The Distribution of Inversion Lengths in Bacteria

David Sankoff[1], Jean-François Lefebvre[2], Elisabeth Tillier[3], Adrian Maler[1], and Nadia El-Mabrouk[2]

[1] Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa K1N 6N5, Canada
sankoff@uottawa.ca

[2] Département d'informatique et recherche opérationnelle, Université de Montréal, CP 6128 succ. Centre-ville, Montréal, Québec H3C 3J7 Canada
{lefebvre,mabrouk}@iro.umontreal.ca

[3] Ontario Cancer Institute, Princess Margaret Hospital, 620 University Avenue, Suite 703, Toronto, Canada
e.tillier@utoronto.ca

Abstract. The distribution of the lengths of genomic segments inverted during the evolutionary divergence of two species cannot be inferred directly from the output of genome rearrangement algorithms, due to the rapid loss of signal from all but the shortest inversions. The number of short inversions produced by these algorithms, however, particularly those involving a single gene, is relatively reliable. To gain some insight into the shape of the inversion-length distribution we first apply a genome rearrangement algorithm to each of 32 pairs of bacterial genomes. For each pair we then simulate their divergence using a test distribution to generate the inversions and use the simulated genomes as input to the reconstruction algorithm. It is the comparison between the algorithm output for the real pair of genomes and the simulated pair which is used to assess the test distribution. We find that simulations based on the exponential distribution cannot provide a good fit, but that simulations based on a gamma distribution can account for both single-gene inversions and short inversions involving at most 20 genes, and we conclude that the shape of latter distribution corresponds well to the true distribution at least for small inversion lengths.

1 Introduction

The study of genome rearrangement has made it clear that the lengths of the chromosomal segments inverted, transposed or reciprocally translocated is not determined simply by a random choice of two breakpoints anywhere in the genome. While this is very-well documented in eukaryotes [2, 10, 13, 5], it is also true that prokaryotic genome rearrangement also operates under a variety of constraints on inversion site and length of inverted segments [12, 17, 11]. Incorporating information on such constraints into procedures for reconstructing genome divergence, e.g. in terms of weights in a parsimony analysis, probabilities in a likelihood analysis or priors in a Bayesian analysis, is a desirable goal for

J. Lagergren (Ed.): RECOMB 2004 Ws on Comparative Genomics, LNBI 3388, pp. 97–108, 2005.

evolutionary methodology. With this motivation, in this paper we study the distribution of lengths of the segments that are inverted in the evolutionary history of bacterial genomes. Inherent in this study are many assumptions, not the least of which is that the distribution in question exists, i.e., represents a tendency relatively fixed over time and across the phylogenetic spectrum of bacteria. While we cannot resolve such a far-reaching question here, our results will provide a measure of confirmatory justification.

Another assumption is that inversion is the dominant process of gene order change in bacteria. Our approach will control for changes in genome size through gene gain and gene loss, but not for the effects of simply transposing segments from one area of the genome to another. This does not seem to be unwarranted; we find no systematic discussion of a transposition process in the literature on bacterial genomes, though transposition of small segments is very common in eukaryotic nuclear genomes [5, 10], and duplication-loss, which has the same effect as transposition, is often cited as an explanation for gene-order change in eukaryotic organelle genomes [3].

In a previous study [11], we analyzed the inversion lengths inferred between each of four pairs of bacterial genomes and discovered an unexpectedly high number of short inversions, single-gene inversions in particular. This contrasted with the null hypothesis that the two breakpoints of an inversion occur randomly and independently within the genome of length n, which predicts a uniform distribution $U[1, \frac{n}{2}]$ of inversion lengths, where the $\frac{n}{2}$ reflects the fact that for a circular genome, an inversion of length l is indistinguishable from the complementary inversion of length $n - l$.

