Estimators of Translocations and Inversions in Comparative Maps

David Sankoff[1] and Matthew Mazowita[1]

Department of Mathematics and Statistics, University of Ottawa,
585 King Edward Avenue, Ottawa ON, Canada, K1N 6N5
{sankoff,mmazo039}@uottawa.ca
http://albuquerque.bioinformatics.uottawa.ca

Abstract. In a comparative map, the number of translocations in the evolutionary history of a chromosome can be estimated solely on the basis of the conserved *syntenies* it contains. This estimate, subtracted from the number of conserved *segments*, then allows the estimation of the number of inversions that have affected the chromosome. Summing these estimates over all chromosomes provides a startlingly accurate estimator (as assessed by simulation) of the total number of rearrangements of each type occurring in the evolutionary divergence of two genomes.

1 Introduction

The quantitative comparative study of whole-genome maps, exemplified by the linkage-based work of Nadeau and Taylor [8] and more recent versions based on gene content of conserved segments [16, 15, 5], makes no formal reference to the processes that create the breakpoints between conserved segments while progressively fragmenting these segments. It only assumes implicitly that the number of breakpoints and segments increases in proportion to the number of rearrangement events affecting either of the two genomes being compared. In contrast, the algorithmic approach to genome rearrangements [4, 19] infers a most parsimonious history of specific inversions and reciprocal translocations. The particulars of this inference are not always reliable due to the highly non-unique nature of the solutions, the characteristic underestimation of parsimony and, especially, the susceptibility of these methods to lose the details, though not the overall trends, in the evolutionary signal [17].

Between these analytical extremes of ignoring the processes giving rise to a comparative map and ambitiously trying to infer them in all their detail, are there any limited aspects of rearrangement history that can be inferred with some confidence? Building on the ideas in [13] and [14], we claim that by analyzing the number of conserved syntenies in a comparative map, we can accurately infer, using a simple estimator, the number of reciprocal translocations involved in generating this map. Furthermore, depending on the degree of resolution of the map, by contrasting the number of conserved segments with the number of conserved syntenies, we can estimate the number inversions or other intra-chromosomal events. Applying this methodology to data compilations on human-mouse maps at differing levels of resolution shows that the estimated number

J. Lagergren (Ed.): RECOMB 2004 Ws on Comparative Genomics, LNBI 3388, pp. 109–122, 2005.

of reciprocal translocations is relatively stable, but that the inferred number of inversions increases with finer resolution.

2 Models

We can model the autosomes of a genome as c linear segments with lengths $p(1), \cdots, p(c)$, proportional to the number of base pairs they contain, where $\sum_{i=1}^{c} p(i) = 1$. To balance realism and simplicity in our model, we:

- Set aside the sex chromosomes, which are largely excluded from inter-chromosomal exchanges.
- Impose a threshold and a cap on chromosome size, rejecting any translocation that results in a chromosome too small or too large. Theories about meosis, e.g. [18], can be adduced for these constraints, though there are clear exceptions, such as the "dot" chromosomes of avian and some reptilian and other vertebrate genomes [2, 1].
- Impose a left-right orientation on each chromosome, such that a left-hand fragment must always rejoin a right-hand fragment. In reality, an inverted left-hand fragment may rejoin another left-hand fragment, but the only statistical effect of our restriction is to ensure that, throughout the simulation, each chromosome always retains a segment, however small it may become, containing its original left-hand extremity. This restriction models the conservation of the centromere without introducing complications such as trends towards or away from acrocentricity.
- Postpone our consideration of chromosome fusion and fission, so that the number of chromosomes is constant throughout the time period governed by the model. Later, we simply assume that the case where fusions or fissions occur will be well approximated by interpolating two models (with fixed chromosome number) corresponding to the two genomes being compared.

We also assume the two breakpoints of a translocation are chosen independently according to a uniform distribution over all autosomes, conditioned on their not being on the same chromosome. There is no statistical evidence [11] that translocational breakpoints cluster in a non-random way on chromosomes, except in a small region immediately proximal (within 50-300Kb) to the telomere in a wide spectrum of eukaryote lineages [7].

