Bioinformatics, 32(7), 2016, 1109–1111
doi: 10.1093/bioinformatics/btw022
Advance Access Publication Date: 21 January 2015
Applications Note

OXFORD

Systems biology

The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries

Lars Barquist[1,2], Matthew Mayho[1], Carla Cummins[1], Amy K. Cain[1], Christine J. Boinett[1], Andrew J. Page[1], Gemma C. Langridge[1], Michael A. Quail[1], Jacqueline A. Keane[1] and Julian Parkhill[1,*]

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK and [2]Institute for Molecular Infection Biology, University of Würzburg, Würzburg D-97080, Germany

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on 15 September 2015; revised on 11 December 2015; accepted on 13 January 2016

Abstract

Summary: Transposon insertion sequencing is a high-throughput technique for assaying large libraries of otherwise isogenic transposon mutants providing insight into gene essentiality, gene function and genetic interactions. We previously developed the Transposon Directed Insertion Sequencing (TraDIS) protocol for this purpose, which utilizes shearing of genomic DNA followed by specific PCR amplification of transposon-containing fragments and Illumina sequencing. Here we describe an optimized high-yield library preparation and sequencing protocol for TraDIS experiments and a novel software pipeline for analysis of the resulting data. The Bio-Tradis analysis pipeline is implemented as an extensible Perl library which can either be used as is, or as a basis for the development of more advanced analysis tools. This article can serve as a general reference for the application of the TraDIS methodology.

Availability and implementation: The optimized sequencing protocol is included as supplementary information. The Bio-Tradis analysis pipeline is available under a GPL license at https://github.com/sanger-pathogens/Bio-Tradis

Contact: parkhill@sanger.ac.uk

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

1 Introduction

Steady improvements in high-throughput sequencing technologies have resulted in an increasing number of sequenced bacterial genomes, revealing extensive genetic diversity both within and between species. Associated sequencing-based technologies, such as RNA-seq, ChIP-seq and RIP-seq provide insight into the effects of this variation on gene expression and regulation; however, none provides direct information on cell survival, and hence how this genetic variation may impact the fitness of the bacterium (Gray *et al.*, 2015). Transposon insertion sequencing (TIS) bridges this gap between sequence and fitness by allowing for direct measurement of survival dynamics within a population of single transposon mutants, by using sequencing reads flanking transposon insertions as a read-out of mutant frequency within the population (Barquist *et al.*, 2013a; Van Opijnen and Camilli, 2013). We previously developed a method for this purpose, called Transposon Directed Insertion Sequencing (TraDIS; Langridge *et al.*, 2009). TraDIS uses fragmentation of genomic DNA followed by specific PCR amplification of transposon-containing fragments to selectively enrich for transposon-flanking sequences, and can be adapted for any transposon of interest through a simple redesign of sequencing primers. TraDIS has since been applied to a variety of target organisms and transposons in a wide variety of both *in vivo* and *in vitro* growth conditions. These include Tn*5*-based libraries in *Salmonella* (Barquist *et al.*, 2013b; Chaudhuri *et al.*, 2013; Langridge *et al.*, 2009) and *Escherichia* (Dziva *et al.*, 2013; Eckert *et al.*, 2011) and Mariner-based libraries in *Clostridia* (Dembek *et al.*, 2015) and *Mycobacteria* (Weerdenburg *et al.*, 2015).

2 Library preparation and sequencing

We have made a number of refinements to the TraDIS sequencing protocol since its initial publication (Langridge *et al.*, 2009), described in more detail in the supplement. We have redesigned TraDIS adapters and primers using a splinkerette approach (Devon *et al.*, 1995; Rad *et al.*, 2015; Uren *et al.*, 2009), which increases enrichment of genuine transposon-chromosome junctions by preventing hybridization of the reverse primer until the transposon-specific forward primer has generated a complementary strand. We have substituted a magnetic bead-based fragment size selection for gel-based size selection to increase yield and allow for easier automation (Bronner *et al.*, 2013). Finally, we have substituted Kapa Hifi DNA polymerase for Taq polymerase, as this enzyme has been shown to have minimal amplification biases (Quail *et al.*, 2012), and reduced the number of cycles of PCR amplification to provide a more accurate representation of input.

TraDIS sequencing primers are designed to begin sequencing within the transposon sequence, so as to provide a short 8–10 base 'transposon tag' at the beginning of each read to verify that each read originates from a genuine transposon-chromosome junction. This poses a challenge for Illumina sequencing machines, as the base-calling algorithms assume a complex sample for the purposes of calibration. We have developed HiSeq and MiSeq recipes that use 'dark cycles' during which chemistry is run but no imaging is performed to read through this transposon tag, before imaged sequencing commences on the complex chromosomal DNA (see supplement). Once the first read is completed, the DNA is denatured and the transposon-specific sequencing primer is re-annealed for a separate short 10–12 cycle transposon read. This requires a PhiX (or other complex library) spike-in of 5–10% to prevent sequencing failure due to a lack of fluorescence in some channels. Using this protocol we routinely achieve results of $>90\%$ of sequencing reads both containing an intact transposon tag and mapping uniquely to the source genome. We have applied this method to Tn*5*-, Tn*917*-, *Himar1*- and Mu-based mutant libraries, and it should be adaptable to any transposon of interest assuming a suitable priming site exists (see supplement for design parameter details).

