The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/111011

Please be advised that this information was generated on 2017-09-22 and may be subject to change.
ESSENTIALS: Software for Rapid Analysis of High Throughput Transposon Insertion Sequencing Data

Aldert Zomer1,2, Peter Burghout1, Hester J. Bootsma1, Peter W. M. Hermans1, Sacha A. F. T. van Hijum2,3,4,5

1 Laboratory of Pediatric Infectious Diseases, Radboud University Medical Centre, Nijmegen, The Netherlands, 2 Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Medical Centre, Nijmegen, The Netherlands, 3 NIZO Food Research, Kluyver Centre for Genomics of Industrial Fermentation, Ede, The Netherlands, 4 TI Food and Nutrition, Wageningen, The Netherlands, 5 Netherlands Bioinformatics Centre, Nijmegen, The Netherlands

Abstract

High-throughput analysis of genome-wide random transposon mutant libraries is a powerful tool for (conditional) essential gene discovery. Recently, several next-generation sequencing approaches, e.g. Tn-seq/INseq, HITS and TraDIS, have been developed that accurately map the site of transposon insertions by mutant-specific amplification and sequence readout of DNA flanking the transposon insertions site, assigning a measure of essentiality based on the number of reads per insertion site flanking sequence or per gene. However, analysis of these large and complex datasets is hampered by the lack of an easy to use and automated tool for transposon insertion sequencing data. To fill this gap, we developed ESSENTIALS, an open source, web-based software tool for researchers in the genomics field utilizing transposon insertion sequencing analysis. It accurately predicts (conditionally) essential genes and offers the flexibility of using different sample normalization methods, genomic location bias correction, data preprocessing steps, appropriate statistical tests and various visualizations to examine the results, while requiring only a minimum of input and hands-on work from the researcher. We successfully applied ESSENTIALS to in-house and published Tn-seq, TraDIS and HITS datasets and we show that the various pre- and post-processing steps on the sequence reads and count data with ESSENTIALS considerably improve the sensitivity and specificity of predicted gene essentiality.


Editor: Nicholas James Provart, University of Toronto, Canada

Received April 14, 2012; Accepted July 17, 2012; Published August 10, 2012

Copyright: © 2012 Zomer et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the European Commission FP7 Marie Curie IEF Action [274586 to AZ] and by the Netherlands Genomics Initiative Horizon Breakthrough [93518023 to PB]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have the following interest: Sacha A.F.T. van Hijum is employed by NIZO food research. There are no patents, products in development or marketed products to declare. This does not alter the authors’ adherence to all the PLoS ONE policies on sharing data and materials, as detailed online in the guide for authors.

* E-mail: a.zomer@cuukz.umcn.nl

Introduction

Discovery of microbial genes essential for growth, survival and/or pathogenesis has frequently been used for gene functional analysis, determining the minimal functional genome and identification of therapeutic targets [1–3]. Traditionally, this approach involves extensive testing of fitness defects of single mutants during relevant in vivo or in vitro conditions with methods that are far from high-throughput.

A first step towards high-throughput screening for essential genes was made with Signature Tagged Mutagenesis (STM) [4]. STM screens single mutants in pools of up to 96 tagged mutants, and readout of mutant pools before and after conditional challenge occurs by Southern hybridization detection of mutant-specific DNA tags. Microarray-based methods for detection of transposon mutants, e.g. Transposon Site Hybridization (TraSH) and Genomic Array Footprinting (GAF) [5–7], further improved the throughput of microbial essential gene discovery. However, these approaches potentially suffer from cross-hybridization and lack of resolution hampering the identification of the exact location of the transposon insertion site [8,9]. These problems were alleviated by high-throughput transposon insertion sequencing analysis methods such as Tn-seq or TraDIS or variants thereof [10–13].

With transposon sequencing analysis the presence of each unique mutant within a defined or random transposon mutant library is determined by amplification of DNA flanking the transposon insertion site followed by massively parallel sequencing. Sequence reads from DNA flanking transposon insertion sites are mapped on the reference genome and summarized for each insertion and gene, generating a measurement of fitness for every knockout in comparison to the expected values based on mutant library size, number of possible unique insertion sites per gene and number of sequence reads. Similarly, to identify conditionally essential genes, these libraries are exposed to a challenge condition that will induce loss of mutants of genes essential for survival in these conditions. A measurement of fitness for every knockout comparison between challenge and control condition can then be calculated. In summary, with this technique (i) the decreased fitness of mutants can be detected, (ii) the ubiquity of a specific mutant can be counted and compared to all other mutants in the mutant library and (iii) the exact location of the transposon insertion can be determined.

