Bayesian Estimation of Genomic Distance

Richard Durrett[*†], Rasmus Nielsen[†], and Thomas L. York[†]

[*]Mathematics and [†]Biological Statistics
and Computational Biology, Cornell U., Ithaca NY 14853

Abstract. We present a Bayesian approach to the problem of inferring the number inversions and translocations separating two species. The main reason for developing this method is that it will allow us to test hypotheses about the underlying mechanisms, such as the distribution of inversion track lengths or rate constancy among lineages. Here, we apply these methods to comparative maps of eggplant and tomato, human and cat, and human and cattle with 170, 269, and 422 markers respectively. In the first case the most likely number of events is larger than the parsimony. In the last two cases the parsimony solutions have very small probability.

1. Introduction

Understanding the relationship between the organization of two genomes is important for transferring information between species. For example, for finding animal models of human diseases or locating genes of agricultural importance (O'Brien et al, 1999). Inferences concerning genome evolution have primarily used parsimony methods. In the biology literature, experts have compared chromosome banding patterns to detail the evolution of primate genomes (Dutrillaux 1979) or have used comparative gene mapping and cross-species chromosome painting to analyze mammalian genome evolution (Chowdary et al 1998, Haig 1999, Murphy, Stanyon, and O'Brien 2001). In a separate and independent effort computer scientists have developed algorithms for computing the minimum number of events needed to transform one genome into another (Hannenhalli and Pevzner 1995a, 1995b).

In an earlier paper (York, Durrett, and Nielsen 2002) we developed a Bayesian method to infer the history of inversions separating homologous chromosomes from two different species. This method can be applied to mitochondrial genomes, X chromosomes, and to chromosome arms of species of genus *Drosophila* (Ranz, Casals, and Ruiz 2001) and *Anopheles* (Sharakhov et al. 2002), where pericentric inversions and translocations are rare. In this paper we will extend those methods to the problem of genomic distance, i.e., the number of inversions, and translocations, needed to transform one genome into another. Here, fissions and fusions are included as a special case of translocations in which one of the input or output chromosomes is empty.

There are several reasons for preferring our statistical approach to parsimony methods. The first and simplest is that there is no guarantee that nature took the shortest path. Indeed, we will see in two of our examples that the parsimony solution is not in the 95% credible interval. A second important reason is that our approach allows for estimation of rates of evolution and testing of hypotheses: Are all inversions equally likely? Are rates the same in different lineages? Answering the first question is important for a proper analysis of the second. Otherwise we will not know if the perceived higher rate of rearrangements in rodents compared to carnivores is due to higher density of markers in the first case.

In this paper we will concentrate on estimating the number of events needed to change one genome into the other leaving hypothesis testing issues for later work. In Section 2 we will describe existing parsimony methods. Section 3 explains our new approach. In Section 4 we analyze three data sets.

2. Parsimony Methods

Hannenhalli and Pevzner (1995a, 1995b) developed a polynomial algorithm for computing the distance between two genomes, i.e., what is the smallest number of inversions and translocations needed to transform one genome into another? They accomplished this in a two steps. In the first paper they solved the problem of computing the reversal distance, i.e., the smallest number of inversions needed to transform one chromosome or mitochondrial genome into another. In the second paper they showed how to reduce the genome distance problem to the one for reversal distance by concatenating chromosomes into one unit. That approach is convenient for proving the

existence of a polynomial algorithm, but not for our implementation so we will take a different approach.

To illustrate the method we will consider part of the data of Doganlar et al. (2002) who constructed a comparative genetic linkage map of eggplant (Solanum melongena) based on 233 markers for tomato. Using the first letter of the common name to denote the species they found that the marker order on T1 and E1 and on T8 and E8 were identical, while in four other cases (T2 vs. E2, T6 vs. E6, T7 vs. E7, T9 vs. E9) the collections of markers were the same and the order became the same after a small number of inversions was performed (3, 1, 2, and 1 respectively).

In our example we will compare of the remaining six chromosomes from the two species. The first step is to divide the chromosomes into *conserved segments* where the adjacency of markers has been preserved between the two species, allowing for the possibility of the overall order being reversed. When such segments have two or more markers we can determine the relative orientation. However as the HP algorithm assumes one knows the relative orientation of segments we will have to assign orientations to conserved segments consisting of single markers in order to minimize the distance. In the case of the tomato-eggplant comparison there are only 5 singleton segments so one can easily consider all $2^5 = 32$ possibilities. In the human-cat and human-cattle comparisons in Section 4 there are 21 and 75 single marker segments respectively, so the amount of work is considerable ($2^{21} = 2{,}097{,}152$) or impossible ($2^{75} = 3.77 \times 10^{22}$).

The first part of Figure 1 shows the two genomes with an assignment of signs to the single marker segments that minimizes the distance. The first step in preparing to use the HP algorithm is to double the markers. When segment i is doubled we replace it by two consecutive numbers 2i-1 and 2i, e.g., 6 becomes 11 and 12. A reversed segment -i is replaced by 2i and 2i-1, e.g., -5 is replaced by 10 and 9. The second step is to add ends to the chromosomes and enough empty chromosomes to make the number of chromosomes equal. In this example, no empty chromosomes are needed. We have labeled the ends in the first genome by 1000 to 1011 and in the second genome by 2000 to 2011.

The second part of Figure 1 shows the result of the first two preparatory steps. Commas indicate separations between two segments or between a segment and an end. The next step is to construct the breakpoint graph which results when the commas are replaced by edges that connect vertices with the corresponding numbers. We did not draw the graph since we only need to know the connected components of the graph. Since each vertex has degree two, these are easy to find: start with a vertex and follow the connections. The resulting component will either be an path that connects two ends or a cycle that consists of markers and no ends. In our example there are five paths of length three that connect ends. These "short paths" tell us that end 1000 in genome 1 corresponds to end 2000 in genome 2, etc. The other correspondences between ends will be determined after we compute the distance.

The other "long paths" in the breakpoint graph are listed in Figure 1. At this point there are no cycles in the breakpoint graph. To compute a lower bound for the distance now we start with the number of commas seen when we write out one genome. In this example that is 33. We subtract the number of connected components in the breakpoint graph. In this example that is 5 + 7 = 12, and then add the number of paths that begin and end in the same genome, which in this case is 0. The result which is 21 in this case is a

lower bound on the distance since any inversion or translocation can at most reduce this quantity by 1, and it is 0 when the two genomes are the same.