A reciprocal translocation between two chromosomes h and k consists of breaking each one, at some interior point, into two segments, and rejoining the four resulting segments such that two new chromosomes are produced, each containing a left-hand part of one of the original chromosomes and the right-hand part of the other. We label each new chromosome according to which left-hand it contains, but for each of its constituent segments, we retain the information of which ancestral chromosome it derived from.

With further translocations, if a breakpoint falls into a previously created segment on chromosome i, it divides that segment into two new segments, the left-hand one remaining in chromosome i, while the right-hand one, and all

the other segments to the right of the breakpoint, are transferred to the other chromosome involved in the translocation,

In contrast to translocations, the two breakpoints of an inversion cannot be considered to be independently positioned on the chromosome. According to Kent et al. [6] the median length of an inversion, of which there are many thousands in the mouse-human comparison, is less than 1 Kb, so that on a chromosomal scale most inversions have their two breakpoints very close together.

There are no definitive data on the distribution of inversion lengths. The best that exist, based on human-mouse comparisons, were published in Table 2 of [6], which estimates about 8000 inversions (with or without partial or complete duplications) with median length from 300-800 bp (depending on the subcategory of inversion), including 160 inversions longer than 100 Kb within longer syntenic blocks (*intrablock* inversions). In [10] it is estimated that there are 150 inversions longer than 1 Mb; these each involve at least one whole syntenic block (*suprablock* inversions) and hence do not overlap with the 160 previously mentioned. These data (median = 600 with 310 inversions out of 8000 longer than 1 Mb) determine a gamma distribution with shape parameter $\alpha = 6.539$ and scale parameter (on a logarithmic scale) $\beta = 0.447$. This distribution has a median of 600 bp and a 0.05 tail > 100 Kb, to include 0.02 intrablock, 0.02 suprablock and a generous 0.01 for those inversions falling through the cracks between the definitions of 100 Kb intrablock and 1Mb suprablock inversions.

In our simulations, we will test our models with a much more disruptive distribution of inversion lengths. We generate these lengths according to a gamma distribution with shape parameter $\alpha = 3$ and scale parameter (on a logarithmic scale) $\beta = 1.127$.

One inversion breakpoint is chosen at random on the genome, as with translocations, and the second is chosen equiprobably to the left or right and according to the specified gamma. If the second breakpoint exceeds the end of the chromosome, both breakpoints are disregarded and a substitute inversion is generated. With this truncation protocol, keeping in mind the logarithmic scale, only about 7 % of the inversions are larger than 1 Mb. For any accounting of rearrangements which allows 8000 or more inversions of all sizes, seven percent still represents a generous number of very large inversions (> 1 Mb). We do this to provide as severe as possible a test of the formulae we will develop, which are more susceptible to fail if there are many large inversions.

The total number of segments on a human chromosome i is

$$n^{(i)} = t^{(i)} + 2u^{(i)} + 1, \tag{1}$$

where $t^{(i)}$ is the number of translocational breakpoints on the chromosome, and $2u^{(i)}$ is the number of inversion breakpoints.

3 Prediction and Estimation

We assume that our random translocation process is temporally reversible, and to this effect we show in Figure 1 and Section 4.1 that the equilibrium state of

our process well approximates the observed distribution of chromosome lengths in the human genome. This assumption allows us to treat the mouse genome as ancestral and the human as derived (or vice versa), instead of considering them as diverging independently from a common ancestor.

To start with, we consider a process that involves translocations but no inversions. At the outset, assume the first translocation on the human lineage involves ancestral chromosome i. The assumption of a uniform density of breakpoints across the genome implies that the "partner" of i in the translocation will be chromosome j with probability $p_i(j) = \frac{p(j)}{1-p(i)}$. Thus the probability that the new chromosome labelled i contains no fragment of ancestral chromosome j, where $j \neq i$, is $1 - p_i(j)$. For small $t^{(i)}$, after chromosome i has undergone $t^{(i)}$ translocations, the probability that it contains no fragment of the ancestral chromosome j is approximately $(1 - p_i(j))^{t^{(i)}}$, neglecting second-order events, for example, the event that j previously translocated with one or more of the $t^{(i)}$ chromosomes that then translocated with i, and that a secondary transfer to i of material originally from j thereby occurred.

