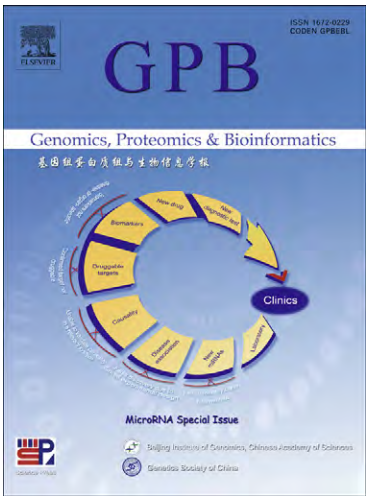
Accepted Manuscript



Global Genomic Arrangement of Bacterial Genes Is Closely Tied with the Total Transcriptional Efficiency

Qin Ma, Ying Xu

PII: S1672-0229(13)00008-9
DOI: http://dx.doi.org/10.1016/j.gpb.2013.01.004
Reference: GPB 52

To appear in:

Please cite this article as: Q. Ma, Y. Xu, Global Genomic Arrangement of Bacterial Genes Is Closely Tied with the Total Transcriptional Efficiency, (2013), doi: http://dx.doi.org/10.1016/j.gpb.2013.01.004

Global Genomic Arrangement of Bacterial Genes Is Closely Tied with the Total Transcriptional Efficiency

Qin Ma[1], Ying Xu[1,2,3,*]

[1]*Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA*

[2]*BioEnergy Science Center (http://bioenergycenter.org/), USA*

[3]*College of Computer Science and Technology, Jilin University, Changchun 130012, China*

[*] Corresponding author.

E-mail: xyn@bmb.uga.edu (Xu Y)

Running title: *Ma Q and Xu Y/Genomic Arrangement Analysis of Bacterial Genes*

Abstract

The availability of a large number of sequenced bacterial genomes allows researchers not only to derive functional and regulation information about specific organisms but also to study the fundamental properties of the organization of a genome. Here we address an important and challenging question regarding the global arrangement of operons in a bacterial genome: why operons in a bacterial genome are arranged in the way they are. We have previously studied this question and found that operons of more frequently activated pathways tend to be more clustered together in a genome. Specifically, we have developed a simple sequential distance-based pseudo energy function and found that the arrangement of operons in a bacterial genome tend to minimize the clusteredness function (C value) in comparison with artificially-generated alternatives, for a variety of bacterial genomes. Here we extend our previous work, and report a number of new observations: (a) operons of the same pathways tend to group into a few clusters rather than one; and (b) the global arrangement of these operon clusters tend to minimize a new "energy" function (C^+ value) that reflects the efficiency of the transcriptional activation of the encoded pathways. These observations provide insights into further study of the genomic organization of genes in bacteria.

Keywords: Global genomic arrangement; Bacterial genome; Chromosomal supercoils; Biological pathways; Gene expression

Introduction

Up till very recently, our understanding about the global arrangement of operons in a bacterial genome has been very limited, mainly because of the lack of relevant data. The availability of both genomic and transcriptomic data for a large number of bacterial organisms (http://www.ncbi.nlm.nih.gov/genome/browse/) provides unprecedented opportunities for researchers to derive novel and deeper understanding about what determines the organization of the genomic information (such as operons and associated regulatory systems) in a genome. Understanding such an issue not only satisfies our scientific curiosity but also provides useful information that can guide the synthesis of artificially-designed chromosomes and organisms in the emerging field of synthetic biology [1-3]. A number of computational studies have been carried out aimed at understanding the genomic organization of genes. These include the studies of gene expression periodicities along a bacterial genome [4-8], transcriptional regulation [9-11], functional relatedness [5, 12, 13] and co-evolution relationship *versus* genomic co-locations [14]. While substantial efforts have been invested in studying the genome organization since early 90's [15-18], our understanding about the factors and rules that may determine the global organization of a genome is still fragmented.

We have previously proposed the first model aimed to explain the organizational principle of a bacterial genome [19]. Our main finding was that the global arrangement of operons in a bacterial genome is tightly determined by the activation frequencies of the biological processes encoded in the genome, including metabolic and transcriptional regulation pathways [19, 20]. This relationship can be captured to a large degree using a simple mathematical function, based on the sequential relationships among the relevant operons [19].