The present paper builds on the previous work in two ways. First, we greatly expand our sample of genome pairs, from four to 32, deliberately picked to represent the range between closely-related and phylogenetically distant pairs, and we use a more systematic method than in the previous paper for validating relations of orthology within each pair. Second, rather than just reject the uniform null hypothesis, we attempt to pin down aspects of the probability distribution of inversion length in bacterial evolution. More precisely, we focus on the shape of this distribution only where the inversions are short, namely single-gene inversions and inversions of at most 20 genes. This rather restrained ambition is warranted by the discovery in [11], summarized in Section 2 below, that in genomes that have been even moderately rearranged by the accumulation of inversions, parsimonious methods such as that Hannenhalli-Pevzner (HP) algorithm [7], can only recover the details of very short inversions. Simulations in [11] showed that the longer inversions "recovered" by such algorithms are overwhelmingly different from those used to generate the genomes.

In the next section of this paper we recap only the part of [11] which deals with signal decay. In Sections 3 and 4 we describe our methods and data. Section 5 contains our results, which show that a two-parameter distribution function, such as the gamma distribution, is necessary to reasonably fit the numbers of short inversions observed in the 32 pairs of genomes, but that a one-parameter distribution, such as that of the negative exponential distribution, is inadequate. These results are further discussed in Section 6.

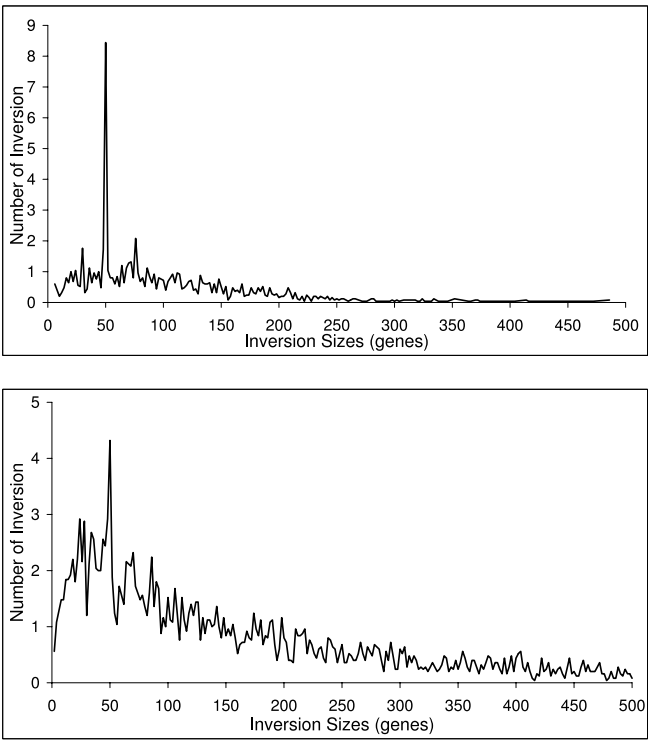

Fig. 1. Frequency of inversion sizes (or lengths) inferred by the algorithm for random genomes obtained by performing i inversions of length $l = 50$. The figure on the top is for $i = 80$ and the bottom one is for $i = 200$

2 Decay of Evolutionary Signal with Inversion Length

Consider two genomes containing the same set of genes but in different orders, where this difference is generated by evolutionary operations of a given type, such as inversions. We first ask to what extent the evolutionary histories reconstructed by the HP type of algorithm [7] actually reflect the true events. It is well-known that past a threshold of θn, where n is the number of genes and θ is in the range of $\frac{1}{3}$ to $\frac{2}{3}$, the *number* of operations begins to be underestimated by edit operation-based inferences (e.g., [8, 9]). Before that threshold, the total number may be accurately estimated but whether any signal is conserved as to the actual individual operations themselves, and which ones, is a different question.

In [11], we carried out the following test: For a genome of size $n = 1000$, we generated i inversions of length $l = 5, 10, 15, 20, 50, 100, 200$ at random, and then reconstructed the optimal inversion history, for a range of values of i. Typically, for small enough values of i, the algorithm reconstructs the true inversion history. Depending on l, however, above a certain value of i, the reconstructed inversions manifest a range of lengths, as illustrated in Figure 1 (reproduced from [11]).