3 The Bio-Tradis analysis pipeline

To support the use of this improved TraDIS protocol, we have developed a portable processing and analysis pipeline implemented in the Perl and R languages. The functionality provided is similar to that in other recently published TIS analysis pipelines (DeJesus *et al.*, 2015; Solaimanpour *et al.*, 2015), however our command-line driven approach has been designed with a production environment in mind, where many sequencing libraries may be processed simultaneously. We provide tools for each step of analysis from the raw unaligned fastq files produced by the sequencer, through to predictions of gene essentiality and fitness effects. The main pipeline script, **bacteria_tradis**, filters reads in fastq format for transposon tags, removes these tags, then maps the modified reads using the SMALT short read mapper (https://www.sanger.ac.uk/resources/software/smalt/), with support for multiple contigs and/or replicons, such as plasmids. Default k-mer, step size and percent identity parameters are set depending on input read length, though these can be manually specified by the user. The mapped bam file is then processed to produce plot files, containing insertion counts per nucleotide, suitable for visualization in the Artemis genome browser (Carver *et al.*, 2012) and for further analysis. The mapping, processing, and data manipulation steps are implemented as self-contained Perl modules that could be easily used as a foundation for the development of more sophisticated analyses.

Additional scripts are provided to process these plot files in conjunction with genome annotations in EMBL-Bank format to produce annotated tab-delimited files containing various statistics including read counts and unique insertion sites per gene. Two basic analysis scripts for this gene-level data written in R are available. One, **tradis_essentiality.R**, produces predictions of gene essentiality within a high-density transposon library based on the empirically observed bimodal distribution of insertion sites over genes when normalized for gene length (Barquist *et al.*, 2013b; Langridge *et al.*, 2009). The second, **tradis_comparisons.R**, applies the edgeR package (Robinson *et al.*, 2010) to identify significant differences in read counts, and hence mutant frequencies, between experimental conditions (Dembek *et al.*, 2015) providing insight into the relative contribution of all mutagenized genes to fitness under the assayed condition.

4 Summary

We have described recent refinements to the TraDIS method for the sequencing and analysis of dense transposon libraries. These include an optimized sequencing protocol, and processing and analysis tools that can rapidly provide insight into the contribution of genomic regions to organismal fitness. It is our hope that making these tools more accessible will accelerate their application to an ever wider variety of bacteria and experimental conditions.

Acknowledgements

The authors would like to thank Mark Gibbs and Isabelle Rasolonjatovo (Illumina FAS) for expert assistance in developing the sequencing recipes described here, and Elizabeth Huckle and Tristram Keith Bellerby for proofreading the manuscript.

Funding

This work was supported by the Wellcome Trust, grant number WT098051. LB is supported by a research fellowship from the Alexander von Humboldt Stiftung/Foundation. AKC and CJB were supported by the Medical Research Council, grant number G1100100/1.

Conflict of Interest: none declared.

References

Barquist,L. *et al.* (2013a) Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biol.*, **10**, 1161–1169.

Barquist,L. *et al.* (2013b) A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium. *Nucleic Acids Res.*, **41**, 4549–4564.

Bronner,I.F. *et al.* (2013) Improved Protocols for Illumina Sequencing. *Curr. Protoc. Hum. Genet.*, **80**, 18.2.1–18.2.42.

Carver,T. *et al.* (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, **28**, 464–469.

Chaudhuri,R.R. *et al.* (2013) Comprehensive assignment of roles for *Salmonella* Typhimurium genes in intestinal colonization of food-producing animals. *PLoS Genet.*, **9**, e1003456.

DeJesus. *et al.* (2015) TRANSIT – a software tool for Himar1 TnSeq analysis. *PLoS Comp. Biol.*, **11**, e1004401.

Dembek,M. *et al.* (2015) High-throughput analysis of gene essentiality and sporulation in *Clostridium difficile*. *MBio*, **6**, 02383–14.

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

Dziva,F. *et al*. (2013) Sequencing and functional annotation of avian pathogenic *Escherichia coli* serogroup O78 strains reveal the evolution of *E. coli* lineages pathogenic for poultry via distinct mechanisms. *Infect. Immun*., **81**, 838–849.,

Eckert,S.E. *et al*. (2011) Retrospective application of transposon-directed insertion site sequencing to a library of signature-tagged mini-Tn5Km2 mutants of *Escherichia coli* O157:H7 screened in cattle. *J. Bacteriol*., **193**, 1771–1776.

Gray,A.N. *et al*. (2015) High-throughput bacterial functional genomics in the sequencing era. *Curr. Opin. Microbiol*., **27**, 86–95.

Langridge,G.C. *et al*. (2009) Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. *Genome Res*., **19**, 2308–2316.

Van Opijnen,T. and Camilli,A. (2013) Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat. Rev. Microbiol*., **11**, 435–442.

Quail,M.A. *et al*. (2012) Optimal enzymes for amplifying sequencing libraries. *Nat. Methods*, **9**, 10–11.

Rad,R. *et al*. (2015) A conditional piggyBac transposition system for genetic screening in mice identifies oncogenic networks in pancreatic cancer. *Nat. Genet*., **47**, 47–56.

Robinson,M.D. *et al*. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Solaimanpour. *et al*. (2015) Tn-Seq Explorer: a tool for the analysis of high-throughput sequencing data of transposon mutant libraries. *PLoS One*, **10**, e0126070.

Uren,A.G. *et al*. (2009) A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nat. Protoc*., **4**, 789–798.

Weerdenburg,E.M. *et al*. (2015) Genome-wide transposon mutagenesis indicates that *Mycobacterium marinum* customizes its virulence mechanisms for survival and replication in different hosts. *Infect. Immun*., **83**, 1778–1788.