Here we extend our previous work to provide new and deeper understanding about the global arrangement of operons. Specifically we developed an improved scoring function which makes the current genomic arrangement of operons stand out more substantially against the alternative arrangements. This allows us to discover that operons in the same metabolic or transcriptional regulation pathway tend to group into a few, rather than one, sequentially nearby clusters, and the global arrangement of these operon clusters tends to make the transcriptional activation of the

encoded pathways as efficient as possible.

Results

We have carried out our study on *E. coli* K-12, which has 4144 protein-encoding genes [21] and 2462 predicted operons [22]. The 347 biological pathways encoded in *E. coli* K12 are retrieved from EcoCyc [23] and expression data for each gene under 466 growth conditions from the M3D database [24] are used to estimate the activation frequency of each pathway.

Operons encoding the same pathways tend to be clustered locally, but not globally, along the genome

We have previously used the following function to measure the global clusteredness of the component operons of the i^{th} pathway, c_i [19],

$$c_i = \sum_{j=1}^{M_i} d_{ij} \qquad (1)$$

Where M_i represents the number of operons in the i^{th} pathway and d_{ij} is the distance between the j^{th} operon and $[M_i/2]^{th}$ operon (*i.e.*, median operon) in the i^{th} pathway. Then the overall clusteredness of the operons encoding all the pathways, C, can be measured using the following function:

$$C = \sum_{i=1}^{N} c_i \qquad (2)$$

where N is the number of pathways under consideration (N = 347 in this study). A key discovery that we made was that the current genomic locations of all the operons in *E. coli* K12 tend to minimize this function in comparison with artificially-generated alternatives [19].

Our recent study indicates that while our initial one-cluster model per pathway provides an informative approximation, it is too crude. Our current analysis suggests that each pathway, particularly the one encoded by a large number of operons, tends to consist of a few operon clusters along the genome, as illustrated in **Figure 1** (taking the four largest pathways in EcoCyc as examples). Based on this realization and analysis of the M3D gene expression data, we hypothesize that all the operons are arranged at specific genomic locations to facilitate the most efficient transcription of the operons encoding each pathway during the life cycle of the bacteria.

We aim to use the following C^+ function (to distinguish with previous C value, C^+ was used here) to assess the validity of this hypothesis by identifying a partition of the genome into set of genomic segments such that the function is minimized:

$$C^+ = \sum_{i=1}^{N} f_i \times (t_i + \alpha \sum_{j=1}^{M} \omega_{ij}) \qquad (3)$$

where N is the number of known pathways encoded in the target genome; M is the number of partitioned segments of the genome to be determined; f_i represents the activation frequency of the i^{th} pathway estimated from the gene expression data in M3D following the method described previously [19] (with default value 1 for each pathway if such information is not available); t_i is the number of partitioned segments containing operons of the i^{th} pathway; s_j is the number of operons in the j^{th} partitioned segment ($j \in [1, M]$), and s_{ij} is the number of operons of the i^{th} pathway covered in the j^{th} partitioned segment; ω_{ij} is defined to be 0 if $s_{ij} = 0$; otherwise $\omega_{ij} = (s_j - s_{ij})/s_j$; and α is a scaling factor whose default value is set as 1. Intuitively, this function measures the total number of partitioned segments that encode the operons of each pathway across all the pathways, along with a measurement of the density of such operons among all the operons encoded in each partitioned segment. A detailed dynamic programming procedure for calculating an optimal partition of a target genome, along with the associated M value, is given in Materials and methods section.

We have randomly reshuffled the genomic locations of X% (X= 10, 20, …, 100) of all the *E. coli* K12 operons with the locations of the remaining operons fixed, and calculated the C^+ value of each reshuffled genome. We reshuffled the genomic locations of selected operons 100,000 times for each X, as we did in our original study [19]. **Figure 2A** shows the C^+ value distribution for all the selected Xs *versus* the current genome of *E. coli* K12. Clearly we see that the C^+ value of the current genome tends to be smaller than the C^+ values of the alternatives, *i.e.*, the artificially-generated genomes. To compare the current scoring function with our previous scoring function C, we did the same calculation using the previous function C (shown in **Figure 2B**). We have estimated the tail probability of "X < the dash line" (dash line denotes the current genomic arrangement), *i.e.*, a *P*-value, for each score distribution in Figure 2, which is summarized in **Table 1**.

