FiTnEss - a novel statistical method for identification of essential genes in bacteria from Tn-Seq data

RUI YANG[1], BRAD POULSEN[1,2,3], TIANTIAN WHITE[3], NOAM SHORESH[1], DEBORAH HUNG[1,2,3]

1 Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, Massachusetts 02142, United States
2 Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, United States
3 Department of Molecular Biology and Center for Computational and Integrative Biology, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, United States

Overview

Background: Transposon insertion sequencing (Tn-Seq) is a high-throughput technique to generate and culture a large random collection of genetically perturbed bacteria, followed by sequencing to measure the frequency of each perturbation in the culture.

Mutants carrying insertions in essential genes are expected to be greatly depleted in the growth culture.

Aim: Identify essential genes using transposon insertion sequencing data.

Method: *FiTnEss* (Finding Tn-Seq Essential Genes) – a novel statistical method with two global parameters to identify gene essentiality using Tn-Seq data.

Application: Implemented *FiTnEss* on large scale cross-sectional data of *Pseudomonas aeruginosa,* and successfully characterized 321 core essential genes, validated by single-gene deletion experiments.

Transposon Insertion Sequencing

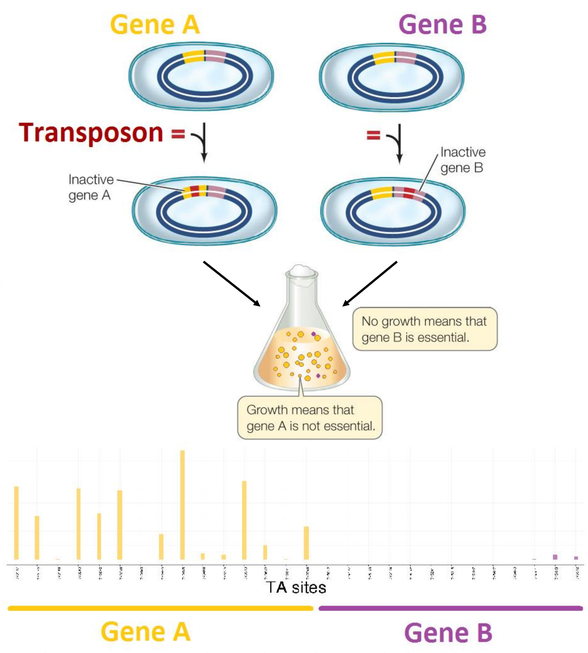

Figure 1. Transposon inserts in random TA site in each gene (from Hillis, Principles of Life, Figure 12.8)

Method

Approach 1
Threshold by average reads per gene

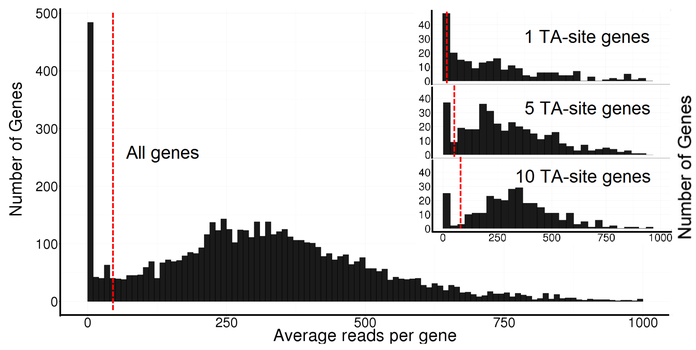

Figure 2. Average reads per gene

Problem: distribution is gene-size dependent
Unable to use a single threshold across all genes

Approach 2
Reads per sampled 1 TA site from each gene

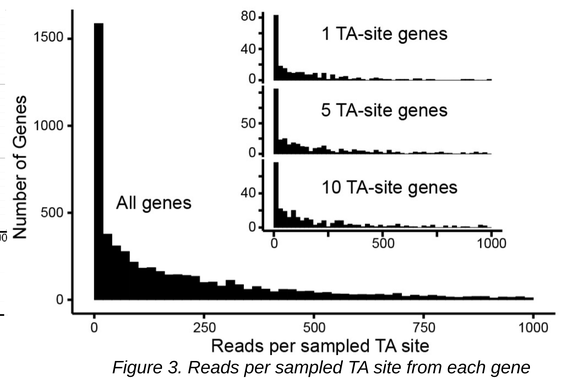

Figure 3. Reads per sampled TA site from each gene

Advantage: gene-size independent
Consistent distribution across all genes

Model Design

Assumptions and Model Design

The read counts at the different TA sites of a gene are assumed to be all drawn from the same distribution, whose mean reflects the fitness cost of deleting the gene (high cost → less reads):

$$x_{g,i} \sim Geo(p_g),$$
$$\text{for a specific gene } g, \text{ for } i = 1, \ldots, N_{TA}.$$

