BMC Bioinformatics

BioMed Central

Research article

Dependence of paracentric inversion rate on tract length

Thomas L York*[1], Rick Durrett[2] and Rasmus Nielsen[1,3]

Address: [1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, USA, [2]Department of Mathematics, Cornell University, Ithaca, USA and [3]Center for Bioinformatics, University of Copenhagen, Copenhagen, Denmark

Email: Thomas L York* - tly2@cornell.edu; Rick Durrett - rtd1@cornell.edu; Rasmus Nielsen - rasmus@binf.ku.dk

* Corresponding author

Published: 3 April 2007

BMC Bioinformatics 2007, **8**:115 doi:10.1186/1471-2105-8-115

Received: 18 January 2007

Accepted: 3 April 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/115

Abstract

Background: We develop a Bayesian method based on MCMC for estimating the relative rates of pericentric and paracentric inversions from marker data from two species. The method also allows estimation of the distribution of inversion tract lengths.

Results: We apply the method to data from *Drosophila melanogaster* and *D. yakuba*. We find that pericentric inversions occur at a much lower rate compared to paracentric inversions. The average paracentric inversion tract length is approx. 4.8 Mb with small inversions being more frequent than large inversions.

If the two breakpoints defining a paracentric inversion tract are uniformly and independently distributed over chromosome arms there will be more short tract-length inversions than long; we find an even greater preponderance of short tract lengths than this would predict. Thus there appears to be a correlation between the positions of breakpoints which favors shorter tract lengths.

Conclusion: The method developed in this paper provides the first statistical estimator for estimating the distribution of inversion tract lengths from marker data. Application of this method for a number of data sets may help elucidate the relationship between the length of an inversion and the chance that it will get accepted.

Background

Reconstructing the history of inversions and/or translocations separating two chromosomes or genomes is a classical problem in computational biology dating back as far as early work by the pioneers of genetic research from the 1930's (eg. [1]). In many applications, this problem has been treated as a problem of finding the minimum number of events required in the evolutionary history of the two genomes. The computational problem involved is known as sorting by reversal (e.g. [2-4]). An alternative approach is to estimate the number of events using statistical estimators that take into account that more inversions (and translocations) may have occurred than the minimum possible number. Larget et al. [5] and York et al. [6] have developed Bayesian methods based on Markov Chain Monte Carlo (MCMC) for estimating the history of inversions separating two chromosomes. The following description is based on the method of York et al. [6]. In brief, a Markov chain is established that has, as its stationary distribution, the posterior distribution of inversion paths (possible histories of inversions).

The likelihood function is calculated assuming inversions occur according to a Poisson process and assuming a uni-

form prior over all possible inversions paths of a fixed length. The inversion path is then represented explicitly in the computer memory and updates are proposed according to a proposal kernel, allowing exploration of the posterior distribution. The update kernel is guided by the parsimony distance computed from the breakpoint graphs developed for solving the sorting by reversal problem [4,7]. Using the parsimony distance to guide updates greatly increases convergence rates of the Markov chain. Point estimates of the number of inversions, with associated measures of statistical confidence are then obtained from the posterior distribution. The method of [6] was extended in [8] to the case of multiple chromosomes differing by an unknown number of translocations and inversions. Similarly, [9] extends this type of approach to rearrangements due to transpositions and inverted transpositions in addition to inversions. The advantage of these Bayesian approaches is that they use all of the information in the marker data to obtain a statistical estimate of the number of inversions and translocations. However, so far these approaches have assumed that a long chromosomal segment is as likely to be inverted as a short one, and have lumped together pericentric and paracentric inversions rather than distinguishing between them. Pericentric inversions appear to be rarer than paracentric ones, and there is evidence for a length-dependent effect also [10,11], with selection related to recombination in the inverted region of inversion heterozygotes as a possible cause. Another simplification made hitherto is that only the order of markers in a set (and their orientations, in the case of signed data) has been used, so no account is taken of the uneven spacing of the markers. The objective of this paper is to modify the previous methods to take into account these factors. This will allow us to estimate the relative frequency of pericentric and paracentric inversions and to estimate the distribution of inversion tract lengths. We apply the new method to genomic data from *D. melanogaster* and *D. yakuba*.

The assumption of tract-length independence is relaxed also in [12], which considers the problem of finding the optimal inversion path when the cost of an inversion depends on its tract-length; they do not address determining that dependence from data.

Results and Discussion

Results

We have analyzed a set of 388 markers on the three major chromosomes, 2, 3 and X, using distance information from *D. yakuba* but only marker order information from *D. melanogaster*. The positions of the centromeres are also used. For chromosome 3 we found 163 markers in 23 conserved blocks as shown in figure 1. For the purposes of finding the inversion distance between these two marker arrangements this may be represented as the following signed permutation:

1, -4, 3, -2, 5, -7, 6, -8, 9, -10, 11, 20, -13, 14, -18, 16, -17, -15, 19, 12, 21, -22, 23.

The inversion distance is 13. The centromere is between blocks 10 and 11. For chromosome X we found 84 markers in 14 blocks as shown in figure 2. The signed permutation is:

1, -2, 3, -10, -4, 11, -9, -13, -7, 5, 12, 8, 14.

The inversion distance is 7. The centromere is between block 14 and the chromosome end. Similarly for chromosome 2 we found 141 markers in 19 blocks as shown in figure 3. The signed permutation is:

1, -4, 6, -2, 5, -3, 7, 12, -8, -11, -13, -9, 10, -14, 15, -18, -16, -17, 19.

The inversion distance is 11 including one pericentric inversion. The centromere is in the middle of block 11. Thus, it takes at least 31 inversions, including at least one pericentric inversion, to turn the *D. yakuba* marker arrangement on chromosomes 2, 3 and X into the *D. melanogaster* arrangement.

A run of 1.1×10^6 updates was performed, taking 90 cpu hours on a 1.8 GHz Athlon processor. The first 1.1×10^5 updates were discarded as burn-in. Figures 4 through 6 show histograms of quantities of interest using the remainder of the MCMC output; agreement among the four replicate chains (shown with dotted and dashed lines) is very good, indicating good MCMC convergence.

Table 1 lists 95% credible intervals and maximum a posteriori (MAP) estimates of the number of parametric inversions, Lpa, the rate parameters for paracentric, and pericentric, λpa and λpe, and the parameter β which describes the strength of the tract-length dependent effect in our model. These parameters are more fully defined in the methods section.

From figure 4 it is clear that the number of inversions is compatible with the parsimony estimate of 31 inversions. However, the most likely number of inversions is 33. The credible interval (Table 1) excludes more than 37 inversions at the 95% level. The 95% credible interval for the rate parameter λ_{pa} is [0.028, 0.094] Mb^{-2}.