Through the above analysis, we have demonstrated that the current genomic arrangement of operons in the genome of *E. coli* K12 tends to minimize the C^+ value compared to those of the alternative genomic arrangements of operons; and the C^+ function makes the actual genome stands out more significantly than previous C function against the alternatives. The effect of increasing the impact of ω_{ij} defined in (3) or estimating the actual pathway activation frequency based on M3D microarray data is presented in Figure S1 and Table S1.

Pathway-based analysis of the genomic partition

The above calculation also gives rise to an optimal partition of the *E. coli* K12 genome into 104 segments, which optimizes the C^+ value (see details in Table S2). We have examined how the operons encoding the 347 biological pathways are located in the 104 partitioned segments of the *E. coli* genome. For each of the randomly-shuffled genomes and the original genome of the *E. coli* K12 shown in Figure 2, we retrieved the corresponding partitioned segments and calculated the total number of partitioned segments covered by each pathway across all the 347 pathways under consideration. From **Figure 3**, we can see that this total number of the current genome of *E. coli* K12 is clearly smaller than any of the randomly-reshuffled genomes.

In a separate study (Ma et al., unpublished data), we provided strong evidence that each of the partitioned segments corresponds to a supercoiled domain in the folded chromosome of *E. coli* K12[25-28], each being independently foldable. Clearly when a pathway needs to be activated, the (folded) segments containing operons of the pathway need to unfold first to make all the relevant operons transcriptionally accessible, a process that consumes energy. The data shown in Figure 3 suggests that the organism has evolved to minimize the total number of segments that need to be unfolded in order to activate each pathway, hence making this part of the operating cost for the living bacteria as small as possible.

Concluding remarks

We have previously demonstrated that operons encoding more frequently activated pathways tend to be more clustered together in a genome [19]. We have also noted that what prevents such operons from forming one cluster is that many individual operons may each be involved in multiple pathways. Hence they form multi-operon clusters as we have demonstrated in this work that they tend to form a few operon clusters each contained in a folded genomic segment. To

activate a pathway, energy is required to unfold the genomic segments that contain the operons encoding the pathway, so that the relevant operons are transcriptionally accessible. By putting all these results together, we speculate, with strong evidence, that operons encoding various pathways are specifically located in a given genome such there is a tendency to minimize the overall energy needed to have the relevant pathways transcribed, during the life cycle of the bacteria.

Materials and methods

A dynamic programming method for genomic partition

The optimization problem defined in (3) can be solved using a dynamic programming approach. The following recurrence equation gives rise to a simple dynamic programming algorithm for finding the optimal $C^*(S,E)$ value, which generates a partition of a target genome and hence the number M of the partitioned segments, where S and E represent the first and the last inter-operonic regions of the genome, respectively.

$$C^*(S,t) = \min_{j \in [t-U, t-L]} [C^*(S,j) + C^*(j,t)] \qquad (4)$$

where L and U are set at 10k bp and 100k bp, defining the lower and the upper bounds of a partitioned segment [10-12].

Authors' contributions

YX conceived the basic idea and planned the project. QM developed the algorithm, analyzed the data and carried out the computational experiments. Both authors edited the manuscript and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

Acknowledgements

This research was supported in part by National Science Foundation (Grant No. NSF DEB-0830024 and NSF MCB-0958172) and by grant from the BioEnergy Science Center (BESC) of

US Department of Energy through the Office of Biological and Environmental Research.

