Supplementary Information: Detection of genomic islands via segmental genome heterogeneity

Aaron J. Arvey, Rajeev K. Azad, Alpan Raval, Jeffrey G. Lawrence [*]

1 Probability distribution of D^m

Consider a set of probability distributions $\{p_j\},\ j = 1, \ldots, l$, over sequences of $m+1$ letters (as described in the main text), with each distribution p_j carrying a weight factor π_j. It is convenient to view the weights $\{\pi_j\}$ as prior probabilities assigned to the distributions $\{p_j\}$. We can then write all relevant distributions in conditional form as follows.

$$\begin{aligned} \pi_j &\equiv p(j), \\ p_j(x|\mathbf{z}) &\equiv p(x|\mathbf{z}, j), \\ p_j(x, \mathbf{z}) &\equiv p(x, \mathbf{z}|j). \end{aligned} \tag{1}$$

The MJSD D^m can then be expressed in the simple form

$$D^m = \sum_{j=1}^{\ell} \sum_{\mathbf{z}} \sum_{x} p(x, \mathbf{z}, j) \log_2 \left(\frac{p(x|\mathbf{z}, j)}{p(x|\mathbf{z})} \right). \tag{2}$$

Now consider a symbol sequence that is broken up into ℓ fragments, and that each conditional distribution $p(\cdot|j)$, $j = 1, \ldots, \ell$ represents the probability of occurrence of symbols within the jth fragment. As mentioned above, the prior probabilities $p(j)$ are then usually taken to be proportional to the length of the jth fragment. Each probability in the above equation then represents a relative frequency or ratio of counts. The fundamental count of interest here is $N_{x,\mathbf{z},j}$, which denotes the number of times the word $\mathbf{z}$ is followed by symbol x in fragment j. From this, we obtain

$$\begin{aligned} p(x, \mathbf{z}, j) &= N_{x,\mathbf{z},j}/N, \\ p(x|\mathbf{z}, j) &= N_{x,\mathbf{z},j}/N_{\mathbf{z},j}, \\ p(x|\mathbf{z}) &= N_{x,\mathbf{z}}/N_{\mathbf{z}}, \end{aligned} \tag{3}$$

where $N = \sum_{x,\mathbf{z},j} N_{x,\mathbf{z},j}$, $N_{\mathbf{z},j} = \sum_x N_{x,\mathbf{z},j}$, $N_{x,\mathbf{z}} = \sum_j N_{x,\mathbf{z},j}$, and $N_{\mathbf{z}} = \sum_x N_{x,\mathbf{z}}$.

We will now proceed to show that D^m can generally be interpreted as a log likelihood ratio, just as in the $m = 0$ case [1].

[*] Corresponding author.
Email addresses: aarvey@cs.ucsd.edu (Aaron J. Arvey), rka5@pitt.edu (Rajeev K. Azad), araval@kgi.edu (Alpan Raval), jlawrenc@pitt.edu (Jeffrey G. Lawrence).

1.1 Log-likelihood interpretation of D^m

Consider the following two hypotheses about the behavior of the Markov source –

Null hypothesis H_0: The true, unknown joint distribution $q(x, \mathbf{z}, j)$ of $x, \mathbf{z}$, and j is of the separable form $q(x|\mathbf{z})q(\mathbf{z}, j)$, i.e., the distribution of a symbol only depends on the previous m symbols preceding it, not on the particular sequence fragment it lies in.

Alternative hypothesis H_1: The joint distribution $q(x, \mathbf{z}, j)$ is arbitrary.

Note that both hypotheses above are composite because they do not actually specify values of probabilities. However, H_1 has $k^{m+1}l - 1$ free parameters while H_0 has $k^m(k-1) + k^m l - 1$ free parameters.

The likelihood of the observed sequence under H_0 is given by

$$L(H_0) = \prod_{x,\mathbf{z},j} [q(x|\mathbf{z})q(\mathbf{z}, j)]^{N_{x,\mathbf{z},j}}, \tag{4}$$

while the likelihood under H_1 is given by

$$L(H_1) = \prod_{x,\mathbf{z},j} q(x, \mathbf{z}, j)^{N_{x,\mathbf{z},j}}. \tag{5}$$

Since both hypotheses are composite, we can now ask for what values of the parameters $q(\cdot)$ are the likelihoods maximized. Maximizing $L(H_0)$ under the constraints $\sum_x q(x|\mathbf{z}) = 1$ and $\sum_{\mathbf{z},j} q(\mathbf{z}, j) = 1$ leads to the maximum likelihood estimates

$$\begin{aligned} \hat{q}(x|\mathbf{z}) &= \frac{\sum_j N_{x,\mathbf{z},j}}{\sum_{x,j} N_{x,\mathbf{z},j}} = p(x|\mathbf{z}), \\ \hat{q}(\mathbf{z}, j) &= \frac{\sum_x N_{x,\mathbf{z},j}}{\sum_{x,\mathbf{z},j} N_{x,\mathbf{z},j}} = p(\mathbf{z}, j). \end{aligned} \tag{6}$$

Similarly, the maximum likelihood estimate of $q(x, \mathbf{z}, j)$ under H_1 turns out to be $p(x, \mathbf{z}, j)$. Thus defining the likelihood ratio λ as

$$\lambda = \frac{L_{\max}(H_0)}{L_{\max}(H_1)}, \tag{7}$$

we obtain

$$-\ln \lambda = N \sum_{j,\mathbf{z},x} p(j, \mathbf{z}, x) \ln \left(\frac{p(x|\mathbf{z}, j)}{p(x|\mathbf{z})} \right), \tag{8}$$

where we have used $N_{x,\mathbf{z},j} = Np(x, \mathbf{z}, j)$ and $p(x, \mathbf{z}, j) = p(x|\mathbf{z}, j)p(\mathbf{z}, j)$. Therefore, the MJSD $D^m = -\ln \lambda/(N \ln 2)$, which shows that D^m is proportional to a log likelihood ratio.

