**JOURNAL OF COMPUTATIONAL BIOLOGY**
**Volume 12, Number 6, 2005**

**Pp. 812–821**

# Chromosomal Breakpoint Reuse in Genome Sequence Rearrangement

DAVID SANKOFF[1] and PHIL TRINH[2]

## ABSTRACT

**In order to apply gene-order rearrangement algorithms to the comparison of genome sequences, Pevzner and Tesler bypass gene finding and ortholog identification and use the order of homologous blocks of unannotated sequence as input. The method excludes blocks shorter than a threshold length. Here we investigate possible biases introduced by eliminating short blocks, focusing on the notion of breakpoint reuse introduced by these authors. Analytic and simulation methods show that reuse is very sensitive to the proportion of blocks excluded. As is pertinent to the comparison of mammalian genomes, this exclusion risks randomizing the comparison partially or entirely.**

**Key words:** comparative genomics, rearrangements, Hannenhalli–Pevzner algorithm, breakpoints, synteny blocks.

## 1. INTRODUCTION

UNTIL RECENTLY, ALGORITHMS FOR STUDYING the evolution of gene order could be applied only to small genomes (mitochondria, chloroplasts, prokaryotes), the difficulty with mammalian and other larger eukaryotic nuclear genomes lying not so much in their much greater length but rather in the absence of comprehensive lists of genes and their orthologs. Pevzner and Tesler (2003a, b, c) have suggested a way to bypass gene finding and ortholog identification by using the order of syntenic blocks constructed solely from sequence data as input to a genome rearrangement algorithm. The method focuses on major evolutionary events by discarding blocks smaller than a threshold length. This use of large blocks only, however, may radically change the events that are inferred to explain the differences between the two genomes. We model the effects of eliminating short blocks, concentrating on the summary statistic of "breakpoint reuse" introduced by Pevzner and Tesler. They did not propose this as an evolutionary distance, but in the context of their protocol it effectively measures to what extent genomes have diverged in becoming random permutations of blocks with respect to each other. We use analytic and simulation methods to investigate breakpoint reuse as a function of the proportion of blocks deleted.

## 2. FROM GENOME SEQUENCE TO GENOME REARRANGEMENT

Algorithmic inference of genome rearrangement, based on "sorting by reversals (inversions)," "sorting by reversals and translocations," or similar methods, as reviewed by Sankoff and El-Mabrouk (2001), has

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada, K1N 6N5.
[2]School of Mathematics and Statistics, Carleton University, Ottawa, Canada, K1G 2L7.

been predicated on the representation of the genome as a signed permutation of $(12\cdots n)$, in the case of unichromosomal organisms, or as a fragmented signed permutation in the case of multichromosomal organisms. Each of the $n$ terms in this representation of a genome corresponds to a unique term in the other, possibly in a different position or with a different sign. Each term represents a gene or other marker that has been mapped by genetic or molecular biological techniques or abstracted from the underlying sequence data. In large measure, genome rearrangement algorithms have been developed in the context of organellar or other small genomes, where gene finding and ortholog identification have not been major obstacles. Recent improvements in efficiency (Bader *et al.*, 2001; Bergeron, 2001; Tesler, 2002) enable the algorithmic approach to genome rearrangement to handle many thousands of genes in reasonable computing time. With large nuclear genome sequences, particularly from the higher eukaryotes, however, uncertainties in how to align one genome to another, lack of complete consensus inventories of genes, and the difficulties of distinguishing among paralogs widely distributed across the genome constitute apparently insurmountable impediments to the direct application of the algorithms.

### *2.1. The Pevzner–Tesler protocol*

In comparing drafts of the human and mouse genomes, Pevzner and Tesler (2003a, b, c) adopted an ingenious strategem to leap-frog the global alignment, gene finding, and ortholog identification steps. They analyzed almost 600,000 relatively short (average length 340 bp) anchors of highly aligned sequence fragments as a starting point for building blocks of conserved synteny. They amalgamated neighboring subblocks using a variety of criteria to avoid disruptions due to "microrearrangements"—small inversions, deletions, movements—of segments less than 1 Mb and then discarded from further analysis the blocks remaining below this threshold. This procedure eventually succeeds in inferring a set of 281 blocks larger than 1 Mb, somewhat more than predicted in the original study of Nadeau and Taylor (1984) or the number of "synteny bins" (somewhat more than 200) in the current NCBI Human–Mouse Homology Maps (*www.ncbi.nlm.nih.gov/Homology/ComMapDoc.html*).

Pevzner and Tesler go a step further and use the order of these blocks on the 23 chromosomes as input to an improved version (Tesler, 2002) of the gene order rearrangement algorithms originally devised by Hannenhalli and Pevzner (1995, 1999) and Hannenhalli (1996) in order to reconstruct aspects of the actual sequence of inversions and translocations responsible for the divergent structures of the two genomes. This step takes account neither of the sizes of the blocks nor of the portion of the genome excluded by virtue of the exclusion of small blocks.

