BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 5 2011, pages 720–722
doi:10.1093/bioinformatics/btq722

Sequence analysis
Advance Access publication January 17, 2011

Identifying viral integration sites using SeqMap 2.0

Troy B. Hawkins[1,*], Jessica Dantzer[2], Brandon Peters[2], Mary Dinauer[1,3], Keithanne Mockaitis[4], Sean Mooney[5] and Kenneth Cornetta[1]

[1]Department of Medical and Molecular Genetics, [2]Center for Computational Biology and Bioinformatics, [3]Department of Pediatrics, Herman B Wells Center for Pediatric Research, Indiana University School of Medicine, Indianapolis, IN 46202, [4]Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405 and [5]The Buck Institute for Age Research, Novato, CA 94945, USA

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: Retroviral integration has been implicated in several biomedical applications, including identification of cancer-associated genes and malignant transformation in gene therapy clinical trials. We introduce an efficient and scalable method for fast identification of viral vector integration sites from long read high-throughput sequencing. Individual sequence reads are masked to remove non-genomic sequence, aligned to the host genome and assembled into contiguous fragments used to pinpoint the position of integration.

Availability and Implementation: The method is implemented in a publicly accessible web server platform, SeqMap 2.0, containing analysis tools and both private and shared lab workspaces that facilitate collaboration among researchers. Available at http://seqmap.compbio.iupui.edu/.

Contact: troyhawk@iupui.edu

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

Received on July 30, 2010; revised on December 1, 2010; accepted on December 2, 2010

Retroviruses were first characterized by their ability to cause malignancy. Subsequently, retroviruses were identified that lacked oncogenes but mediated malignancy through a process termed insertional mutagenesis (IM). The molecular mechanisms of IM are varied but most commonly involve upregulation of cellular oncogenes in close proximity to the site of viral integration via *cis*- and *trans*-effects of promoter and enhancer sequences within the viral long terminal repeats (LTRs).

Because of IM effects, the mapping of retroviral integration sites (RISs) has become a powerful tool for identifying cellular oncogenes. Copeland and Jenkins (Buchberg *et al*., 1990; Copeland and Jenkins, 1990) used retroviruses to identify potential oncogenes by determining the site of viral integration in tumor tissues. This work led to the development of a database of cancer-associated genes (Akagi *et al*., 2004).

IM has also been associated with malignancy in the setting of human gene therapy applications. While most gene therapy trials have not been associated with the development of cancer, a notable exception was the treatment of X-linked Severe Combined Immuno-Deficiency (SCID-X1), where several patients developed a T-cell leukemia associated with vector integration near the proto-oncogenes *LMO2*, *BMI1* and *CCND2* (Hacein-Bey-Abina *et al*., 2003, 2008). The US Food and Drug Administration (FDA) now requires assessment of RISs for any human gene therapy trials utilizing integrating vector systems (USDHHS, 2006).

In animal models and human clinical trials, retroviral transduction targets millions of cells. As integration can occur throughout most of the genome, the resulting cell populations can contain extremely large, but unknown, numbers of RISs. Initial methods to identify the RISs utilized PCR-based capture and amplification assays that were inefficient and highly labor intensive. High-throughput next-generation sequencing technologies have facilitated much more efficient identification of RISs, which presents a new bioinformatics challenge.

We (Peters *et al*., 2008) and others (Appelt *et al*., 2009; Giordano *et al*., 2007) had previously developed web-based bioinformatics tools that can facilitate identification of RISs by mapping sequence data obtained from Sanger sequencing technology, but the tools are not sufficient to quickly map and characterize RISs in high-throughput methods. Here we introduce and explain our new methodology for quickly mapping RISs to a reference genome from extremely large datasets.

Depending on the frequency of insertion sites within the cell population, and the number of samples run in parallel, there can be anywhere from 50 to 5000-fold coverage of an individual RIS within the reads generated from a single sequencing run. SeqMap 2.0 provides a scalable method for sequence matching, clustering and alignment, and also addresses challenges specific to 454 pyrosequencing data output, namely base stutter and redundant coverage of each RIS.

The SeqMap 2.0 workflow has three stages: (i) sequence processing, including identification and masking of vector features and distribution of sequence reads into multiplex identifier (MID)/barcode-specific groups; (ii) sequence clustering and alignment; and (iii) data visualization and storage for further analysis (Supplementary Fig. 1B).