Then the probability that the new (i.e., human) chromosome i now contains at least one fragment from j is approximately $1-(1-p_i(j))^{t^{(i)}}$ and the expected number of ancestral chromosomes with at least one fragment showing up on human chromosome i is

$$E(c^{(i)}) \approx 1 + \sum_{j \neq i} [1 - (1 - p_i(j))^{t^{(i)}}], \tag{2}$$

where the leading 1 counts the fragment containing the left-hand endpoint of the ancestral chromosome i itself. We term $c^{(i)}$ the number of *conserved syntenies* on chromosome i.

3.1 The Case of No Saturated Chromosomes

Substituting $c^{(i)}$ for $E(c^{(i)})$ in eqn (2) suggests solving

$$c - c^{(i)} = \sum_{j \neq i} [1 - p_i(j)]^{\widehat{t^{(i)}}}, \tag{3}$$

to provide an estimator of $t^{(i)}$. Newton's method, initialized by the estimator in Section 3.4, converges rapidly for the range of parameters used in our studies, as long as $c^{(i)} \neq c$. This is the empirically interesting case; we know of no comparative map where a chromosome of one genome shares at least one significant syntenic segment with every autosome of the other genome.

Then $\hat{t} = \frac{1}{2} \sum_{i=1}^{c} \widehat{t^{(i)}}$ is an estimator of the total number of translocations intervening between the ancestral and modern genome, since each translocation is counted on two chromosomes.

In the hypothetical case $c^{(i)} = c$, we say chromosome i is saturated and there is no solution to eqn 3. For the sake of completeness, we will also study this case.

3.2 When There Are Some Saturated Chromosomes

If the genome after a certain number translocations contains at least one saturated chromosome, i.e., with $c^{(i)} = c$, our estimator must take on a different form.

Let c^* be the number of saturated chromosomes. Suppose there have been t translocations in the evolutionary history, with

$$t^{(i)} = 2tp(i) \tag{4}$$

affecting autosome i. Since the probability is approximately $1-[1-p_i(j)]^{t^{(i)}}$ that at least one segment from original chromosome j is contained by chromosome i, if these events were independent for all the $c-1$ original chromosomes j, (which they are obviously are not, for small values of t), then the probability that $c^{(i)} = c$, i.e., chromosome i contains segments from all $c-1$ of the others, as well as the original i by default, would be the

$$P(i,t) = \prod_{j \neq i} (1 - [1 - p_i(j)]^{t^{(i)}}). \tag{5}$$

Now, we may assume independence is asymptotically approached with large t, so that the expected number of saturated chromosomes $E(c^*)$ is approximately P_t, where

$$P_t = \sum_{i=1}^{c} P(i,t), \quad \sigma^2 = \sum_{i=1}^{c} P(i,t)(1 - P(i,t)). \tag{6}$$

These quantites can all be calculated from the $p_i(j)$ based on the given $p(i)$. Then, in the presence of c^* saturated chromosomes, we define $\hat{t}(c^*)$ to be the inverse function of P_t applied to c^*.

3.3 The Completely Saturated Case

The estimators in Section 3.2 are defined as long as $c^* < c$. When all c chromosomes are saturated, the *completely saturated* case, the best we can do is to define $\hat{t}(c) = \hat{t}(c-1)$, with the understanding that this may well be a severe underestimate.

3.4 Equal Size Chromosomes

When all chromosomes are of equal size, a situation that can be maintained only by requiring the chromosomal fragments exchanged during translocation to be of the same length,

$$E(c^{(i)}) \approx 1 + (c-1)[1 - (1 - \frac{1}{c-1})^{t^{(i)}}]. \tag{7}$$

In this case [14], eqn (3) is directly solved as:

$$\widehat{t^{(i)}} = \frac{\log(c-1) - \log(c - c^{(i)})}{\log(c-1) - \log(c-2)}. \tag{8}$$

3.5 The Effect of Inversions

Inversions have the effect of fragmenting and changing the order of the segments that are transferred among chromosomes by translocation. An estimate of the number of inversion breakpoints on a chromosome is derived from eqn (1) as

$$2\widehat{u^{(i)}} = n^{(i)} - \widehat{t^{(i)}} - 1, \tag{9}$$

and the number of inversions will be half that. The presence of inversions will have an effect on the estimate of t. Modeling this effect of inversions is not easy; prior inversions on chromosome j can either increase or decrease the effect of a translocation of i and j on $c^{(i)}$. As the inversion rate increases, however, the process becomes a "gossip" process among the c autosomes – after each interaction (i.e., communication) between two chromosomes, they both contain material from every original chromosome in the union of the two chromosomes before the interaction. Here, the chromosomes are saturated at a very rapid rate.