1.2 The χ^2 distribution of D^m

Having shown that D^m is proportional to $-\ln \lambda$, it remains to show that $-\ln \lambda$ has a χ^2 distribution. Indeed, it is a standard theorem in statistics that the asymptotic distribution of $-2\ln\lambda$ is a χ^2 distribution [2] for any maximum likelihood ratio, provided (a) H_0 is nested in H_1, (b) the maximum likelihood is computed from a non-boundary point where the likelihood function is differentiable, and (c) the unknown parameters in the two hypotheses are real numbers that take values on an interval. Since all of these criteria are satisfied for the above hypotheses, the theorem implies that in the limit of large sample size (large N), $-2\ln\lambda$ has a χ^2 distribution with degrees of freedom equal to the difference in the number of undetermined parameters for the two hypotheses. In our case, the number of degrees of freedom is $k^{m+1}l-1-k^m(k-1)-k^ml+1=k^m(l-1)(k-1)$. We thus obtain the result: for large N, $X=D(2N\ln 2)$ has a χ^2_d distribution with

$$d=k^m(k-1)(l-1) \tag{9}$$

degrees of freedom.

2 Probability distribution of $D^m_{\max}$

Consider now the problem of determining the point of maximum heterogeneity in a symbolic sequence. This point is defined as the sequence position such that the two subsequences split by that position have the maximum possible value of D^m, namely, $D^m_{\max}$. As previously shown, D^m follows a χ^2 distribution; we now wish to find an analogous distribution for $D^m_{\max}$.

[Supp Fig. 1 about here.]

As a starting point to determine the distribution of $D^m_{\max}$, we begin with the assumption that each value of D^m obtained by sliding the segmentation point along the sequence is i.i.d. This assumption yields

$$P(D^m_{\max}\le x)=P(D^m\le x)^N \tag{10}$$

where N is the length of the sequence. However, it is well known that the value of D^m sampled at position i along the string is highly correlated with the values of D^m sampled at $i-1$, $i+1$, and other surrounding positions [3,4,1]. An analytic form of the true distribution of $D^m_{\max}$ has been derived for $m=0$ as N approaches ∞ [4]:

$$P(D^0_{\max}\le x) \quad \overset{N\to\infty}{\longrightarrow} \quad \exp\left(-2e^{A-\sqrt{B(2N\ln 2)x}}\right), \tag{11}$$

where

$$A=2\ln\ln N+\frac{d}{2}\ln\ln\ln N-\ln\Gamma\left(\frac{d}{2}\right) \qquad \text{and} \qquad B=2\ln\ln N, \tag{12}$$

with $d=k^m(k-1)$ degrees of freedom. The inaccuracy of Eq. (11) for finite length sequences is depicted in Supplementary Figure 1a. The inaccuracy is especially large when $P(D^m_{\max}\le x)$ is high, which is the most important part of a power curve should high confidence be desired. Following methods originally proposed by [1], we turn to fitting the empirical distribution of $D^m_{\max}$ to a modified form of the χ^2 distribution.

We first use a Monte Carlo algorithm to determine an empirical distribution for $D^m_{\max}$. The significance of $D^m_{\max}$ depends on the length of the string N, so we perform the Monte Carlo analysis for various N and interpolate the fitted parameters. We also perform the analysis for multiple values of m, k. The algorithm is as follows:

(1) For many values of N
   (a) Select m and k.
   (b) Generate a large number (here, 10^5) of random strings of length N.
   (c) Determine $\hat{D}^m_{\max}$ on each of the randomly generated strings.
(2) Fit $P(\hat{D}^m_{\max} \leq x)$ to $\{F_d[2N(\ln 2)x\beta]\}^{N_{\text{eff}}}$ by determining appropriate N_{eff} and β.
(3) Interpolate the values of the fitted parameters N_{eff} and β.

[Supp Fig. 2 about here.]

Let $\hat{D}^m_{\max}$ be the observed maximum value found. We fit the empirical $P(\hat{D}^m_{\max} \leq x)$ to the parametrized form

$$P(\hat{D}^m_{\max} \leq x) = \{F_d[2N(\ln 2)x\beta]\}^{N_{\text{eff}}} \tag{13}$$

where F_d is the cumulative distribution function of the χ^2 distribution with d degrees of freedom, $d = k^m(k-1)(\ell-1)$ as found in Eq. 9, and β and N_{eff} are fitting parameters that reflect the lack of independence between values of D^m sampled at different positions along the sequence.

We find that N_{eff} is linearly related to $\log N$. Also, while β is effectively a constant function of $\log N$ for both $m = 0$ and $m = 1$, it has a weak linear dependence on $\log N$ for $m = 2$ and $k = 4$. The empirical distribution of the $\hat{D}^m_{\max}$ is shown in Supplementary Figure 1b, c, and d for $k = 4$. The fitted values of N_{eff} and β are shown in Supplementary Figure 2 for $k = 4$. The relations between N_{eff} and β with respect to N are given in Supplementary Table 1 for $k = 2, 4$.

[Supp Table 1 about here.]

3 Recursive sequence segmentation with $D^m_{\max}$

In previous sections we have discussed theoretical aspects of the generalized Markovian Jensen-Shannon divergence. In this section, we describe a sequence segmentation method based on the MJSD. The method involves computing D^m for every position along a sequence. If the maximum value, $D^m_{\max}$, is large enough to be considered statistically significant, then the position where $D^m_{\max}$ was found is a segmentation point. The sequence is split in two at the segmentation point and the two resulting subsequences are again candidates for segmentation. If $D^m_{\max}$ is not statistically significant, no further segmentation is carried out. The algorithm is as follows:

(1) The algorithm takes as its argument the parameters:
   - A string **S** of length N.
   - Maximum allowed false positive rate α. The lower the value of α, the lower the number of segmentation points reported.
   - An integer L representing the shortest allowed string (this is necessary because variances of entropy estimates become large for strings that are too short).