SeqMap 2.0 is able to analyze data from the major PCR techniques used in RIS analysis: ligase-mediated PCR (LM-PCR) (Smith, 1992), linear-amplification-mediated PCR (LAM-PCR) (Schmidt *et al*., 2003, 2007) and non-restrictive LAM-PCR (nrLAM-PCR) (Gabriel *et al*., 2009); see Supplementary Material. Each individual sequence read input to SeqMap 2.0 originates from an amplicon with common features. From 5′ to 3′ is a sequencing adaptor, a nucleotide bar code, viral LTR, RIS-flanking genomic sequence,

*To whom correspondence should be addressed.

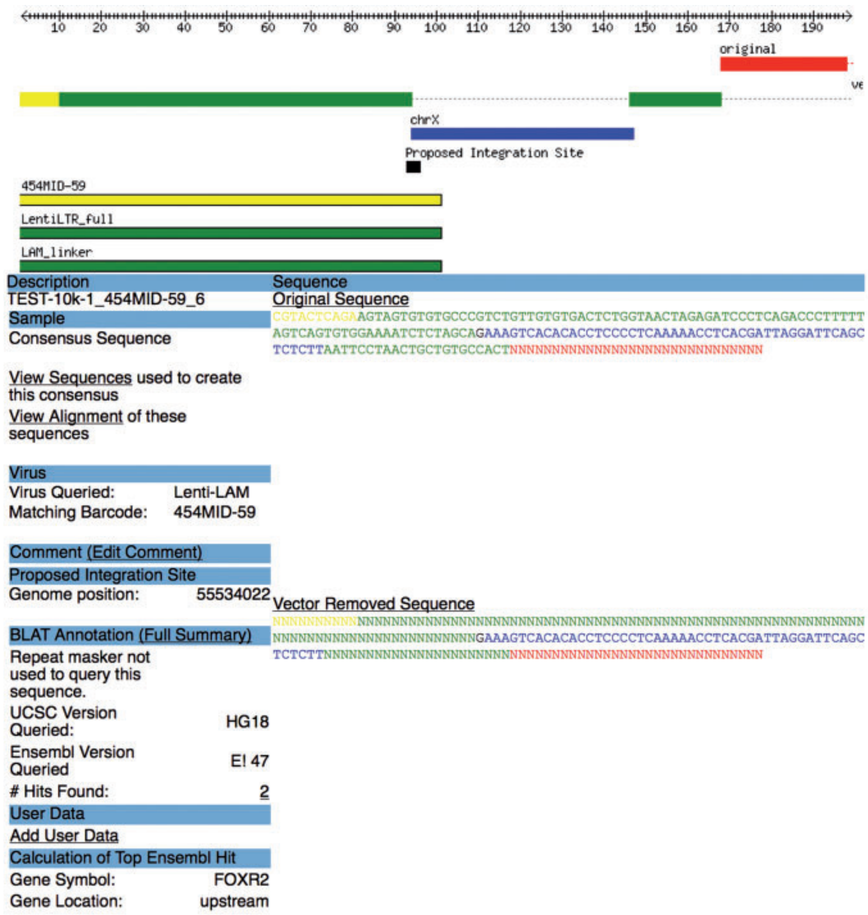

Fig. 1. Graphical representation of mapped integration site in sequence viewer. The consensus sequence for a cluster is shown with glyphs for bar code (yellow), vector feature (green) and genomic alignment (blue) at the top of the page, color-coordinated to the sequences shown below at the right. A specific integration site is proposed (black) when the position flanking the user-defined LTR feature aligns to the genome. Details for the integration are shown at the left, including links to a list and MSA of reads contributing to the consensus sequence for the RIS, and details of the genomic alignment linked to the Entrez entry for the closest identified gene. Users can access expanded graphics of local genomic regions from the batch summary page (data not shown).

linker cassette and another sequencing adaptor (Supplementary Fig. 1A). This sequence processing phase removes these common features to isolate the genomic portion of the read for clustering and mapping, and to group reads belonging to individual samples by bar code. First, each vector feature is matched to a database of input sequence reads by pairwise alignment (Brudno, 2007). Each base position in a vector feature mapping to a read is then masked. Second, direct regular expression matching is used to 'read' the bar code included in each sequence read. At this stage, reads are split into coded groups for further analysis during the clustering and mapping stages.

Redundancy in coverage necessitates the use of clustering to group similar sequence reads before mapping and visualization. Rather than using all-by-all pairwise alignment (Niu *et al*., 2010) or clustering by alignment to dynamically created contiguous sequence fragments, we cluster individual sequence reads by grouping those reads mapped by Blat (Kent, 2002) alignment to an overlapping region in the reference genome of the host cell. Each of the reads mapping to a common genomic region is assigned into a cluster, and all of the reads in each cluster are aligned by MUSCLE (Edgar, 2004). A simple majority-voting algorithm is used to create a consensus sequence of each RIS. This RIS sequence is then Blat aligned back to the reference genome. Once a genomic location is confirmed, the exact position of the RIS is defined by the genomic position flanking proviral LTR in the consensus sequence. Since LTR regulatory regions may influence cellular genes within a large distance of the RIS (Hargrove *et al.*, 2008; Kustikova *et al*., 2005; Lazo *et al*., 1990; Sadat *et al*., 2009), genes located within 300 kb of the RIS are identified and reported.