In general the distance between genomes can be larger than the lower bound from the breakpoint graph and the computation of the exact distance can be quite complex. In the reversal distance problem there can be obstructions called *hurdles* that can prevent us from decreasing the distance and hurdles can be intertwined in a *fortress of hurdles* that takes an extra move to break. In the genome distance problem the situation is even more complex and one must consider six different quantities to compute the distance. Fortunately these complexities rarely arise in biological data sets. Bafna and Pevzner (1995) considered the reversal distance problem for 11 chloroplast and mitochondrial data sets and in all cases they found that the distance is equal to the lower bound. That is also the case for the three examples we will consider in this paper.

To verify that 21 is the minimum distance we construct a sequence of 21 moves that transforms genome1 into genome 2. To explain the procedure we call the edges that result from commas in genome 1 black edges, and those from genome 2 gray edges. On each step we will choose an inversion or translocation involving two gray edges to increase the number of connected components by 1. It is easy to check that this can only be done by picking two edges in the same cycle.

Some of the cycles are easy to deal with. Using dashes to indicate gray edges, flipping segment 17 turns 1007 36-32 33-35 34-2009 into 1007 36-32 33-2009 and 35=34, where = indicates there is a black and a gray edge connecting 34 and 35 now. Flipping -18 -17 now we have 1007 36-2009 and 32=33. Of course we could have also flipped -18 17 to -17 18 and then flipped -17. Similarly, 1009 44-42 43-40 41-2002 is eliminated by changing 21 -22 -20 to -22 -21 -20 with 2 inversions and 1011 54-49 48-52 53-51 50-2011 is eliminated by changing 24 -26 27 25 to 24 25 26 27 with 3 inversions.

To follow the solution on page 1707 of Dongalar et al. (2002) we will also do two inversions to turn 1 -5 2 to 1 2 5, flip the order of 14 11 -15 3, flip -19, and perform a reciprocal translocation of T4 and T10. After these 11 events the situation is given by the arrangement called Step 1at the top of Figure 2. There are now 23 components in the breakpoint graph, so the distance has been reduced to 10. At this point Dongalar et al (2002) invoke a nonreciprocal translocation to put 3 4 in the right place. In our scheme this requires two translocations. Written in terms of the cycles the moves are 29-5 with 4-9 and then 29-9 with 8-2006. In terms of the chromosomes we first make 1 2 3 4 and -6 -5 15 -11 -14 9 and then 1 2 3 4 5 6 and 15 -11 -14 9.

Performing translocations 27-19 with 26-29 and 31-39 with 38-47 then performing two inversions to turn -12 23 to 23 12 leads to the arrangement called Step 2. After these six changes the distance is 4. To finish up we flip -9 then perform the translocation 46-23 with 31-47 to make 23 24 25 26 27 and -18 -17 -16 12 13 14 11 -15. Flipping 11 -15 to 15 -11 and then 12 13 14 15 we have E10 written backwards and we are done.

The methods of Hannenhalli and Pevzner just described give an estimate of the number of events that can be seen with the marker set used, e.g., an inversion that occurs between two markers or reverses the orientation of a single marker cannot be detected. The method of Nadeau and Taylor (1984), which can be used when we know the distances between markers in one of the genomes, avoids this problem. Their computation is based on the length of conserved segments with more than two markers,

enlarges the lengths based on the expected value of the unseen flanking ends and then uses the transformed data to get an estimate of the average size of conserved segments. Using the total size of the genome, this translates into an estimate of the number of conserved segments, S.

Subtracting the number of chromosomes in the genome for which the distances are known from S gives an estimate of the number of disruptions, R, that have occurred. Given R, we divide by 2 to get an estimate of the number of events that have occurred. In the literature this estimate is called the *breakpoint distance*. It is reasonable to use in Nadeau and Taylor's context, since they are considering the situation in which complete information about marker order has revealed all of the disruptions, and if we assume that all breakpoints are equally likely then with high probability no two disruptions will occur between the same pair of nucleotides. On the other hand with a small number of markers, e.g., in the tomato-eggplant comparison considered above the breakpoint distance is not very accurate. For the six chromosomes considered, we have 27-6 = 21 disruptions but the genomic distance is 20.

3. A Bayesian Approach

The aim of the Bayesian approach is to generate the probability distribution of inversions and translocations given the data. We will assume that inversions and translocations arise in the genome at a constant rate and that the process of inversions and translocations forms a Markov chain with state space (Ω_O) given by the set of all possible orderings of the markers on ordered chromosomes. Letting all markers be signed indicating their orientation on the chromosome, there are

$$\left|\Omega_O\right| = 2^N \frac{(M+N-1)!}{(M-1)!}$$

orderings of N signed markers onto M numbered chromosomes, allowing empty chromosomes, i.e. chromosomes containing no markers. In our representation, transitions between neighboring states (states differing by only a single inversion or a single translocation) in Ω_O occur at rate λ_I if the two states differ by an inversion and λ_T if they differ by a translocation. Because of the symmetry of the transition rates, the Markov chain is time reversible and has a uniform stationary distribution, i.e.

$$\pi_O(x) = 2^{-N} \frac{(M-1)!}{(M+N-1)!}, \forall x \in \Omega_O .$$

Because the ordering of chromosomes is not of interest, we also consider a Markov chain with a collapsed state space (Ω_U) on equivalence classes with ordered signed markers arranged on unordered chromosomes. An element of Ω_U, with M_0 empty chromosomes, is an equivalence class of $2^{(M-M_0)} M!/M_0!$ elements of Ω_O. The stationary probability assigned to an element of Ω_U with M_0 empty chromosomes is then

$$\pi_U(x) = 2^{(M-N-M_0)} \frac{M!(M-1)!}{M_0!(M+N-1)!}, \forall x \in \Omega_U$$

,

Transitions between neighboring states in this collapsed Markov chain occur at rates λ_I if the two states differ by an inversion and λ_T if they differ by a translocation, except for translocations involving a fission of two chromosomes in an equivalence class with M_0 empty chromosomes. These transitions occur at rate $2M_0\lambda_T$. Therefore, the transition probabilities of the Markov chain obey the detailed balance equations and the Markov chain is time reversible. In the following we will use the representation based on the Markov chain with state space on Ω_U.