A minimum of one pericentric inversion (on chromosome 2) is needed to rearrange the markers, and the posterior probability of > 1 pericentric inversions is $< 1 \times 10^{-4}$. The

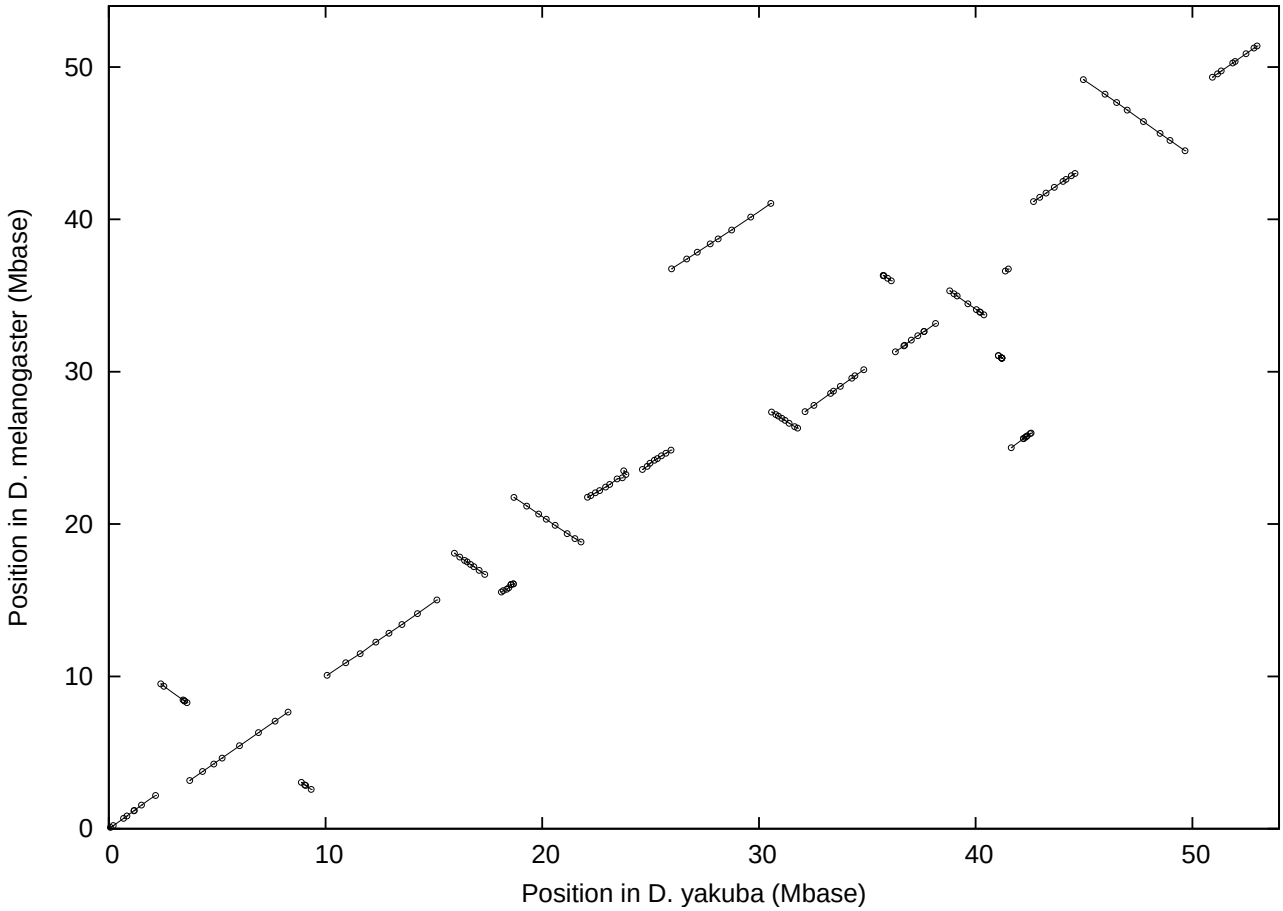

Figure 1
Position in *D. melanogaster* vs. position in *D. yakuba* for chains after filtering, chromosome 3. Lines show blocks.

pericentric rate parameter, λ_{pe}, is correspondingly small, with a MAP estimate of 7.5×10^{-4} Mb^{-2}.

The MAP estimate of the tract length dependence parameter (β) is 0.130 Mb^{-1} with a 95% credible interval of [0.044, 0.22] Mb^{-1} (Table 1 and Figure 6). Using (6), and the chromosome arm lengths (20.7, 22.3, 21.2, 24.2 and 28.7 Mb), we find that $\beta_{MAP} = 0.130$ Mb^{-1} corresponds to a mean tract length of 4.8 *Mb*, compared with 8.1 *Mb* assuming $\beta = 0$.

The fact that β is positive, and that values of β close to zero receive very little support shows that small tract lengths are favored over large tract lengths.

Figure 7 shows the posterior joint distribution of β and λ_{pa}, together with the corresponding MAP estimate $(\beta, \lambda_{pa})_{MAP} = (0.122\ Mb^{-1}, 0.053\ Mb^{-2})$. The observed positive correlation between the two parameters is not surprising. The rate of inversions of tract length τ is proportional to $\lambda_{pa} e^{-\beta\tau}(\ell - \tau)$, which for $\lambda_{pa} > 0$ and $0 < \tau \leq \ell$ is a decreasing function of β. Unless increasing β (favoring short tract lengths more) allows the observed rearrangement to be accomplished with fewer inversions, then λ_{pa} must increase as β does.

Discussion

One drawback of the current method is that, as is the case for many other MCMC methods, it is computationally slow. Nonetheless, the speed of the program is not so slow that it is prohibitive, as illustrated in the analysis of the *Drosophila* data. The existing program should be able to handle somewhat larger data sets (up to perhaps 100 blocks and 800 markers) by some combination of running longer, running replicate chains on separate processors, and, in some cases breaking a multi-chromosome data set down into individual chromosomes or arms and analyzing each one separately. To go beyond that would require substantial work to improve the algorithm. A related issue is the resolution at which we can analyze