4 Simulations

4.1 Equilibrium Distribution of Chromosome Size

Models of accumulated reciprocal translocations for explaining the observed range of chromosome sizes in a genome date from the 1996 study of Sankoff and Ferretti [12]. They proposed a lower threshold on chromosome size in order to reproduce the appropriate size range in plant and animal genomes containing from two to 22 autosomes. A cap on largest chromosome size has also been proposed [18] and shown to be effective [3]. Economy and elegance in explaining chromosome size being less important in the present context than simulating a realistic equilibrium distribution of these sizes, we imposed both a threshold of 50 Mb and a cap of 250 Mb on the process described in Section 2, simply rejecting any translocations that produced chromosomes out of the range. These values were inspired by the relative stability across primates and rodents evident in the data in Table 1.

Simulating the translocation process 100 times up to 10,000 translocations each produced the equilibrium distribution of chromosome sizes in Figure 1. The superimposed distribution of human autosome sizes is very close to the equilibrium distribution.

Table 1. Shortest and longest chromosome, in Mb

genome	shortest	longest
mouse	61	199
human	47	246
rat	47	268
chimp	47	230

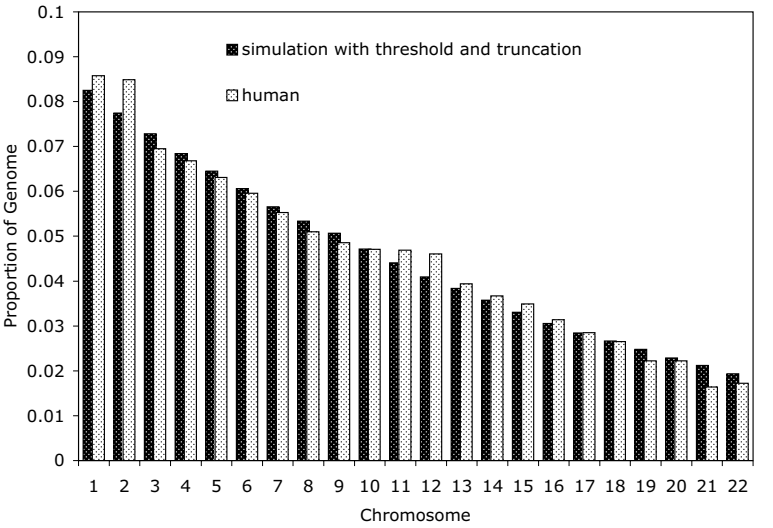

Fig. 1. Comparison of equilibrium distribution of simulated chromosome sizes with human autosome sizes

4.2 Domains of the Estimators

In 1000 runs of the simulation, the smallest t for which any chromosome was found to be saturated was 127. Indeed, up to $t = 348$, no saturated chromosomes were found in over half of the runs. By $t = 552$, however, all thousand runs had at least one saturated chromosome. Figure 2 depicts how the mean number of saturated chromosomes increases as a function of t.

As mentioned in Section 2, the range of empirical interest, at least for human-mouse comparisons, is thus well within the range of the estimator based on eqn (3). It is conceivable, however, that some remotely related genomes might require the estimator based on eqn (5).

In 1000 runs, the smallest t for which a genome was found to be completely saturated was 1747. By $t = 2749$, half of the runs were completely saturated; by $t = 3886$ all were completely saturated.

4.3 Performance of the Estimator When There Are No Saturated Chromsomes

Figure 3 (left) depicts the estimated number of translocations as a function of the true number t in the simulation, when no saturated chromosomes are encountered. For $t < 400$, the estimator $\hat{t}$ appears completely unbiased, with only moderate variance. As explained in Section 4.2 and as can be seen in Figure 2, $t < 400$ is also the range where saturated chromosomes are rarely encountered.