Consider the genome of two organisms. The genomic marker data from one species (x_1) can be transformed into the genomic marker data from another species (x_2) through a sequence of inversions and translocations. Because the process we have defined is time reversible we can write the sampling probability as

$$\Pr(x_1, x_2 \mid \Theta) = \Pr(x_1) \Pr(x_1 \rightarrow x_2 \mid \Theta)$$

where $\Pr(x_1 \rightarrow x_2 \mid \Theta)$ is the transition probability for the transformation from x_1 to x_2, and Θ is the vector of parameters. Because a model parameterized in terms of t, λ_T and λ_I is not identifiable, we arbitrarily set $t = 1$ and define $\Theta = (\lambda_T, \lambda_I)$. Therefore, λ_T and λ_I can also be interpreted as the expected number of translocations and inversions per marker pair in the history of the two species. $\Pr(x_1)$ does not depend on Θ, and the likelihood function is, therefore, simply given by

$$L(\Theta) = \Pr(x_1 \rightarrow x_2 \mid \Theta).$$

Let Ψ be the (countably infinite) set of all possible evolutionary paths from x_1 to x_2. We notice

$$\Pr(x_1 \rightarrow x_2 \mid \Theta) = \sum_{\mathbf{y} \in \Psi} \Pr(\mathbf{y} \mid \Theta).$$

To estimate Θ we establish a Markov chain with state space on $[0,\infty)^2 \times \Psi$ and with stationary distribution given by the joint posterior distribution of parameters and evolutionary path

$$\pi(\mathbf{y}, \lambda_T, \lambda_I) = p(\mathbf{y}, \lambda_T, \lambda_I \mid x_1, x_2), \forall \mathbf{y} \in \Psi, \lambda_T \in [0,\infty), \lambda_I \in [0,\infty).$$

To ensure the posterior is proper and biologically meaningful, the support of λ_T and λ_I are restricted to the intervals $(0, \lambda_{Tmax})$ and $(0, \lambda_{Imax})$, respectively. The Markov chain is simulated using the Metropolis-Hastings algorithm similarly to the method used in York *et al.* (2003). In brief, the posterior distribution is proportional to the product of the likelihood function and the prior distribution

$$p(\mathbf{y}, \lambda_T, \lambda_I \mid x_1, x_2) \propto \Pr(\mathbf{y} \mid \lambda_T, \lambda_I) p(\lambda_T) p(\lambda_I). \quad (1)$$

The Markov chain is then simulated by iteratively proposing new values of **y**, λ_T and λ_I, and accepting these new values according to the appropriate Metropolis-Hastings acceptance probabilities (see York *et al.* 2003). These acceptance probabilities are evaluated using the terms of the right hand side of equation (1). In this expression, the densities for the λ_T and λ_I can trivially be evaluated at the relevant values and the probability of an evolutionary path, $\Pr(\mathbf{y} \mid \lambda_T, \lambda_I)$ can also be calculated. If at time t there are $I_T(t)$ possible translocations that can occur and $I_I(t)$ possible inversions that can occur, the total rate of translocations and inversions are $I_T(t)\lambda_T$ and $I_I(t)\lambda_I$, respectively. Notice that because translocations may change the possible number of inversions and translocations, the rate at which chromosomal rearrangements occur may not be constant in time. However, the chance that any particular inversion or translocation occurs as the first event after time t, is $\lambda_I/(I_T(t)\lambda_T + I_I(t)\lambda_I)$ for inversions and $\lambda_T/(I_T(t)\lambda_T + I_I(t)\lambda_I)$ for translocations and the waiting time to the first event is exponentially distributed with parameter $I_T(t)\lambda_T + I_I(t)\lambda_I$. So the probability assigned to a particular evolutionary path, with S events (inversion and translocations), is

$$\Pr(\mathbf{y} \mid \lambda_T, \lambda_I) = \prod_{i=1}^{S+1} \lambda_E(i) e^{-(I_T(t_i)\lambda_T - I_I(t_i)\lambda_I)t_i} ,$$

where t_i is the time between events $i-1$ and i for $i \le S$, $t_{S+1} = 1 - \sum_{j=1}^{S} t_j$ and

$$\lambda_E(i) = \begin{cases} \lambda_T \text{ if } i \le S \text{ and } i\text{th event is a translocation} \\ \lambda_I \text{ if } i \le S \text{ and } i\text{th event is an inversion} \\ 1 \text{ if } i = S+1 \end{cases} .$$

In praxis, to avoid keeping track of the times between events, we use a method based on uniformization to ensure that the total rate of change is constant in time. In brief, we allow pseudoevents of evolutionary change, which have no effect on marker order, to occur at rate $\max_{(I_T, I_I)}\{I_T\lambda_T + I_I\lambda_I\} - (I_T(t)\lambda_T + I_I(t)\lambda_I)$ at time t. The total rate of evolutionary change is then kept constant at rate $\max_{(I_T, I_I)}\{I_T\lambda_T + I_I\lambda_I\}$. As in York *et al.* (2003), an efficient proposal kernel for **y** is constructed by considering the breakpoint graph of Hannenhali and Pevzner (1995a).

The prior densities for λ_T and λ_I are assumed to be independent uniform on (0, λ_{Tmax}) and (0, λ_{Imax}), respectively. Updates to these parameters are then proposed by simulating a new value of the parameter uniformly in a window around the current value, independently of each other and of updates of **y**. The marginal posterior distributions of the parameters are then estimated by sampling values of λ_T and λ_I from the simulated Markov chain at stationarity. The distribution of values of λ_T and λ_I is proportional to the joint likelihood function for λ_T and λ_I, and the results can be interpreted both in a Bayesian and a likelihood framework. Here we will focus on the Bayesian interpretation since the usual asymptotic properties of the likelihood function may not be satisfied.

The method can also be used to make inferences on **y**. The posterior distribution of the number of inversions (L_I) and the number of translocations (L_T) that have occurred in the history of the two species can be estimated by sampling from the simulated Markov chain. Likewise, the posterior distribution of inversion tract lengths can be estimated.

To improve convergence we used the technique of Metropolis-coupled Markov chain Monte Carlo, in which each chain with stationary distribution π is coupled to a set of "heated" chains with modified stationary distributions, $\pi_i = \pi^{1/T_i}$, where the "temperatures", T_i, are slightly greater than one, resulting in flattened distributions and, consequently, better mixing. Only the unheated chain is used in estimating posterior distributions.

