23

Statistical Evaluation of Genetic Footprinting Data

Gábor Balázsi

Summary

As transposomics is extended to genome scale, appropriate statistical methods need to be developed to assign significance to gene essentiality. In this chapter, the author presents a set of steps that, together with genome-scale insertion data and the complete genome sequence of a prokaryote, can be used to classify the genes of the organism as either "essential" or "nonessential."

Key Words: essentiality; genetics; insertion; mutagenesis; Poisson distribution; significance; transposomics.

1. Introduction

The number of genes in prokaryotes can reach a few thousand ***(1–3)***, but many of these genes are dispensable. Identifying the genes that are essential in various conditions can result in a better understanding of prokaryotic biology, a better functional annotation of gene products, and the development of more efficient antibiotics.

One of the genome-wide gene essentiality screens used a Tn*5*-based transposome mutagenesis system and identified 620 essential genes and 3126 nonessential genes in *Escherichia coli* (***[4]*** and **Chapter 6**). With the extension of transposomics to genome scale, it becomes crucial to develop statistical methods to reliably identify essential genes and assign significance to essentiality calls.

A statistical approach to transposomics is presented in the next section. This approach assumes that insertions are random events that resemble a Poisson process over large portions of the chromosome. The author discusses two biological factors that influence the validity of this assumption: variation of insertion density along the chromosome and the contribution of essential genes to reduce the number of insertions. The possible pitfalls of the technique are discussed briefly at the end of the chapter.

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

2. Materials

In addition to a workstation that can be programmed in a programming language such as C, Perl, or Java, the following data are needed to identify the essential genes of a prokaryote:

1. Transposon insertion locations for the whole genome.
2. A completed genomic sequence of the prokaryote.
3. The most complete annotation of all open reading frames (ORFs) in the genome.

3. Methods

The basic assumption of transposon mutagenesis is that trasposon insertions occur randomly and with uniform density throughout the chromosome. After mapping the insertions along the chromosome, genes without insertions are likely candidates to be essential. However, genes can also be missed by chance, and labeling all genes without insertions as "essential" will generate many false positives. It is therefore necessary to reduce the number of false positives by assigning significance to genes with no insertions.

Intuition tells us that if a gene is very short, or if the insertion density is very low, the gene can easily be missed by insertions. In general, if the insertion density is r, the probability of N insertions occurring within a DNA region of length L is given by the Poisson distribution ***(5)***:

$$P_N(L) = \frac{(rL)^N}{N!} e^{-rL}, \quad (1)$$

and therefore, the probability to have no insertions in a gene of length L (measured in base pairs) is

$$P_0(L) = e^{-rL}. \quad (2)$$

If the insertion density r were known, this formula could be used to determine the significance of essentiality calls. However, r is unknown, and therefore it has to be determined prior to the classification of genes according to their essentiality.

The simplest way to determine the insertion density r might be to divide the total number of insertions N_T mapped around the chromosome by the length of the full chromosome, L_T:

$$r = \frac{N_T}{L_T}. \quad (3)$$

However, this simplistic approach could be misleading for two reasons. First, nothing guarantees that the insertion density along the chromosome is constant (**Note 1** and **Fig. 1A**). Second, since essential genes on the chromosome exclude insertions, **equation 3** will underestimate the insertion density (**Note 2** and **Fig. 1A**).

To avoid the first problem (variation of insertion density along the chromosome), r should be estimated locally instead of globally. To estimate r locally, the number of insertions should be determined within a DNA region surrounding the gene, rather than the whole chromosome. To avoid the second problem (the bias introduced by essential

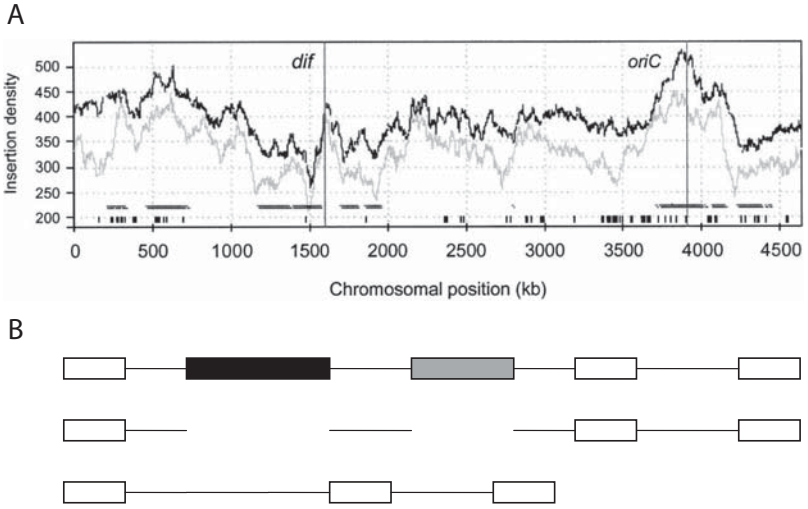

Fig. 1. **(A)** Distribution of transposon insertion densities along the *E. coli* chromosome. Gray lines show the transposon insertion densities calculated as the number of transposition events per 100-kb sliding window over the entire *E. coli* MG1655 chromosome. Values indicated by the black lines were computed in a similar manner, except that all chromosomal regions corresponding with essential and ambiguous genes were excluded from the calculations in order to reconstruct insert distribution prior to selective outgrowth. Gaps in the data (chromosomal regions where transposition events could not be detected due to technical reasons) are indicated by short vertical lines along the x axis. The regions where the distributions of transposition events significantly deviate ($p < 0.01$) from a Poisson process are marked by horizontal double lines. *OriC* shows the origin of chromosomal replication, and *dif* denotes the *dif* locus within the replication termination area. (Reprinted from Ref. ***4*** with permission from American Society for Microbiology.) **(B)** Correcting the bias introduced by essential genes. For the estimation of transposon insertion density within a DNA region, genes with no insertions (or, ideally, all known ORFs) should be left out from the analysis to eliminate the bias of essential genes, which exclude insertions. Shading indicates nonessential genes (white), essential gene (black), and gene with no insertions—a new candidate for essentiality (gray).

genes), the insertion density should be determined only within noncoding regions along the chromosome. This will ensure that essential genes will be excluded and will not cause a bias in the insertion density (**Notes 3** and **4**).

