Bioinformatics, 33(17), 2017, 2781–2783
doi: 10.1093/bioinformatics/btx320
Advance Access Publication Date: 11 May 2017
Applications Note

Systems biology

MAGenTA: a Galaxy implemented tool for complete Tn-Seq analysis and data visualization

Katherine Maia McCoy[†], Margaret L. Antonio[†] and Tim van Opijnen*

Biology Department, Boston College, Chestnut Hill, MA, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on March 31, 2017; revised on May 2, 2017; editorial decision on May 8, 2017; accepted on May 10, 2017

Abstract

Motivation: Transposon insertion sequencing (Tn-Seq) is a microbial systems-level tool, that can determine on a genome-wide scale and in high-throughput, whether a gene, or a specific genomic region, is important for fitness under a specific experimental condition.
Results: Here, we present MAGenTA, a suite of analysis tools which accurately calculate the growth rate for each disrupted gene in the genome to enable the discovery of: (i) new leads for gene function, (ii) non-coding RNAs; (iii) genes, pathways and ncRNAs that are involved in tolerating drugs or induce disease; (iv) higher order genome organization; and (v) host-factors that affect bacterial host susceptibility. MAGenTA is a complete Tn-Seq analysis pipeline making sensitive genome-wide fitness (i.e. growth rate) analysis available for most transposons and Tn-Seq associated approaches (e.g. TraDis, HiTS, IN-Seq) and includes fitness (growth rate) calculations, sliding window analysis, bottleneck calculations and corrections, statistics to compare experiments and strains and genome-wide fitness visualization.
Availability and implementation: MAGenTA is available at the Galaxy public ToolShed repository and all source code can be found and are freely available at https://vanopijnenlab.github.io/MAGenTA/.
Contact: vanopijn@bc.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transposon insertion sequencing (Tn-Seq) combines transposon mutagenesis on a genome-wide scale with massively parallel sequencing to determine, whether a gene, or a specific genomic region, is important for fitness under a chosen experimental condition (Gawronski *et al.*, 2009; Goodman *et al.*, 2009; Langridge *et al.*, 2009; van Opijnen *et al.*, 2009; van Opijnen and Camilli, 2013). Tn-Seq has been successfully used to identify essential genes (Chao *et al.*, 2013; Griffin *et al.*, 2011; Klein *et al.*, 2012; Le Breton *et al.*, 2015; Zomer *et al.*, 2012), however, it is often ignored that it is possible to look further then just a binary outcome (i.e. live versus dead/very sick). Instead we have shown that the fitness of each mutant can be accurately calculated and represented as the growth rate. This approach is so sensitive that it enables identification of mutants that differ by as little as 2 min (5%) in their doubling times (van Opijnen *et al.*, 2009; van Opijnen and Camilli, 2012). To underscore this sensitivity we have confirmed growth rates obtained from Tn-Seq for >300 *in vitro* and *in vivo* phenotypes by individual growth and 1 × 1 competitions. Consequently, besides identifying (conditionally) essential genes, this approach has enabled a range of discoveries including the identification of new leads for gene function, novel small non-coding RNAs (ncRNAs), higher order genome organization and of host-factors that affect bacterial host susceptibility (Carter *et al.*, 2014; Jensen *et al.*, 2016; Mann *et al.*, 2012; van Opijnen *et al.*, 2009, 2016; van Opijnen and Camilli, 2012). Here we present MAGenTA (Microbial Assessment by Genome-Wide Tn-Seq Analysis), a complete Tn-Seq analysis pipeline implemented in Galaxy, but also available as separate scripts. MAGenTA makes sensitive genome-wide fitness (i.e. growth rate) analysis available for most transposon and Tn-Seq associated approaches [e.g. TraDis (Langridge *et al.*, 2009), HiTS (Gawronski *et al.*, 2009), IN-Seq (Goodman *et al.*, 2009)], and includes fitness (growth rate) calculations, bottleneck calculations and corrections, statistical comparisons of conditions or strains, sliding window analysis and genome-wide fitness visualization.

2 Pipeline description

MAGenTA covers the entire process of Tn-Seq data analysis from processing raw-sequence reads up to fitness calculations and data visualization, which is described in detail in Supplementary Material S1. Below we highlight major steps in the analysis that make the approach unique. The analysis is broadly split up into four parts that can be run independently. In part 1 raw data are processed and mapped to the genome, in part 2 bottlenecks are determined and corrected, and fitness is calculated for individual insertions and specified regions, in part 3 statistics are performed, data are analyzed by sliding window, and conditions or strains are statistically compared, while in part 4 data are visualized.

2.1 Read trimming, barcode splitting and mapping

Tn-Seq can be performed with different types of transposons and there are different sample preparation methods that affect what type of final product is sequenced. Some methods require specific ways in which raw reads are processed (described in Supplementary Material S1), which can be implemented here. If reads contain a multiplexing barcode they can be sorted based on this barcode, and finally reads are mapped to a reference genome.