For higher values of t we had to run the simulation many additional thousands of times to accumulate 100 runs without any saturated chromosome, fully

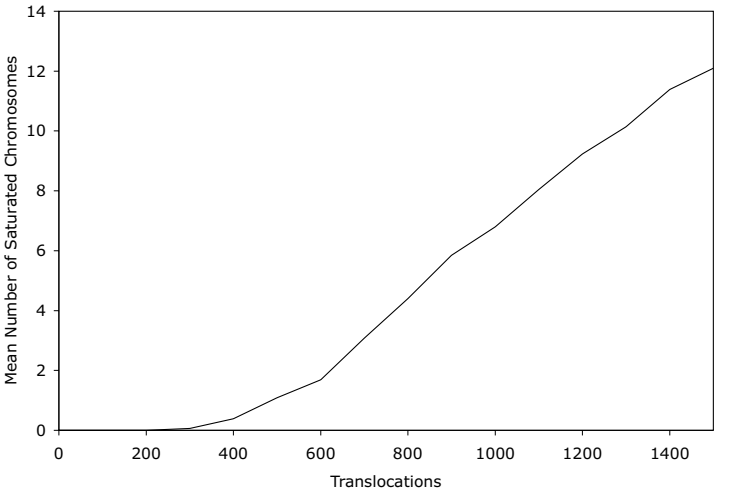

Fig. 2. Average number, over 100 runs, of saturated chromosomes per genome, as a function of the number of translocations. There are very few genomes with saturated chromosomes for $t < 400$

conscious that this atypical sub-sample was unlikely to result in accurate estimates of t. This accounts for the severe bias for large t.

4.4 Performance of the Estimator with Some Saturated Chromsomes

Application of the version of the estimator presented in Section 3.2 to instances where one or more of the chromosomes are saturated gave the results in Figure 3 (right). This has a small (around 70 translocations) constant (hence easily corrected) negative bias over the range from $t = 600$, where there is still a minority of genomes with any saturated chromosomes, to $t = 2000$, when most genomes have many saturated chromosomes. The error rate, however, is much higher than the estimator based on eqn (3) over the range where they can be compared.

For low and high values of t we had to run the simulation many extra times to accumulate 1000 runs with at least one but less than 22 saturated chromosomes. The bias this introduces is evident for $t < 500$.

4.5 The Effect of Inversions on $\hat{t}$

In our simulations, we interspersed v inversions between successive translocations, for $v = 0, 1, 10, 50$ and 100. The effect of this was to bias $\hat{t}$ positively, a proportion of inversions being inferred as translocations. This is depicted in Figure 4. This bias seems quite severe for the larger values of v, but it should be

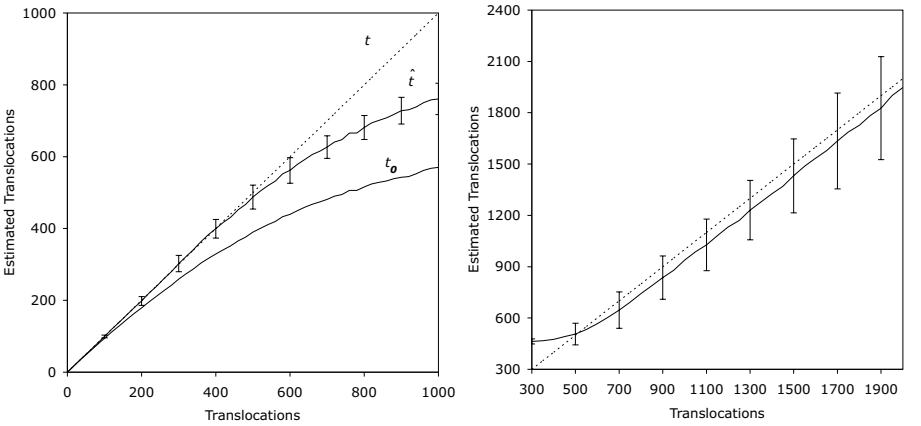

Fig. 3. (Left) Mean value, over 100 runs, of $\hat{t}$ as a function of t where there are no saturated chromosomes. Error bars ± 1 s.d. Estimator t_0 based on equal-length chromosome model, eqn (8). (Right) Mean value, over 1000 runs, of $\hat{t}$ as a function of t for the case of saturated chromosomes, using the version of the estimator in Section 3.2. Note that as in Figure 2, there are very few genomes with saturated chromosomes for $t < 400$

remembered from Section 2 that to detect these effects we are using very exaggerated inversion lengths. When we substitute a more realistic gamma function, no bias is apparent, even for $v = 50$ as indicated in the figure.

