BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 24 2008, pages 2923–2925
doi:10.1093/bioinformatics/btn541

Sequence analysis

iMapper: a web application for the automated analysis and mapping of insertional mutagenesis sequence data against Ensembl genomes

Jun Kong, Fei Zhu, Jim Stalker and David J. Adams*

Experimental Cancer Genetics, The Wellcome Trust Sanger Institute, Hinxton, Cambs, CB10 1HH, UK

Received on September 1, 2008; revised on September 30, 2008; accepted on October 16, 2008

Advance Access publication October 30, 2008

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: Insertional mutagenesis is a powerful method for gene discovery. To identify the location of insertion sites in the genome linker based polymerase chain reaction (PCR) methods (such as splinkerette-PCR) may be employed. We have developed a web application called *iMapper* (Insertional Mutagenesis Mapping and Analysis Tool) for the efficient analysis of insertion site sequence reads against vertebrate and invertebrate Ensembl genomes. Taking linker based sequences as input, *iMapper* scans and trims the sequence to remove the linker and sequences derived from the insertional mutagen. The software then identifies and removes contaminating sequences derived from chimeric genomic fragments, vector or the transposon concatamer and then presents the clipped sequence reads to a sequence mapping server which aligns them to an Ensembl genome. Insertion sites can then be navigated in Ensembl in the context of genomic features such as gene structures. *iMapper* also generates test-based format for nucleic acid or protein sequences (FASTA) and generic file format (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 thousands of DNA sequence reads.

Availability: *iMapper* is web based and can be accessed at http://www.sanger.ac.uk/cgi-bin/teams/team113/imapper.cgi.

Contact: da1@sanger.ac.uk; iMapper@sanger.ac.uk

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

1 INTRODUCTION

Retroviral-based insertional mutagenesis screens in mice have been a valuable tool for the discovery of oncogenes and tumor suppressors in mice (Mikkers and Berns, 2003), and also for gene discovery in cultured cells (Du *et al*., 2005). More recently transposon-based approaches such as the use of the *Tc1*-family transposon *Sleeping Beauty* (Collier *et al*., 2005; Dupuy *et al*., 2005) and the *Trichoplusia*-derived transposon *Piggybac* (Wang *et al*., 2008) have been developed increasing the repertoire of insertional mutagens available as gene discovery tools in mammals. To determine where in the genome an insertional mutagen has inserted the usual approach is to use a linker based polymerase chain reaction (PCR) method, such as vectorette or splinkerette (Devon *et al*., 1995). For any insertional mutagenesis screen to cover a significant proportion of the genome, it is desirable to perform a screen using hundred 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 represents contaminating sequences within pools of insertion site PCR products. Thus without careful processing of DNA sequence data, the direct mapping of insertion site sequence reads to the genome may result in the identification of false-positive insertion sites. To facilitate the analysis of linker mediated insertion site sequences, we have developed a web application called *iMapper* (Insertional Mutagenesis Mapping and Analysis Tool). Using linker based PCR sequence reads as input *iMapper* uses a local sequence alignment algorithm to identify a tag sequence derived from the end of the insertional mutagen (Supplementary Material). *iMapper* then scans the downstream sequence for user defined contaminating sequences, processes the sequence to identify the restriction site sequence used for linker ligation during the insertion site PCR, clips out the genomic sequence between the tag and first restriction enzyme cutting site and presents this sequence to a rapid mapping algorithm called sequence search and alignment by hashing algorithm (SSAHA) (Ning *et al*., 2001). Output is then generated in various formats. The main features of *iMapper* include:

(1) Efficient and accurate processing of insertion site sequence data and analysis against Ensembl human, mouse, rat, zebrafish, *Drosophila* and *Saccharomyces cerevisiae* genomes.

(2) Output of annotated sequence reads in tabular format with links to Ensembl *ContigView* so that insertion sites can be viewed in the context of gene structures and other genomic features.

(3) Output of processed sequence data in test-based format for nucleic acid or protein sequences (FASTA) and generic file format (GFF) allowing insertion site sequence data to be analyzed in any sequence analysis package and displayed as a distributed annotation system (DAS) track against an Ensembl genome.

*To whom correspondence should be addressed.

(4) Output of a graphical chromosome '*KaryoView*' showing insertion sites against an ideogram of each chromosome.

2 METHODS

2.1 Architecture

The *iMapper* interface is web-based (Supplementary Material). The sequence analysis module within *iMapper* uses Perl and computer generated imagery (CGI), and for comparison to an Ensembl genome Perl scripts run against the Ensembl application programming interface (Ensembl-API).

2.2 Input format

Sequence is imported into *iMapper* in FASTA format. Sequence data in this format can either be pasted into the text box or imported using the file upload option (Supplementary Material).

2.3 User defined parameters

After sequence input, a user can define the species against which they would like their insertion site data analyzed from human, mouse, rat, zebrafish, *Drosophila* and *S. 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 from all output formats. The sequence of the mutagen 'tag' sequence can then be specified, or a prevalidated tag sequence can be selected from the drop down menu. We provide tag sequences for the transposons *Sleeping Beauty* and *Piggybac*, and for the U3LTR of the MuLV retrovirus. The sequence of the restriction site can then be specified. 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. Advanced options can then be specified. These include the tag alignment parameters and the sequence of contaminating sequences, which will be highlighted in the tabular format (Supplementary Material). 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. 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 an insertion event in or near a gene. For example, it may be desirable to identify insertion events that mutate in or upstream of a gene, but not downstream. Sequence processing commences when the 'Submit Query' button is selected.