For non-essential genes, we assume that the inverse of $p_g$comes from a log-normal distribution

$$p_g{}^{-1} \sim Lognormal(\mu, \sigma)$$

with parameters μ, σ.

The N_{TA} - dependent total number of reads in each gene is then captured as negative binomial distribution, which is the sum of geometric:

$$\sum x_{g,i} \sim \sum Geo(p_g)$$
$$Total\ reads\ per\ gene \left(\sum x_{g,i}\right) \sim NB(p_g, N_{TA})$$

After obtaining (μ, σ), we are then able to capture the non-essential distributions across all genes.

Fitting Model Parameters

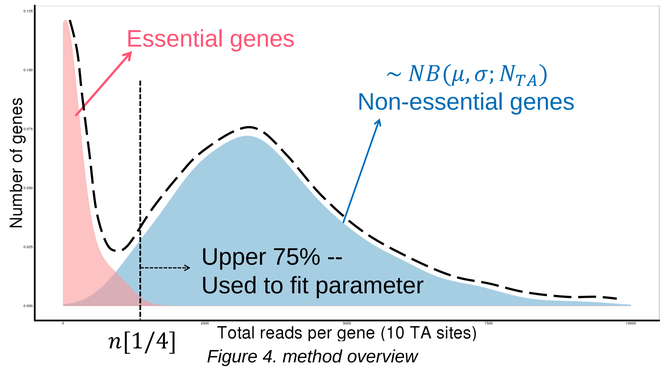

Figure 4. method overview

The empirical distribution gets contributions from essential (red) and non-essential (blue) genes. To fit μ and σ, which parametrize the non-essential part, we ignore the lower 25% of the data

Cramer-von Mises criterion:

$$\omega^2 = \int_{n[1/4]}^{+\infty} [F_n - F_n^*]^2 dF^*$$

Minimizing the distance (ω^2) between the empirical cumulative mass function (F_n) and the simulated one (F_n^*), we obtain estimates of μ and σ.

Finding Essential Genes

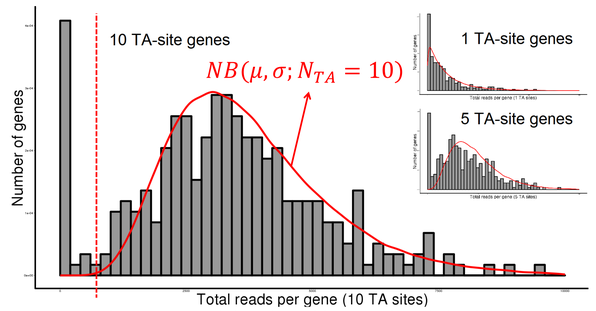

Figure 5. Actual data (gray histogram) and simulated non-essential distribution (red curve) for 10-TA genes

After obtaining optimized parameters, we constructed non-essential distribution for all N_{TA} (number of TA sites in each gene) categories.

Each gene is then tested on this non-essential distribution. Genes with adjusted p-value smaller than 0.05 in both replicates are identified as "confident essential" (FWER) or "candidate essential" (FDR).

Application

We implemented *FiTnEss* on 9 strains of *Pseudomonas aeruginosa* under 5 biological conditions, and successfully characterized its core essential genome (321 genes).

Single-gene deletion experiments for validation:

	FiTnEss Essentiality Prediction		
Actual Growth	Confident (35)	Candidate (15)	Not (65)
Essential	86% (30)	27% (4)	0% (0)
Growth-defective	14% (5)	27% (4)	9% (6)
Non-essential	0% (0)	46% (7)	91% (59)

Table 1. Validation results

Conclusion

In this study, we developed a novel statistical method (*FiTnEss*) to identify essential genes using TnSeq data.

Manuscript on *BioRxiv* (QR Code)

A Bioconductor package for *FiTnEss* is under development.