It is ironic that as while high inversion rate increases the bias in translocations rates, it actually decreases the bias in estimating the inversion rate, as a proportion of the total number of inversions. This follows since a small proportion of the number of inversions will be a large proportion of the number of translocations.

4.6 The Effect of Inversions on Saturation

The addition of inversions to the simulation accelerates the saturation of the chromosomes.

In our simulations, the lowest t's for which we encountered a saturated chromosome ($c^* = 1$) was $t = 127$ with no inversions (the second lowest was $t = 195$), $t = 140$ with one inversion per translocation, $t = 145$ with ten inversions per translocation, $t = 133$ for 50 inversions, $t = 110$ for 100 inversions, and $t = 38$ for the gossip process (1000 runs with zero, one, ten inversions and gossip, 100 runs with 50 and 100 inversions).

The lowest t's for which we encountered a completely saturated case ($c^* = c$) was $t = 1747$ with no inversions, $t = 1653$ with one inversion per translocation, $t = 1177$ with ten inversions per translocation, $t = 729$ for 50 inversions, $t = 450$ for 100 inversions, and $t = 71$ for the gossip process.

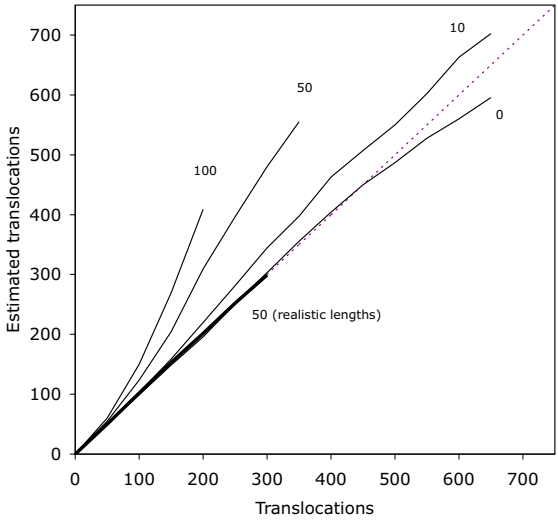

Fig. 4. Increase in number of translocations inferred as a function of intrachromosomal activity (inversions). Curves labelled by number of inversions per translocation. Curves end when computing time considerations make it unfeasible to accumulate 100 runs with no saturated chromosomes. Data are shown for simulations with two alternative models of inversions, one with a realistic inversion length distribution and the other with exaggerated inversion lengths. The heavy line shows the result of using the realistic distribution: virtually no effect on the accuracy of the estimator. The thin line shows what would happen were the inversion lengths unrealistically large

The way in which c' grows in the presence of large numbers of inversions should approach the performance of the gossip process discussed in Section 3.5, so that saturation should be reached at the same time as the gossip process. It is clear from these results that even with 100 inversions per translocation, we are still quite far from the limiting case.

5 Application to the Human-Mouse Comparison

To illustrate the use of the estimators, for the 22 human autosomes, a 100 Kb resolution construction abstracted from the UCSC Genome Browser indicates 237 autosomal segments, while the sum of the $c^{(i)}$ is 107. Solving eqn (3) for each autosome, based on its $c^{(i)}$, summing the 22 values of $\widehat{t^{(i)}}$, and dividing by 2, gives a total of 50 translocations.

By eqn (1), this leaves unaccounted for

$$\begin{aligned} 2\sum \widehat{u^{(i)}} &= \sum n^{(i)} - 2\hat{t} - 22 \\ &= 115 \end{aligned}$$

segments, which must be attributed to local rearrangements such as $\hat{u} \approx 58$ inversions.