This method has been successfully applied to determine gene essentiality and the minimal genome of Streptococcus pneumoniae.
using various settings of the ESSENTIALS tool and analyzed with sequence reads and count data with ESSENTIALS considerably show that the various pre- and post-processing steps of the and various visualizations to examine the results. Additionally we correction, data preprocessing steps, appropriate statistical tests different sample normalization methods, genomic location bias (conditionally) essential genes and offers the flexibility of using transposon insertion sequencing analysis. It accurately predicts data [20,21].

Results and Discussion

To facilitate the analysis of data generated by transposon insertion sequencing by genomics researchers we developed a web-based software tool suitable for researchers in the genomics field utilizing transposon insertion sequencing analysis. It accurately predicts (conditionally) essential genes and offers the flexibility of using different sample normalization methods, genomic location bias correction, data preprocessing steps, appropriate statistical tests and various visualizations to examine the results. Additionally we show that the various pre- and post-processing steps of the sequence reads and count data with ESSENTIALS considerably improves both sensitivity and specificity of predicted gene essentiality.

Removal of Reads Mapping in Repeat Regions or in the 3’ Terminus Improves Essential Gene Detection

Insertion sites that do not have unique flanking sequences cannot be assigned to a single gene and are as such not informative for gene essentiality. For instance in S. pneumoniae R6 more than 5% of the insertion site flanking sequences have a perfect match elsewhere on the genome (results not shown). Failure to remove these reads will result in assigning reads to essential genes, causing these potentially to be detected as non-essential. Reads with a perfect match elsewhere on the genome sequence were excluded, which resulted in removal of 12,006 insertion site flanking sequences, which in turn resulted in an additional 17 essential genes detected for S. pneumoniae R6. ROC analysis of essential versus non-essential genes showed an increased area under the curve (AUC) (Table 1). For P. aeruginosa PA01, after repeat-filtering, five genes were no longer considered essential for tobramycin resistance. Repeat filtering did not result in an increase in predictive power of essential genes, as the AUC did not increase (Fig. 1B).

Likewise, transposons that are inserted in the 3’ terminus of a gene often might not lead to loss of function of that gene. Removal of reads mapped to these insertions resulted in an additional 53 essential genes in S. pneumoniae R6, a near-perfect AUC and consequently a dramatic reduction of the P-value (Table 1; Fig. 1A). For P. aeruginosa PA01 a small decrease in the AUC but a minor improvement of p-value was observed, however the two populations of essential genes and non-essential genes still could not be considered significantly different (p<0.01) (Table 1; Fig. 1B).
Normalization Improves Detection of Essential Genes

In many published transposon insertion sequencing analysis studies, information regarding experimental variability from the replicate experiments is not used in statistical testing for gene essentiality. For instance Langridge and co-workers combined the reads per gene of replicate target and control experiments, added an arbitrary value of 100 to all read counts, and calculated ratio’s followed by testing for deviation from a normal distribution [30]. To determine the experimental variability, replicate experiments of the same mutant library should be performed and compared using appropriate statistical tests. Sequencing a sample to half the read-depth compared to other samples will yield half the number of reads mapping to each gene. To facilitate comparisons between samples of varying read-depth, scaling of the samples to their total number of reads is therefore required. Simple total read count scaling is not always appropriate; the number of reads mapping to each gene. To facilitate comparisons

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Applied processing step</th>
<th>Essential genes detecteda</th>
<th>Predictive value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essential S. pneumoniae R6</td>
<td>Scaling</td>
<td>288</td>
<td>0.9517</td>
</tr>
<tr>
<td></td>
<td>Repeatfiltering</td>
<td>305</td>
<td>0.9588</td>
</tr>
<tr>
<td></td>
<td>Gene truncation</td>
<td>358</td>
<td>0.9996</td>
</tr>
<tr>
<td></td>
<td>Quantile normalization</td>
<td>359</td>
<td>0.9996</td>
</tr>
<tr>
<td></td>
<td>L TMM normalization</td>
<td>342</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>Genomic location bias correction</td>
<td>339</td>
<td>1.0000</td>
</tr>
<tr>
<td>Essential for tobramycin resistance P. aeruginosa PAO1</td>
<td>Scaling</td>
<td>185</td>
<td>0.6774</td>
</tr>
<tr>
<td></td>
<td>Repeatfiltering</td>
<td>180</td>
<td>0.6747</td>
</tr>
<tr>
<td></td>
<td>Gene truncation</td>
<td>173</td>
<td>0.6667</td>
</tr>
<tr>
<td></td>
<td>Quantile normalization</td>
<td>174</td>
<td>0.6640</td>
</tr>
<tr>
<td></td>
<td>L TMM normalization</td>
<td>190</td>
<td>0.6640</td>
</tr>
<tr>
<td></td>
<td>Genomic location bias correction</td>
<td>121</td>
<td>0.7634</td>
</tr>
<tr>
<td>Literature</td>
<td></td>
<td>117</td>
<td>0.7406</td>
</tr>
</tbody>
</table>