### *2.2. The reuse statistic r*

One of the key results reported by Pevzner and Tesler pertains to the "reuse" of the breakpoints between each contiguous pair of syntenic blocks they used as input to their rearrangement algorithms. Basic to the combinatorial optimization approach to inferring genome rearrangements are the (Watterson *et al.*, 1982) bounds

$$\frac{b}{2} \le d \le b, \tag{1}$$

where $b$ is the number of breakpoints and $d$ the number of rearrangements necessary to convert one genome into another. (Note that without loss of generality in the unichromosomal case, one of the two genomes may be represented as the identity permutation $12\cdots n$, and we consider there to be a breakpoint if the first block in the other genome is not "1" or the last block is not "$n$." For multichromosomal genomes, analogous but more complicated conditions define breakpoints at the beginning or end of chromosomes.)

We define breakpoint reuse as

$$r = \frac{2d}{b}. \tag{2}$$

Then

$$1 \le r \le 2. \tag{3}$$

The lower value $r = 1$ is characteristic of an evolutionary trajectory where each inversion or translocation breaks the genome at two sites specific to that particular rearrangement; no other inversion or translocation

breaks the genome at either of these sites. High values of $r$, near $r = 2$, are characteristic of evolutionary histories where each rearrangement after the first one breaks the genome at one new site and at one previously broken site. Pevzner and Tesler found that $r = 1.9$ in their comparison of the human and mouse genomes, so that on the average, each breakpoint was involved in about two rearrangement events. They argued that this was evidence that evolutionary breakpoints are concentrated in fragile regions, i.e., regions susceptible to recurrent breakage, covering a relatively small proportion of the genome.

Now, it is also an observed property of purely random permutations of length $n$, where it can be shown that $b \approx n$, that the number of inversions needed to sort them ($d$) is very close to $n$, and thus breakpoint reuse is close to 2. We discuss this in Section 5 below. Without disputing the substantive claim about fragile regions, for which there may be independent evidence (e.g., Kent *et al.*, 2003), we may ask what breakpoint reuse in empirical genome comparison really measures: a bonafide tendency for repeated use of breakpoints or simply the degree of randomness of one genome with respect to the other at the level of synteny blocks, in the sense that the evolutionary history has been partially or totally obscured by random noise, such as the discarding of small blocks. We will show here how this randomness may be an artifact of the Pevzner–Tesler protocol for constructing the synteny blocks.

## 3. SIMULATING INVERSION WITH A BLOCK-SIZE THRESHOLD

To see whether a high inferred rate of breakpoint reuse necessarily reflects a high rate when the genome was derived, we will generate a genome with NO breakpoint reuse ($r = 1$), then mimic the Pevzner–Tesler imposition of a block-size threshold and calculate $r$ for the remaining configuration of blocks. We generate a permutation of length $n = 1{,}000$ or $n = 100$ by applying $d$ "two-breakpoint" inversions to the identity permutation $(12 \cdots n)$. A two-breakpoint inversion is one that disrupts two hitherto intact adjacencies in the starting (i.e., identity) permutation. At each step, the two breakpoints are chosen at random among the remaining original adjacencies. This represents the extreme hypothesis of no breakpoint reuse at all during evolution.

Of course, our "blocks" are just elements in the permutation and have no associated size, and indeed the Hannenhalli–Pevzner procedures do not involve any concept of block size. Thus, to imitate the effect of imposing a block-size threshold, we simply delete a fixed proportion of the blocks at random, the same blocks from both the starting and derived genomes, relabel the remaining blocks according to their order in the starting (identity) genome, and apply the Hannenhalli–Pevzner algorithm.

It can be shown that before any deletions, the Hannenhalli–Pevzner algorithm will recover exactly $d$ inversions. At each step, it will find a configuration of form

$$\cdots gh \mid -(i-1), \ldots, -(h+1) \mid ij \cdots, \tag{4}$$

where the "|" represents a breakpoint and will "undo" the inversion between $h$ and $i$, removing two breakpoints. There being $b = 2d$ breakpoints, breakpoint reuse is 1.0.

What happens when we delete blocks before we apply the Hannenhalli–Pevzner algorithm? Suppose $j \neq i + 1$ in the above example, and we delete $i$. Then the two-breakpoint inversion from $-(i-1)$ to $-(h+1)$ is no longer available to undo. An inversion that erases the breakpoint between $h$ and $-(i-1)$ will not eliminate a second breakpoint. So while the distance $d$ drops by 1, the number of breakpoints $b$ also only drops by 1, and from Equation (2), $r$ increases.

The probability that one, two, or more two-breakpoint inversions are "spoiled" in this way depends on the number of blocks deleted.

Figure 1 shows how $r$ increases with $\theta$, the proportion of blocks deleted, for different values of $d$, for $n = 100$, and $n = 1{,}000$.