- An alphabet $\mathcal{A}$ that should be used.
- The order of the Markov source m.

(2) At each position i, $L \leq i \leq N - L$, the value $D^m(p_1, p_2)$ is computed, with p_1 derived from the subsequence $s_1, \ldots, s_i$ and p_2 derived from the subsequence $s_{i+1}, \ldots, s_N$.

(3) Once $D^m(p_1, p_2)$ has been computed at the $N - 2L$ locations, the maximum, $D^m_{\max}$, and the location at which this maximum is attained, are found.

(4) The cumulative distribution function of $D^m_{\max}$, given by Eq. 13 must be greater than $1 - \alpha$ for i to be considered a valid segmentation point. This function is computed using the appropriate fitted values of β and N_{eff}. If i is a valid segmentation point, the sequence $\mathbf{S}$ is segmented into two subsequences at i and the entire algorithm is iterated using each subsequence as an input.

This recursive algorithm is related to several existing methods. The $m = 0$ case is derived in [1]. A similar $m = 0$ algorithm in a theoretical statistics context is presented in [4,3]. An alternative criterion for determining significance is proposed in [5]. The criterion is based on a model selection argument, utilizing the Bayesian and Akaike information criteria to trade off segmentation complexity and likelihood of the segmentation. The algorithm is also closely related to the two-level block approach of [6], which can also be generalized to Markov sources.

4 Size Effects in Concatenation Experiments

In this *in silico* experiment, we concatenate different pieces of different genomes and test whether the segmentation procedure can determine the concatenation point accurately for different values of the Markov order m. We select two genomes and extract a contiguous piece of DNA with length N bases from a random location in both genomes. We then create a new sequence which is composed of the two pieces concatenated together. This new sequence therefore has length $2N$. We find the point of maximum D^m in this new sequence. We then repeat this entire procedure 10^5 times (by selecting different random pieces of DNA of size N from the two genomes). If the segmentation algorithm were perfectly reliable, it would consistently give a very high value for $D^m_{\max}$ at position N.

We evaluate this method (Supplementary Figure 3) by providing the upper and lower 90% confidence intervals for the true location of $D^m_{\max}$. The size of this interval depends on 1) the genomic differences between the two genomes being assessed, and 2) the size of the segments.

We also use a control to check whether the algorithm is biased towards report $D^m_{\max}$ occurring at the central position. We find the location corresponding to $D^m_{\max}$ for a contiguous stretch of DNA of length $2N$ from a *single* genome. We do not find any tendency for $D^m_{\max}$ to occur at the central position in this control simulation (Supplementary Figure 3 (c)).

[Supp Fig. 3 about here.]

Naturally, the segmentation works much better when the genomes have different distributions of nucleotides. This can be seen in Supplementary Figure 3 (a) where *S. enterica* (G+C is 52.1%) with *M. loti* (G+C is 62.7%). It may seem that the segmentation can detect differences in the genomes of *S. enterica* and *E. coli*, as depicted in Supplementary Figure 3 (b). However, in Supplementary Figure 3 (b), this same pattern can be seen when only the *S. enterica* genome is used to select random sequences from perhaps disparate sections of the genome. Finally, taking a contiguous region of length $2N$ from the *Salmonella* genome yields a control distribution for checking that our results are not a result of bias. Many other experiments and controls confirm the conclusions that can be drawn from Supplementary Figure 3 (data not included), namely that, when segments from distinct genomes are concatenated, the MJSD accurately picks the true segmentation point, and that this accuracy increases with increasing Markov order.

The recursive algorithm's inability to distinguish *S. enterica* from *E. coli* is plausible given their phylogentic proximity. We show a similar result in the main text in the context of finding horizontally transfered genes in artificial chimeric genomes.

5 Genomes used

We use several genomes to demonstrate the efficacy of the recursive MJSD method (Supplementary Table 2). Known islands in *S. enterica* Typhi CT18 are given in Supplementary Table 3.

[Supp Table 2 about here.]

[Supp Table 3 about here.]

6 Artificial horizontal gene transfer

Supplementary Table 4 shows similar results to those in the paper for windows of size 5kb.

[Supp Table 4 about here.]

7 Expression profile overlaps MJSD predictions

As discussed in the main text, the expression profile during *S. enterica* host infection indicates that many known virulence genes are differentially regulated during infection. We use the microarray data from [10] and examine genes that are upregulated more than 2-fold during any phase of infection. The genes that are upregulated are enriched for those that are horizontally aquired ($P < 0.02$, Fisher's exact test).

The recursive MJSD method has particularly high sensitivity to these known virulence genes (see main text). Thus we searched for other regions where the recursive MJSD algorithm overlaps upregulated genes. Using this criterion, we find another 30 regions (encompassing an additional 8% of the genome) that do not overlap existing annotations. These annotations are visualized in Supplementary Figure 4.

[Supp Fig. 4 about here.]

The overexpressed genes and locations of islands can be found in Supplementary Table 5. This list contains several house-keeping genes (e.g., STY2301) that may be up-regulated due to increased metabolism. In addition, there are regions that harbor known virulence and drug resistant genes which may have originated from foreign ancestry (e.g., STY1076, STY2632, STY2341). Furthermore, there are several genes contained in islands that overlap or are adjacent to known pathogenicity islands (e.g., STY1076, STY2632, STY1990).

[Supp Table 5 about here.]

[Supp Table 6 about here.]

References

[1] Grosse, I., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J., and Stanley, H. E. (March, 2002) Analysis of symbolic sequences using the Jensen-Shannon divergence. *Physical Review E,* **65**, 1–15.

[2] Wilks, S. (1962) Mathematical Statistics, Wiley, New York, .