Our estimates of posterior distributions use every 8th update, after excluding some number of updates at the beginning of the chain as a burn-in phase. To determine an appropriate length of the burn-in phase, we run multiple chains and compare the between chain variance of the evolutionary path length L (for example), B_L, to the within chain variance, W_L, requiring that B_L/W_L become small. This is essentially the method of Gelman and Rubin (1992). We similarly consider B/W for L_I, L_T, λ_I, and λ_T, and define the burn-in phase to end when all five of these ratios are < 0.1.

4. Analysis of Three Data Sets

Tomato vs. Eggplant. Doganlar et al. (2002) constructed a comparative map of tomato and eggplant consisting of 233 markers. Thinning their data set to choose one marker from each group in which the order was not resolved, with leads to a data set with 170 markers. Based on this map and comparisons with maps for potato (Tanksley et al. 1992) and pepper (Livingstone et al. 1999) they concluded: "Overall, eggplant and tomato were differentiated by 28 rearrangements, which could be explained by 23 paracentric inversions and five translocations during evolution from the species' last common ancestor."

Doganlar's solution postulates a nonreciprocal translocation that inserts a small piece of the short arm of T5 into the middle of the short arm of T3, which would take two moves in our analysis. On the other hand as demonstrated in Section 2, the parsimony solution for this comparison is 21 events for the six chromosomes and 7 inversions for the other six or also a total of 28 using only ordinary translocations. An advantage of the HP method is that it provides a systematic method for finding shortest paths and in addition establishes that there are no shorter paths.

Our Bayesian analysis produced 95% credible intervals of [5,7], [21,31], and [28,37] for the number of inversions, translocations, and total number of events separating tomato and eggplant. These results are from 6 unheated chains, each coupled to 4 heated chains. Each of these 30 chains was updated 459,000 times, using 75,000 seconds of CPU time on a 1.5GHz processor. The first 14,000 updates are excluded as burn-in. The posterior distribution for the number of translocations assigns probability 0.08172 to 5, 0.55407 to 6, 0.32137 to 7, 0.03832 to 8, and 0.00453 to 9 or more. Figure 3 gives the posterior distribution of the number of inversions, which has a mode of 25.

Thus even in the case of these two closely related species, the mostly likely number of inversions and translocations are somewhat higher than the parsimony estimates.

Figure 4 gives the posterior joint distribution of the rates of inversions λ_t and inversions λ_i. The mode of the distribution occurs at $\lambda_t = 0.000219$ and $\lambda_i = 0.0194$. To interpret these numbers we note that in the two genomes being compared there is an average of 30,271 possible translocations and 1,335 possible inversions. Multiplying we see that the total rate is 6.629 for translocations and is 25.899 for inversions. This rate assumes that the total evolutionary time between the two genomes has been scaled to be 1. Taking 12 million years as an estimate of the divergence between tomato and eggplant we arrive at rates of 0.276 and 1.078 per genome per million years for translocations and inversions.

Human vs. Cat. Murphy et al. (2000) created a radiation hybrid map of the cat genome integrating 424 genes with 176 microsatellite markers. Using resources on the NCBI home page, http://www.ncbi.nlm.nih.gov, we were able to locate the position of 281 of the genes in the human genome. From this set we deleted 12 singleton disruptions, i.e., isolated genes that map to a chromosome different from their neighbors. We do this because it is our belief that these genes are the result of duplications of small segments which recent studies (Bailey et al. 2002a) have shown to occur frequently in the human genome. In support of this practice we note that none of the regions deleted are observed in the chromosome painting experiments of Weinberg et al (1997) and those techniques are thought to be capable of detecting segments as small as 5 megabases.
Parsimony analysis shows that the marker order in the human genome can be transformed into the cat genome in 78 moves (14 translocations and 64 inversions). Our Bayesian analysis gives 95% credible intervals of [12,15], [71,89], and [85,102] for the number translocation, inversions, and total number of events respectively. These results are from 6 unheated chains, each coupled to 3 heated chains. Each of these 24 chains was updated 2.2 million times, using 790,000 seconds of CPU time. The first 306,000 updates are excluded as burn-in.

Note that the parsimony distance is not in the 95% credible interval for the total number of events. In fact, the posterior probability of a total of 78 events is approximately 2×10^{-5}. The posterior distribution for the number of translocations assigns probability 0.5967 to 12, 0.3451 to 13, 0.0531 to 14, and 0.0050 to 15 or more, so this time the smallest value is the most likely. The posterior distribution of the number of inversions is given in Figure 5, with the mode being 79.

The extensive conservation of gene order between human and cat is a widely cited result, see e.g., O'Brien et al (1999). However, in some cases this fact is exaggerated based on a misunderstanding: Bailey et al. (2002b) say "Even human comparisons to distantly related mammals demonstrate strong conservation with the estimated number of rearrangements varying from 17 in felines to 180 in mice." Unfortunately, here they are comparing the number of inversions and translocations in felines based on FISH data which cannot detect many inversions with the number of inversions and translocations in mice based on a fairly dense comparative map.

Murphy et al. (1999a) mapped 25 markers on the feline X chromosome and found that the marker order in humans is identical. In the radiation hybrid map of Murphy et al. (2000), complete conservation of marker order holds for the eight markers on HSA 9 and

FCA D4 but most other chromosomes show a number of inversions. Our estimate for human-cat comparison is about ½ of the one for human-cattle given below. It would be interesting to use three species comparisons to partition the events between lineages but here and in most other cases there are not enough shared markers to make inferences. Bourque and Pevzner (2002) compared human, cat and rat. They found 193 markers shared by the three species but when they discarded the ones for which they could not determine the orientation only 114 were left.

It is interesting to compare the number of translocations estimated with the result of chromosome painting experiments of Weinberg et al (1997). The *syntenic segments* (i.e., contiguous sets of markers that all map to the same chromosome) from chromosome painting correspond to those in the comparative map with one notable exception: HSA 12 has a segment homologous to part of FCA D3 which was not revealed by the original chromosome painting but was identified by the follow-up study of Murphy et al. (1999b). If we add this segment to the chromosome painting then one needs 12 translocations and 2 inversions to transform one genome into the other.