We note

- the initial rate of increase of $r$ depends only on $d/n$; the larger $d/n$, i.e., the more rearranged the genome, the greater the increase in $r$.
- the increase in $r$ levels off below $r = 2$ and then descends sharply. The maximum level attained increases with $n$.

In the next two sections, we explain these two observations.

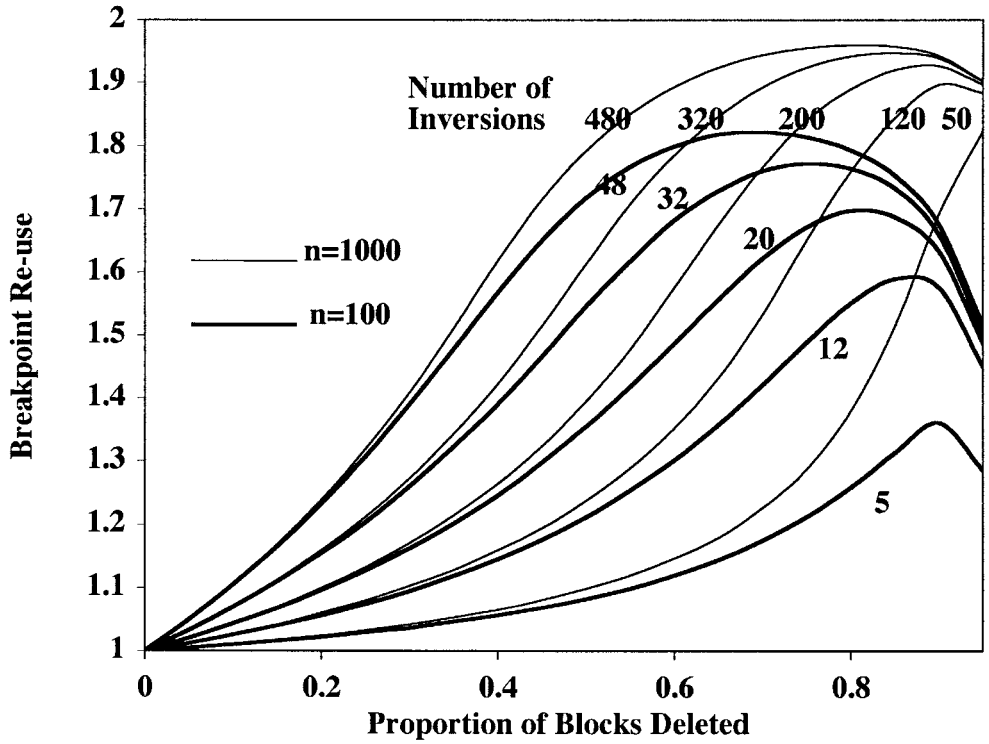


**FIG. 1.** Effect of deleting random blocks on breakpoint reuse, as a function of proportion of blocks deleted, for various levels of rearrangement of the genome.

## 4. A MODEL FOR BREAKPOINT REUSE

In this section, we explain the detailed shape of the $r$ versus $\theta$ curves in Fig. 1, in particular the increasing portions of these curves before they level off.

The combinatorial relationships among breakpoints as they are "undone" by the Hannenhalli–Pevzner algorithm are very specific and intricate. But by neglecting all but the most immediate of these relationships, we can develop a simplified probabilistic model, based partly on the original random process we used to generate the genomes. We can then calculate the expected change of $b$ and $d$ as the result of deleting one "small" block at each step. This change motivates a (deterministic) recurrence whose behavior we compare to our simulation results in order to verify that our simple model captures the quantitative essentials of the dependence of $r$ on $\theta$. This includes the observation about the pertinence of $d/n$ for the initial shape of the curves.

Suppose a genome $G$ has $b$ breakpoints with respect to $(12\cdots n)$ and the inversion distance is $d = d_2 + d_1$, where $d_1$ and $d_2$ represent the number of one-breakpoint inversions and two-breakpoint inversions required to sort $G$ optimally. (We disregard the small number of cases that require an inversion removing no breakpoints.) Then

$$b = 2d_2 + d_1, \tag{5}$$

since no breakpoints are introduced during the sorting algorithm. Note that the decomposition of $b$ in Equation (5) is unique, though there may be many alternative sequences of $d = d_1 + d_2$ inversions that can sort $G$. We focus on one particular such sequence $\mathcal{S}$. For each breakpoint $x$ in $G$, we can identify exactly one inversion in $\mathcal{S}$ that removes $x$, although $x$ may previously have acted as a breakpoint for one or more other inversions that did not remove it, in which case $x$ is being "reused" by the inversion that removes it. This identification relation between inversions and breakpoints then affects only one of the breakpoints of each one-point inversion and both of the breakpoints of a two-point inversion.

