapper* working parameters. After filling out the form, users can press 'Submit Query' to submit their sequence data to a back-end server and start mapping sequence reads using *iMapper*. More detailed descriptions on how to use *iMapper* can be found in the online help page.



**iMapper**  
Insertional Mutagenesis Mapper

[About](#)

iMapper is a sequence analysis tool designed specifically for large-scale analysis of insertional mutagenesis tag sequences against vertebrate and invertebrate genomes. It trims real genomic segments from linker-based PCR sequence input and automatically maps insertion sites onto an assembled genome. [Learn more...](#)

[Submit your sequence to iMapper](#)

[Hide advanced options](#)

Alignment criteria:

Alignment threshold:  %

Gap penalty:

Match score:

Mismatch score:

Contaminating sequences:

seq1:

seq2:

Alignment percentage:  %

Search in first  residues

Overlapping genes:

A gene overlaps a hit if it falls within the following window

5' flanking bp:

3' flanking bp:

SSAHA mapping parameters:

A unique mapping will be determined if SSAHA mapping score >

and

The score for best hit >  fold of other scores

Species:

Sequencing from:

Output option:

☒ Output all sequences

☐ Output good sequences (containing tag sequence)

Mutagen tag sequence:

Or choose from preset:

Restriction site:

Advanced options: [Show advanced options](#)

se upload your sequence file:

Or paste here in FASTA format:

**Figure 2-3. The interface of *iMapper***

The interface of *iMapper* is web-based with advanced options (smaller window on the left) expanded to illustrate the parameters that are adjustable within the web tool. More instructions on the usage of *iMapper* interface are described below.

#### **2.4.1.1 Basic setup window**

The basic setup window allows users to readily use *iMapper* to process their sequencing reads by submitting their sequence data and define a few simple working parameters. Sequence data is imported into *iMapper* in FASTA format. Sequence data in this format can either be pasted into a text box provided or imported using the file upload option. After sequence input a user can define the species against which they would like their insertion site data analysed from Human, mouse, rat, zebrafish, *Drosophila*, and *Saccharomyces cerevisiae*. The orientation of the tag in the sequence can then be chosen, and the output option selected. Selecting 'output good sequences' will exclude those sequences that do not contain the tag sequence, sequences that do not map to the genome and sequences that are identified to contain contaminating sequences. The sequence of the mutagen 'tag' sequence can then be specified, or a pre-validated tag sequence can be selected from the drop down menu. We provide tag sequences for *Sleeping Beauty* and *piggyBac* transposons, and for the U3 long terminal repeat (U3LTR) of the MuLV retrovirus. The sequence of the restriction site can then be specified. Alternatively the sequence of the linker could be entered. At this point the user has defined the boundaries of the sequence which will be mapped to the genome as the sequence between the tag and the restriction site or linker. These basic setups will allow users to work with their sequences submitted to *iMapper*.

#### **2.4.1.2 Advanced options**

In addition, advanced options for *iMapper* can be specified in the unfolded window 'Show Advanced Options'. These include the tag alignment parameters and the sequence of contaminating sequences in a tabular format. It is also possible to specify the parameters used by the SSAHA algorithm for matching the genomic sequence between the tag and the restriction site or linker, to the genome. These parameters can make a dramatic difference for SSAHA mapping especially when for analyzing short sequences such as those generated by 454/Roche sequencing. Finally, it is possible to specify the criteria for 'gene overlaps'. By specifying 'gene overlaps' it is possible to vary the spatial criteria for defining what constitutes a transposon insertion event in or near to a gene. For example it may be desirable to identify insertion events that mutate in or upstream of a gene, but not downstream.