References

1. Peccoud J, Isalan M. The PLOS ONE synthetic biology collection: six years and counting. PLoS One 2012;7:e43231.
2. Képès F, Jester BC, Lepage T, Rafiei N, Rosu B, Junier I. The layout of a bacterial genome. FEBS Lett 2012;586:2043-8.
3. Juhas M, Eberl L, Church GM. Essential genes as antimicrobial targets and cornerstones of synthetic biology. Trends Biotechnol 2012;30:601-7.
4. Campbell AM. Genome organization in prokaryotes. Curr Opin Genet Dev 1993;3:837-44.
5. Sukhodolets VV. Principles of prokaryotic genome organization. Genetika 1992;28:28-37.
6. Pavitt GD, Higgins CF. Chromosomal domains of supercoiling in *Salmonella typhimurium*. Mol Microbiol 1993;10:685-96.
7. Kolsto AB. Dynamic bacterial genome organization. Mol Microbiol 1997;24:241-8.
8. Lim HN, Lee Y, Hussein R. Fundamental relationship between operon organization and gene expression. Proc Natl Acad Sci U S A 2011 28;108:10626-31.
9. Mathelier A, Carbone A. Chromosomal periodicity and positional networks of genes in *Escherichia coli*. Mol Syst Biol 2010;6:366.
10. Wright MA, Kharchenko P, Church GM, Segre D. Chromosomal periodicity of evolutionarily conserved gene pairs. Proc Natl Acad Sci U S A 2007;104:10559-64.
11. Képès F. Periodic transcriptional organization of the *E.coli* genome. J Mol Biol 2004;340:957-64.
12. Jeong KS, Ahn J, Khodursky AB. Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. Genome Biol 2004;5:R86.
13. Janga SC, Salgado H, Martinez-Antonio A. Transcriptional regulation shapes the organization of genes on bacterial chromosomes. Nucleic Acids Res 2009;37:3680-8.
14. Kolesov G, Wunderlich Z, Laikova ON, Gelfand MS, Mirny LA. How gene order is influenced by the biophysics of transcription regulation. Proc Natl Acad Sci U S A 2007;104:13948-53.
15. Elati M, Nicolle R, Junier I, Fernández D, Fekih R, Font J, et al. PreCisIon: PREdiction of CIS-regulatory elements improved by gene's positION. Nucleic Acids Res 2012;doi:

10.1093/nar/gks1286.

16. Fang G, Rocha EP, Danchin A. Persistence drives gene clustering in bacterial genomes. BMC Genomics 2008;9:4.
17. Junier I, Hérisson J, Képès F. Periodic pattern detection in sparse boolean sequences. Algorithms Mol Biol 2010;5:31.
18. Ling X, He X, Xin D. Detecting gene clusters under evolutionary constraint in a large number of genomes. Bioinformatics 2009;25:571-7.
19. Yin Y, Zhang H, Olman V, Xu Y. Genomic arrangement of bacterial operons is constrained by biological pathways encoded in the genome. Proc Natl Acad Sci U S A 2010;107:6310-5.
20. Zhang H, Yin Y, Olman V, Xu Y. Genomic arrangement of regulons in bacterial genomes. PLoS One 2012;7:e29496.
21. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, et al. *Escherichia coli* K-12: a cooperatively developed annotation snapshot--2005. Nucleic Acids Res 2006;34:1-9.
22. Mao F, Dam P, Chou J, Olman V, Xu Y. DOOR: a database for prokaryotic operons. Nucleic Acids Res 2009;37:D459-63.
23. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, et al. The EcoCyc Database. Nucleic Acids Res 2002;30:56-8.
24. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, et al. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic Acids Res 2008;36:D866-70.
25. Browning DF, Grainger DC, Busby SJ. Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. Curr Opin Microbiol 2010;13:773-80.
26. Dillon SC, Dorman CJ. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nature Rev Microbiol 2010;8:185-95.
27. Dame RT, Espeli O, Grainger DC, Wiggins PA. Multidisciplinary perspectives on bacterial genome organization and dynamics. Mol Microbiol 2012;86:1023-30.
28. Rocha EP. The organization of the bacterial genome. Annu Rev Genet 2008;42:211-33.

Figure legends

Figure 1 An illustration of the clustering property of operons in the same pathway

The four rings from outside to inside, respectively, represent the genomic location of each operon in four metabolic pathways, namely phosphoribosyl pyrophosphate (PRPP) pathway (P1), *de novo* purine 2 pathway (P2), PWY0-781 pathway (P3) and glycolysis TCA glyox bypass pathway (P4), labeled by the black bars in the corresponding yellow rings.

