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 Tcl-family transposon Sleeping Beauty (Collier et al., 2005; Dupuy et al., 2005) and the Trichoplusia-derived transposon Piggybac (Wang et al., 2008) have been developing 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 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. Thus without careful processing of DNA sequence data, the direct mapping of insertional 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 insertional site PCR, clips 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.

© 2008 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
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 integration sites (Wang et al., 2008). Because PiggyBac integrations invariably occur at ‘TTAA’ sites a precisely identified tag sequence will 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...
A web application for the automated analysis

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