The consensus sequence is used as the basis for visualizing each RIS. We map the location of each vector feature, the bar code and the genomic alignment to the sequence using BioPerl graphics. The names of and distances to the closest genes in both Ensembl and UCSC genome builds are reported, and the raw multiple sequence alignment (MSA) of reads contributing to the RIS is linked (Fig. 1).

SeqMap 2.0 allows a user to: (i) upload full sets of 454 pyrosequencing reads, (ii) create savable lists of bar codes and identifiers, (iii) create savable lists of vector features to mask from each read and (iv) identify the appropriate reference genomes to which RISs should be mapped. The rest of the process is completely automated and data are returned to the user through secure login to a saved workspace or by email. Investigators are also able to use SeqMap 2.0 as a collaborative research tool by creating lab workspaces accessible to multiple users. SeqMap 2.0 is available at http://seqmap.compbio.iupui.edu/.

Funding: This work was supported by the National Institutes of Health (P40 RR024928, R01LM009722, T32 HL007910, P01 HL53586); Indiana Clinical and Translational Sciences Institute Bioinformatics and Advanced Information Technology Cores (U54 RR025761).

Conflict of Interest: none declared.

REFERENCES

Akagi,K. *et al.* (2004) RTCGD: retroviral tagged cancer gene database. *Nucleic Acids Res.*, **32**, D523–D527.

Appelt,J.U. *et al.* (2009) QuickMap: a public tool for large-scale gene therapy vector insertion site mapping and analysis. *Gene Ther.*, **16**, 885–893.

Brudno,M. (2007) An introduction to the Lagan alignment toolkit. *Methods Mol. Biol.*, **395**, 205–220.

Buchberg,A.M. *et al.* (1990) Evi-2, a common integration site involved in murine myeloid leukemogenesis. *Mol. Cell Biol.*, **10**, 4658–4666.

Copeland,N.G. and Jenkins,N.A. (1990) Retroviral integration in murine myeloid tumors to identify Evi-1, a novel locus encoding a zinc-finger protein. *Adv. Cancer Res.*, **54**, 141–157.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Gabriel,R. *et al.* (2009) Comprehensive genomic access to vector integration in clinical gene therapy. *Nat. Med.*, **15**, 1431–1436.

Giordano,F.A. *et al.* (2007) New bioinformatic strategies to rapidly characterize retroviral integration sites of gene therapy vectors. *Methods Inf. Med.*, **46**, 542–547.

Hacein-Bey-Abina,S. *et al.* (2003) LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science*, **302**, 415–419.

Hacein-Bey-Abina,S. *et al.* (2008) Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J. Clin. Invest.*, **118**, 3132–3142.

Hargrove,P.W. *et al*. (2008) Globin lentiviral vector insertions can perturb the expression of endogenous genes in beta-thalassemic hematopoietic cells. *Mol. Ther.*, **16**, 525–533.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Kustikova,O. *et al*. (2005) Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. *Science*, **308**, 1171–1174.

Lazo,P.A. *et al*. (1990) Long distance activation of the Myc protooncogene by provirus insertion in Mlvi-1 or Mlvi-4 in rat T-cell lymphomas. *Proc. Natl Acad. Sci. USA*, **87**, 170–173.

Niu,B. *et al.* (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*, **11**, 187.

Peters,B. *et al.* (2008) Automated analysis of viral integration sites in gene therapy research using the SeqMap web resource. *Gene Ther.*, **15**, 1294–1298.

Sadat,M. *et al*. (2009) Retroviral vetor integration in post-transplant hematopoiesis in mice conditioned with either sub-\myeloablative or ablative conditioning. *Gene Ther.*, **16**, 1452–1464.

Schmidt,M. *et al.* (2003) Efficient characterization of retro-, lenti-, and foamyvector-transduced cell populations by high-accuracy insertion site sequencing. *Ann. N. Y. Acad. Sci.*, **996**, 112–121.

Schmidt,M. *et al.* (2007) High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). *Nat. Methods*, **4**, 1051–1057.

Smith,D.R. (1992) Ligation-mediated PCR of restriction fragments from large DNA molecules. *PCR Methods Appl.*, **2**, 21–27.