Finally we would like to compare our estimates with those of Nadeau and Taylor. There are 83 conserved segments with 2 or more markers with an average length of 16.3 megabases, which results in an estimated average length of 18.2 megabases. Using 3.2 gigabases for the length of the human genome yields an estimate of 175.8 segments. Subtracting 22 chromosomes in the human genome, we arrive at an estimate of 155.8 disruptions or an estimate of 77.4 events. The reader should note that not only is this lower than the parsimony and Bayesian estimates but also that the method of Nadeau and Taylor includes events that are not visible with our set of markers.

Figure 6 gives the posterior distribution for posterior joint distribution of the rates of inversions λ_t and inversions λ_i . The mode of the distribution occurs at $\lambda_t = 0.000161$ and $\lambda_i = 0.0350$. To interpret these numbers we note that in the two genomes being compared there are an average of 79,650 possible translocations and 2,370 possible inversions. Multiplying we see that the total rate is 12.82 for translocations and is 82.95 for inversions. Taking 100 million years as an estimate for the divergence between humans and cats we arrive at rate estimates of 0.0641 and 0.415 per genome per million years for translocations and inversions.

Human vs. cattle. Band et al. (2000) constructed a radiation hybrid map of cattle (*Bos taurus*) with a total of 638 genes. For the data see http://bos.cvm.tamu.edu/htmls. Using resources on the NCBI home page, we were able to locate the position of 448 of the genes in the human genome. Deleting 24 singleton disruptions for the reasons indicated above results in a map with 422 markers. Again the reduced map is consistent with the results of chromosome painting (see Hayes 1995 and Chowdhary et al 1996). Parsimony analysis shows that the marker order in the human genome can be transformed into the cattle genome in 155 moves (20 translocations and 135 inversions). Our Bayesian approach experienced convergence problems in this example. Four unheated chains were each coupled to 8 heated chains. Two of these sets of chains ran for 1.3 million updates, the other two sets for 1.5 million updates, using a total of 2.6 million CPU seconds. Our usual criterion for the end of burn-in was not attained; a burn-in of 600,000 updates is used for the plots shown.

Figures 7 and 8 give the four posterior distributions for the number of translocations and inversions. In figure 7 there is a considerable difference between the chains indicating convergence problems. The qualitative differences between chains in the number of inversions are not as great as in the case of translocations. The modes are all in the range 185-191 but the variance differs considerably from run to run. We cannot make statements with much confidence about the number of inversions and translocations. However two things are clear: (a) the number of events is roughly twice that in the human cat comparison even though the divergence times are similar and (b) our conclusions differ considerably from those of Band et al (2000) say that their "comparative map suggests that 41 translocation events and a minimum of 54 internal rearrangements have occurred." They do not explain how they reached this conclusion. However, we would require a larger number of translocations if we had not deleted the singletons and they would underestimate the number of inversions if they used the breakpoint distance.

To estimate distances using the methods of Nadeau and Taylor we note that there are 125 conserved segments with 2 or more markers with an average length of 7.19 megabases, which results in an estimated average length of 7.57 megabases. Using 3.2 gigabases for the length of the human genome yields an estimate of 422.7 segments. Subtracting 22 chromosomes in the human genome gives an estimate of 400.7 disruptions or an estimate of 200 events. This is somewhat larger than our Bayesian estimate, but that is consistent with the fact that our estimate is restricted to the events that can be detected by the 422 markers in our map.

5. Conclusions

Elucidating the evolutionary history of genomes is one of the major goals of comparative genomics. We have shown that a full probabilistic approach to the problem is feasible, but the method experiences difficulty with data sets that have experienced a large number of events (inversions and translocations). For the human-cat and human-cattle comparisons, the probability that the true number of events is close to the minimum number is very small, so in these cases the parsimony estimate is not reliable. The breakpoint distance which is the number of disruptions/2 is even worse.

Our analysis of the three data sets has revealed a consistent pattern in which inversions are 4-7 times more frequent than translocations. Our ratio is larger than the conclusion of Murphy et al (2000) that "the ordered gene maps of the cat and cow genomes show approximately twice as many intrachromosomal (inversions) as interchromosomal changes." There are three reasons for this discrepancy: (i) biologists use the breakpoint distance, which for these maps underestimates the minimum number of inversions, (ii) our Bayesian estimate is typically considerably larger than the minimum distance, and (iii) as genetic maps become more dense more inversions can be seen.

Our methods can be extended to a model in which inversion probabilities depend upon track length. This generalization is important if we are going to accurately extrapolate from the number of inversions seen by a collection of markers to the number that have actually occurred in the history of a genome. A second extension we would like

to pursue is to incorporate gene duplications in order to treat the entire data set and obtain rate estimates for small segmental duplications.

REFERENCES

Bafna, V. and Pevzner, P. (1995) Sorting by reversals: Genome rearrangement in plant organelles and evolutionary history of X chromosome. *Mol. Biol. Evol*. 12, 239-246

Bailey, J.A. et al. (2002a) Recent segmental duplications in the human genome. *Science* 297, 1003-1007

Bailey, J.A. et al. (2002b) Human-specific duplication and mosaic transcripts: The recent paralogous structure of chromosome 22. *Am. J. Hum. Genet*. 70, 83-100

Bourque, G., and Pevzner, P.A. (2002) Genome-scale evolution: reconstructing gene orders in ancestral species. *Genome Research*. 12, 26-36

Chowdhary, B.P., Fronicke, L., Gustavsson, I., Scherthan, H. Comparative analysis of cattle and human genomes: detection of ZOO-FISH and gene mapping-based chromosomal homologies. *Mammalian Genome.* 7 (1996), 297-302

Doganlar, S., Frary, A., Daunay, M.C., Lester, R.N., and Tanksley, S.D. (2002) A comparative genetic linkage map of eggplant (*Solanum melongea*) and its implications for genome evolution in the Solanaceae. *Genetics.* 161, 1697-1711

Dutrillauz, B. (1979) Chromosomal evolution in primates: Tentative phylogeny from *Microcebus murinus* (prosimian) to man. *Hum. Genet*. 48, 251-314

Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistic. Sci.* **7**, 457-511

Haig, D. (1999) A brief history of human autosomes. *Phil. Trans. Roy. Soc. London Series B*. 354, 1447-1470

Hannenhalli, S., and Pevzner, P.A. (1995a) Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). Pages 178--189 in *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing*. Full version in the *Journal of the ACM*. 46, 1-27