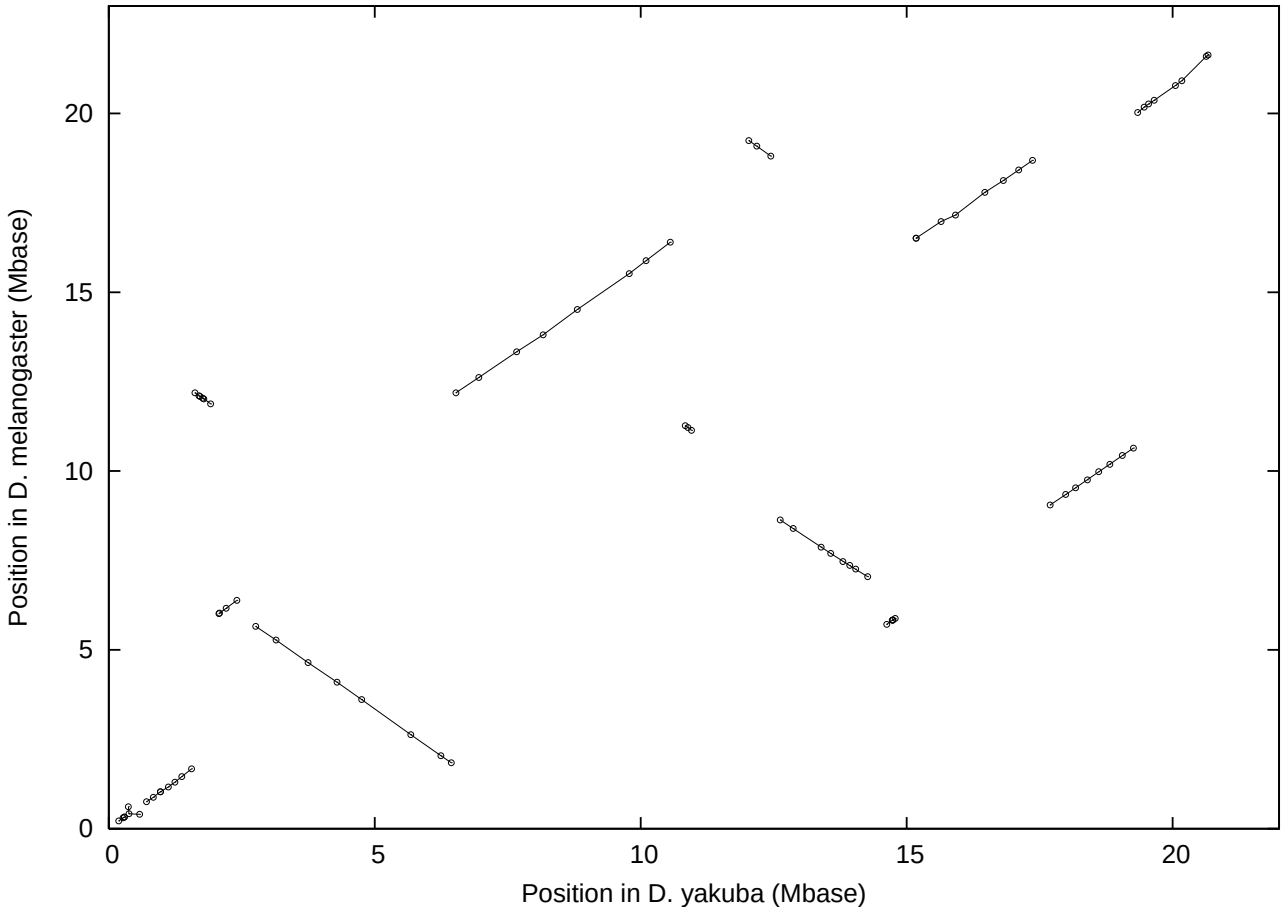

Figure 2
Position in *D. melanogaster* vs. position in *D. yakuba* for chains after filtering, chromosome X. Lines show blocks.

genome rearrangements. In addition to the increased computational burden of analyzing more and shorter blocks, the assumption implicit in our method that the observed rearrangement is due solely to inversions (and translocations) becomes more problematic for smaller scale rearrangements. The method is, therefore, more suitable for making statements regarding inversions occurring at the scale of hundreds of kilobases or megabases than at the scale of a few kilobases. Nonetheless, with several blocks less than 200 kilobases long in our data, and 5 chromosome arms totaling 117 megabases, we are apparently sensitive to inversion tract-lengths down to about 1% of the chromosome arm length.

Conclusion

The method developed in this paper provides the first statistical estimator for estimating the distribution of inversion tract lengths from marker data. Application of this method to a number of data sets may help elucidate the relationship between the length of an inversion and the chance that it will get accepted.

Methods

The model

Using marker order information only

Previously [6,8] we have considered models of rearrangements of M markers on C chromosomes, in which only the order of the markers is used. Two arrangements of a set of markers are considered to be the same if and only if every pair of markers adjacent in one arrangement is also adjacent in the other, and every marker adjacent to a chromosome end in one arrangement is adjacent to a chromosome end in the other. In the case of a single chromosome the markers divide it into $M + 1$ segments and we can distinguish $N_I = M(M + 1)/2$ inversions corresponding to unordered pairs of distinct segments. Assuming a Poisson

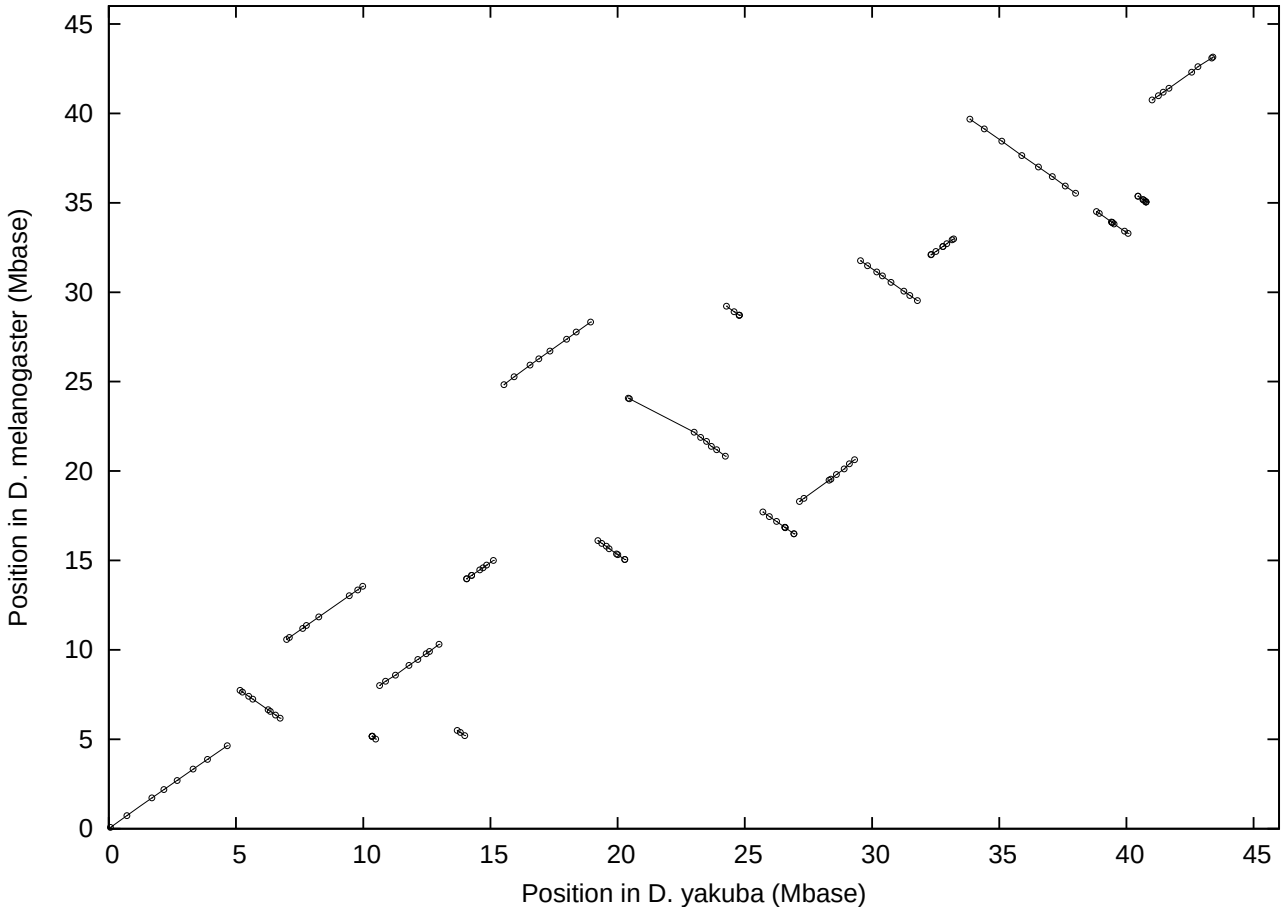

Figure 3
Position in *D. melanogaster* vs. position in *D. yakuba* for chains after filtering, chromosome 2. Lines show blocks.