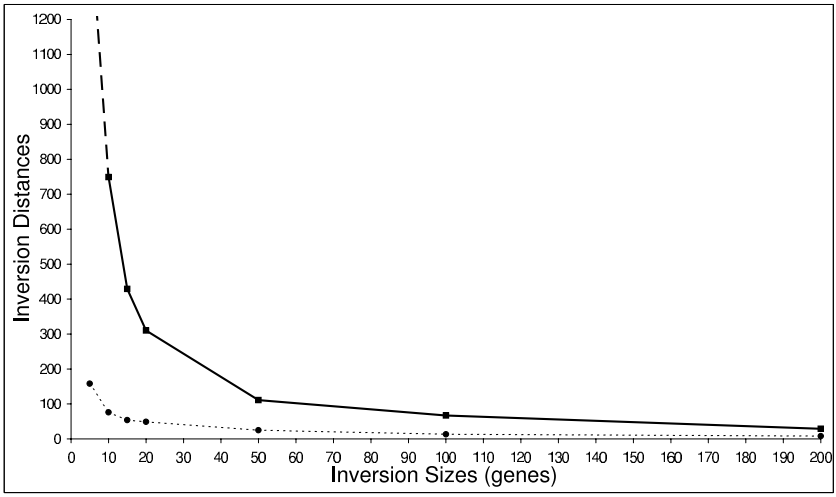

Fig. 2. The solid line plots s, and the dotted line plots r (see text above)

For each l, we calculated

$r_l = \min\{i|\text{reconstruction has at least 5\% error}\}$

and

$s_l = \max\{i|\text{reconstruction has at most 95\% error}\}$,

where any inversion having length different from l is considered to be an error. Figure 2 (reproduced from [11]) plots r and s as a function of l and shows how quickly the detailed evolutionary signal decays for large inversions. Nevertheless, we note that for very small inversions, there is a clear signal preserved long after longer ones have been completely obscured.

3 Method

In our quest for the distribution of inversion lengths in bacteria, there are three steps applied to each pair of genomes in our sample:

- We use a carefully validated method for establishing orthologies between the two genomes, based on both sequence and genomic context [4].
- We calculate the inversion distance between the two genomes, as well as a number of detailed evolutionary scenarios exemplifying this distance
- We simulate a matching pair of genomes whose divergence is based on whatever distribution we are testing.

These steps are detailed in the following paragraphs.

3.1 Orthology

In the new method developed in [4], potential orthologs are evaluated according to a number of criteria:

- status of BLAST match; whether it is the best match in both directions
- quality of BLAST match; in terms of statistical significance
- scope of BLAST match with respect to the total length of the gene
- presence or absence of contextual markers conserved in both genomes
- whether there are near optimal competing genes in either genome

This enabled us to construct a matched set of orthologous genes in both genomes with a maximum of confidence. Of course, some of the matches are less clear than others, and the matches in closely related genomes tend to be less ambiguous than in distant pairs. Nevertheless these matches represent a systematic, multi-criterion, best estimate.

Once the matches are established, we constructed reduced genomes of equal length by deleting those genes not identified as being in an orthologous match. This paper reports on the analysis of these reduced genomes only, though we have also analyzed the full genomes using an inversion/insertion/deletion procedure [6]. Results from the latter were generally less clear, though they did not conflict with the results reported here.

Note that our use of reduced genomes means that our characterizations of inversions as "single-gene" or "1-20 genes" in the comparison of the reduced genomes may sometimes refer to somewhat larger inversions when the deleted genes from the unreduced genomes are restored.

3.2 Algorithm

The results of genome rearrangement algorithms are highly non-unique; many different evolutionary scenarios have the same, minimal, number of steps.

In a previous publication [1] we developed a general method that allows a choice among equally optimal solutions (i.e., the same minimal number of operations) generated by a HP type of algorithm, based on any one of many possible secondary criteria. This takes advantage of the many equally valid choices that may be available at each step of the algorithm.