Suppose that we delete one block $i$ at random and relabel blocks $j = i + 1, \ldots, n$ as $j = i, \ldots, n - 1$, respectively. The number of breakpoints $b$ changes only if the original block $i$ was flanked by two breakpoints, i.e., if its left-hand neighbor is not $i - 1$ and its right-hand neighbor is not $i + 1$. (Or, for a block with negative polarity $-i$, its left-hand neighbor is not $-(i + 1)$ and its right-hand neighbor is not $-(i - 1)$.) Otherwise, the deletion of $i$ and the relabeling would not remove a breakpoint.

Because the breakpoints are distributed by our construction uniformly across the genome, the probability of this bilateral flanking event is

$$\phi = b(b - 1)/n(n - 1). \tag{6}$$

And this remains true of $\phi(t)$ as $n(t) = n - t$ and $b(t)$ decrease through successive deletion of randomly chosen blocks at "time" steps $t = 1, \ldots$.

Suppose now block $i$ is flanked by two of the $b = 2d_2 + d_1$ breakpoints in $G$. These two breakpoints are removed by two inversions in one of three major ways in $\mathcal{S}$. To construct our simplified probabilistic model, we first calculate the probabilities of each case by considering the two inversions involved as independent choices from $d_1$ one-breakpoint inversions and/or $d_2$ two-breakpoint inversions, as appropriate.

1. The left-hand and right-hand breakpoints are each removed by a separate one-breakpoint inversion. There are $d_1(d_1 - 1)$ events of this kind possible.
2. Each is removed by a separate two-breakpoint inversion. Since either end of each two-breakpoint inversion may be the one adjacent to $i$, there are $2d_2(2d_2 - 1) - d_2 = 2d_2(2d_2 - 2)$ possible events of this kind. The $-d_2$ accounts for the fact that once one end of a two-breakpoint inversion is chosen to be adjacent to $i$, the other end is generally not also adjacent to $i$.
3. One is removed by a one-breakpoint inversion and one by a two-break point inversion. There are $4d_1d_2$ events of this kind, taking into account the facts that the two-breakpoint inversion can be on either side of $i$ and that either end of it may be adjacent to $i$.

There are other ways in which two breakpoints on either side of $i$ may be removed in $\mathcal{S}$, such as by a single two-breakpoint inversion, but these are relatively rare for large $n$ and so we do not include them in our probabilistic model.

Then the total number of the three kinds of events is $(2d_2 + d_1)(2d_2 + d_1 - 1) - d2 = b(b - 1) - d_2$, and the probabilities of each of the three kinds are the following.

$$\text{Case 1.} \quad p_1 = \frac{d_1(d_1 - 1)}{b(b - 1) - d_2} \tag{7}$$

$$\text{Case 2.} \quad p_2 = \frac{2d_2(2d_2 - 2)}{b(b - 1) - d_2} \tag{8}$$

$$\text{Case 3.} \quad p_3 = \frac{4d_1d_2}{b(b - 1) - d_2} \tag{9}$$

When $i$ is deleted by one of the three kinds of events, what is the effect on $b, d_1$, and $d_2$ in $\mathcal{S}$? In Case 1, one breakpoint remains where $i$ used to be, and $b$ is diminished by 1. Normally, neither of the two inversions flanking $i$ will remain in a new optimal sequence of inversions $\mathcal{S}'$ on the $n - 1$ remaining blocks, since neither removes any breakpoints, but we can assume that some optimal $\mathcal{S}'$ exists where $d_1$ is diminished only by 1, since $b = 2d_2 + d_1$, although other scenarios are possible. We will discuss these later in this section.

In Case 2, on the other hand, if we retain all the inversions in $\mathcal{S}$ in constructing $\mathcal{S}'$, the first of the original two-breakpoint inversions to apply will remove only one breakpoint (not the breakpoint where $i$ used to be), but the second will remove two. Thus, removing $i$ diminishes both $b$ and $d_2$ by one, but increases $d_1$ by one.

Similarly, in Case 3, if the two-breakpoint inversion applies first in $\mathcal{S}$, the one-breakpoint inversion will still apply, both $b$ and $d_2$ will drop by one, and $d_1$ will increase by one. In the other half of the instances

TABLE 1. PROBABILITIES AND USUAL EFFECTS OF DISCARDING GENE $i$, GIVEN IT IS FLANKED BY TWO BREAKPOINTS[a]

| *Case* | *Configuration* | *Probability* | *Net effect on* $d_1$ | *Net effect on* $d_2$ |
|---|---|---|---|---|
| 1 | $g \mid -(i-1), \ldots, h \mid i \mid j, \ldots, -(i+1) \mid k$ | $\frac{d_1(d_1-1)}{b(b-1)-2d_2}$ | $-1$ | 0 |
| 2 | $g \mid -(i-1), \ldots, -(g+1) \mid i \mid -(k-1), \ldots, -(i+1) \mid k$ | $\frac{2d_2(2d_2-2)}{b(b-1)-2d_2}$ | $+1$ | $-1$ |
| 3 | $g \mid -(i-1), \ldots, -(g+1) \mid i \mid j, \ldots, -(i+1) \mid k$ | $\frac{4d_1d_2}{b(b-1)-2d_2}$ | $+1$ | $-1$ |

[a]Probabilities include those of inverted or nested versions (not listed) of configurations shown.

of Case 3, where the one-breakpoint inversion applies first in $\mathcal{S}$, after $i$ is deleted this inversion can no longer remain in an optimal sequence $\mathcal{S}'$, and the two-breakpoint inversion becomes a one-breakpoint inversion. Since $b$ and $d_2$ cannot both drop by one with no net change in $d_1$, we assume that in some optimal sequence $\mathcal{S}'$, there is an increase of one in $d_1$, though other scenarios are possible.