[3] Horváth, L. (October, 1989) The limit distributions of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Multivariate Analysis,* **31**, 148–159.

[4] Horváth, L. and Csorgo, M. (2002) Limit Theorems in Change Point Analysis, Probability and StatisticsWiley, .

[5] Li, W. (2001) DNA segmentation as a model selection process. *RECOMB '01: Proceedings of the fifth annual international conference on Computational biology,* pp. 204–210.

[6] Szpankowski, W., Ren, W., and Szpankowski, L. (2005) An optimal DNA segmentation based on the MDL principle. *Int. J. Bioinformatics Research and Applications,* **1**(1), 3–17.

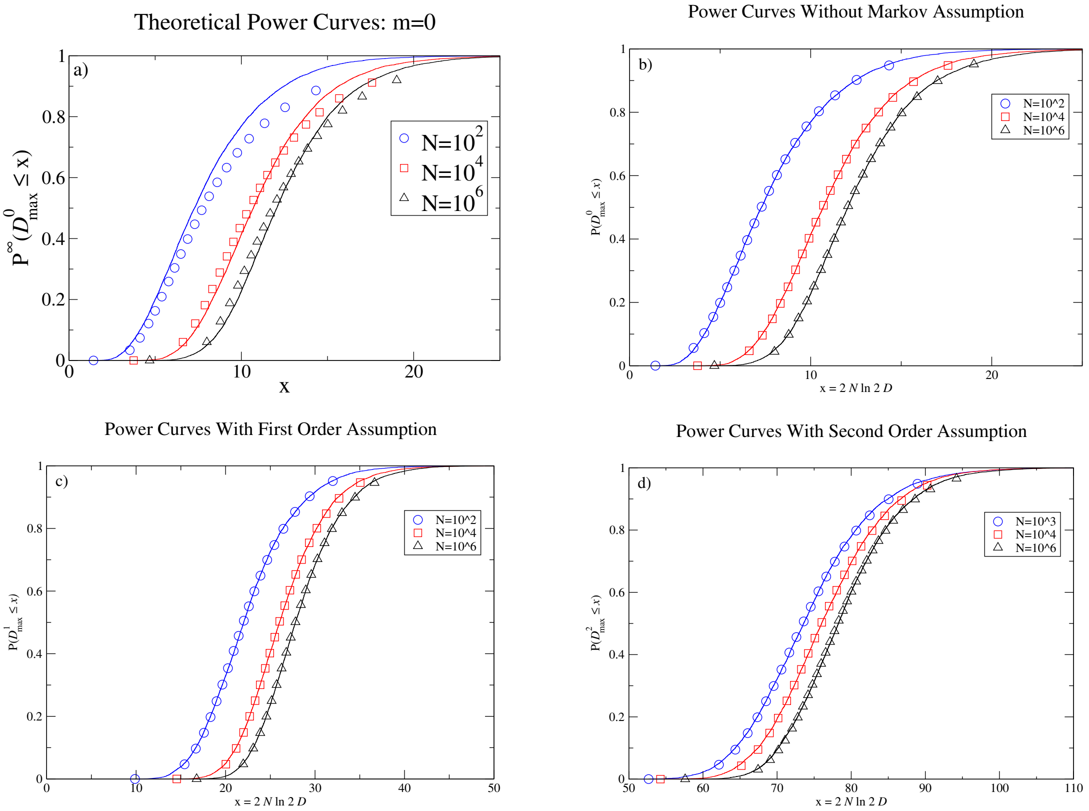
[7] Bernaola-Galván, P., Román-Roldán, R., and Oliver, J. L. (May, 1996) Compositional segmentation and long-range fractal correlations in DNA sequences. *Physical Review E,* **53**(5), 5181–5189.

[8] Azad, R. K., Rao, J. S., Li, W., and Ramaswamy, R. (September, 2002) Simplifying the mosaic description of DNA sequences. *Physical Review E,* **66**, 031913.

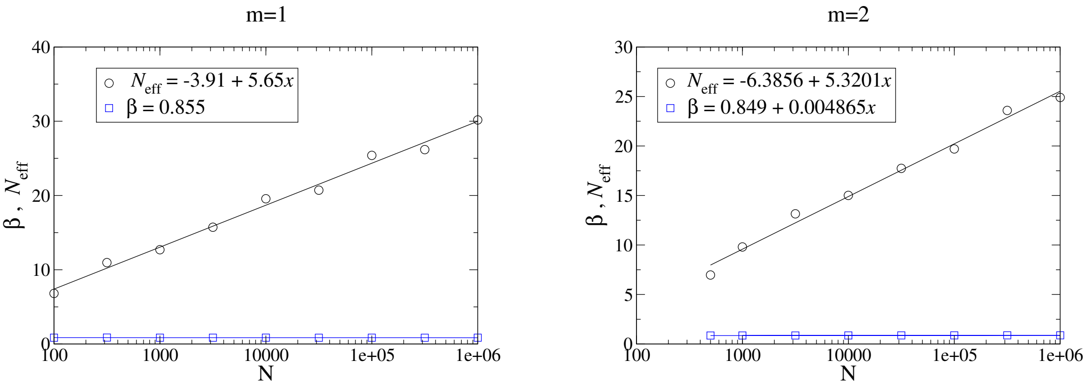
[9] Bernardi, G., Olofsson, B., Filipski, J., Salinas, M. Z. J., Cuny, G., Meunier-Rotival, M., and Rodier, F. (May, 1985) The mosaic genome of warm-blooded vertebrates. *Science,* **228**, 953–958.

[10] Faucher, S., Porwollik, S., Dozois, C., McClelland, M., and Daigle, F. (Feb, 2006) Transcriptome of salmonella enterica serovar typhi within macrophages revealed through the selective capture of transcribed sequences. *Proceedings of the National Academy of Sciences,* **103**, 1906–1911 10.1073/pnas.0509183103.