process with rate Λ, and assuming the N_I inversions to be equiprobable, the probability of a particular path X consisting of L inversions is:

$$P(X \mid \Lambda) = \frac{e^{-\Lambda}\Lambda^L}{L!} N_{\mathrm{I}}^{-\mathrm{L}} = \frac{e^{-\Lambda}}{L!}\lambda^L$$

where $\lambda = \Lambda/N_I$. The posterior probability density is then

$$p(X, \lambda \mid D) \propto P(D \mid X)\frac{e^{-\Lambda}}{L!}\lambda^L p(\lambda).$$

The prior $p(\lambda)$ is taken to be uniform between 0 and λ_{max}, and zero elsewhere. The data in this case are the marker orders D_1 and D_2 observed in two taxa. A path X starting at D_1 either ends up at D_2, in which case $P(D|X) = 1$, or it ends up at some order other than D_2, in which case $P(D|X) = 0$.

We construct an initial path by starting with D_1, and performing inversions and translocations until D_2 is obtained. Using the Hannenhalli-Pevzner breakpoint graph theory of sorting by reversal (i.e. by inversions) [4,7], which has been used in all the MCMC sampling approaches to the inversion problem [5,6,9,8], we preferentially choose rearrangements that lead to short paths. Proposed updates are constructed by choosing two points along the existing path and constructing a path between them in the same way, thus guaranteeing $P(D|X) = 1$.

Starting from a particular marker order there are N_I distinct inversions, each occurring with rate λ, i.e., the probability of a particular inversion occurring in a short time t is λt where time is scaled such that the whole rearrangement process takes unit time.

In order to handle multiple chromosomes and translocations [8], we require distinct parameters λ_I and λ_T for the rates of inversions and translocations respectively. For

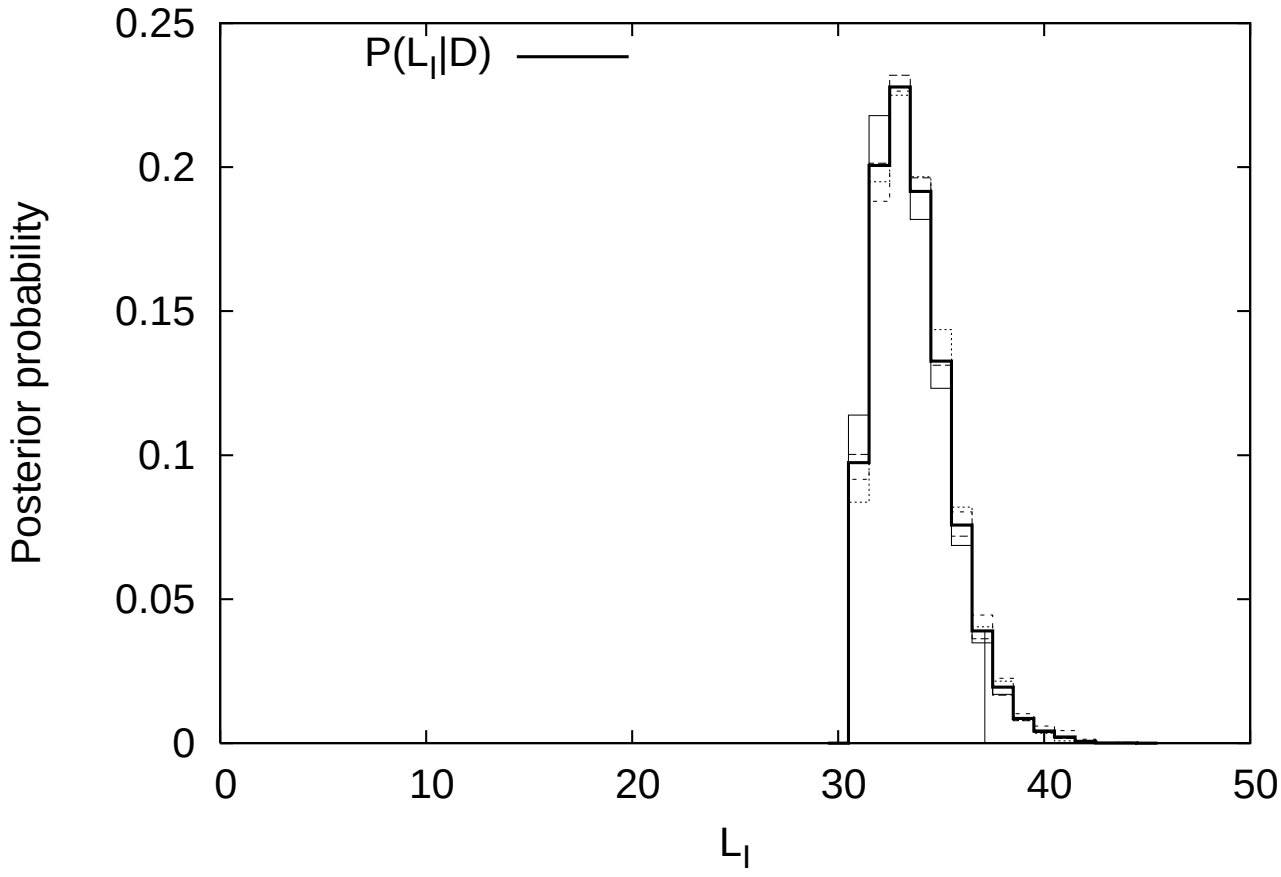

Figure 4
Posterior distribution of the number of inversions L_I. Vertical lines indicate the 95% credible interval.

each arrangement of markers there will be some number of inversions N_I and some number of translocations N_T; both of these depend on how many markers are on the various chromosomes, and therefore can change along a path. For this reason we now uniformize the process by defining a total event rate $\Lambda(\lambda_I, \lambda_T)$ which is guaranteed to be at least as great as the sum of the total inversion and total translocation rates,

$\Lambda(\lambda_I, \lambda_T) > \Lambda_{real} \equiv \Lambda_I + \Lambda_T \equiv N_I\lambda_I + N_T\lambda_T$, with "dummy" rearrangments (which have no effect on the genome) occurring with rate $\lambda_d = \Lambda - \Lambda_{real}$. Now $\Lambda(\lambda_I, \lambda_T)$ is fixed along the path and we may write

$$P(X \mid \lambda_I, \lambda_T) = \frac{e^{-\Lambda}}{L!}\lambda_I^{L_I}\lambda_T^{L_T}\prod_k \lambda_{d_k}$$

where the product is over the dummy events on the path, indexed by k. Note that the path X here is a sequence of inversions, translocations and dummies.

