Published in final edited form as:

Biotechniques. 2011 February ; 50(2): 96–97. doi:10.2144/000113600.

Identifying insertion mutations by whole-genome sequencing

Harold E. Smith

National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA

Abstract

Insertion mutagenesis via mobile genetic element is a common technique for the analysis of gene function in model organisms. Next-generation sequencing offers an attractive approach for localizing the site of insertion, but alignment-based mapping of mobile genetic elements is challenging. A computational method for identifying insertion sites is reported herein. The technique was validated by mapping transposons in both bacterial and nematode species. The approach should be extensible to other systems that employ mobile genetic elements to generate mutations.

Keywords

Next-generation sequencing; whole-genome resequencing; mutation identification; insertion mapping

Next-generation sequencing provides an attractive means to map mutations on a genome-wide scale, and various bioinformatics tools have been developed for this application (reviewed in Reference 1). Alignment of the sequenced reads to a reference genome allows the identification of mutations. Mobile genetic elements (referred to hereafter as transposons) pose a particular alignment challenge. Transposon insertion generates novel sequence junctions that are absent from the reference. Furthermore, the reference either lacks the transposon sequences (for introduced transposons) or contains multiple copies (if endogenous) that are typically masked. Those problems can be addressed by constructing mate-pair libraries of sufficient insert size to span the transposon. Insertions produce read-end alignments separated by less than the insert size, although the exact nature and position of the insertion is unknown.

Herein we describe split-end alignment, which is an alternative approach for transposon insertion mapping. It utilizes independent alignment of the two ends of short sequence reads from standard (non–mate-pair) libraries to the genome reference and a transposon-specific reference. The rationale for this strategy is straightforward, and is analogous to the split-read

Address correspondence to Harold E. Smith, National Institutes of Health, 8 Center Drive, Room 1A11, Bethesda, MD, USA. smithhe2@niddk.nih.gov.

Supplementary material for this article is available at www.BioTechniques.com/article/113600.

Competing interests

The author declares no competing interest.

method used for mRNA splice junctions (2). If the transposon junction is located near the center of the read, then the different ends can be aligned independently to the different references.

To test this approach, sequence libraries were constructed from genomic DNA samples from *Escherichia coli* strain M5964, a derivative of strain GC4468 (3), and 11 mutant lines generated from M5964 (Judah L. Rosner and Robert G. Martin, personal communication). Multiplexed samples were prepared according to the manufacturer's protocol (Illumina, San Diego, CA, USA). Thirty-six–cycle sequencing of the starting strain (one lane) and pooled mutant libraries (three lanes) yielded 11,552,833 and 55,334,250 high-quality reads, respectively. The multiplexed samples were deconvoluted into the 11 constituent mutant data sets using CASAVA (version 1.6.0; Illumina). Depth of coverage was 89× for the starting strain and 25–59× for the mutants.

The genome sequence of *E. coli* strain MG 1655 (4) was used as the reference for alignment. A second reference file (designated IS) was constructed from sequences (length 20 nucleotides) derived from the 5′ and 3′ ends of known *E. coli* insertion elements. Seven deletion mutations had been engineered into the starting strain using a FRT-flanked kanamycin-resistance cassette and flippase recombination (5); therefore, 20-nucleotide sequences from each end of that cassette were also included in the IS reference file. Note that those sites are conceptually equivalent to insertions of an introduced transposon.

Sixteen nucleotides from the 5′ and 3′ end of each sequence read were aligned to the genome and IS reference sequences using the program ELAND (version 2; Illumina). Because the algorithm allows up to two mismatches, it creates a window of eight nucleotides (read length of 36, less 14 from each end) where the novel junction can fall without affecting the end-segment alignments. Starting strain M5964 was analyzed for detection of known insertion elements by screening for reads that aligned to the genomic reference at one end and the IS reference at the other end. Five of the 46 insertion elements present in the reference genome MG1655 lie within a large deletion of the starting strain (see Supplementary Materials for details). All of the remaining 41 insertion elements were readily identified, as were all seven of the engineered deletion mutations (data not shown). No additional insertion elements were discovered. Each insertion element was represented by 44 reads on average (range of 17–91).

Split-end alignment was repeated for each of the 11 samples from mutation lines. Nine differed from the starting strain by the novel transposition of a single insertion element (Table 1). Three lines contained the identical mutation at position 3,411,632 between *envR* and *acrE*, and likely arose from a single transposition event that occurred prior to mutant screening. Three additional lines contained independent transposon insertions within the same interval. Two other lines contained independent insertion elements in the *hns* gene, while another mutant line contained an insertion in *yeiT*. Transpositions were not detected in the two remaining lines.

All of the samples were further analyzed for additional variants using the programs BFAST (6) and SAMtools (7) to permit detection of SNPs as well as small insertions and deletions.