Hannenhalli, S., and Pevzner, P. (1995b) Transforming men into mice (polynomial algorithm for the genomic distance problem. Pages 581—592 in *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, New York

Hayes, H. Chromosome painting with human chromosome-specific DNA libraries reveals the extent and distribution of conserved segments in bovine chromosomes. *Cytogenet. Cell. Genetics*. 71 (1995), 168-174

Livingstone, K.D., Lackney, V.K., Blauth, J.R., van Wijk, R., and Jahn, M.K. (1999) Genome mapping in Capsicum and the evolution of genome structure in the Solanaceae. *Genetics* 152, 1183-1202

Murphy, W.J., Shan, S., Chen, Z.Q., Pecon-Slattery, J., and O'Brien, S.J. (1999a) Extensive conservation of sex chromosome organization as revealed by parallel radiation hybrid mapping. *Genome Research*. 9, 122-1230

Murphy, W.J., Menotti-Raymond, M., Lyons, L.A., Thompson, M.A., and O'Brien, S.J. (1999b) Development of a feline whole genome radiation hybrid panel and comparative mapping of human chromosome 12 and 22 loci. *Genomics*. 57, 1-8

Murphy, W.J., Shan, S., Chen, Z.Q., Pecon-Slattery, J., Yukhi, N., Hirschmann, D., Menotti-Raymond, M., and O'Brien, S.J. (2000) A radiation hybrid map of the cat genome implications for comparative mapping. *Genome Research*.10, 691-702

Murphy WJ, Stanyon, R, and O'Brien, SJ. Evolution of mammalian genome organization inferred from comparative mapping. Genome Biology 2 (2001), 1-8

Nadeau, J.H., and Taylor, B.A. (1994) Lengths of chromosomal segments conserved since the divergence of man and mouse. *Proc. Natl. Acad. Sci*. 81, 814-818

O'Brien, SJ, Menotti-Raymond, M, Murphy, WJ, Nash, WG, Wienberg, J, Stanyon, R, Copeland, NG, Jenkins, NA, Womack, JE, Graves, JAM. (1999) The promise of comparative genomics in mammals. *Science.* 286 458-465

Ranz, J.M., Casals, F., and Ruiz, A. (2001) How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the Genus *Drosophila. Genome Research*. 11, 230-239

Sharakhov, I.V., and 11 coauthors. (2002) Inversions and gene order shuffling in *Anopheles gambiae* and *A. funestrus*. *Science.* 298, 182-185

Tanksley, S.D., and 18 coauthors. (1992) High density molecular linakge maps of the tomato and potato genomes. *Genetics* 132, 1141-1160

Weinberg, J., Stanyon, R., Nash, W.G., O'Brien, P.C.M., Yang, F., O'Brien, S.J., and Ferguson-Smith, M.A. (1997) Conservation of human vs. feline genome organization revealed by reciprocal chromosome painting. *Cytogent. Cell. Genet*. 72, 211-217

York, T.L., Durrett, R., and Nielsen, R. (2002) Bayesian estimation of inversions in the history of two chromosomes. *J. Comp. Bio*.

Eggplant
E3. 1 2 3 4 5 6
E4. 7 8
E5. 9 10
E10. 11 12 13 14 15 16 17 18
E11. 19 20 21 22
E12. 23 24 25 26 27

Tomato
T3. 1 -5 2 6
T4. 21 -22 -20 8
T5. -4 14 11 -15 3 9
T10. 7 16 -18 17
T11. -19 24 -26 27 25
T12. -12 23 13 10

Eggplant Doubled, Ends Added
1000, 1 2 , 3 4 , 5 6 , 7 8 , 9 10 , 11 12 , 1001
1002, 13 14 , 15 16 , 1003
1004, 17 18 , 19 20 , 1005
1006, 21 22 , 23 24 , 25 26 , 27 28 , 29 30 , 31 32 , 33 34 , 35 36 , 1007
1008, 37 38 , 39 40 , 41 42 , 43 44 , 1009
1010, 45 46 , 47 48 , 49 50 , 51 52 , 53 54 , 1011

Tomato Doubled, Ends Added
2000, 1 2 , 10 9 , 3 4 , 11 12 , 2001
2002, 41 42 , 44 43 , 40 39 , 15 16 , 2003
2004, 8 7 , 27 28 , 21 22 , 30 29 , 5 6 , 17 18 , 2005
2006, 13 14 , 31 32 , 36 35 , 33 34 , 2007
2008, 38 37 , 47 48 , 52 51 , 53 54 , 49 50 , 2009
2010, 24 23 , 45 46 , 25 26 , 19 20 , 2011

Short Paths in Breakpoint Graph = 5 (One for each end that agrees.)
1000-1-2000, 1001-12-2001, 1003-16-2003, 1006-13-2002, and 1011-20-2005.

Long Paths in Breakpoint Graph = 7
1004 17 6 7 27 26 19 18 2005
1006 21 28 29 5 4 11 10 2 3 9 8 2004
1007 36 32 33 35 34 2007
1008 37 47 46 25 24 2010
1009 44 42 43 40 41 2002
1010 45 23 22 30 31 14 15 39 38 2008
1011 54 49 48 52 53 51 50 2009
1011 54 50 51 53 52 48 49 2009

Figure 1. Example of genome distance computation

Step 1
1 2 5 6
7 8
-22 -21 -20 16 17 18
-4 -3 15 -11 -14 9
19 24 25 26 27
-12 23 13 10

Short Paths in Breakpoint Graph = 9 (ends 1, 6, 7, 8, 10, 18, 19, 22, 27)

Long Paths and Cycles in Breakpoint Graph = 14
1004 17 27 26 19 18 2005
1006 21 28 29 5 4 9 8 2006
1010 45 23 22 30 31 39 38 47 46 25 24 2010
2=3, 6=7, 10=11, 14=15, 32=33, 34=35, 40=41, 42=43, 48=49, 50=51, 52=53

Distance = 33 – 23 = 10

Step 2
1 2 3 4 5 6
7 8
19 20 21 22
-9 10
23 12 13 14 11 -15
-18 -17 -16 24 25 26 27

Short Paths in the Breakpoint Graph = 10 (ends 1, 6, 7, 8, 10, 18, 19, 22, 23, 27)