In actual application of the Hannenhalli–Pevzner algorithm, there are usually many optimal $\mathcal{S}$ and many optimal $\mathcal{S}'$. For the purposes of constructing our model, however, summarized in Table 1, we have assumed a minimum of changes between $\mathcal{S}$ and $\mathcal{S}'$.

The quantities in the last three columns of Table 1 suggest the deterministic model:

$$d_2(t+1) = d_2(t) + \frac{b(t)(b(t)-1)}{(nt)(n-t-1)}(-p_2(t) - p_3(t)) \tag{10}$$

$$= d_2(t) - \frac{2d_2(t)(2d_2(t)-2) + 4d_1(t)d_2(t)}{(1-\epsilon(t))(n-t)(n-t-1)} \tag{11}$$

$$d_1(t+1) = d_1(t) + \frac{b(t)(b(t)-1)}{(n-t)(n-t-1)}(p_2(t) + p_3(t) - p_1(t)) \tag{12}$$

$$= d_1(t) + \frac{2d_2(t)(2d_2(t)-2) + 4d_1(t)d_2(t) - d_1(t)[d_1(t)-1]^+}{(1-\epsilon(t))(n-t)(n-t-1)} \tag{13}$$

where $\epsilon(t) = d_2(t)/[b(t)(b(t)-1)]$ and $t$ ranges from 0 to $n$, with initial conditions $b(0) = 2d_2(0) = 2d(0)$ and $d_1(0) = 0$. The term $[d_1(t)-1]^+$ is zero for noninteger values of $d_1(t) < 1$.

Figure 2 shows how the recurrence in Equations(10)–(13) models closely the average evolution of $r$ as the number of blocks randomly deleted increases, particularly at the outset, before there are large numbers of one-breakpoint inversions in the Hannenhalli–Pevzner reconstruction.

To explain the initial coincidence between the curves for $n = 100$ and $n = 1{,}000$ in Fig. 1, let $\theta = t/n$, the proportion of blocks deleted. From the last two columns of Table 1, we can see that $b$ diminishes by 1 each time a block with flanking breakpoints is deleted, i.e., each time $n$ diminishes by 1, which itself occurs with probability given by (6). From Equations (11) and (13),

$$d(t+1) = (t) - \frac{d_1(t)[d_1(t)-1]^+}{(1-\epsilon(t))(n-t)(n-t-1)}. \tag{14}$$

For a continuous-time approximation of our model, we could calculate from Equations (2), (14), and (6),

$$\frac{dr}{d\theta} = n\left(\frac{2}{b(t)}\frac{-d_1(t)[d_1(t)-1]^+}{(1-\epsilon(t))(n-t)(n-t-1)} - \frac{2d(t)}{b(t)2}\frac{-b(t)(b(t)-1)}{(n-t)(n-t-1)}\right). \tag{15}$$