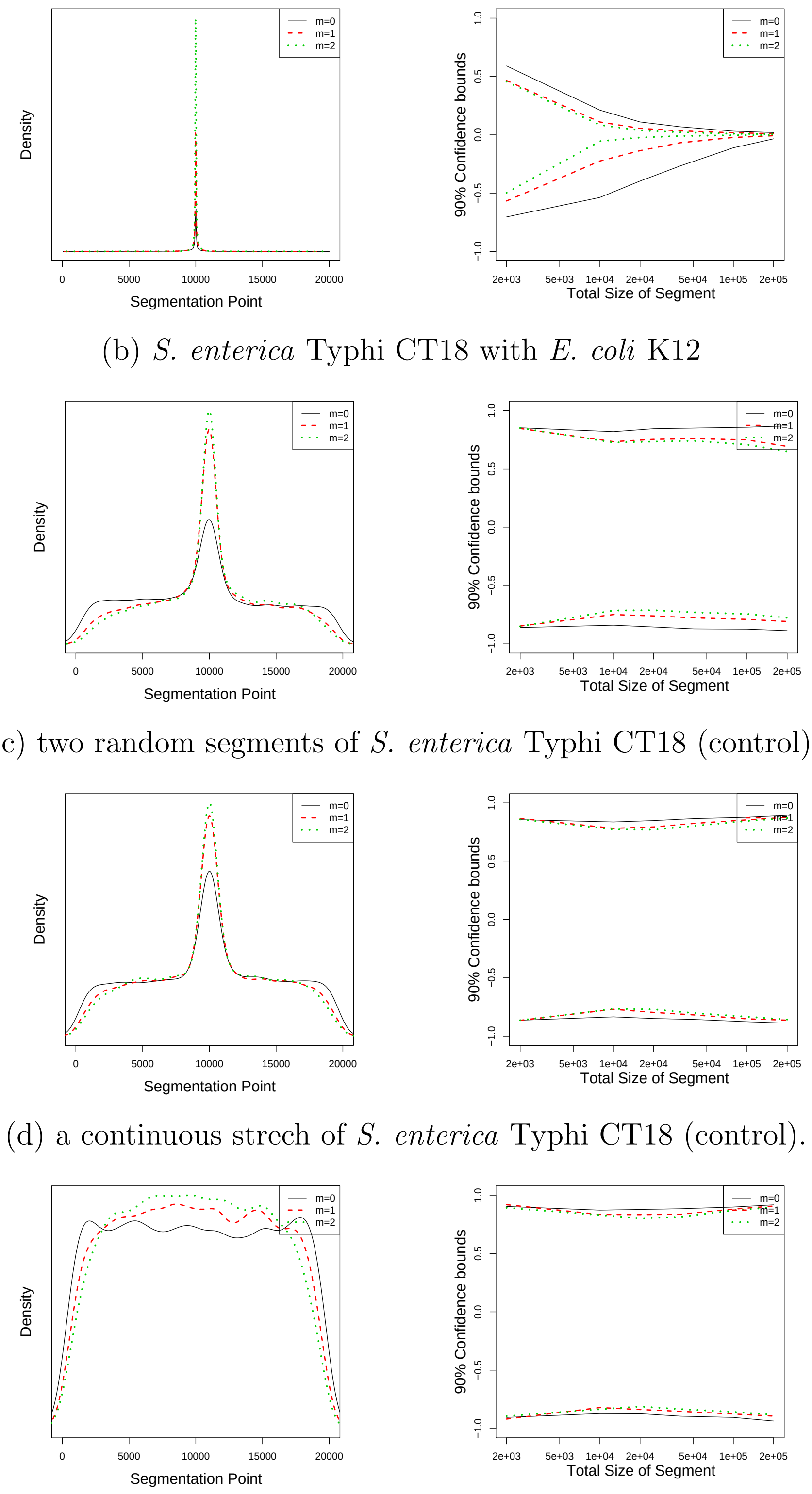
List of Figures



Supp Fig. 1. Cumulative distribution function of $D^m_{\max}$ for $k = 4$. The solid curves represent the empirical distribution function (as found through Monte Carlo simulations) and the symbols represent the values given by Eq. 11 in panel a) and by Eq. 13 with fitted parameters in panels b), c), and d). a) Plot demonstrating that the theoretical cumulative distribution function of $D^0_{\max}$, as defined by Eq. 11, is not a good fit to the true empirical distribution. b,c,d) Fitted curves of the empirical distribution $P(D^m_{\max} \leq x)$ as determined through Monte Carlo simulation for orders $m = 0$ (b), $m = 1$ (c), and $m = 2$ (d). The solid lines represent the empirical distribution and the symbols represent the values given by a modified χ^2 distribution with fitted parameters N_{eff} and β, which are determined by minimizing the Kolmogorov-Smirnov distance between the empirical $P(D^m_{\max} \leq x)$ and the cumulative distribution function of the modified χ^2 distribution.



Supp Fig. 2. N_{eff} and β as a function of $\log_{10} N$ for $m = 1, 2$ and $k = 4$ (a similar plot for $m = 0$ can be found in Figure 5 of [1]). Solid lines represent regression lines obtained by least squares fitting.