Long Paths and Cycles in Breakpoint Graph = 19
1004 17 19 18 2005
1006 21 28 29 2006
46 23 22 30 31 47
2=3, 4=5, 6=7, 8=9, 10=11, 14=15, 24=25, 32=33, 34=35, 38=39, 40=41,
42=43, 48=49, 50=51, 52=53

Distance = 33 – 29 = 4

Figure 2. Intermediate points in our version of the path of Dongalar et al. (2002), which does may of the inversions at the beginning

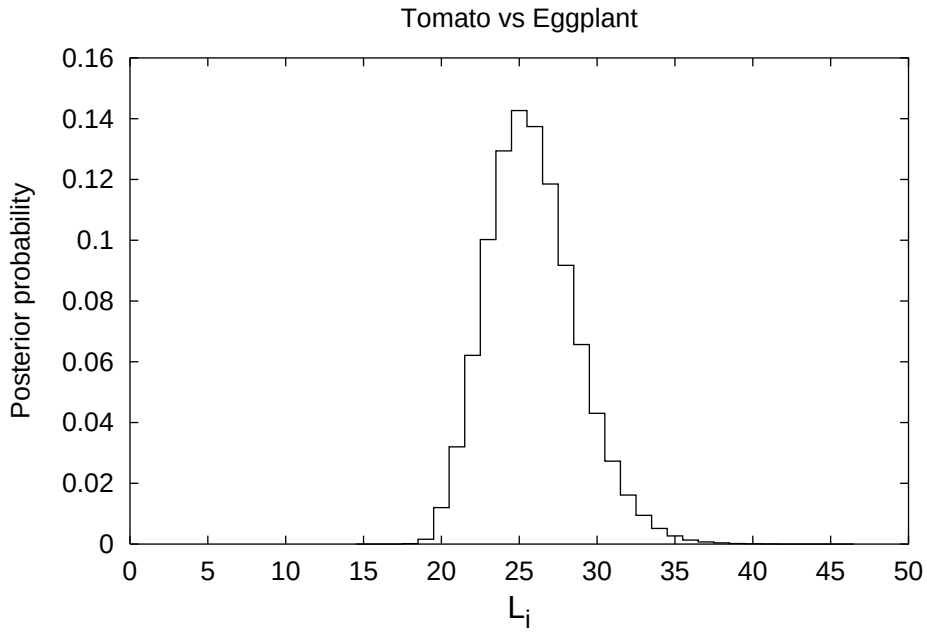

Figure 3. Posterior distribution of the number of inversions for the tomato-eggplant comparison.

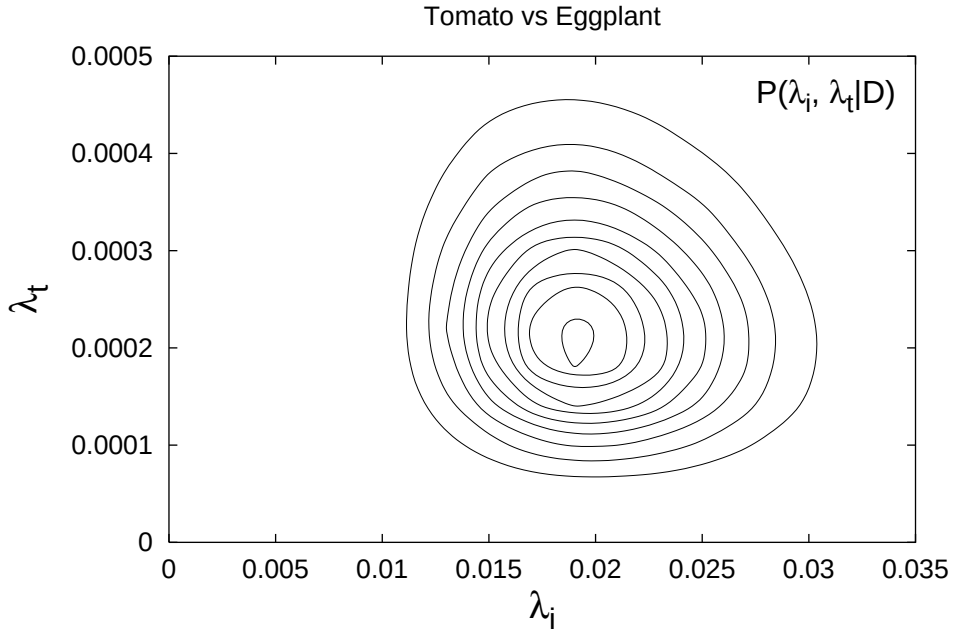

Figure 4. Posterior joint distribution of inversion and translocation rates for the tomato-eggplant comparison

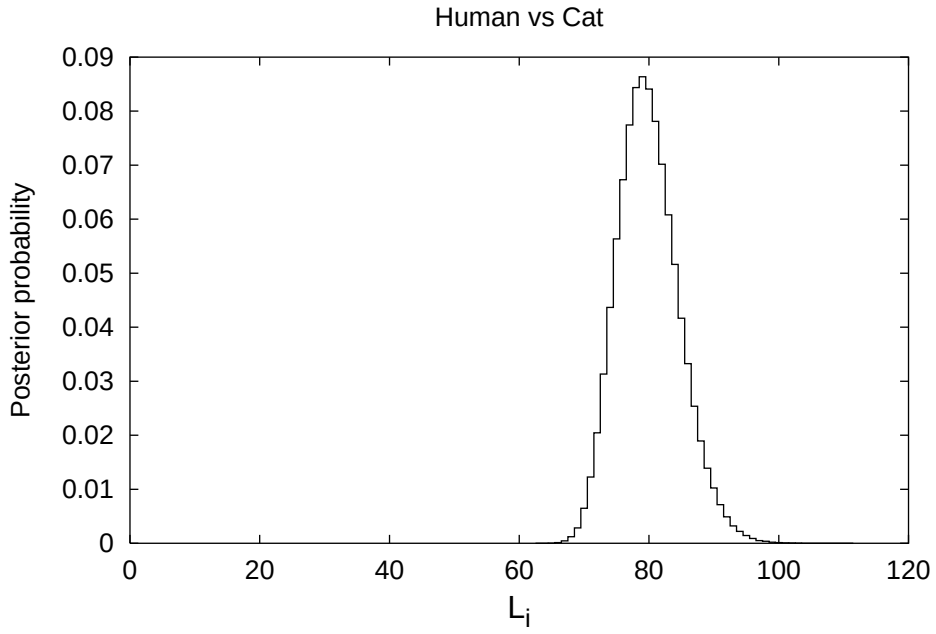

Figure 5. Posterior distribution of the number of inversions for the human-cat comparison