## 2.4.2 Program Output

### 2.4.2.1 Output page for annotated sequence

After submitting the sequence data, *iMapper* generates a html-based output of the analyzed sequence data in tabular format (**Figure 2-4 A**). Sequences such as the tag sequence, the restriction site and genomic sequence in between are highlighted in this view. If a unique genomic mapping is identified, links to the Ensembl *ContigView* are provided to view the insertion site against Ensembl genome structures (**Figure 2-4 B**). In addition, *iMapper* also identifies if the insertion site overlaps with any Ensembl known genes (within 10kb up or downstream by default), and if so a link to gene pages is also provided. *iMapper* uses a real-time display algorithm and annotated sequences are streamed to the browser as they are generated, allowing users to view their results before whole sequences are processed. After analyzed all the input sequence data, *iMapper* summarises the results to show the total number sequences that have been analyzed and then number of sequences in each category, including how many reads contain the tag sequence, contaminating sequence and how many can be mapped to genome or overlap with any Ensembl known genes. On the bottom of the output page, *iMapper* provides three useful links to display additional information. First is a link to generate clipped genomic sequences in FASTA format for all the processed reads containing tag sequences. The second link is for generating a GFF file for all the sequence reads that can be mapped onto a genome, including genomic locations, length, whether the sequence overlaps with a gene and the gene name. The gene feature list is recognised by Excel which can be used for further data processing and sorting. The third link will generate a chromosome side view graph for all the mapped sequences using Ensembl *KaryoView* browser (**Figure 2-4 C**).



**Figure 2-4. The output of *iMapper***

(A) Example of the tabular output for one sequence read processed by *iMapper*. The sequence is annotated in different colours representing tag sequence (green), clipped genomic sequence (yellow) and restriction site (brown). Other information such as the position of the tag sequence and restriction site, the location on the genome where this sequence read maps to, and the links to Ensembl *ContigView* and gene information page are all provided. (B) Detailed view of an insertion site in relation to gene structure in Ensembl *ContigView*. The red bar represents the position of the insertion site. (C) *KaryoView* picture of insertion sites generated by *iMapper* via Ensembl *KaryoView*.

#### **2.4.2.2 FASTA and GFF formats**

When the analysis run is complete, links to a FASTA file of the processed traces and a GFF file of the data are provided on the bottom of the output page. The link to the FASTA file generates clipped genomic sequences between the tag sequence and the first restriction site, or the full length sequence after the tag sequence if no restriction site can be identified. These processed genomic sequences are particularly useful for analysis by other sequence processing packages or mapping tools. The second link generates a GFF file for all the sequence reads which can be mapped onto an Ensembl genome. GFF is a file format used for describing genes and other features associated with DNA sequences such as the genomic location, length, whether the sequence overlaps with a gene and the gene names. More information on GFF files can be found at <http://www.sanger.ac.uk/resources/software/gff/>. This gene feature list is recognised by Microsoft Excel which can be used for further data processing and sorting. Furthermore, the GFF file format can be uploaded and displayed as a DAS track against an Ensembl genome.

#### **2.4.2.3 KaryoView graph**

To obtain a global overview of the sequence data *iMapper* has a link to an Ensembl *KaryoView* to generate a side view graph of the data against a chromosomal ideogram. This graph gives a clear and direct view of the distribution of insertion sites on different chromosomes in the genome. For instructions on how to generate a chromosome side view graph such as in **Figure 2-4 C** please refer to section 2.2 Materials and Methods.

### **2.4.3 Performance of *iMapper***

#### **2.4.3.1 Determining the optimal working parameters for *iMapper***

*iMapper* uses the Smith-Waterman local sequence alignment (111) algorithm to identify the mutagen tag sequences. The specificity of tag identification depends on the length of the tag sequence entered, and the pre-defined thresholds specified for sequence tag identification including the percentage alignment threshold, gap penalty, match and mis-match score. Longer tag sequence inputs, higher alignment percentages and more stringent gap and mis-match scores will result in more accurate tag sequence matching. We have tested the optimal tag sequence length and percentage threshold using a dataset of 1920 *piggyBac* insertion site sequence reads (109). Because *piggyBac* integrations invariably occur at 'TTAA' sites a

precisely identified tag sequence will always be preceded by the sequence TTAA. As shown in **Figure 2-5 A** the minimal advisable tag sequence length is 15 bp. Next we determined the optimal percentage threshold for sequence tag identification to be used as the default, and determined this to be 80 % (**Figure 2-5 B**). Finally we optimized the SSAHA sequence parameters to be used as the default and found that for sequences from splinkerette PCR reactions that contain genomic junction fragments of, on average, 200 bp in length, the optimal SSAHA score is 35. This score should be ideal for insertion site sequences generated by capillary read sequencing but may need to be lowered to 20 for shorter reads such as those generated by 454 sequencing. It is advisable to optimize the SSAHA mapping score for each dataset and select a score that generates the highest number of uniquely mapped reads. This is important because the default mapping parameters used by *iMapper* are stringent and will return only those reads that map to unique, unambiguous genomic locations.