Table 2. Inference of inter- and intra-chromosomal rearrangements based on number of conserved segments and number of segment-sharing autosome pairs in the two genomes. Calculated separately from data on segments on 22 human autosomes (H) and on 19 mouse autosomes (M). Sources: UCSC Genome Browser May 2004, builds Mm5 and Hg17, level = 1, alignments $\geq$ 100 Kb, GRIMM 300 Kb and 1 Mb blocks, from http://nbcr.sdsc.edu/GRIMM/HMR_Aug2003, NCBI human build 34.3 and mouse build 32.1

resolution of comparative map	autosomal segments $\sum n^{(i)}$ H/M	segment-sharing chromosome pairs $\sum c^{(i)}$	inter-chromosomal $\frac{1}{2}\sum \widehat{t^{(i)}}$ H/M/**mean**	intra-chromosomal $\sum \widehat{u^{(i)}}$ H/M/**mean**
100 Kb (UCSC)	237 254	107	50 51 **50**	58 67 **62**
300 Kb (GRIMM)	377 377	109	50 52 **51**	127 127 **127**
1 Mb (GRIMM)	268 268	104	47 49 **48**	76 76 **76**
n/a (NCBI)	379 381	110	51 53 **52**	128 128 **128**

Table 2 shows the results of these calculations for this and a number of other maps of various levels of resolution. Of interest is the relative stability of the estimates of the number of reciprocal translocations versus the dependence of local rearrangements on resolution.

6 Discussion

We have documented the behaviour of an estimator of the number of translocations intervening between two rearranged genomes, based only on the numbers of conserved syntenies on each chromosome, the lengths of the chromosomes and a simplified random model of interchromosomal exchange. In the absence of inversions, this estimator has undetectable bias up to 400 translocations, and has a rather moderate variance. Addition of high rates of inversion introduces some bias, though to detect this in our simulations we had to use many thousands of unrealistically large, hence maximally disruptive, inversions. If we also know the number of conserved segments, we can infer the number of inversions with corresponding accuracy.

The good properties of this estimator even after hundreds of translocations are remarkable given that it only explicitly takes into account the first-order effects of interchromosomal exchange. The fact that the introduction of chromosome sizes improves the estimator to such a degree compared to the previous version in [14], is somewhat surprising in that these sizes fluctuate greatly during the simulation.

Though this estimator is almost always well-defined and accurate (i.e., unbiased) for over 400 translocations in the human case, even in the presence of considerable intrachromosomal activity, we also explored a version of the estimator applicable to the situation with saturated chromosomes. This situation would be of some practical interest between 300 to 1000 translocations or more, though we are not aware of such cases being discussed in the biological liter-

ature. We found the estimator to be accurate after about 600 translocations, with a small but constant bias. We looked into this estimator because it was derived differently from eqn (3), but there are other, perhaps better, possibilities. For example, instead of solving eqn (3) for each chromosome separately, we are currently investigating the incorporation of eqn (4) into (3), so that we can solve

$$c^2 - \sum c^{(i)} = \sum_i \sum_{j \neq i} [1 - p_i(j)]^{2\hat{t}p(i)}, \tag{10}$$

for $\widehat{t}$ directly for the entire genome, dispensing with the distinction between models for genomes having no saturated chromosomes and those having some.

In this paper, we applied our estimate to the human-mouse comparison at various levels of resolution. This showed that translocation estimates are extremely stable, while variability in the number of inversions inferred accounted for all the variation in the number of conserved segments due to differing levels of resolution. This reflects the discovery of high numbers of smaller-scale local arrangements recognizable from genomic sequence [6].

Our estimates of the number of translocations and inversions in the evolutionary divergence of man and mouse are only about a half of what has been published by Pevzner and Tesler [9, 10] who have attempted to reconstruct algorithmically the details of this history. Our model assumes each translocation and inversion creates two new segments, but the algorithms require a number of rearrangements almost equal to the number of segments to account for how the segments are ordered on the chromosomes. This accounts for the difference between the two sets of results. The reason the algorithms require one rearrangement per segment instead of one rearrangement per two segments is either

- Rearrangements almost always use at least one previously used breakpoint per rearrangement instead of two new ones, because breakpoints are largely confined to a small number of *fragile* regions on each chromosome, so that there is no parsimonious analysis of the segment orders which involves all or mostly two-breakpoint rearrangements, or
- Our two breakpoint per rearrangement model is correct, but the neglect, in the algorithmic approach, of segments smaller than a certain threshold value obscures the history and presents the algorithm with an effectively randomized order of segments along the chromosome [17]. Genomes with randomly ordered chromosomal segments tend to require one rearrangement per segment.

In any case, we also note that the proportion of inversions and translocations, if not their absolute numbers, is the same in our approach as in the results of the algorithmic approach.