Using distance information

Often, in addition to knowing the order of markers, we have some form of distance information, such as recombination distance or number of nucleotides between markers. We use this information by generating proposed paths which start from one of the genomes (call it genome 1) specified in the data, with not only the marker order being as specified, but also the distances. The distance information for genome 2 is ignored. A path is then constructed which has distance information at every step, but only the marker *order* at the end of the path is required to agree with genome 2. If distance information is available for both genomes, we can choose to use the distance information from either genome but not from both. We would like to be able to use the distance information from both genomes, where available, but we don't know how to construct a path which ends not only with a specified marker order, but also with (or close to) a specified set of inter-marker distances. This is particularly difficult because in reality the sum of these distances is not conserved.

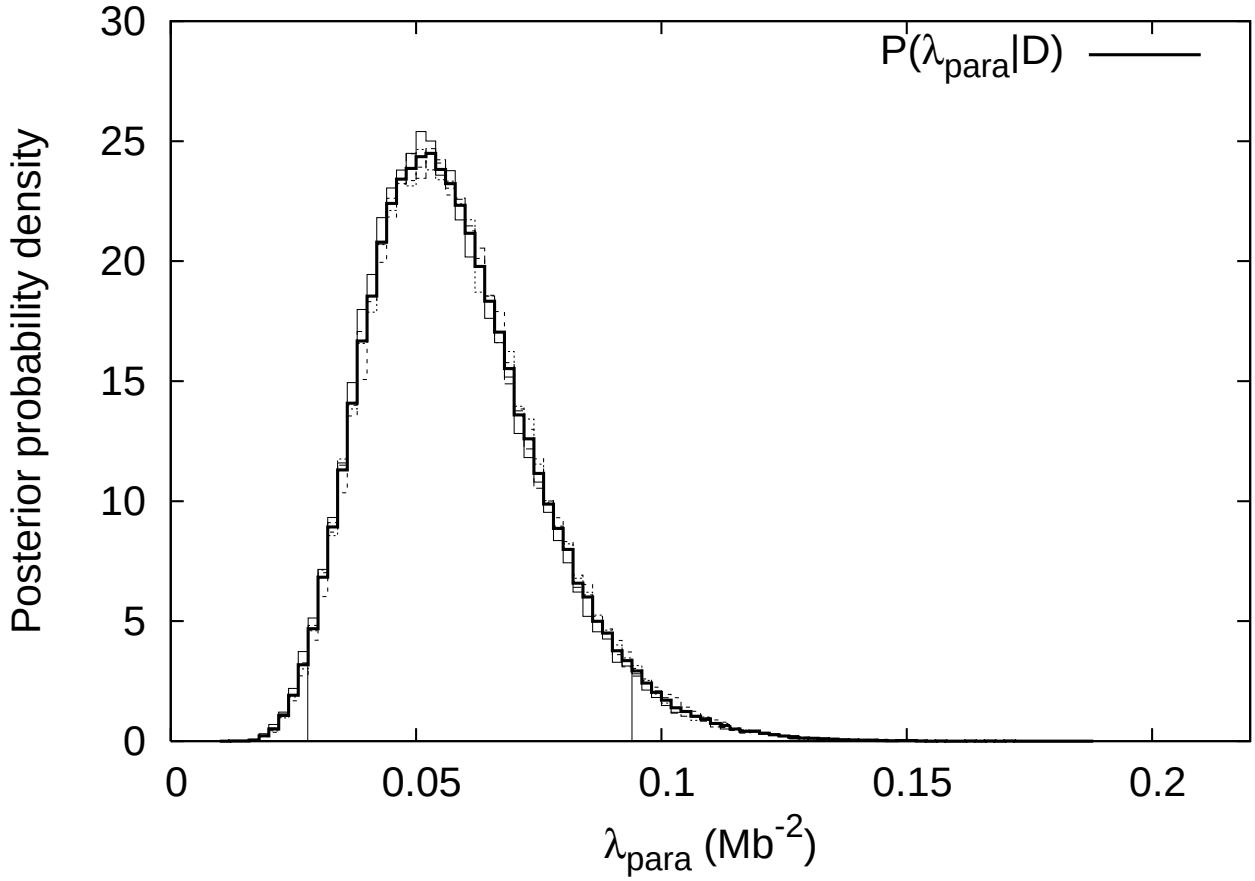

Figure 5
Posterior distribution of λ_{pa}. Vertical lines indicate the 95% credible interval.

Consider again for the moment a single chromosome, with M markers and length ℓ. When using only marker order information, we distinguished $N_I = M(M + 1)/2$ inversions and assumed equiprobability. Now, using distance information, an inversion is specified by the distances x_1 and x_2 of the breakpoints from one end of the chromosome, with (x_1, x_2) lying in the triangle $0 < x_1 < x_2 < \ell$, and we assume (for now) a uniform distribution over this region. The total rate of inversions is then $\lambda_I \ell^2/2$ including inversions of segments containing zero markers. If we exclude these the rate is $\Lambda_I = \lambda_I(\ell^2 - \sum_{i=0}^{M} s_i^2)/2$ where the s_i are distances separating adjacent markers. In the multiple chromosome case this becomes:

$$\Lambda_I = \frac{\lambda_I}{2}\sum_{j=1}^{C}\left(\ell_j^2 - \sum_{i=0}^{m_j} s_{ij}^2\right)$$

where j now indexes chromosomes, and m_j is the number of markers on chromosome j. In the corresponding expression for translocations:

$$\Lambda_T = \lambda_T F \sum_{j=1}^{C}\sum_{k=j+1}^{C} \ell_j \ell_k$$

the factor F is the number of allowed translocations for each choice of breakpoints. After breaking two chromosomes into four pieces, there are 2 ways to put them back together (in addition to the initial configuration); if both of these are allowed then $F = 2$, but if we require every chromosome to always have exactly one centromere (as we will do later) then one of these is disallowed, and $F = 1$. Now instead of (3) we have

$$p(X \mid \lambda_I, \lambda_T) = \frac{e^{-\Lambda}}{L!}\lambda_I^{L_I}\lambda_T^{L_T}\prod_k \lambda_{d_k},$$