Given our interest in short inversions, we adopt inversion length as our secondary criterion. Thus a solution can be obtained by selecting, at random, one of the shortest allowable inversions at each step of the HP procedure. Running the algorithm several times gives rise to several possible solutions. We can then tabulate how many times inversions of a particular length recur in the set of solutions. In [11], we showed that this length-based strategy enhanced the difference between pairs of real genomes and simulated pairs where the inversion lengths were sampled from the $U[1, \frac{n}{2}]$ distribution. The number of reconstructed single-gene and other short inversions, already higher in the real genome comparison than in the simulations, based on HP with no secondary strategy, increased markedly under the length-based strategy. There was little increase in the number of reconstructed single-gene and other short inversions for the genomes created with uniformly generated inversions. In other words, the increased number of short inversions inferred by length-based strategy was not simply an artifact of this strategy since it had little if any effect on the simulated genomes. Rather we

attributed it to better detection of bonafide short inversions whose signal we know to be conserved despite extensive genome rearrangement.

3.3 Simulations

To estimate the shape of the probability distribution of inversion lengths l, we explored

- a single parameter distribution, namely a negative exponential distribution

$$p(l) = \lambda e^{-\lambda l}. \tag{1}$$

- a two-parameter distribution, namely a gamma distribution

$$p(l) = \frac{l^{\alpha-1} e^{-l/\beta}}{\beta^{\alpha} \Gamma(\alpha)}. \tag{2}$$

For each each distribution p with cumulative P, we derived simulated pairs of genomes to compare with each of the 32 real ones as follows. For a given pair of bacterial genomes, let n be the length of the reduced genome, and let i be the number of inversions necessary to derive one from the other, as measured by the HP algorithm. We sampled somewhat more than i inversions (to compensate for the bias introduced by parsimonious reconstruction in a later step) from the probability distribution and used these to evolve a new circular genome starting from $1, 2, \cdots, n$. One of the breakpoints for each inversion was located randomly on the genome, and the second was located according to the sample inversion length. If an inversion was longer than $n/2$, we discarded it and did not count it, since and inversion of length l is the same as an inversion of length $n - l$ for a circular genome. So the effective length distribution was actually

$$p^*(l) = \frac{p(l)}{P(\frac{n}{2})}$$

for

$$0 < l \leq \frac{n}{2} \tag{3}$$

and zero elsewhere.

4 The Pairs of Bacterial Genomes

We informally sampled 32 pairs of genomes from those treated in [4], choosing some that are as phylogenetically distant as possible, and some that are relatively closely related. These are listed in Table 1, which also lists i, the minimum number of inversions necessary to convert one (reduced) genome to another, the size of the reduced genome n, i.e., the number of orthologous gene pairs in the two genomes as determined by the method in [4], the normalized inversion distance i/n, and the number of single-gene and 1-20 gene inversions.

Table 1. Pairs of bacterial genomes in this study. n is the number of orthologous genes identified in the two genomes, i is the inversion distance. Pairs ordered from least to highest values of the normalized inversion distance i/n. Last two columns give the number of inversions, out of a total of i, that involve just one gene and twenty or fewer genes, respectively