2.2 Fitness calculations and statistics

Most Tn-Seq analyses use some form of the competition index as a proxy for fitness, however an arguably more sensitive and quantitative approach is one we developed to calculate fitness representing the growth rate (van Opijnen and Camilli, 2013). This requires sampling of two time points for sequencing and determination of the expansion or 'contraction' (van Opijnen and Camilli, 2012) of the population. This will generate insertion specific fitness values, and by averaging over all the insertions found within a gene or region aggregate fitness is calculated, resulting in single fitness values and standard deviations for each gene in the genome (van Opijnen *et al.*, 2009; van Opijnen and Camilli, 2013). Importantly, this enables robust statistical analysis (see below).

2.2.1 Bottleneck calculations and corrections

A challenge that can affect Tn-Seq data, especially from *in vivo* experiments, is that mutants may be lost stochastically, for instance while establishing an infection. Such bottlenecks are problematic because it becomes difficult to distinguish between insertion mutants that disappear due to chance or because they are less fit. However, we have shown that by determining a set of neutral insertions (e.g. in degenerate or pseudo genes), that do not effect fitness and should therefore not disappear from the library, it is possible to estimate such bottlenecks and correct for it in the analysis (van Opijnen and Camilli, 2012).

2.3 Genome-wide sliding window analysis and fitness comparisons

Data can be further explored by performing statistics to determine which fitness changes are significant and a sliding window approach based on the methods by Zhang *et al.* (2012) enables the user to screen the genome for regions that have a significant fitness effect. Additionally, fitness can be compared across experiments and environmental conditions, and even between strains, with the goal to reveal strain-specific fitness effects (Jensen *et al.*, 2016; van Opijnen *et al.*, 2016). N.B. Two tools are specifically geared towards transposons that insert into TA sites and are marked within Supplementary Material S1.

2.4 Genome-wide fitness visualization of single insertions

Besides visualization of individual insertions across a genome, users can visualize multiple fitness data tracks simultaneously as well as multiple different genomes (e.g. strains) for visual inspection of the data and generating publishable figures.

3 Conclusions

MAGenTA provides complete Tn-Seq analysis from raw read processing to fitness calculations, and data visualization. MAGenTA has been used in multiple studies and enables the ability to calculate a highly quantitative measure of fitness, identify bottlenecks, perform robust statistical analyses and visually inspect the data. Moreover, we have provided example and practice files at Github as well as a detailed step-by-step manual (Supplementary Material S1) to ensure that MAGenTA is accessible to users with variable levels of bioinformatics experience.

Acknowledgements

We thank Jon Anthony and Tony Schreiner for bioinformatics and hardware support, and Derek Thibault, and Matthew Crum for walking through beta-versions of the step-by-step manual and giving helpful suggestions for improvements.

Funding

This work has been supported by NIAID (grants R01-AI110724, R21-AI117247, U01-AI124302) and by Boston College URFs for K.M.M. and M.M.A.

Conflict of Interest: none declared.

References

Carter,R. *et al*. (2014) Genomic analyses of pneumococci from children with sickle cell disease expose host-specific bacterial adaptations and deficits in current interventions. *Cell Host Microbe*, **15**, 587–599.

Chao,M.C. *et al*. (2013) High-resolution definition of the *Vibrio cholerae* essential gene set with hidden Markov model-based analyses of transposon-insertion sequencing data. *Nucleic Acids Res*., **41**, 9033–9048.

Gawronski,J.D. *et al*. (2009) Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. *Proc. Natl. Acad. Sci. USA*, **106**, 16422–16427.

Goodman,A.L. *et al*. (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe*, **6**, 279–289.

Griffin,J.E. *et al*. (2011) High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathogens*, **7**, e1002251.

Jensen,P. *et al*. (2016) Network analysis links genome-wide phenotypic and transcriptional stress responses in a bacterial pathogen with a large pan-genome. bioRxiv 071704; doi: https://doi.org/10.1101/071704.

Klein,B.A. *et al*. (2012) Identification of essential genes of the periodontal pathogen Porphyromonas gingivalis. *BMC Genomics*, **13**, 578.

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

Le Breton,Y. *et al*. (2015) Essential Genes in the Core Genome of the Human Pathogen Streptococcus pyogenes. *Sci. Rep*., **5**, 9838.

Mann,B. *et al*. (2012) Control of virulence by small RNAs in Streptococcus pneumoniae. *PLoS Pathogens*, **8**, e1002788.

van Opijnen,T. *et al*. (2009) Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods*, **6**, 767–772.

van Opijnen,T. and Camilli,A. (2012) A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome Res*., **22**, 2541–2551.

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.

van Opijnen,T., Dedrick,S., and Bento,J. (2016) Strain dependent genetic networks for antibiotic-sensitivity in a bacterial pathogen with a large pan-genome. *PLoS Pathogens*, **12**, e1005869.

Zhang,Y.J. *et al*. (2012) Global assessment of genomic regions required for growth in Mycobacterium tuberculosis. *PLoS Pathogens*, **8**, e1002946.

Zomer,A. *et al*. (2012) ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PloS One*, **7**, e43012.