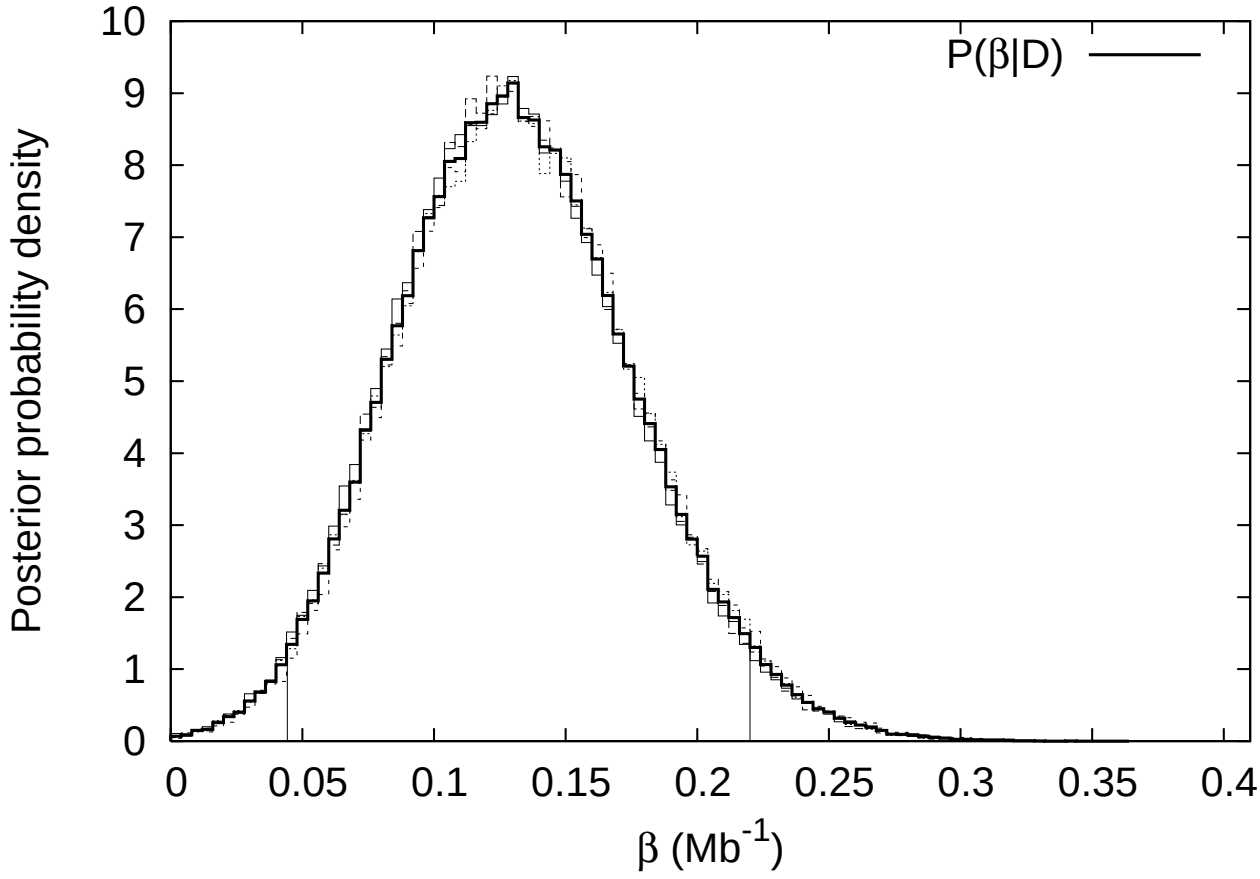

Figure 6
Posterior distribution of exponential tract-length dependence parameter β. Vertical lines indicate the 95% credible interval.

which differs from (3) in that it is a *density* and because the dummy event rates, λ_{d_k}, now depend on the continuous breakpoint positions.

An earlier version of our software, implementing this method of using distance information, but ignoring tract-lengths, was used in a comparative analysis of *Arabidopsis thaliana*, *Arabidopsis lyrata* and *Capsella* [13].

Inversion tract lengths

We are interested in investigating how the rate at which inversions occur depends on the inversion tract length, i.e., the distance between inversion breakpoints. To make this question more precise, we note that if the two breakpoints defining an inversion are distributed uniformly and independently along a chromosome, then the tract length, $\tau \equiv |x_2 - x_1|$, is distributed as $p(\tau) \propto (\ell - \tau)$, $0 < \tau < \ell$, and the mean tract length is $\ell/3$. Now let us consider a joint distribution of the breakpoints which falls exponen-

Table 1:

	95% credible interval	MAP estimate	units
L_{pa}	[30,36]	32	
λ_{pa}	[0.028, 0.094]	0.053	Mb^{-2}
λ_{pe}	[0, 0.0041]	7.5×10^{-4}	Mb^{-2}
β	[0.044, 0.22]	0.13	Mb^{-1}

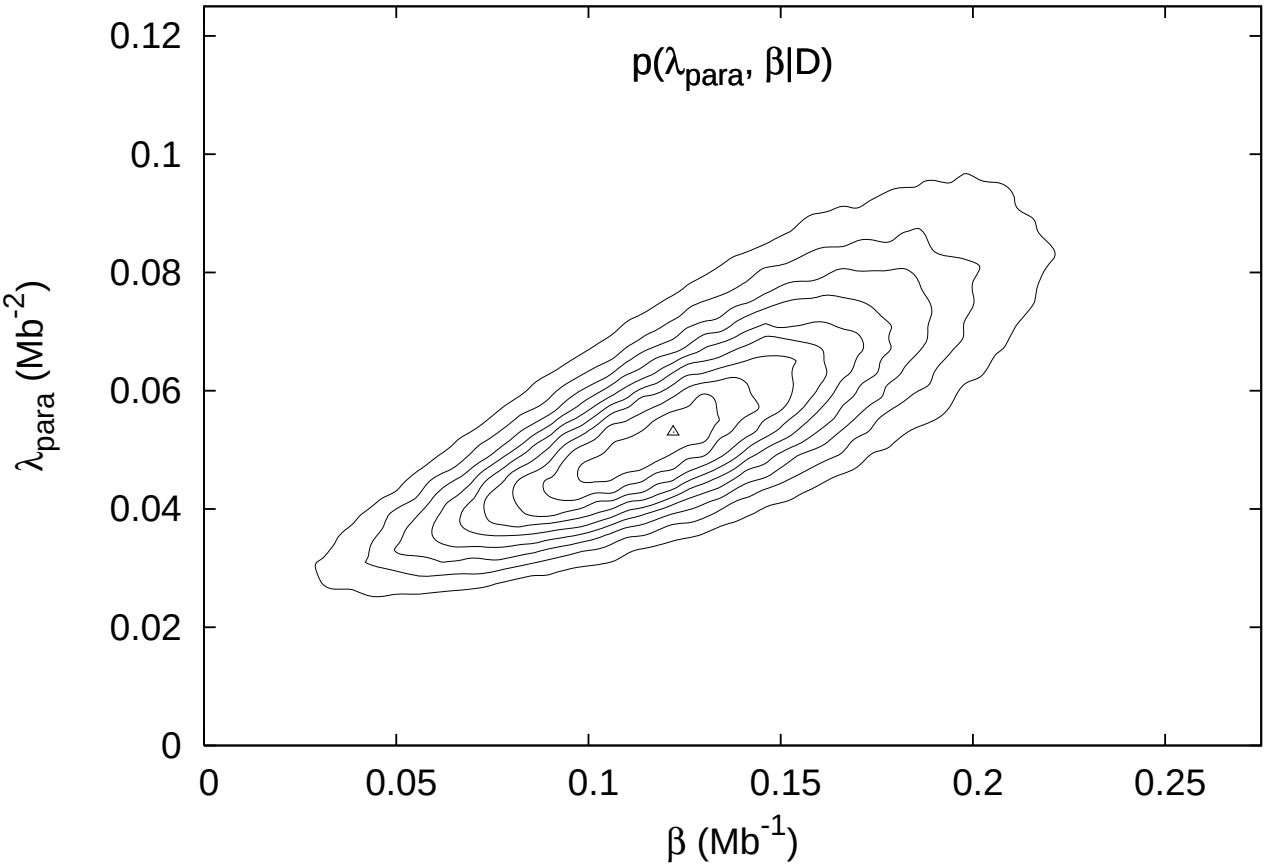

Figure 7
Posterior joint distribution of λ_{pa} and β. The triangle marks the mode.