genome A	genome B	i	n	i/n	1	≤ 20
Neisseria meningitidis MC58	Neisseria meningitidis Z2491	53	1606	0.03	0	3
Salmonella typhi	Shigella flexneri 2a	196	2801	0.07	4	16.4
Escherichia coli CFT073	Salmonella typhimurium LT2	244	3145	0.08	8.6	25.6
Mycobacterium leprae	Mycobacterium tuberculosis CDC1551	109	1367	0.08	4	14.8
Staphylococcus aureus Mu50	Staphylococcus epidermidis ATCC 12228	148	1805	0.08	0.2	15.8
Streptococcus agalactiae 2603	Streptococcus pyogenes	201	1156	0.17	5	28.2
Streptococcus mutans	Streptococcus pyogenes	211	1046	0.20	6	19
Agrobacterium tumefaciens C58 Uwash Circ	Sinorhizobium meliloti	347	1705	0.20	14.6	50.4
Escherichia coli CFT073	Yersinia pestis CO92	642	2363	0.27	13	71
Pseudomonas aeruginosa	Pseudomonas putida KT2440	996	3189	0.31	34	117.2
Corynebacterium glutamicum	Mycobacterium tuberculosis CDC1551	380	1087	0.35	11.4	39.4
Bacillus halodurans	Bacillus subtilis	748	1912	0.39	31	91.6
Salmonella typhi	Vibrio cholerae ChI	584	1479	0.39	20	109
Listeria innocua	Staphylococcus aureus Mu50	471	1085	0.43	22.4	69.6
Escherichia coli K12	Vibrio cholerae ChI	717	1648	0.44	29	108
Bacillus halodurans	Listeria innocua	518	1186	0.44	22.6	61.2
Bacillus halodurans	Oceanobacillus iheyensis	818	1856	0.44	18	105
Listeria monocytogenes	Staphylococcus epidermidis	486	1080	0.45	15	78
Clostridium acetobutylicum	Clostridium perfringens	564	1211	0.47	22.6	70.2
Salmonella typhimurium LT2	Shewanella oneidensis	741	1474	0.50	33.6	100
Clostridium perfringens	Thermoanaerobacter tengcongensis	427	841	0.51	23	60
Oceanobacillus iheyensis	Thermoanaerobacter tengcongensis	452	853	0.53	18.4	61.8
Pseudomonas putida KT2440	Vibrio cholerae ChI	639	1160	0.55	28.4	90.4
Staphylococcus epidermidis ATCC 12228	Thermoanaerobacter tengcongensis	359	623	0.58	17	56.6
Listeria innocua	Thermoanaerobacter tengcongensis	450	770	0.58	21.8	64.2
Clostridium perfringens	Staphylococcus epidermidis ATCC 12228	391	640	0.61	19.4	60
Bacillus halodurans	Clostridium perfringens	619	944	0.66	28	81
Mycobacterium tuberculosis CDC1551	Mycoplasma penetrans	90	137	0.66	6	16
Bacillus subtilis	Streptococcus agalactiae 2603	533	806	0.66	27.8	65
Bacillus halodurans	Streptococcus pneumoniae R6	509	737	0.69	21.2	90.4
Staphylococcus aureus Mu50	Streptococcus pyogenes MGAS8232	649	939	0.69	25	127
Streptococcus mutans	Thermoanaerobacter tengcongensis	432	598	0.72	18	76

We note that parts of the evolutionary history separating many of the gene pairs are shared; perhaps the most obvious example is the *E.coli – Vibrio* and *Salmonella – Vibrio* comparisons, since these reflect a largely similar historical divergence, *E.coli* and *Salmonella* having a relatively recent common ancestor. This kind of dependence, which in general increase measures of dispersion but not bias, is not as great among our other pairs of genomes, and is in any case virtually impossible to avoid in a phylogenetic context.

5 Results

Applying our algorithm to the 32 pairs of bacterial genomes, repeating each comparison ten times with different random choices of shortest allowable inversion at each step, we counted the average number of single-gene inversions and the average number of inversions of length 20 or less. These were normalized by n and plotted against the normalized inversion distance i/n in Figure 3. We also

plotted on Figure 3 the result of our simulations based on the negative exponential and gamma distributions. For the negative exponential, it can be seen that a value of λ that allows the curve obtained from $p(l) \leq 1$ to fit the real data on single-gene inversions does not allow the curve obtained from $p(l) \leq 20$ to fit the real data on inversions of length 20 and less, and vice-versa. For the gamma distribution, on the other hand, values of α and β can be found that fit both sets of data, although for 1-20 gene inversions, the fit breaks down when $i > n/2$.