Our model includes a feature that approximates the principle of conservation of the centromere. This principle prohibits translocations that result in one chromosome with no centromere and the other with two centromeres. In ongoing work we are attempting to drop this feature, since it is not always operative on the evolutionary time scale, taking into account such mechanisms as cen-

tromere inactivation and neocentromere activation, or chromosome fusion and chromosome fission.

Acknowledgements

Thanks to Aleksander Lenert and Phil Trinh for guidance with the genome browsers and other tools, to Adrian Maler for calculating the parameters of the inversion-size distribution, to David Kempe, Jon Kleinberg and David Liben-Nowell for discussions on the gossip problem, and to Nabil Benabbou for help with running our simulations on the High Performance Computing Virtual Laboratory facilities. Research supported by a Discovery grant to DS and and an undergraduate summer research scholarship to MM from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS 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

1. Bed'hom, B. (2000). Evolution of karyotype organization in *Accipitridae*: A translocation model. In Sankoff, D. and Nadeau, J. H. (eds) *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families*. Dordrecht, NL, Kluwer, 347–56.
2. Burt, D.W. (2002). Origin and evolution of avian microchromosomes *Cytogenetic and Genome Research*, **96,** 97–112.
3. De, A., Ferguson, M., Sindi, S. and Durrett, R. (2001). The equilibrium distribution for a generalized Sankoff-Ferretti model accurately predicts chromosome size distributions in a wide variety of species. *Journal of Applied Probability*, **38,** 324–34.
4. Hannenhalli, S. and Pevzner, P. A. (1995). Transforming men into mice (polynomial algorithm for genomic distance problem). *Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science*. 581–92.
5. Housworth, E. A. and Postlethwait, J. (2002).Measures of synteny conservation between species pairs. *Genetics*, **162,** 441–8.
6. 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.
7. Mefford, H.C. and Trask, B.J. (2002). The complex structure and dynamic evolution of human subtelomeres. *Nature Reviews in Genetics*, **3,** 91-102; 229.
8. Nadeau, J. H. and Taylor, B. A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences, USA*,**81,** 814-8.
9. Pevzner, P. A. and Tesler, G. (2003). Genome rearrangements in mammalian genomes: Lessons from human and mouse genomic sequences. *Genome Research*, **13,** 37-45
10. Pevzner, P. A. and Tesler, G. (2003). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences, USA*, **100,** 7672-7

11. Sankoff, D., Deneault, M., Turbis, P. and Allen, C.P. (2002) Chromosomal distributions of breakpoints in cancer, infertility and evolution. *Theoretical Population Biology*, **61,** 497–501.
12. Sankoff, D. and Ferretti, V. (1996). Karotype distributions in a stochastic model of reciprocal translocation. *Genome Research*, **6,** 1–9.
13. Sankoff, D., Ferretti, V, and Nadeau, J.H. (1997) Conserved segment identification. *Journal of Computational Biology*, **4,** 559–65.
14. Sankoff, D., Parent, M.-N. and Bryant, D. (2000). Accuracy and robustness of analyses based on numbers of genes in observed segments. In Sankoff, D. and Nadeau, J. H. (eds) *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families*. Dordrecht, NL, Kluwer, 299–306.
15. Sankoff, D., Parent, M.-N., Marchand, I. and Ferretti, V. (1997). On the Nadeau-Taylor theory of conserved chromosome segments. In Apostolico, A. and Hein, J. (eds) *Combinatorial Pattern Matching. Eighth Annual Symposium*. Lecture Notes in Computer Science **1264,** Springer Verlag, 262–74.
16. Sankoff, D. and Nadeau, J.H. (1996). Conserved synteny as a measure of genomic distance. *Discrete Applied Mathematics*, **71,** 247–57.
17. Sankoff, D. and Trinh, P. (2004). Chromosomal breakpoint re-use in the inference of genome sequence rearrangement. *Proceedings of RECOMB 04, Eighth International Conference on Computational Molecular Biology*. New York: ACM Press, 30–5.
18. Schubert, I. and Oud, J.L. (1997). There is an upper limit of chromosome size for normal development of an organism. *Cell*, **88,** 515–20.
19. Tesler, G. (2002). Efficient algorithms for multichromosomal genome rearrangements, *Journal of Computer and System Sciences*, **65,** 587–609.