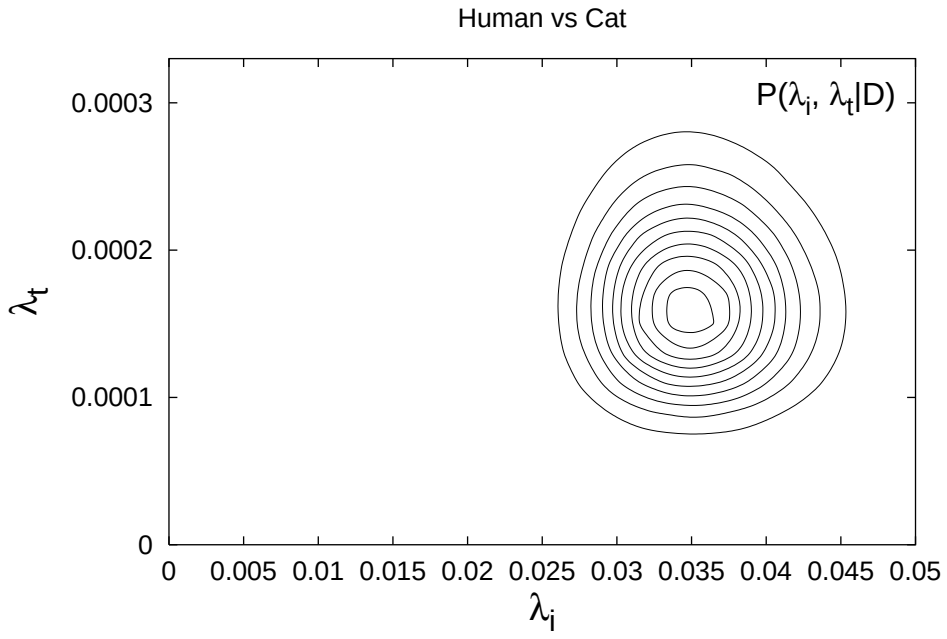

Figure 6. Posterior joint distribution of inversion and translocation rates for the human-cat comparison.

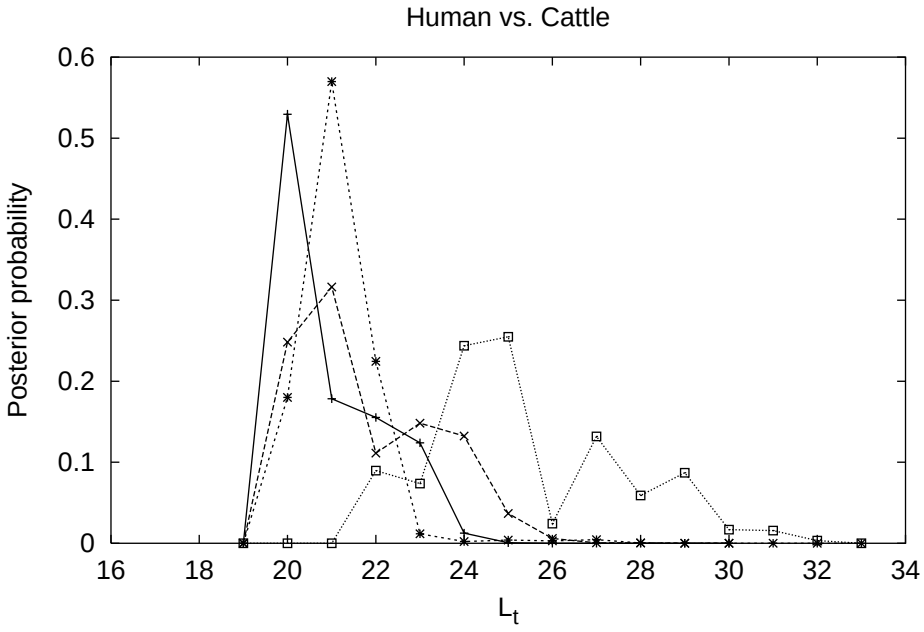

Figure 7. Posterior distributions for the number of translocations in four runs of the human-cattle comparison.

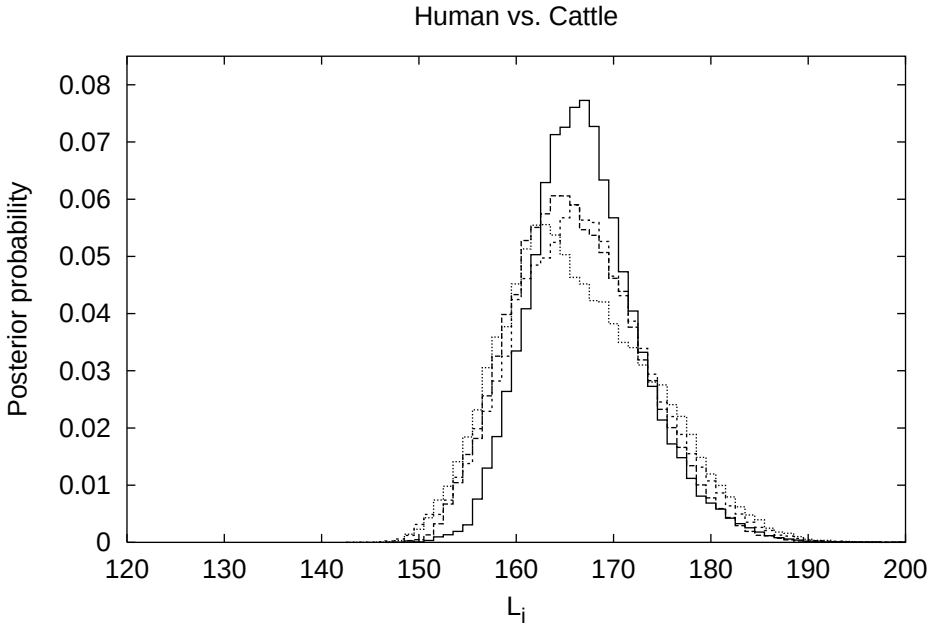

Figure 8. Posterior distributions for the number of inversions in four runs of the human-cattle comparison.