We found such values of the parameters of the gamma distribution by minimizing the sum of squared differences, between each real pair of genomes and the corresponding simulated pair, of the normalized number of single-gene inversions in a minimal inversion scenario in plus the analogous difference for the normalized number of 1-20 gene inversions. The latter differences were weighted by a factor of 0.1, since the number of short inversions was approximately 10 times as large as the number of single-gene inversions. We iterated by fixing each parameter in turn and searching for the minimizing value of the other parameter.

6 Discussion

To what extent do our results bear on the question of whether there is a universal distribution of inversion lengths across the bacterial domain? After all, this distribution is the result of numerous mechanistic mutational processes at the chromosomal level as well as selective processes operating on cell form and function, both of which can be expected to vary among genomes.

The generality of the distribution can be assessed in part by the deviation of the sample points from the overall trend in Figure 3. While it is true there is a degree of statistical fluctuations, our results are thoroughly compatible with the hypothesis that all the pairs are following a common tendency. That the more distantly related genome pairs have fewer 1-20 gene inversions than the corresponding simulated pairs indicates some tendency for the signal from the short inversions to be lost for reasons other than genome rearrangement, which should affect the simulated and real pairs in the same way. The observed shortfall in the number of short inversions for normalized distances greater than about 0.45 is partly due to an greater incidence of undetectable orthology in the more distant pairs, and partly to our way of treating unequal gene complements, of accumulated gene gain and loss for these pairs. Neither of these problems affect the simulated genome pairs. Whichever the explanation, the fact remains that all the distant pairs manifest the same shortfall, and there is no idiosyncratic behaviour from genome pair to genome pair evident at the aggregate level. Note that overall inversion *frequency* is not addressed in our analysis, since we are using no external time measure to calibrate the genomic distances, but this is not pertinent to our results.

Recently, attention has been drawn to the prevalence and significance of short inversions, albeit more in eukaryotes [2, 10, 16, 13, 5] than in prokaryotes [15, 11]. Here we have advanced our approach to the study of short inversions, taking advantage of the greatly elevated persistence in their evolutionary signal, compared