The predictive value of each method was assessed using ROC curves and a Welch T-test.

aCut-offs for S. pneumoniae R6 were automatically detected by ESSENTIALS while for P. aeruginosa PAO1 a cut-off of 2.5 fold underrepresentation of reads per gene in the challenge condition was used to facilitate comparison with the literature data from Gallagher et al.

doi:10.1371/journal.pone.0043012.t001

Genomic Location Insertion and Read Count Bias can be Corrected by LOESS Regression

Conceivably, because of genomic replication during growth and the resulting increase in available DNA close to the origin of replication (ORI), read counts increase as they are closer to the ORI, especially if multiple replication forks are present [32]. Additionally, more genomic DNA close to the ORI is available for transposon mutagenesis, resulting in a higher number of transposon insertions closer to the ORI. These two factors produce a substantial bias, showing up as a typical V-shape in the read counts per gene relative to the genomic location, with average read counts at the ORI being at least 3 times higher than those near the terminus of replication (Fig. 2A). Gallagher et al. [10] noticed a similar bias in their study when P. aeruginosa PAO1 was exposed to tobramycin. They corrected for this bias by calculating the local read density within a 100-kb window and normalizing the number of reads at that position relative to the average local read density for that window.

Because the local read density window needs to be optimized for genome size and mutant library size, we opted instead to correct for this bias by using locally weighted scatterplot smoothing (LOESS) on read counts per insertion site and per gene as function of genomic position. By applying the ratio between the LOESS curve and a straight line of the average signal to the read counts, the read count bias per insertion site

Relative Log Expression (RLE) trim an upper and lower fraction of the data and use the remaining data to calculate normalization factors. Application of the TMM normalization method to the S. pneumoniae R6 datasets resulted in better separation of essential and non-essential genes judged from the increase in AUC values (Table 1; Fig. 1A). Application of the RLE method gave near-identical results compared to TMM normalization (results not shown).
and per gene (Fig. 2B) was alleviated without the requirement of parameter optimization. Using this approach, a better separation of non-essential and essential genes was observed (Table 1, Fig. 1A and B) for both S. pneumoniae R6 and P. aeruginosa PAO1, reflected in decreased p-values and a perfect AUC for S. pneumoniae R6 and a higher AUC than the prediction by Gallagher et al. [10]. Significant difference between the essential and non-essential gene distributions is shown by *(p<0.01).

doi:10.1371/journal.pone.0043012.g001

ESSENTIALS can be Used for different Transposon Mutagenesis Sequencing Methods

In order to validate ESSENTIALS as a tool for general use we have analyzed several transposon sequencing literature datasets obtained from SRA at EBI [33] that were based on different transposon insertion sequencing technologies. We compared the
results of ESSENTIALS on gene essentiality with the results presented in the respective studies using either the optimal fold change under-representation cut-off predicted by ESSENTIALS or the same fold change cut-off as applied in the study [34–36] and a FDR corrected p-value <0.05 cut-off. Because the results obtained with the reference sets (Table 1) produced, in the case of

Figure 2. Read count as a function of genomic position per 1 kb. Read count of a single Tn-seq experiment of S. pneumoniae R6 gene essentiality as a function of the genomic position before (A) and after (B) genomic location correction using Loess. Each dot represents 1 kb of sequence. Regression on the data was performed using Loess as implemented in the loess R package and plotted on the graph as a black line. doi:10.1371/journal.pone.0043012.g002
S. pneumoniae R6, a 100% correct prediction, and for P. aeruginosa PAO1, a better prediction than presented by the original authors [10] we are confident in suggesting ESSENTIALS performs equally well or better in detecting (conditionally) essential genes using the optimal settings (Table 1, in bold). Although a direct comparison of the number of false positives and false negatives cannot be made because of the arbitrary fold change and P-value cut-offs or different analysis algorithms used in the studies describing these datasets, ESSENTIALS allowed determination of the vast majority of the previously reported conditionally essential genes (Table 2). Additionally, the flexibility of ESSENTIALS in selecting how to process the various read file formats allows the analysis of all known transposon insertion sequencing analysis methods.