**Figure 2-5. Performance of *iMapper*: analysis of 1920 *piggyBac* traces**

(A) Determination of the optimal length of the mutagen tag sequence. (B) Determination of the optimal percentage threshold for sequence tag identification. The 'Accuracy' is determined by the presence of the *piggyBac* 'TTAA' signature sequence at the end of the tag sequence that has been identified, the 'Coverage' is calculated by using the number of sequences containing tag sequences divided by the total number of sequences. (C) The data analyzed by *iMapper* using optimized parameters (analyzed using tag sequence TATCTTTCTAGGGTTAA (17 bp), percentage threshold = 80 %).

### 2.4.3.2 Timing and capacity

*iMapper* is a highly efficient program for the annotation of large numbers of insertion sites when compared with manual annotation. Generally, it takes the program less than half a second to process each sequence read and another second to map the sequences using the SSAHA server. Parameters that affect *iMapper*'s processing time include the input tag sequence length, whether it is used to search for tag sequences in a defined orientation or whether to exclude contaminating sequences. **Table 2-1** lists the time required for *iMapper* to process two different datasets using a variety of settings. The capacity of *iMapper* to process sequence reads in one submission depends on the web browsers capacity and computer memory. We have analysed over 10,000 traces in one submission using a Mac OS system or PC. However to guarantee the stability of the operating system, we have limited the submission scale to 10,000. More sequence traces can be processed in multiple batches or using a command version *iMapper* which is available upon request.

**Table 2-1. Time required for *iMapper* to analyze two datasets with different settings.**

	Tag length	Define orientation?	Contaminating sequence	SAHHA mapping	Time
582 PB sequences	17	Yes	-	No	2min30s
	17	Yes	-	Yes	8min50s
	17	No	-	Yes	11min25s
4032 SB sequences	17	No	-	No	32min
	17	No	-	Yes	1h35min
	17	No	2	Yes	1h55min

### 2.4.3.3 Sequence analysis of real insertion site data using *iMapper*

We used the optimal length for a *Sleeping Beauty* (SB) tag sequence (‘TTCCGACTTCAACTGTA’, 17 bp) and the optimal percentage threshold (80 %) to test a *Sleeping Beauty* dataset containing 4032 sequence reads generated from mouse tumours using shotgun cloning (**Figure 2-5 C**). After *iMapper* processing, SB tag sequences were identified in 3576 sequence reads (89 %), of which nearly half of the reads (1531, 44 %) could be mapped to the Ensembl mouse genome using SSAHA (mapping score = 35), and two-thirds (1020 reads) of which overlapped to within 10 kb of a known Ensembl gene. An additional 128 reads could be mapped to the genome when using a score of 20 for SSAHA mapping rather than the default score of 35. The majority of these reads were short sequences of 25-45 bp, indicating that a smaller SSAHA mapping score could potentially benefit short sequence analysis such as those derived from 454 or Illumina-Solexa sequencing. These results demonstrate that *iMapper* is extremely efficient for the analysis of large-scale insertion data large-scale and provides a useful overview of the insertion sites as well as detailed information about each insertion site identified.

## 2.5 Discussion

Large-scale insertional mutagenesis screens require thousands, if not tens of thousands of sequencing reads to be processed and mapped in an accurate and efficient way. This has always been a time-limiting step for insertional mutagenesis experiments even with the extensive involvement of computational biologists. The fact that sequencing reads that are generated by linker-based PCR need to be carefully processed to get rid of contaminating sequences, chimeric sequences and tag sequences before sending the real genomic fragments for mapping, have complicated the analysis of insertion site sequence data. Despite the great power and popularity of insertional mutagenesis as a tool for screening candidate genes in model organisms and cell culture systems, there is no tailor-made software to facilitate the analysis and processing of insertion site data. As a result, each lab has developed their own approach and methods for analysing insertion site sequence data using different software, mapping algorithms and parameters. This has complicated the comparison and sharing of results.