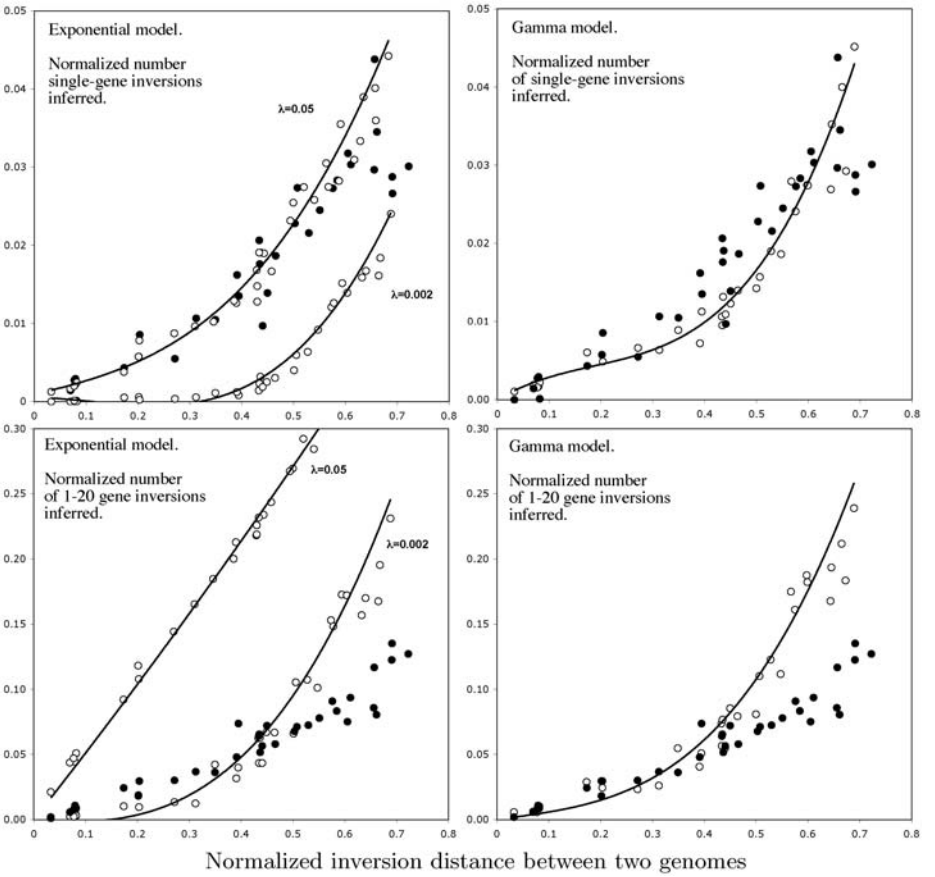

Fig. 3. Fit of exponential and gamma models (open dots and trend lines) to data on single gene inversions and 1-20 gene inversions (filled dots). Exponential parameter $\lambda = 0.002$ or 0.05, gamma parameters $\alpha = 0.60, \beta = 1200$

to that of longer inversions. We found that the distribution of *inferred* inversion lengths could be accounted for by a gamma distribution for the *generating* inversions, with a high proportion of single-gene and other short inversions and a rapid but non-exponential initial decline. The initial 30 values of the gamma distribution with parameters $\alpha = 0.60, \beta = 1200$ are depicted in Figure 4. Note that we do not consider any but this first few values of l. The upper tail of the gamma distribution is not relevant to this study; indeed our generation procedure truncates most all of the domain of the distribution greater than $\frac{n}{2}$. In any case, we are using the gamma as a descriptive device and are not suggesting it is theoretically privileged in being mathematically derived from some mutational or selective model for the inversion process. Note that in [11] we ruled out a uniform distribution as descriptively inadequate, and in the present paper we also ruled out the exponential distribution.

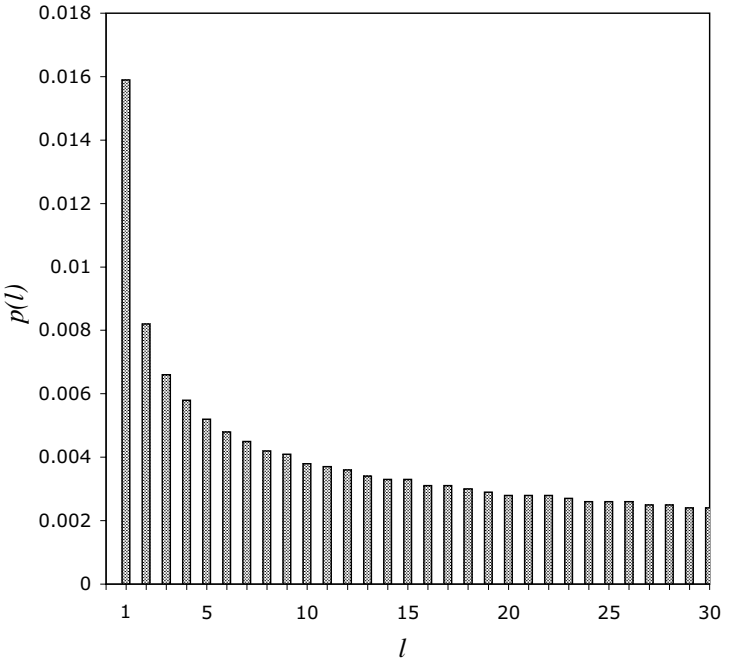

Fig. 4. Gamma distribution with parameters $\alpha = 0.60, \beta = 1200$

How can the preference for short inversions be explained? We suggest that it is a combination of factors:

- Single-gene inversions may represent a particular evolutionary mechanism with selective functional consequences. They may allow a gene to obtain transcriptional independence from its erstwhile operon, or to otherwise change its expression pattern, or to take advantage of new or altered functionality, or to participate in a different pathway through a more appropriate genomic positioning (cf genomic hitchhiking [14]).
- Single-gene inversions may simply be the clearest manifestation of a universal tendency towards short inversions as the least disruptive of the gene proximity configuration, and attendant functionality, of a genome. In [15], we argued that a predisposition for such inversions in small genomes might explain the prevalence of internally-shuffled "gene clusters" found across many sequenced genomes in microorganisms, in contrast to the "conserved segments", including fixed gene order, pattern characteristic of the higher eukaryotes.
- Mechanistic process that favour mutational processes operating over short distances.