Figure 2 Comparison of the C and C^+ value distributions for the actual and reshuffled genomes

A. Distributions of C^+-value for the *E. coli* K-12 genome. **B**. Distributions of C -value for the *E. coli* K-12 genome. In each panel, the *x* axis represents C or C^+ value and the *y* axis represents the density of reshuffled genomes with a specific C or C^+ value. For each distribution, *X*% (*X* = 10, 20,…, 100) of all the operons were randomly reshuffled, arranged from left to right. The C or C^+ value for the current arrangement of the operons in *E. coli* K-12 is represented by a vertical dash line.

Figure 3 Comparison of the number of segments required to activate each pathway for the actual and the reshuffled genomes

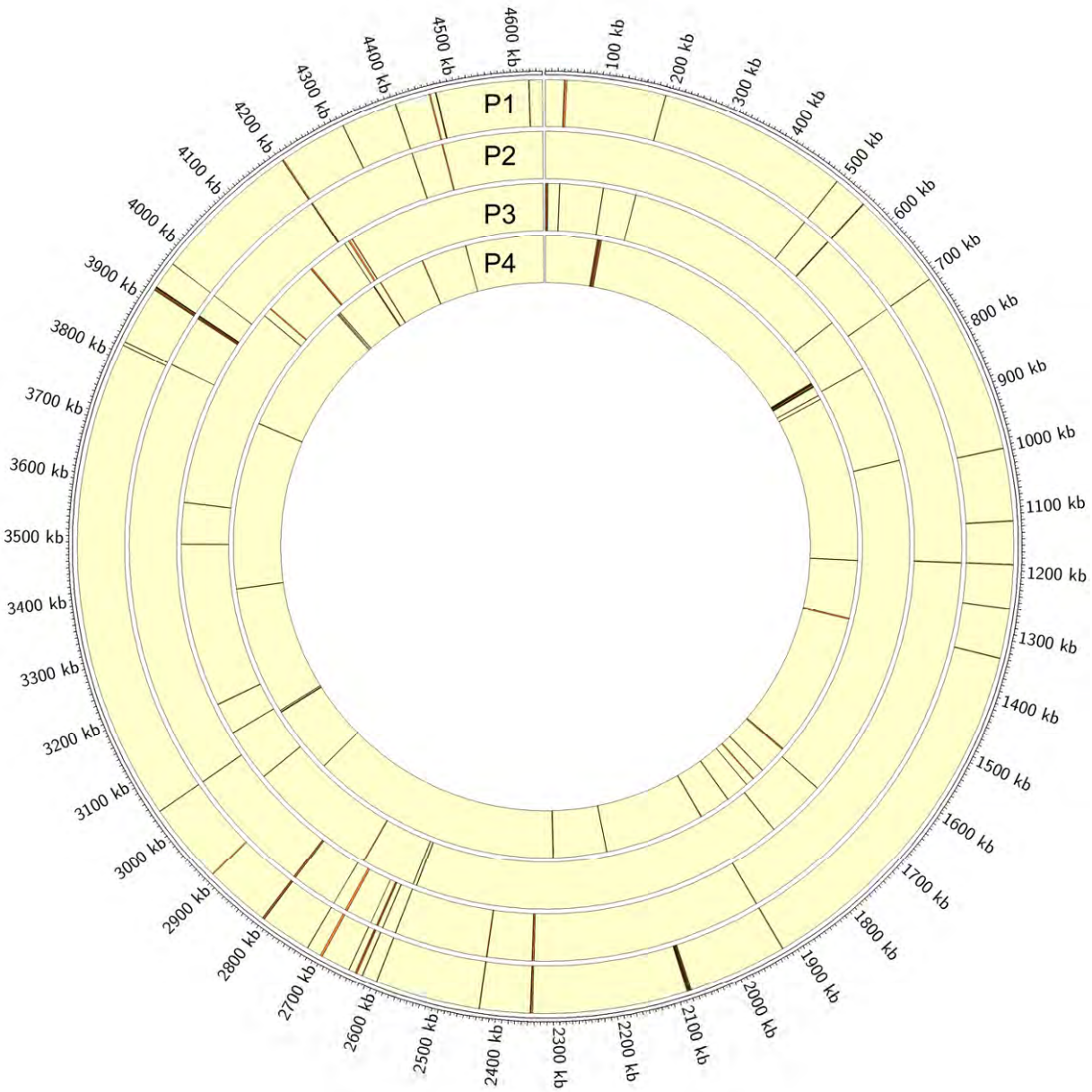
Ten boxplots for the *E. coli* K-12 genome, representing *X*% (*X* = 10, 20,…, 100) of all the operons that were randomly reshuffled, are arranged from left to right. The number of partitioned segments for the actual genome is represented by a horizontal dash line.