As a freely accessible web tool, *iMapper* provides a simple solution for the processing of insertional mutagenesis data by experimental biologists who do not have a computational

background. *iMapper* also provides the user with the possibility to optimise the analysis of their data by defining some useful parameters during sequence processing and mapping, such as the length of mutagen tag sequence, the parameters used for local sequence alignment and SSAHA mapping. The quick and easy-to-use *iMapper* software could also provide different labs from around the world with a standard solution for insertional mutagenesis data processing, making it possible to compare and share mutagenesis screen results. Furthermore, by working closely with the Ensembl genome browser, *iMapper* also provides additional information and options for downstream analysis, such as viewing the genomic structure surrounding the insertion site, generating a *KaryoView* picture of the insertion sites and obtaining information on the insertion site gene. The fact that *iMapper* is maintained by the Wellcome Trust Sanger Institute, one of the largest bioinformatics centres in Europe will provide regular updates on sequence information and technical support.

One of the most obvious advantages for *iMapper* is the speed at which sequences are analyzed. It takes the software approximately 1.5 seconds to process and analyze each sequence read, the majority of which is spent on mapping the clipped genomic sequence using SSAHA (see results section). However, different setup parameters also affect the processing speed, for example, using longer tag sequences would result in a larger matrix in the local sequence alignment which, could slow down the speed for tag sequence identification, whereas searching for the tag sequence in a single orientation could speed up data processing since otherwise the program would search both the forward and reverse sequences. Other setups such as searching for contaminating sequence could also affect the speed for sequence processing using *iMapper*.

*iMapper* is a convenient tool with many useful functions for insertion site data analysis. The output page of *iMapper* lists the annotated sequences in a user-friendly tabular format, where detailed information and a link to Ensembl *ContigView* page are provided for each mapped sequence. *iMapper* also summarises the analysed data for downstream analysis, for example, *iMapper* is able to generate a *KaryoView* picture of all the mapped insertion sites which could give a direct overview of the results. In addition, *iMapper* generates clipped genomic sequences in FASTA format which could be used for genomic mapping by other softwares. The GFF file for all the mapped sequencing reads generated by *iMapper* provides useful information for further analysis of the insertion site data and can be uploaded as a DAS track against the Ensembl genome.

In *iMapper*, only sequences containing the tag sequence are processed for analysis and mapping on to the genome. Therefore the parameters for tag sequence identification are critical for the performance of *iMapper*. Following a performance test using different tag sequence lengths and percentage thresholds, the optimal length for tag sequence identification was found to be  $\geq 15$  bp and the percentage threshold,  $\geq 80$  % (**Figure 2-4 A and B**). These optimal values were determined when the accuracy and coverage curves reached a plateau, indicating that tag sequence identification was saturated and that further increasing the value would not result in a dramatic improvement in accuracy or coverage. To obtain sequencing reads that are at least 15 bp in length at the tag sequence end, the PCR primers and the sequencing primers should be more than 15 bp away from the end of the tag sequence. Appendix A lists the primer sequences that are recommended for splinkerette PCR on *piggyBac* and Sleeping Beauty samples. The same principles should be followed when designing PCR primers for splinkerette PCR on retrovirus samples. In addition, the parameters for tag sequence identification and genomic sequence mapping should also be determined using real sequence data. For example, if the sequencing is of poor quality or experimental noise is introduced to the tag sequence region, which is at the end of the sequencing reads, it is recommended that a longer tag sequence is used and a lower percentage threshold for tag sequence identification. Furthermore, if the sequence data are predominantly short sequences from 454 or Illumina-Solexa sequencing, it is recommended that a smaller SSAHA mapping score is used, which favours the identification of shorter sequences between 25-45 bp.

In conclusion, *iMapper* is designed for high-throughput applications and can efficiently process tens of thousands of DNA sequence reads in a short time. The web-based design of *iMapper* allows world-wide, free access to this software and the friendly user interface and operation functions allow users without a bioinformatics background to easily use *iMapper* to process their insertion site data and to change essential parameters to optimise their experimental results. In addition, *iMapper* provides each annotated sequence in a tabular format and links the mapped insertion sites to the Ensembl genome browser for further downstream analysis.

Availability: *iMapper* is web based and can be accessed at <http://www.sanger.ac.uk/cgi-bin/teams/team113/iMapper.cgi>. The code and algorithms are also available from this website.