How long should the chromosome region be for a reliable local estimation of the insertion density? Insertion density is estimated by counting the number N of insertions and dividing it by the length L of the DNA in which they occur:

$$r_{est} = \frac{N}{L}. \tag{4}$$

As one would expect, the average of r_{est} is

$$\langle r_{est} \rangle = \frac{\langle N \rangle}{L} = \frac{1}{L} \sum_{N=0}^{\infty} N P_N(L) = r. \tag{5}$$

However, even if the rate of insertions is constant along the chromosome, the number of insertions in DNA segments of identical length L will fluctuate around rL because of the random nature of insertions events. As a consequence, there will be an error in determining r_{est}. The magnitude of this error can be measured by the variance:

$$\langle r_{est}^2 \rangle - \langle r_{est} \rangle^2 = \frac{\langle N^2 \rangle - \langle N \rangle^2}{L^2} = \frac{r}{L}. \quad (6)$$

According to this formula, the error committed in the estimation of r is higher for short DNA regions. Therefore, the DNA region should be as long as possible without being influenced by regional fluctuations of the insertion density along the chromosome (**Note 4**).

To summarize, for a proper assessment of gene essentiality, the following steps should be taken:

1. Select a gene with no insertions.
2. Exclude all the known ORFs from the DNA (or all genes with no insertions) surrounding the gene to minimize the bias introduced by essential genes, which reduce insertion density (**Note 4** and **Fig. 1B**).
3. Paste together the DNA fragments remaining after the exclusion of all coding regions until the desired length L is reached. The region used to determine the local density should be as long as possible without being affected by fluctuations of insertion density along the chromosome.
4. Using the noncoding DNA, determine the local density of insertions around the gene.
5. Use **formula 2** and the local insertion density r to determine the probability for the gene to be missed by chance alone.
6. Establish a cutoff (**Note 5**). If $P_0(L) < c$ (the probability of being missed by chance is below the cutoff) label the gene as "essential." Otherwise, label the gene as "nonessential."
7. Repeat **steps 1** to **4** for all genes and for various values of L and c (**Note 5**).

Notes

1. DNA replication is a known factor that could result in a location-dependent insertion density. In exponential growth, bacteria are known to initiate a new round of replication before the previous round has terminated ***(6)***. Therefore, it is possible to have 2, 4, 8, or even 16 copies of the origin of replication compared with the terminus. As a result, a higher amount of DNA is available for insertion around the origin, and therefore insertion density is expected to be highest around the origin and decreasing toward the terminus. This has indeed been observed in the genome-scale footprinting study (***[4]*** and **Chapter 6**).
2. Comparing the insertion density along the *E. coli* chromosome with the insertion-free coding regions included and excluded reveals that r is higher for the latter throughout the chromosome ***(4)***. The difference between the two estimates of the insertion density is highest near the origin and lowest near the terminus, which could be explained by the higher density of essential genes near the origin of replication ***(7, 8)***.
3. The percentage of coding DNA is much higher in prokaryotes than in higher organisms, and therefore excluding all known ORFs from the DNA might reduce the remaining amount of DNA too much and might lead to poor statistics. An alternative could be to exclude only the ORFs with no insertions from the DNA, but this could artificially increase the local insertion density.

4. The density of genes in some chromosomal regions is higher. In this case, by excluding the coding regions and pasting together the noncoding DNA, the distance from the assessed gene might increase too much. To avoid this problem, a critical distance could be established that cannot be exceeded when estimating insertion density around a gene. This will also result in a maximum limit of L, the number of base pairs used for the estimation.
5. The value of the cutoff c used to classify genes as "essential" or "nonessential" and the length of the DNA region used to determine the insertion density are somewhat arbitrary. Essentiality calls should be confirmed by alternate experimental methods to find the optimal value of c and L. Typically, L = 10,000 base pairs and c = 0.01 are acceptable values to start the analysis.

References

1. Deng, W., Burland, V., Plunkett, G. 3rd, Boutin, A., Mayhew, G. F., Liss, P., et al. (2002) Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.* **184**, 4601–4611.
2. Tettelin, H., Nelson, K. E., Paulsen, I. T., Eisen, J. A., Read, T. D., Peterson, S., et al. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae. Science* **293**, 498–506.
3. Blattner, F. R., Plunkett, G. 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474.
4. Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., et al. (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
5. Harris, J. W., and Stocker, H. (1998) *Handbook of Mathematics and Computational Science*, 1st ed., New York: Springer-Verlag.
6. Donachie, W. D. (1968) Relationship between cell size and time of initiation of DNA replication. *Nature* **219**, 1077–1079.
7. Rocha, E. P. (2004) The replication-related organization of bacterial genomes. *Microbiology* **150**, 1609–1627.
8. Rocha, E. P. (2004) Order and disorder in bacterial genomes. *Curr. Opin. Microbiol.* **7**, 519–527.