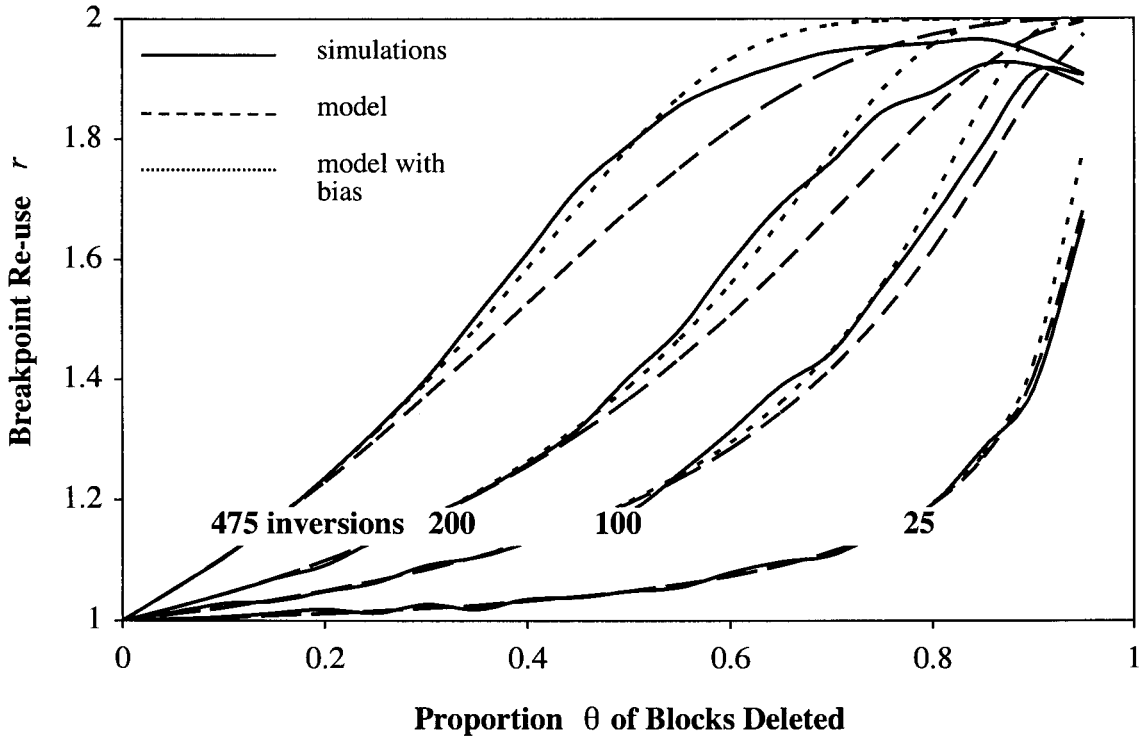


**FIG. 2.** Re-use in simulated genome and deterministic models with equal and with biased choice of inversion.

Because the first term in the parenthesis is identically zero near $\theta = 0$, we can see that

$$dr/d\theta|_{\theta=0} \approx 2d/n. \tag{16}$$

This is why the initial shape of the the curves for $n = 100$ with $I$ inversions is indistinguishable from that for $n = 1{,}000$ with $10I$ inversions in Fig. 1.

We can also see from Equations (2) and (15), how $r$ in our model approaches the asymptote $r = 2$ when $d_1(t) \nearrow d(t)$ and $d_1(t) \nearrow b(t)$ simultaneously, since then

$$\frac{dr}{d\theta} \approx \frac{2d(t)}{n-t}\left(1 - \frac{d_1(t)2}{b(t)d(t)}\right) \nearrow 0. \tag{17}$$

As $d_1(t)$ increases, however, the model first underestimates the increase in reuse rates but then continues to increase toward $r = 2$ while the the value of $r$ in the simulations actually decreases.

The underestimate can be traced to our assumption that when a block is deleted, an optimal sequence $\mathcal{S}'$ for the reduced genome can be derived from $\mathcal{S}$ by just deleting or changing the inversions flanking the block, as in Table 1. In many cases, however, the new optimal sequence $\mathcal{S}'$ will tend to be more dramatically reorganized than that and will often contain a new two-breakpoint inversion not foreseen in our model. The effect of this is that two-breakpoint inversions will be more numerous after a proportion $\theta$ of the blocks are deleted than our model predicts. This effect may be incorporated in our model by assuming that a two-breakpoint inversion will be sampled $2\alpha d_2(t)/d_1(t)$ times more frequently than a one-breakpoint inversion, for some $\alpha > 1$, instead of just $2d_2(t)/d_1(t)$ times more frequently. Figure 2 also shows how the fit of the model to the simulation data is improved with bias parameter $\alpha = 2$. Different values of $\alpha$ will improve the fit for different ranges of $\theta$.

As more and more blocks are dropped from a permutation, it loses its "structure"; i.e., for each surviving breakpoint, its partner in the original inversions creating the permuted genome has been deleted, and it participates instead with various other breakpoints in a number of other one-breakpoint inversions in an essentially unpredictable and random way. We may consider that as a curve in Fig. 1 attains its maximum,

it is entering into the "noisy" region where the historical signal, i.e., the combinatorial relationships among the breakpoints of the inversions producing the initial permutation, becomes thoroughly obscured.

## 5. THE NOISY REGION

In the previous section, we explored the initial shape of the $r$ versus $\theta$ curves in Fig. 1 and explained how this depended on $2d/n$. In this section, we discuss

- why these curves reach a maximum $r < 2$ before $\theta = 1$, and
- why the maximum is less when $n$ is smaller.

The explanation for these features lie the greater prevalence, proportionately speaking, of two-breakpoint inversions in random permutations with small $n$ than in those with large $n$. These permutations are produced by randomly permuting the blocks $12\cdots n$ and then randomly and independently assigning positive or negative polarity to each block, in contrast to the signed permutations produced by the random two-breakpoint inversions in Section 3. In the latter, by construction, $r = 1$, but for purely random signed permutations the reuse rate $r$ approaches 2 for large $n$. It is not hard to see why. Neglecting "end effects" due to the first block not being 1 and/or the last block not being $n$, in these permutations, each of the $n - 1$ possible adjacencies of form $i$, $i + 1$ (or $-(i + 1)$, $-i$) has probability $1/(n - 1)$ of occurring. Thus we can expect only $n - 1 \times 1/(n - 1) = 1$ such adjacencies and $n - 2$ breakpoints. For each of these breakpoints, say of form $i|j$, the probability that there will be a "partner" breakpoint, i.e., of form $-(i+1)|-(j-1)$, either to the left or the right, thus allowing a two-breakpoint inversion, is about $1/4n$, based on the asymptotic independence of the adjacency events for large $n$, so that the expected number of two-breakpoint inversions is about 1/8, taking into account the double counting of each such inversion. This is in contrast with the initial permutations of Section 3, where all the breakpoints had such partners.