tially with tract length, i.e., of the form $p(x_1, x_2) \propto e^{-\beta|x_2 - x_1|}$. With this distribution of breakpoints, the tract-length distribution is $p(\tau) \propto (\ell - \tau)e^{-\beta\tau}$, and the mean tract length is

$$\bar{\tau} = \frac{e^{-\beta\ell}(2 + \beta\ell) + \beta\ell - 2}{\beta(e^{-\beta\ell} + \beta\ell - 1)}.$$

Now, defining

$$A(\ell, \beta) = \int_0^\ell \int_0^{x_2} e^{-\beta|x_2 - x_1|} dx_1 dx_2 = \ell^2(e^{-\beta\ell} + \beta\ell - 1)/(\beta\ell)^2,$$

the total rate of inversions is $\lambda_I A(\ell, \beta)$ including inversions of segments containing no markers. Excluding these and summing over chromosomes:

$$\Lambda_I = \lambda_I \sum_{j=1}^{C} \left(A(\ell_j, \beta) - \sum_{i=0}^{m_j} A(s_{ij}, \beta) \right).$$

We will analyze unsigned data, i.e., we use the positions of the markers but not the orientations of individual markers. In this case inversions containing one marker are undetectable. However, since finding the shortest inversion path is hard for the unsigned case, we study the unsigned problem by working with signed arrangements of markers, and sampling from the set of all (signed) paths consistent with the unsigned data, as described in [6]. This means our paths may include one or more 1-marker inversions.

Paracentric and pericentric inversions

We want to allow pericentric and paracentric inversions to occur at different rates. Under the uniform independent breakpoint distribution assumption the mean tract length of pericentric inversions is $\bar{\ell}_{pe} = \ell/2$ independent of the centromere position. For paracentric inversions on a chromosome arm of length q the mean tract length is $q/3$ and the inversion rate is proportional to q^2. For a chromosome

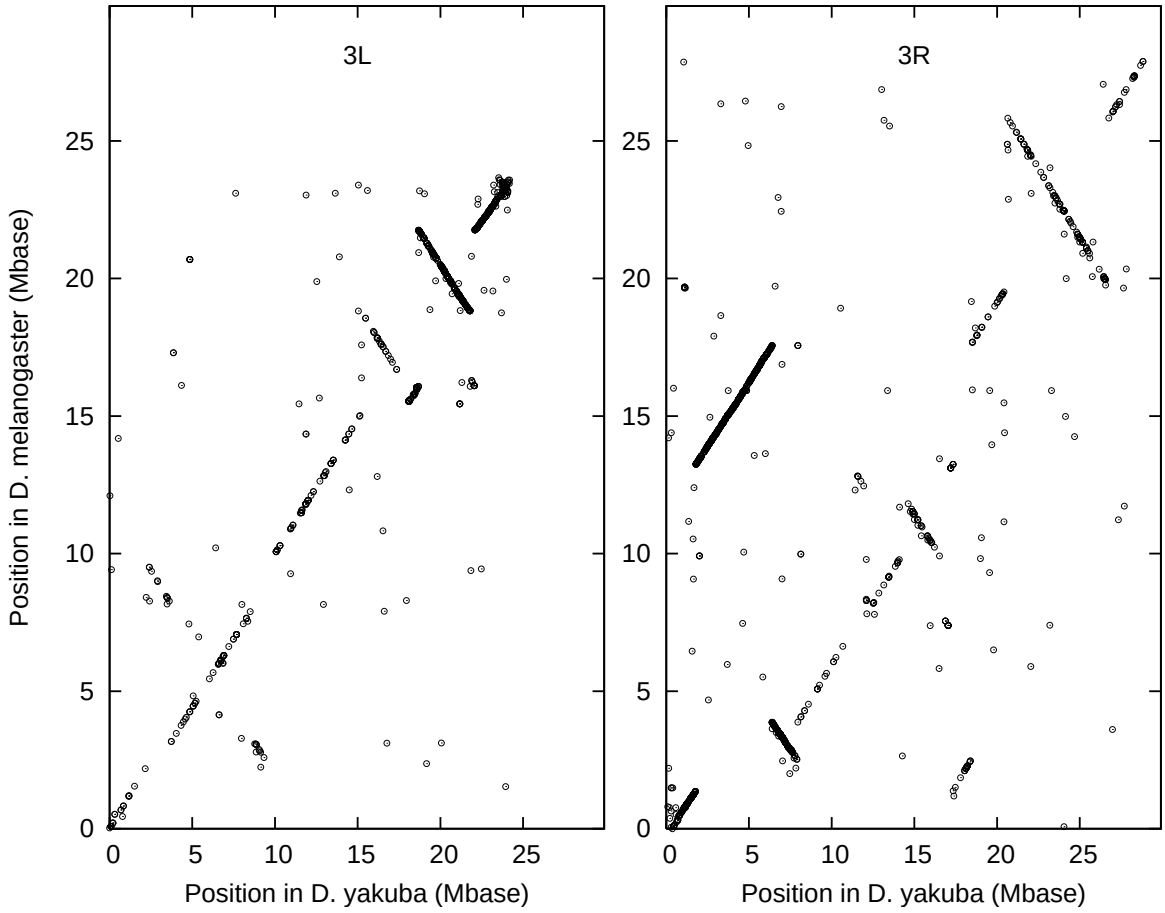

Figure 8
Position in *D. melanogaster* vs. position in *D. yakuba* for chains identified as on the same arm of chromosome 3 in both species.

with arm lengths $\xi\ell$ and $(1 - \xi)\ell$, this leads to a mean paracentric tract length of

$$\bar{\ell}_{pa} = \left(\frac{\xi^2 + (1-\xi)^3}{\xi^2 + (1-\xi)^2} \right) \ell / 3.$$

Depending on ξ, $\bar{\ell}_{pa}$ lies between $\ell/3$ and $\ell/6$, so for any centromere position $\bar{\ell}_{pa} < \bar{\ell}_{pe}$. This means that either a tract-length dependent effect ($\beta > 0$), or an effect which distinguishes only between paracentric and pericentric inversions, can have the effect of suppressing longer tract-length inversions. In order to know whether there is a tract-length effect independent of a possible paracentric/pericentric effect, we keep track of the two kinds of inversions separately, and assume a tract-length dependent paracentric rate

$$\Lambda_{pa} = \lambda_{pa} \sum_{j=1}^{2C} \left(A(\ell_j, \beta) - \sum_{i=0}^{m_j} A(s_{ij}, \beta) \right),$$

where the sums are now over chromosome arms and ℓ_j and m_j are the length and number of markers for the j^{th} arm. We assume a tract-length independent pericentric rate

$$\Lambda_{pe} = \lambda_{pe} \sum_{j=1}^{C} (\ell_{2j-1} \ell_{2j}),$$

where here ℓ_{2j-1} and ℓ_{2j} are the lengths of the two arms of chromosome j.

Now that we distinguish between paracentric and pericentric inversions and allow for a tract-length dependent rate, (5) becomes

$$p(X \mid \lambda_{pa}, \lambda_{pe}, \lambda_T, \beta) = \frac{e^{-\Lambda}}{L!} \lambda_{pa}^{L_{pa}} \lambda_{pe}^{L_{pe}} \lambda_T^{L_T} \prod_k \lambda_{d_k} e^{-\beta \sum_l \tau_l},$$

where now $\Lambda(\lambda_{pa}, \lambda_{pe}, \lambda_T) > \Lambda_{real} = \Lambda_{pa} + \Lambda_{pe} + \Lambda_T$, and $\lambda_d = \Lambda - \Lambda_{real}$ as before, and τ_l is the tract length of the l^{th} paracentric inversion.