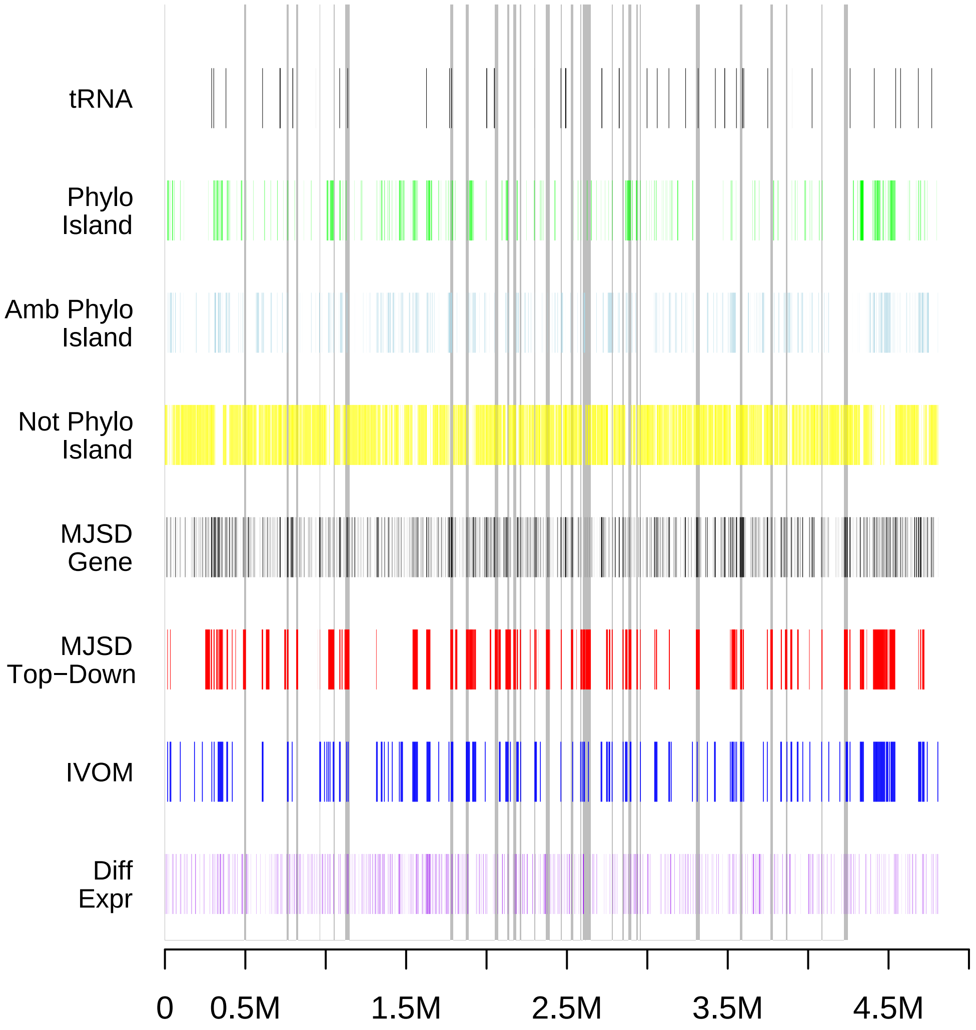
(a) *S. enterica* Typhi CT18 with *M. loti*

(b) *S. enterica* Typhi CT18 with *E. coli* K12

(c) two random segments of *S. enterica* Typhi CT18 (control)

(d) a continuous strech of *S. enterica* Typhi CT18 (control).

Supp Fig. 3. The figures on the left show the probability density (smoothed histograms in R) of the location of the segmentation point for $N = 10000$ and the figures on the right show 90% confidence intervals about the central value N as a fraction of N. The two lines for each value of m represent the upper and lower confidence bounds (determined empirically).



Supp Fig. 4. A set of 30 novel predictions based on the recursive MJSD algorithm and overexpressed genes during infection. These predictions (vertically spanning gray bars) are overlaps between regions found by MJSD top-down (red bars) and differentially expressed genes (purple bars).

List of Tables

	$k = 2$				$k = 4$			
m	a	b	c	d	a	b	c	d
0	7.42	12.41	0	0.80	6.45	-7.38	0	0.80
1	6.84	11.45	0	0.80	5.65	-3.91	0	0.855
2	6.15	10.34	0	0.80	5.32	-6.385	0.004865	0.85

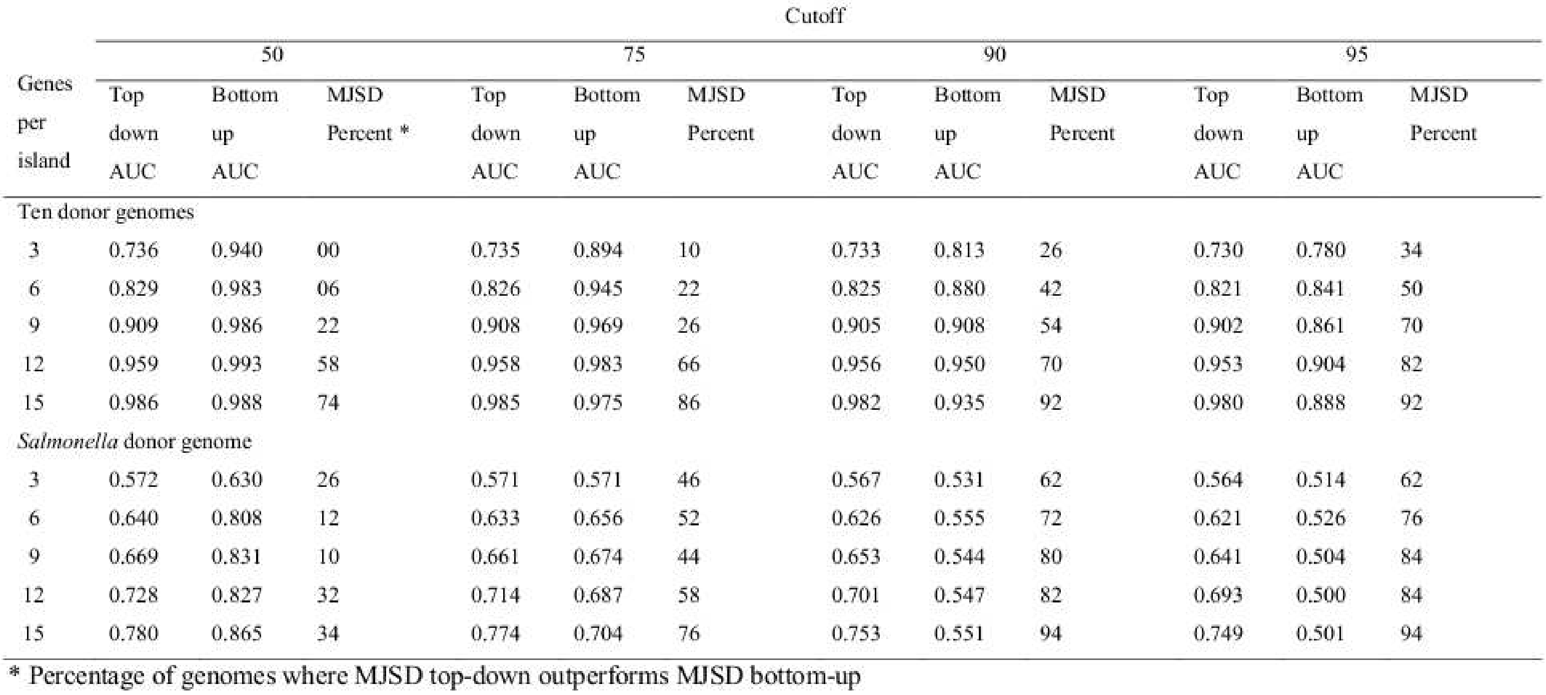
Supp Table 1
Parameters for fitting the power curves obtained by the Monte Carlo algorithm on sequences with an alphabet size of $k = 2, 4$. The empirical distributions can be found in Supplementary Figure 1. The rows correspond to different Markov orders. The values for N_{eff} and β are found according to the relations $N_{\text{eff}} = a \log_{10} N + b$ and $\beta = c \log_{10} N + d$. Note that β is constant when $k = 2$.