Other two-breakpoint inversions may occasionally be created as the rearrangement algorithm "undoes" the one-breakpoint inversions. Indeed, the final inversion to be undone is necessarily a two-breakpoint inversion, since the existence of one breakpoint in a permutation implies the existence of at least one other. Nevertheless, because of the absence, or small number, of two-breakpoint inversions "visible" in the initial permutation, a random signed permutation with large $n$ will usually require almost $b$ inversions to sort, so $r$ will be almost 2. This was noticed in simulation experiments more than 10 years ago (Kececioglu and Sankoff, 1994).

What about smaller values of $n$? Here there will no longer be independence of the adjacency events, so that if $i|j$ is a breakpoint, the probability that $-(i+1)|-(j-1)$ will also be a breakpoint increases. In addition, the "end effects" are no longer negligible. Thus, for $n = 1$, the only two genomes are 1 or $-1$, so that $r = 1$. With $n = 2$, the optimal sort of 6 of the 8 possible signed permutations involves a two-breakpoint inversion, and the average $r$ is only 4/3. The average value $r$ for simulated random genomes of small to moderate size is displayed in Table 2. These results explain the downturn in the curve for $r$ in Fig. 1. As $\theta$ increases, the number of blocks decreases, and they are increasingly randomized, with few two-breakpoint inversions left to undo. The rearrangement algorithm gives results that increasingly resemble those of random permutations. At some $\theta$, the increase in $r$ due to breakpoint reuse is outweighed by the tendency for smaller random permutations to have low $r$, and the curve reaches a maximum and starts to decrease. Where the initial genome size $n$ is 1,000, the trajectory of $r$ can exceed 1.9 before there are few enough blocks left for the random permutation effect to predominate and to halt and reverse the approach to $r = 2$. For initial genome size 100, this occurs earlier and has a greater effect, since the nearly randomized genomes are in the lower size range of Table 2.

TABLE 2. EXPECTED RE-USE AS A FUNCTION OF $n$[a]

| $n$ | 5 | 25 | 50 | 100 | 250 |
|---|---|---|---|---|---|
| $r$ | 1.53 | 1.83 | 1.90 | 1.94 | 1.97 |

[a]Average $r$ for 500 random signed permutations of size $n$.

## 6. CONCLUSIONS

In this paper, we have investigated the effect of deletion threshold size on $r$, albeit indirectly by varying the rate of random deletion of blocks. We have shown in some detail how this increases the reuse rate $r$. For example, for genomes with close to 1,000 syntenic blocks, if a half to two-thirds are deleted, leaving only a few hundred blocks, reuse will rise above 1.8.

We have thus found that we can produce reuse rates of magnitude comparable to that observed by Pevzner and Tesler, without postulating the kind of genomic inhomogeneities they inferred from such rates.

How pertinent our results are to the Pevzner–Tesler analysis depends on two assumptions. One is that it is meaningful to choose the two breakpoints of each inversion involved in generating one of the genomes at random and independently. The second is that the random deletion of blocks simulates the discarding of all blocks below a threshold size.

The first assumption is not strictly justified with respect to inversions; we know from the work of Kent *et al.* (2003) that the distribution of inversion lengths is concentrated on small values, i.e., around 1 Kb. However, these small inversions are not included in the Pevzner and Tesler analysis; they are considered "microrearrangements" that do not disrupt the definition of the larger synteny blocks in which they are embedded. The extent to which the conditioned inversion length distribution, thus truncated, departs from that predicted by independent choice of breakpoints, depends on the threshold and has not been systematically investigated. Whatever the case with inversions, a uniform distribution of translocation breakpoints seems reasonable. Though we did not carry out multichromosomal simulations, the mathematical similarities between multichromosomal and unichromosomal genome rearrangement analysis (Hannenhalli and Pevzner, 1995, 1999) suggest that our simulations based on $n = 100$ blocks before deletion, where we achieve reuse rates of 1.7 and 1.8, is pertinent to interchromosomal exchange. Most important is that where Pevzner and Tesler argue that genomic inhomogeneity must be the cause of high breakpoint reuse, we have shown that in at least one model with no inhomogeneity whatsoever, we can achieve the same levels of reuse by mimicking their procedure of discarding small blocks.