Any knowledge about the distribution of inversion lengths would be invaluable to the inference of genome rearrangements. It is very difficult to obtain suitable data, however, so that the approach offered here is an example of the

indirect methods that must be developed in order to eventually home in on the true distribution.

Acknowledgments

Research supported by grants from the Natural Sciences and Engineering Research Council (NSERC), and the *Fonds québécois de recherche sur la nature et les technologies*. DS holds the Canada Research Chair in Mathematical Genomics. NE-M, ET and DS are affiliated with the Evolutionary Biology Program of the Canadian Institute for Advanced Research. We thank the anonymous referees for extensive comments and suggestions on an earlier version of this paper.

References

1. Ajana, Y., Lefebvre, J.F. , Tillier, E. and El-Mabrouk, N. (2002). Exploring the set of all minimal sequences of reversals - An application to test the replication-directed reversal hypothesis. *Algorithms in Bioinformatics, Second International Workshop, WABI*, Guigó,R. and Gusfield,D. Eds., Lecture Notes in Computer Science, **2452,** 300–15. Springer Verlag.
2. Bennetzen, J.L. and Ramakrishna, W. (2002) Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. *Plant Molecular Biology*, **48,** 821–7.
3. Boore, J.L (2000) The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals, *Comparative Genomics*, Sankoff,D. and Nadeau,J.H. Eds., 133–47. Dordrecht, NL: Kluwer Academic Press
4. Burgetz, I.J., Shariff, S., Pang, A. and Tillier, E.R.M. (2004). Positional homology in bacterial genomes. Submitted ms.
5. Coghlan, A. and Wolfe, K.H. (2002) Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. *Genome Research*, **12,** 857-67.
6. El-Mabrouk, N. (2001) Sorting signed permutations by reversals and insertions/deletions of contiguous segments. *Journal of Discrete Algorithms*, **1,** 105–122.
7. Hannenhalli, S. and Pevzner, P. A. (1999). Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *Journal of the ACM*, **48,** 1–27.
8. Kececioglu, J. and Sankoff, D. (1994) Efficient bounds for oriented chromosome-inversion distance. *Combinatorial Pattern Matching. Fifth Annual Symposium*, Crochemore, M. and Gusfield, D., Eds., Lecture Notes in Computer Science **807,** 307–25. Springer Verlag.
9. Kececioglu, J. and Sankoff, D. (1995) Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica*, **13,** 180-210.
10. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003). Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences, USA*, **100,** 11484–9.
11. Lefebvre, J.-F., El-Mabrouk, N., Tillier, E. and Sankoff, D. (2003). Detection and validation of single-gene inversions. *Bioinformatics*, **19,** i190–6.

12. Mahan, M.J. and Roth, J.R. (1991) Ability of a bacterial chromosome segment to invert is dictated by included material rather than flanking sequence. *Genetics*, **129,** 1021-32.
13. McLysaght, A., Seoighe, C. and Wolfe, K.H. (2000). High frequency of inversions during eukaryote gene order evolution. *Comparative Genomics*, Sankoff,D. and Nadeau,J.H. Eds.,47–58. Dordrecht, NL: Kluwer Academic Press
14. Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A. and Koonin, E.V. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Research*, **30,** 2212–23.
15. Sankoff, D. (2002). Short inversions and conserved gene clusters. *Bioinformatics*, **18,** 1305–1308.
16. Thomas, J. W, and Green, E. D. (2003). Comparative sequence analysis of a single-gene conserved segment in mouse and human. *Mammalian Genome*, **14,** 673–8.
17. Tillier, E.R.M. and Collins, R. (2000) Genome rearrangement by replication-directed translocation. *Nature Genetics*, **26,** 195–7.