Genome	GenBank ID	GC%
E. Coli K12	NC_000913.fna	50.79
M. loti	NC_002678.fna	62.75
S. Typhi CT18	NC_003198.fna	52.09
S. Typhimurium LT2	NC_003197.fna	52.22
S. flexneri	NC_004337.fna	50.89

Supp Table 2
Genomes used in the main text, including their GenBank ID and GC content.

Island	Start	End
SPI-6	302172	361067
SPI-16	605515	609992
SopD2 Islet	964573	965531
Prophage 10	1008747	1051266
SPI-5	1085156	1092735
Bacteriophage	1538899	1572919
SPI-2	1625084	1664823
Novel Island	1776200	1791815
Bacteriophage	1887450	1933558
SPI-17	2460793	2465914
SPI-CS54	2597945	2615052
SPI-9	2743495	2759190
Bacteriophage 27	2759733	2782364
SPI-1	2859262	2899034
SPI-15	3053654	3060017
SPI-8	3132606	3139414
Bacteriophage	3515397	3549055
SPI-3	3883111	3900458
SPI-4	4321943	4346614
SPI-7	4409511	4543072
SPI-10	4683690	4716539

Supp Table 3
Known islands in the *S. enterica* Typhi CT18 genome.



	Cutoff											
	50			75			90			95		
Genes per island	Top down AUC	Bottom up AUC	MJSD Percent *	Top down AUC	Bottom up AUC	MJSD Percent	Top down AUC	Bottom up AUC	MJSD Percent	Top down AUC	Bottom up AUC	MJSD Percent
Ten donor genomes												
3	0.736	0.940	00	0.735	0.894	10	0.733	0.813	26	0.730	0.780	34
6	0.829	0.983	06	0.826	0.945	22	0.825	0.880	42	0.821	0.841	50
9	0.909	0.986	22	0.908	0.969	26	0.905	0.908	54	0.902	0.861	70
12	0.959	0.993	58	0.958	0.983	66	0.956	0.950	70	0.953	0.904	82
15	0.986	0.988	74	0.985	0.975	86	0.982	0.935	92	0.980	0.888	92
Salmonella donor genome												
3	0.572	0.630	26	0.571	0.571	46	0.567	0.531	62	0.564	0.514	62
6	0.640	0.808	12	0.633	0.656	52	0.626	0.555	72	0.621	0.526	76
9	0.669	0.831	10	0.661	0.674	44	0.653	0.544	80	0.641	0.504	84
12	0.728	0.827	32	0.714	0.687	58	0.701	0.547	82	0.693	0.500	84
15	0.780	0.865	34	0.774	0.704	76	0.753	0.551	94	0.749	0.501	94

* Percentage of genomes where MJSD top-down outperforms MJSD bottom-up

Supp Table 4
Accuracy comparison of the top-down and bottom-up (5KB window) MJSD averaged over 50 artificial genomes.

Gene(s)	Start	End
STY0488	492176	503741
STY0759	759014	768486
STY0820	816412	828382
STY0971	964447	965613
STY1076	1049674	1053289
STY1162	1121634	1147125
STY1871	1774880	1792000
STY1990	1872445	1890333
STY2218	2051120	2070614
STY2301	2129017	2138707
STY2341	2166670	2182481
STY2378	2207487	2214049
STY2469	2296986	2301957
STY2542	2369339	2392979
STY2632	2461635	2467344
STY2695	2525426	2538406
STY2748	2583567	2587773
STY2758	2598512	2614007
STY2784	2615728	2648589
STY2900	2778589	2784301
STY2974	2845604	2852028
STY3030	2883036	2900870
STY3073, STY3076	2932201	2941344
STY3091	2954512	2958420
STY3469	3301838	3325920
STY3739	3577130	3590607
STY3906	3764873	3780180
STY3996	3862580	3870240
STY4217	4083086	4089075
STY4343	4222598	4246642

Supp Table 5
The overexpressed gene(s) and locations of 30 predicted pathogenecity islands. A region is predicted to be a putative island if it contains a gene that is overexpressed during infection *and* it is predicted by the recursive MJSD method.