The only additional difference observed between the starting strain and any of the mutant lines was a single SNP in one of the two lines that lacked a novel transposition (Table 1). The SNP was a G→T missense mutation in *hns* that caused an alanine–to–glutamic acid substitution at amino acid 18. Note that the same gene was identified by two of the transposon insertions. The mutation in the remaining line has yet to be identified.

To determine if split-end alignment was applicable to a larger, more complex genome, the endogenous transposons of *Caenorhabditis elegans* were mapped from short-read sequence data. Seven active transposons have been reported for *C. elegans* (8). BLAST alignment of the transposon sequences to the reference genome (Wormbase release WS190; Reference 9) identified 91 intact (containing both terminal inverted repeats) transposons. Libraries were constructed from an N2 Bristol strain derivative, in which transposition is quiescent (10). One lane of 76-cycle sequencing yielded 20,810,145 high-quality reads (~16× coverage). Split-end alignment was performed as above to the genome and transposon reference files (see Supplementary Materials for details). Eleven transposons lay in repeated or low-complexity sequences, typically near gene-poor chromosome ends, which precluded unambiguous alignment to the reference genome; as expected, those transposons were not identified. All of the remaining 80 transposons were successfully recovered (Table 2), demonstrating the suitability of this strategy for model organisms with larger genomes.

Identification of transposon insertions by split-end alignment offers several advantages over current methods. No new software is required, and most alignment programs can be adapted to this purpose without major modification of existing data analysis pipelines. The primary requirement is construction of a transposon-specific reference file. Both novel, introduced transposons as well as endogenous, multi-copy transposons can be identified. The construction of mate-pair libraries is obviated, and the nature and position of the insertion is known with base-pair resolution. The data set of paired genome/transposon reads generated by this method is sufficiently small to be analyzed by bench scientists using desktop computers and/or web-based tools.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Thanks to Lee Rosner, Bob Martin, and Kevin O'Connell for providing DNA samples, and to Michael Krause for comments on the manuscript and fruitful discussions. This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Institute of Diabetes and Digestive and Kidney Diseases. This paper is subject to the NIH Public Access Policy.

References

1. Medvedev P, Stanciu M, and Brudno M. 2009 Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods 6:S13–S20. [PubMed: 19844226]
2. Ameur A, Wetterbom A, Feuk L, and Gyllensten U. 2010 Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol 11:R34. [PubMed: 20236510]
3. Carlioz A and Touati D. 1986 Isolation of superoxide dismutase mutants in *Escherichia coli*: is superoxide dismutase necessary for aerobic life? EMBO J 5:623–630. [PubMed: 3011417]

4. Blattner FR, Plunkett G, III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, et al. 1997 The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462. [PubMed: 9278503]

5. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, et al. 2006 Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol 2:2006.0008.

6. Homer N, Merriman B, and Nelson SF. 2009 BFAST: an alignment tool for large scale genome resequencing. PLoS One 4:e7767. [PubMed: 19907642]

7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, et al. 2009 The Sequence Alignment/Map format and SAM tools. Bioinformatics 25:2078–2079. [PubMed: 19505943]

8. Bessereau JL 2006 Transposons in *C. elegans In* The *C. elegans* Research Community (Ed.), WormBook doi/10.1895/wormbook.1.7.1, http://www.wormbook.org.

9. Chen N, Harris TW, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Canaran P, Chan J, Chen CK, et al. 2005 WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. Nucleic Acids Res 33(Database issue):D383–389. [PubMed: 15608221]

10. Collins J, Saari B, and Anderson P. 1987 Activation of a transposable element in the germ line but not the soma of *Caenorhabditis elegans*. Nature 328:726–728. [PubMed: 3039378]

Table 1.

Summary of *E. coli* mutations.

Mutant line	Mutation	Position	Locus
#2	IS5 insertion	3,411,792	*envR-acrE* interval
#3	IS1 insertion	3,411,760	*envR-acrE* interval
#6	IS5 insertion	3,411,632	*envR-acrE* interval
#9	IS5 insertion	3,411,632	*envR-acrE* interval
#10	IS5 insertion	3,411,632	*envR-acrE* interval
#11	IS1 insertion	3,411,715	*envR-acrE* interval
#13	??	N/A	N/A
#16	Ala18Glu missense	1,292,093	*hns*
#18	IS5 insertion	2,233,123	*yeiT*
#20	IS1 insertion	1,291,827	*hns*
#24	IS5 insertion	1,291,883	*hns*

N/A, not available.

Table 2.

Summary of *C. elegans* transposon mapping.

	Number of transposons identified/total, by chromosome					
Transposon	**chrI**	**chrII**	**chrIII**	**chrIV**	**chrV**	**chrX**
Tc1	3/3	8/9	2/2	3/4	8/10	2/2
Tc2		0/1			1/3	
Tc3	4/5	7/7	2/2	3/3	3/3	1/1
Tc4/Tc4v	3/3	1/1	1/1	1/1	1/1	3/3
Tc5		2/2		3/3		
Tc7		1/1		1/1	2/2	5/7
cemaT1	3/3			3/3	1/2	2/2