Methods

ESSENTIALS is implemented in Perl v5.8.8 and R v2.14.1. Its web interface is generated by the FG-web framework (van Hijum et al., https://trac.nbic.nl/fgweb/). The ESSENTIALS algorithm downloads and preprocesses sequencing read files, matches sequenced transposon flanking reads to insertion sites on the genome and then performs various processing steps and statistics on the resulting transposon insertion count data. Via email the progress of the run is reported.

There are three major sections in the web-interface: (i) configuration file upload and genome selection; (ii) parameter settings; and (iii) displaying the results. The web tool works with major web browsers such as Internet Explorer, Firefox, Safari and Opera. It can be tuned to the needs of a researcher by modifying several parameters controlling the alignments, normalization, statistical tests and visualization.

Input Data

A simplified flow chart of the procedure followed by ESSENTIALS is shown in Figure 3. The genome of the organism that was used to create the knockout library for the Tn-seq can be selected either by (i) selecting from the available daily updated Genbank sequences or, (ii) uploading a Genbank file. A tab-delimited configuration file should be uploaded or can be generated using the ESSENTIALS web-interface. The configuration file should contain the following information: (i) a hyperlink to the sequence reads, (ii) the barcode sequence, if used, (iii) transposon sequence (if used), (iv) condition, (v) knockout library, (vi) sequence file format (FASTA, FASTQ, EXPORT, SCARF, CsfASTA, BAM, SAM or a custom tab delimited file) and (vii) used compression (none, gzip, gz, bzip2). After supplying ESSENTIALS with the sample descriptions and additional information the pre- and post-processing steps of the algorithm can be set by the user and the analysis can be started.

Algorithm

The ESSENTIALS algorithm matches sequenced transposon flanking reads to insertion sites on the genome and then performs statistics on the resulting counts data. A stepwise explanation of the procedure follows below. When a BAM, SAM or a tab delimited file containing the read counts per insertion site is uploaded, the filtering and aligning procedures are skipped and the count data per gene is generated from the user-supplied mappings.

Step 1. All (unique) putative insertion sites (either random or mariner transposon TA insertion sites) are detected on the genome of interest and sequences adjacent to each insertion site are entered into a FASTA file.

Step 2. Sequence read files are downloaded, uncompressed and transformed to FASTA using a modified version of fq_all2fast from MAQ (http://maq.sourceforge.net/). Read files are split on barcodes using a modified version of fastx_split from the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) that uses Levenshtein distance instead of Hamming distance to allow fuzzy matching of barcodes of different lengths. When necessary sequence reads are filtered on the presence of transposon repeats using PASS v1.2 [37] or fastx_split for long (>12) or short transposon ends (= <12), respectively from either the 5’ or 3’ end.

Step 3. The transposon-flanking genomic sequence reads are aligned to the insertion sites using PASS and count data is generated per unique insertion site and per gene. Reads that map on more than one place on the genome can be removed from the analysis. Reads that map in the 3’ end of a gene can also be filtered out to remove transposon insertions that do not cause loss of function. Genomic location insertion and read bias is corrected by LOESS regression on read counts per insertion site and per gene relative to genomic location.

Step 4. Count data is combined for analysis with EdgeR [38] and normalization is performed using either scaling, trimmed mean of M-values (TMM), Relative Log Expression (RLE) [20] or Quantiles [39].

Step 5. Read counts per gene or per insertion site of the control and target samples are tested for significant difference to

Table 2. The use of ESSENTIALS on data generated by various transposon sequencing techniques.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Condition</th>
<th>Number of essential genes/Log2 FC cut-off</th>
<th>Literature</th>
<th>Method</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ESSENTIALS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td>FC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td>FC</td>
<td>N</td>
</tr>
<tr>
<td>S. pneumoniae TIGR4</td>
<td>essential</td>
<td>396</td>
<td>423</td>
<td>-4.1</td>
<td>357</td>
</tr>
<tr>
<td>S. typhi Ty2</td>
<td>essential</td>
<td>356</td>
<td>335</td>
<td>-3.71</td>
<td>323</td>
</tr>
<tr>
<td></td>
<td>bile salt</td>
<td>356</td>
<td>333</td>
<td>-4.00</td>
<td>161</td>
</tr>
<tr>
<td>H. influenzae Rd</td>
<td>in vivo</td>
<td>358</td>
<td>383</td>
<td>-3.2</td>
<td>344</td>
</tr>
</tbody>
</table>

*Optimal fold change (FC) underrepresentation cut-offs detected by ESSENTIALS; N: number; NA: Not available, a different method was used to determine gene essentiality in these studies.