Supplementary Table 6. Performance of MJSD genome segmentation in finding islands predicted by three other methods.

			Percentage of Islands predicted by MJSD (bp predicted/total bp in class)*					
Genome	**Accession**	**SIGI**	**IslandPick**	**IslandPath**	**SIGI & IPick**	**SIGI & IPath**	**IPick & IPath**	**All Three Methods**
Anaeromyxobacter sp. K	NC_011145.	34.4% (15261/44422)	55.2% (11645/21093)	0.0% (0/5719)	100.0% (41760/41760)	0.0% (0/0)	0.0% (0/0)	100.0% (36408/36408)
Burkholderia cenocepacia MC0-3	NC_010512.	0.0% (0/12466)	0.0% (0/13191)	0.0% (0/0)	0.0% (0/15346)	100.0% (6820/6820)	0.0% (0/18690)	100.0% (39690/39690)
Burkholderia cenocepacia MC0-3	NC_010515.	21.8% (6782/31096)	0.0% (0/32367)	54.1% (17361/32118)	100.0% (11558/11558)	95.5% (124300/130109)	100.0% (12459/12459)	85.5% (64514/75425)
Burkholderia multivorans ATCC 17616	NC_010084.	49.5% (13128/26545)	0.0% (0/0)	0.0% (0/6131)	0.0% (0/0)	71.0% (52165/73481)	0.0% (0/0)	100.0% (70596/70596)
Burkholderia pseudomallei 1710b	NC_007435.	62.2% (32442/52159)	100.0% (28466/28466)	0.0% (0/20967)	0.0% (0/0)	26.4% (11927/45230)	100.0% (14463/14463)	100.0% (70223/70223)
Burkholderia pseudomallei 1710b	NC_007434.	42.5% (29792/70103)	0.0% 0/0)	100.0% (18721/18721)	63.9% (26713/41807)	91.3% (77215/84542)	0.0% (0/38)	100.0% (66450/66450)
Burkholderia vietnamiensis G4	NC_009256.	0.0% (0/40424)	2.1% (209/9863)	61.9% (8847/14295)	96.8% (46813/48373)	90.4% (80592/89143)	100.0% (14870/14870)	95.9% (142576/148619)
Escherichia coli 536	NC_008253.	54.6% (94472/173132)	0.0% (0/0)	66.0% (6977/10577)	0.0% (0/0)	70.0% (94631/135117)	0.0% (0/0)	100.0% (103354/103354)
Pseudomonas aeruginosa PA7	NC_009656.	85.0% (49077/57731)	100.0% (39823/39823)	94.9% (72823/76739)	90.6% (47307/52244)	78.0% (47982/61479)	79.8% (28596/35850)	99.2% (37415/37721)
Pseudomonas putida KT2440	NC_002947.	68.0% (49276/72501)	55.8% (79162/141912)	0.0% (0/33037)	100.0% (4362/4362)	100.0% (105091/105091)	100.0% (284/284)	100.0% (151003/151003)
Pseudomonas putida GB-1	NC_010322.	90.5% (42171/46606)	16.7% (15673/93719)	37.1% (14580/39270)	100.0% (8688/8688)	64.2% (35467/55236)	0.0% (0/0)	95.3% (129119/135548)
Pseudomonas syringae pv. phaseolicola 1448A	NC_005773.	75.8% (75971/100197)	19.7% (7431/37722)	41.2% (29950/72723)	0.0% (0/0)	71.1% (58777/82639)	95.6% (18393/19235)	100.0% (113640/113640)
Rhizobium etli CFN 42	NC_007761.	17.9% (5931/33185)	0.0% (0/33858)	0.0% (0/0)	0.0% (0/13861)	100.0% (23822/23822)	0.0% (0/0)	91.2% (83235/91217)
Rhodobacter sphaeroides 17025	NC_009428.	0.0% (0/0)	26.5% (34760/131097)	0.0% (0/32851)	100.0% (1122/1122)	76.4% (17209/22538)	25.5% (22323/87515)	35.6% (25499/71627)
Rhodopseudomonas palustris BisB5	NC_007958.	67.0% (10385/15498)	47.8% (28379/59381)	0.0% (0/4694)	0.0% (0/0)	100.0% (12945/12945)	16.2% (11761/72417)	100.0% (68355/68355)
Rhodopseudomonas palustris HaA2	NC_007778.	15.2% (3338/21930)	17.0% (28440/167558)	0.0% (0/12462)	61.3% (16494/26890)	0.0% (0/18386)	100.0% (28318/28318)	98.3% (90742/92302)
Shewanella baltica OS185	NC_009665.	70.1% (43750/62398)	52.3% (38838/74296)	75.3% (24822/32972)	79.6% (26955/33869)	83.4% (31456/37702)	92.0% (37953/41255)	100.0% (50846/50846)
Shewanella sp. ANA-3	NC_008577.	19.2% (5126/26669)	0.0% (0/0)	0.0% (0/0)	79.4% (37825/47657)	71.3% (32252/45204)	100.0% (414/414)	100.0% (38947/38947)
Sinorhizobium meliloti 1021	NC_003047.	0.0% (0/5846)	0.0% (0/19934)	0.0% (0/23869)	0.0% (0/13450)	100.0% (7003/7003)	0.0% (0/0)	100.0% (67918/67918)
Vibrio parahaemolyticus RIMD 2210633	NC_004603.	38.7% (30355/78448)	23.8% (13009/54607)	0.0% (0/9624)	0.0% (0/89)	90.8% (78465/86427)	0.0% (0/0)	98.3% (35515/36113)
Summary		**52.2% (507257/971356)**	**34.0% (325835/958887)**	**43.4% (194081/446769)**	**74.7% (269597/361076)**	**80.0% (898119/1122914)**	**54.9% (189834/345808)**	**94.9% (1486045/1566002)**

* IPick = IslandPick; IPath = IslandPath