2.4 Sequence processing

The procedure used by the code for sequence processing is shown in the Supplementary Material.

2.5 Tabular format

The tabular format is an html-based output of the analyzed sequence data (Fig. 1). Sequences such as the tag sequence, the restriction site and contaminating sequence are highlighted in this view. Links to Ensembl gene pages and to the Ensembl *ContigView* are also provided from this page.

2.6 FASTA and GFF formats

A FASTA file of the processed traces and a GFF file of the data are provided when the analysis run is complete.

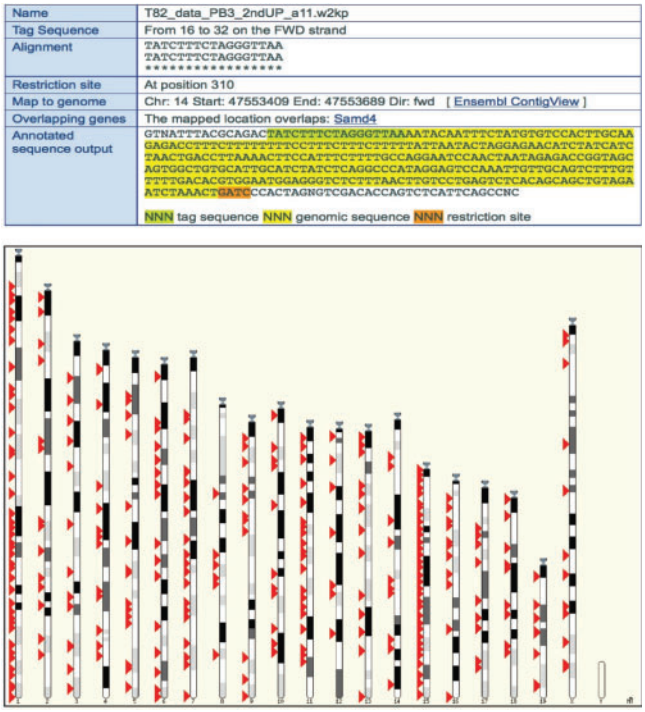

Fig. 1. *iMapper* maps and analyzes insertion site sequence data. *iMapper* generates output in several formats including a tabular format with links to Ensembl *ContigView* and Ensembl gene pages (upper panel), GFF file format, FASTA format and a *KaryoView* format (lower panel) which provides a global overview of all insertion site sequences that were mapped to the genome. Sequence traces displayed in the tabular format are annotated to show the location of the tag sequence (green), the restriction enzyme site (orange) and the mapped genomic sequence (yellow). An alignment of the tag sequence from the trace and the insertion site mapping location is also shown in the tabular format. In the *KaryoView* format, each red triangle indicates an independent insertion site displayed against a genome ideogram.

2.7 *KaryoView*

To obtain a global overview of the sequence data, *iMapper* has a link to an Ensembl *KaryoView* providing a graphical view of the data against a chromosomal ideogram (Fig. 1).

2.8 Performance

The specificity of tag identification depends on the length of the tag sequence entered, and the predefined thresholds specified for sequence tag identification including the percentage alignment threshold, gap penalty, match and mismatch score. Longer tag sequences, higher alignment percentages and more stringent gap and mismatch scores will result in more accurate tag sequence identification. We have tested the optimal tag sequence length and percentage threshold using a dataset of 1920 *PiggyBac* insertion site sequence reads (Wang *et al.*, 2008). Because *PiggyBac* integrations invariably occur at 'TTAA' sites a precisely identified tag sequence will always be followed by the sequence TTAA. As shown in the Supplementary Material, the minimal advisable tag sequence length is 15 bp. We determined the optimal percentage threshold for sequence tag identification, to be used as the default, and determined this to be 80% (Supplementary Material). Finally, we optimized the SSAHA sequence mapping parameters to be used as the default finding that for sequences from splinkerette-PCR reactions

containing 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 selecting 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.

We have used *iMapper* to analyze up to 20 000 DNA sequence traces. It takes, on average, 1–2 s for *iMapper* to analyze each DNA sequence trace (Supplementary Material) and to return the analyzed data in tabular, *ContigView*, GFF, FASTA and *karyoView* formats.

3 SUMMARY

iMapper is a web-based freely accessible solution for the analysis of insertional mutagenesis datasets and should facilitate the many insertional mutagenesis screens that are ongoing worldwide.

Funding: Cancer Research-UK (C20510/A6997) and the Wellcome Trust (76943 to D.J.A.); Wellcome Trust Sanger Institute PhD programme (to J.K. and F.Z.).

Conflict of Interest: none declared.

REFERENCES

Collier,L.S. *et al*. (2005) Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature*, **436**, 272–276.

Devon,R.S. *et al*. (1995) Splinkerettes–improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Res.*, **23**, 1644–1645.

Du,Y. *et al*. (2005) Insertional mutagenesis identifies genes that promote the immortalization of primary bone marrow progenitor cells. *Blood*, **106**, 3932–3939.

Dupuy,A.J. *et al*. (2005) Mammalian mutagenesis using a highly mobile somatic sleeping beauty transposon system. *Nature*, **436**, 221–226.

Mikkers,H. and Berns,A. (2003) Retroviral insertional mutagenesis: tagging cancer pathways. *Adv. Cancer Res.*, **88**, 53–99.

Ning,Z. *et al*. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.

Wang,W. *et al*. (2008) Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **105**, 9290–9295.