Now we can write down the posterior probability:

$$p(X, \lambda_{pa}, \lambda_{pe}, \lambda_T, \beta \mid D) \propto P(D \mid X) \frac{e^{-\Lambda}}{L!} \lambda_{pa}^{L_{pa}} \lambda_{pe}^{L_{pe}} \lambda_T^{L_T} \times \prod_k \lambda_{d_k} e^{-\beta \sum_l \tau_l} p(\lambda_{pa}) p(\lambda_{pe}) p(\lambda_T) p(\beta).$$

The λ's priors are all uniform between 0 and λ_{max} and zero elsewhere. We assume $\beta \geq 0$ with a uniform prior. We assume that chromosomes always have exactly one centromere. In the computer code the breakpoint graph only considers marker-marker adjacencies, not marker-centromere adjacencies, and this means the way proposed rearrangement paths are constructed does not guarantee that centromeres end up in the right place. The centromeres are just passively carried along by the inversions and rearrangements dictated by the breakpoint graph. If the centromeres do not end up in the right place, the proposed path is rejected in the MCMC updating step, leading to loss of efficiency, but not loss of correctness. If the centromere lies within a region of conserved marker order its probability of ending up in the right place will typically be high, but if it lies between conserved regions this probability may be quite low, contributing to a low MCMC acceptance probability.

Data processing

We chose *D. yakuba* to compare with *D. melanogaster*. This choice was dictated by the need to have sufficiently many inversions that the biological problem is interesting, but not so many inversions that computational complexity becomes too large. We started with chained and netted alignments as described in [14]. We used the "net" file droYak2.dm2.net.gz, downloaded from the UC Santa Cruz website. This file contains information on chained alignments ('chains"), organized into hierarchies called "nets". These alignments are based on the Nov. 2005 WUSTL version 2.0 *D. yakuba* assembly and the Apr. 2004, BDGP v. 4/DHGP v. 3.2 *D. melanogaster* assembly.

Figure 8 shows the position in *D. melanogaster* of each chain, plotted versus it's position in *D. yakuba*. The figure shows the 982 chains located on chromosome arm 3*L* in both species and the 1,322 chains located on 3*R* in both species. Many points lie along lines with slopes close to ± 1, as expected for markers rearranged by inversions and translocations. There are, however, many other points scattered about, requiring further processing. First, chains not labeled in the net file as of type "syn" (i.e., syntenic) are eliminated. The chains left after some additional processing will be the markers used by the analysis program; from here on we refer to markers rather than chains.

Remaining markers are further processed by defining blocks within which adjacency is conserved. Two markers which are adjacent in both species are in the same block; if adjacent in just one species they are in different blocks. Blocks containing only a single marker are discarded, and blocks shorter than a minimum length are replaced by a single marker at the block's average position. This procedure is then repeated and the number of blocks may decrease, both directly because of discarding one-marker blocks, and also because when a block is discarded, or when a block is shortened to one marker, neighboring blocks will often join into one block. This procedure is repeated several times while the minimum block length is gradually increased from 100 bases to some final value L_{min}. Thus, a long block can emerge from a set of short blocks as some are eliminated and others join together. In some cases a block which ideally would be retained and incorporated into a long block may be lost during this process, if it is shorter than L_{min} and doesn't join another block soon enough. This can cause gaps in the spacing of markers on the resulting long block or the shortening of the block at an end. Neither of these is a big problem, although shortening at the ends of blocks means breakpoints are less well localized. The set of blocks generated is insensitive to L_{min} over a broad range: for our data, any value of L_{mim} between 25 kilobases and 115 kilobases gives the set of blocks that we analyzed.

Finally, markers are thinned from blocks containing many markers, until no block has more than 8 markers. Markers at the ends of blocks are kept, and the thinning of the others is done so as get a fairly even spacing. This reduces the time and memory requirements of the program, while having little effect on posterior distributions, according to our studies.

Applied to chromosomes *X*, 2, and 3, this procedure gives the 388 markers in 56 blocks shown in figures 1, 2, and 3.

Authors' contributions

RN had the idea of using MCMC methods to study chromosomal rearrangements and of using distance information to study tract lengths. RD contributed key mathematical techniques. TY wrote the MCMC code, analyzed the data, and wrote most of the manuscript. RN wrote parts of the manuscript and all authors participated in revising it. All authors have read and approved the final manuscript.

Acknowledgements

This work was supported by joint NSF/NIGMS grant DMS-02-01037.

References

1. Sturtevant AH, Dobzhansky T: **Inversions in the third chromosome of wild races of Drosophila pseudoobscura, and their use in the study of the history of the species.** *Proceedings of the National Academy of Sciences USA* 1936, **22:**448-450.
2. Watterson GA, Ewens WJ, Hall TE, Morgan A: **The chromosome inversion problem.** *Journal of Theoretical Biology* 1982, **99:**1-7.
3. Kececioglu J, Sankoff D: **Exact and approximation algorithms for the inversion distance between two permutations.** *Algorithmica* 1995, **13:**180-210.
4. Hannenhalli S, Pevzner PA: **Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals).** *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing* 1995:1-27.
5. Larget B, Simon DL, Kadane JB: **Bayesian phylogenetic inference from animal mitochondrial genome arrangements (with discussion).** *J R Stat Soc Ser B* 2002, **64:**681-693.
6. York TL, Durrett R, Nielsen R: **Bayesian Estimation of the Number of Inversions in the History of Two Chromosomes.** *Journal of Computational Biology* 2002, **9(6):**805-818.
7. Bafna V, Pevzner PA: **Genome rearrangements and sorting by reversals.** *SIAM Journal of Computing* 1996, **25:**272-289.
8. Durrett R, Nielsen R, York TL: **Bayesian Estimation of Genomic Distance.** *Genetics* 2004, **166:**621-629.
9. Miklos I: **MCMC genome rearrangement.** *Bioinformatics* 2003, **19(Suppl 2):**130-137.
10. Caceres M, Barbadilla A, Ruiz A: **Inversion Length and Breakpoint Distribution in the Drosophila buzzatii Species Complex: Is Inversion Length a Selected Trait?** *Evolution* 1997, **51(4):**1149-1155.
11. Brehm A, Krimbas C: **Inversion polymorphism in Drosophila obscura.** *Journal of Heredity* 1991, **82(2):**110-117.
12. Pinter R, Skiena S: **Sorting with length-weighted reversals.** *Proceedings of the 13th international conference on genome informatics* 2002:103-111.
13. Yogeeswaran K, Frary A, York TL, Amenta A, Lesser AH, Nasrallah JB, Tanksley SD, Nasrallah ME: **Comparative genome analyses of Arabidopsis spp.: inferring chromosomal rearrangement events in the evolutionary history of A. thaliana.** *Genome Res* 2005, **15(4):**505-515.
14. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes.** *Proc Natl Acad Sci USA* 2003, **100(20):**11484-11489.