*Although the authors state in their methods that a -2 log2 fold change and a p<1*10^-5, adj. p<.25E-4 cut-off was used, only the p-value cut-offs were applied, resulting in a -1.4 fold change cut-off (personal communication Julian Parkhill). doi:10.1371/journal.pone.0043012.t002
Obligate gene essentiality is determined by comparing the expected number of reads per gene (based on the number of insertion sites per gene, the mutant library size, and the sequencing depth) and the measured number of reads per gene. Significantly underrepresented genes are considered (conditionally) essential. ESSENTIALS uses the negative binomial distribution statistical model in EdgeR, an exact test or a General Linearized Model.
likelihood ratio test and estimates the dispersion in the data with quantile-adjusted conditional maximum likelihood (qCML) or Cox-Reid profile-adjusted likelihood (CR). The qCML method is applied to experiments with a single mutant library and the CR method with experiments with multiple libraries. The log2 transformed ratio of target over control or measured over expected and signal of (conditionally) essential genes or insertion sites is then used to generate kernel density plots using a Gaussian model with stepwise increasing bandwidth and 2048 bins until a single (in the case of essential genes) or four (in the case of conditional essential genes) local minima are found. Local minima are detected by calculating the first derivative of the density and by locating the position where it traverses from values below to values above zero. This fold change value corresponds to a value closest to the minimum between the peaks of essential and non-essential genes and can be used as a cut-off to determine whether a gene is essential or not.

Output of ESSENTIALS
Results of ESSENTIALS include: (i) Principal Component Analysis (PCA) plots, which can be used to explore the similarities between the target and control samples based on the number of reads per gene or per insertion site; (ii) Signal to log2-ratio (MA) plots which can be used to visually inspect the fold ratio of target over control or measured over expected and signal of (conditionally) essential genes or insertion sites; (iii) Density plots of the ratio of target over control or measured over expected and signal of (conditionally) essential genes or insertion sites with detected putative fold change cut-offs for selection of essential genes; (iv) tab delimited tables containing the raw counts, normalized counts, output from the statistical tests and genomic information and; (v) Links to visualize the results in MINOMICS [40], also implemented in the FG-web framework.

Availability
The web-interface of ESSENTIALS, the output from the various analyses presented in this paper and an optional demo mode, analyzing a subset of the *S. pneumoniae* R6 dataset, can be accessed freely at http://bamics2.cmbi.ru.nl/software/essentials/. Source code is available via http://trac.nbic.nl/essentials/.

References

Conclusions
Transposon insertion sequencing analysis is becoming the default method of high-throughput fitness screening in prokaryotes. Emergence of next-generation sequencing based approaches such as transposon insertion sequencing analysis and high-throughput random RNAI interference screens in eukaryotic organisms [41,42] will lead to similar data types. ESSENTIALS provides an easy to use and automated method to rapidly analyze these datasets. Prediction of gene essentiality by ESSENTIALS is comparable or possibly better than that reported by the original authors because ESSENTIALS applies data filtering, normalization and suitable statistical tests that are optimized to recover as many as possible essential genes. ESSENTIALS will greatly benefit researchers performing these studies saving both time and providing robust, yet sensitive detection of essential genes from transposon insertion sequencing analysis experiments.

Supporting Information
Table S1 Input data for ROC analysis and statistical tests for performance evaluation of *S. pneumoniae* R6 gene essentiality prediction. (XLSX)
Table S2 Input data for ROC analysis and statistical tests for performance evaluation of *P. aeruginosa* PAO1 gene essentiality prediction. (XLSX)

Acknowledgments
We thank Stefan de Vries and Jeroen Langereis for thoroughly testing the implementation.

Author Contributions
Conceived and designed the experiments: AZ, PB HJB PVMH SAFTvH. Performed the experiments: PB HJB. Analyzed the data: AZ. Contributed reagents/materials/analysis tools: PWMH SAFTvH. Wrote the paper: AZ PB HJB PVMH SAFTvH. developed the FGweb framework used in this study: SvH.