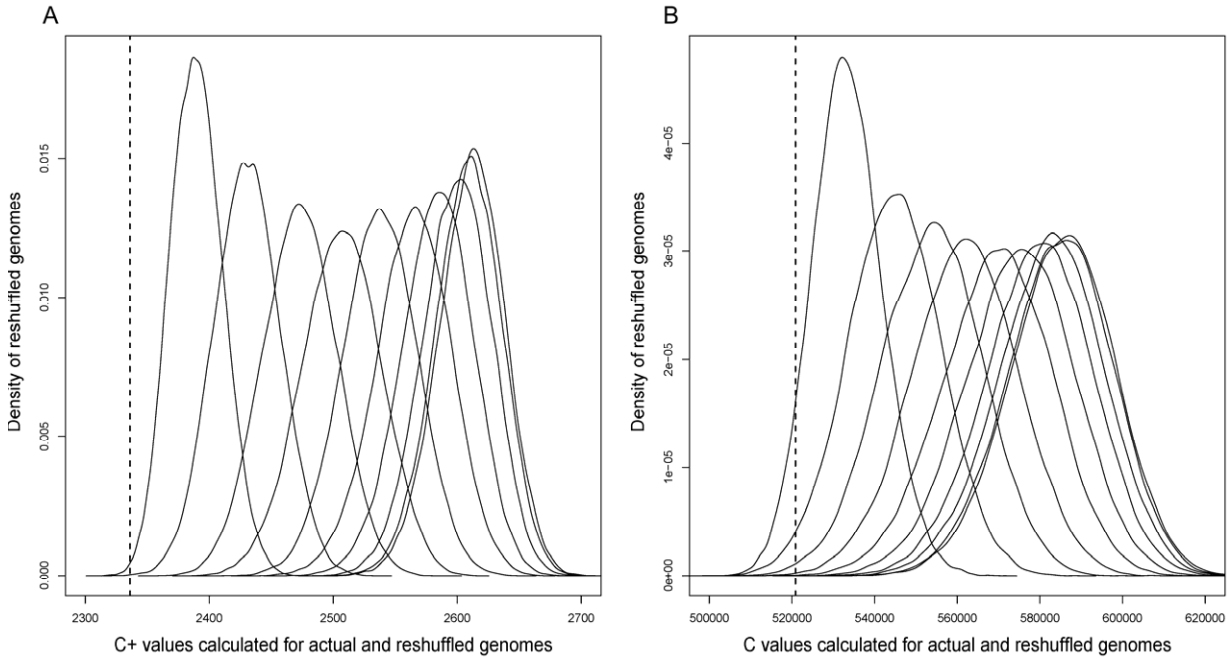
.

Supplementary material

Figure S1 C^+ value distributions for two special considerations

The distributions for actual and reshuffled genomes of C^+ values, considering α =3 (**A**) and actual values for pathway activation frequency (calculated by M3D microarray data) (**B**). In each panel, the *x* axis represents the C^+ value, and the *y* axis represents the density of reshuffled genomes with a specific C^+ value.