The second assumption is that blocks eliminated through a threshold criterion will be randomly distributed throughout the genome. If in the comparison of real genomes, these small blocks are clustered together, our model might not capture the effect on $r$ of deleting them, depending on whether they are created solely by local rearrangements (which may little affect $r$) or by translocation and inversion of large chromosomal segments (in which case, $r$ would be sensitive to deletion). Unfortunately, the spatial distribution of small versus large blocks has not been systematically studied.

In the conference version of this paper (Sankoff and Trinh, 2004), we also asked what is the effect of the Pevzner–Tesler repairing of small rearrangements that disrupt longer syntenic blocks? Though we ran simulations in which segment length was explicitly accounted for, and for which both amalgamation of adjoining short blocks and deletion of isolated short blocks seemed to increase $r$, we had to assume random breakpoints for long inversions, fixed size for short inversions, a threshold for amalgamation, another threshold for deletion, and a criterion for polarity assignment to amalgamated segments, all of which were somewhat arbitrary. Furthermore, Pavel Pevzner (personal communication) has pointed out likely errors in our simulation procedure. Subsequent experiments showed that with realistic sizes and numbers of short inversions, unrealistically large numbers of long inversions were necessary for the amalgamation process to have an effect, though the deletions process continued to augment $r$. It became clear that until some empirically based parametrized distribution of inversion lengths is known, simulations of the effect of amalgamation are premature.

This work was motivated by Pevzner and Teslers introduction of the reuse statistic. Though they used it to infer relative susceptibility of genomic regions to rearrangement, in our analysis it serves rather to measure the loss of signal of evolutionary history, due to the imposition of a threshold for retaining or discarding syntenic blocks. However, we take issue here neither with the imposition of thresholds, which seem to be methodologically reasonable, nor with the suggestions about inhomogeneities of genomes in their susceptibility to breakpoints. Nevertheless, we do not consider the elevated reuse rates found by these authors to be evidence for fragile regions. We have shown that breakpoint reuse of the same magnitude may very well be artifacts of the use of a threshold in a context where NO reuse actually occurred. Indeed, while this may not have been their goal, Pevzner and Tesler have invented a statistic that is a measure of the

noise affecting a genomic rearrangement process at the sequence level. The reuse rate tells us whether we can have confidence in a reconstructed evolutionary signal, whether it must be considered largely random, or whether we are in the "twilight" zone between the two.

## ACKNOWLEDGMENTS

Research was supported by grants from the Natural Sciences and Engineering Research Council (NSERC). D.S. holds the Canada Research Chair in Mathematical Genomics and is a Fellow in the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

## REFERENCES

Bader, D.A., Moret, B.M., and Yan, M. 2001. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comp. Biol.* 8, 483–491.

Bergeron, A. 2005. A very elementary presentation of the Hannenhalli–Pevzner theory. *Dis. Appl. Math.* 146, 134–145.

Hannenhalli, S. 1996. Polynomial-time algorithm for computing translocation distance between genomes. *Dis. Appl. Math.* 71, 137–151.

Hannenhalli, S., and Pevzner, P.A. 1995. Transforming men into mice (polynomial algorithm for genomic distance problem). *Proc. IEEE 36th Ann. Symp. on Foundations of Computer Science*, 581–592.

Hannenhalli, S., and Pevzner, P.A. 1999. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *J. ACM* 48, 1–27.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler. D. 2003. Evolution cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* 100, 11484–11489.

Nadeau, J.H., and Taylor, B.A. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA* 81, 814–818.

Pevzner, P.A., and Tesler, G. 2003a. Genome rearrangements in mammalian genomes: Lessons from human and mouse genomic sequences. *Genome Res.* 13, 37–45.

Pevzner, P.A., and Tesler, G. 2003b. Transforming men into mice: The Nadeau-Taylor chromosomal breakage model revisited. *Proc. RECOMB '03, 7th Int. Conf. on Computational Molecular Biology*, 247–256.

Pevzner, P.A., and Tesler, G. 2003c. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. USA* 100, 7672–7677.

Sankoff, D., and El-Mabrouk, N. 2002. Genome rearrangement. In Jiang, T., Smith, T., Xu, Y., and Zhang, M., eds., *Current Topics in Computational Biology*, 135–155. MIT Press, Cambridge, MA.

Sankoff, D., and Trinh, P. 2004. Chromosomal breakpoint re-use in the inference of genome sequence rearrangement. *Proc. RECOMB '04, 8th Int. Conf. on Computational Molecular Biology*, 30–35.

Tesler, G. 2002. GRIMM: Genome rearrangements web server. *Bioinformatics* 18, 492–493.

Waterston, R., *et al.* 2002. Initial sequencing and analysis of the mouse genome. *Nature* 420, 520–562.

Watterson, G., Ewens, W., Hall, T., and Morgan, A. 1982. The chromosome inversion problem. *J. Theoret. Biol.* 99, 1–7.

Address correspondence to:
*David Sankoff*
*Department of Mathematics and Statistics*
*University of Ottawa*
*585 King Edward Avenue*
*Ottawa, Canada, K1N 6N5*

*E-mail:* sankoff@uottawa.ca