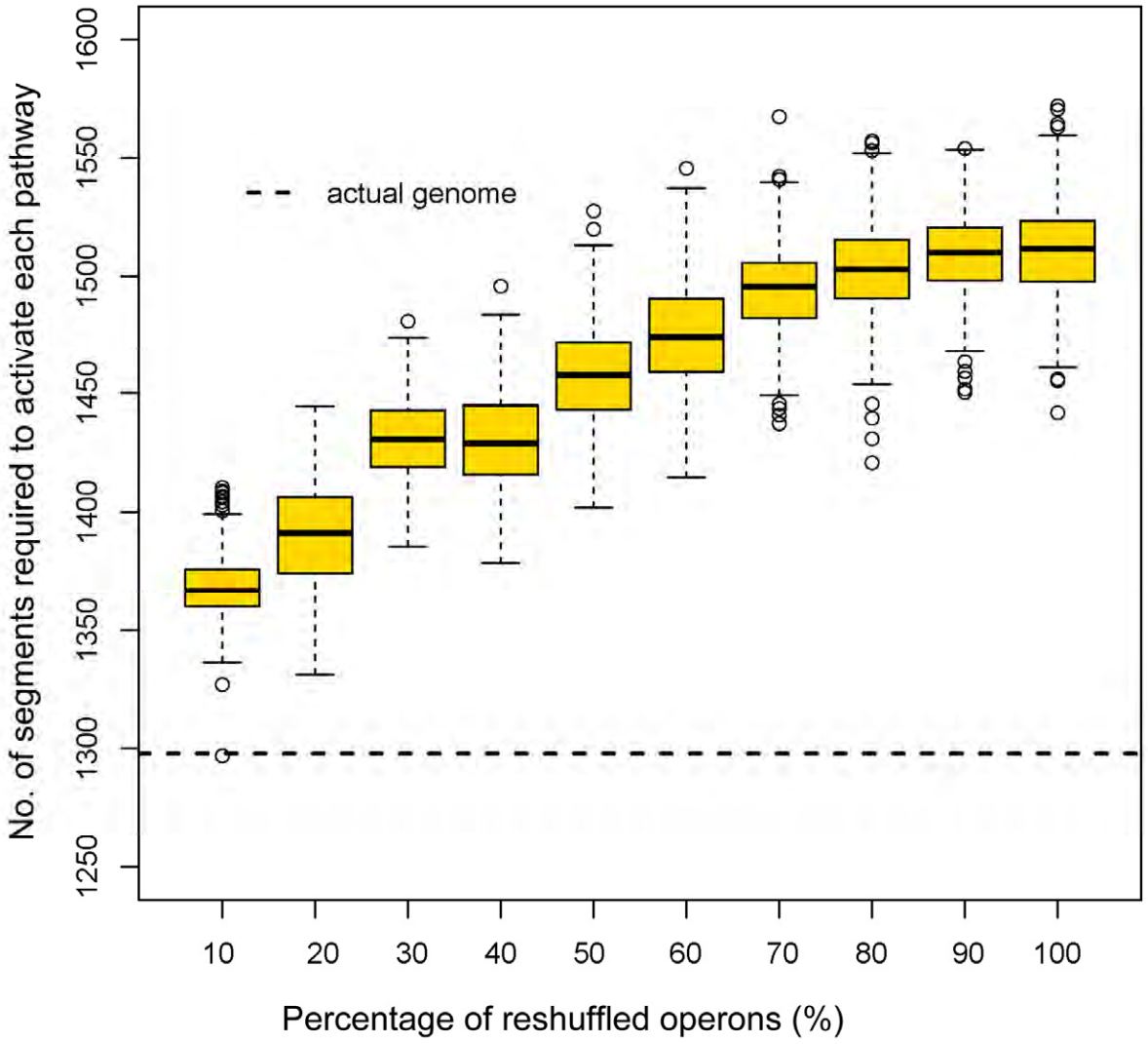


Table 1 Statistical tests of curves in Figure 2

% of reshuffling	$P(C$ value)	$P(C^+$ value)	$P(C$ value) / $P(C^+$ value)
10	0.06947282	0.004927	14.09904
20	0.01698723	0.000201	84.51098
30	0.003641975	2.09E-06	1740.591
40	0.000705596	3.44E-08	20492.59
50	0.000120973	1.42E-11	8523197
60	1.75E-05	2.72E-14	6.43E+08
70	2.77E-06	4.31E-18	6.42E+11
80	7.47E-07	6.50E-22	1.15E+15
90	2.28E-07	1.01E-24	2.25E+17
100	1.45E-07	1.09E-25	1.33E+18

Note: P (X) is the tail probability (P-value) of "X < the dash line" (dash line denotes the current genomic arrangement) for each score distribution. The *P*-values are calculated based on the observations that all the score distributions follow a normal distribution by the skewness and kurtosis test [19]