Genome Rearrangement: Recent Progress and Open Problems

by Rick Durrett*, Cornell U.

April 2, 2003

Abstract

Genomes evolve by chromosomal fissions and fusions, reciprocal translocations between chromosomes, and inversions that change gene order within chromosomes. For more than a decade biologists and computer scientists have studied these processes by parsimony methods, i.e., what is the minimum number of events needed to turn one genome into another? We have recently begun to develop a stochastic approach to this and related questions, which has the advantage of producing confidence intervals for estimates and allowing tests of hypotheses concerning mechanisms. Our efforts are closely related to earlier work on card shuffling by Diaconis et al. However now we are not interested in how many shuffles are needed to obtain randomness, but instead want to look at the deck of cards and guess how many shuffles have been performed. This leads to interesting new questions and some surprising answers. This survey is a snapshot of work in progress. The bad news is that the treatment is far from definitive. The good news is that there are still a number of mysteries to solve.

1. Inversions. We begin with the simplest problem of the comparison of two chromosomes where the genetic material differs due to a number of inversions. This occurs for mitochondrial DNA, mammalian X chromosomes and chromosome arms in some insect species (e.g., *Drosophila* and *Anopheles*). To explain the problem, we begin with an example. The table below gives a small subset of the current comparative map between human and mouse X chromosomes. We give the names of 22 genes in the order they are found in the human genome and the location of the homologue in the Jackson Labs map of the mouse genome. The number represents the location along the mouse in centiMorgans, a measure of distance that corresponds to a 1% chance of recombination per generation.

In this comparison the human chromosome can be divided into seven segments where the adjacency of genes has been preserved, except possibly for a reversal of their overall order. Using a minus to denote a segment with a reversed orientation we can write the mouse genome in terms of the human as -4, 3, 7, -2, 6, -5, -1. To explain: in human segment 4 we find the smallest mouse numbers 1.9 and 1.5 which appear in reverse order. The next smallest numbers appear in 3 in correct orientation, etc.

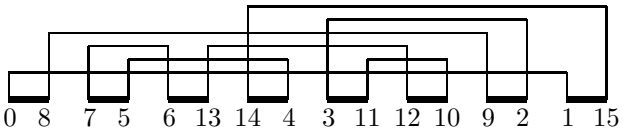
*Partially supported by NSF grants from the probability program (0202935) and from a joint DMS/NIGMS initiative to support research in mathematical biology (0201037).

segment	gene name	mouse (cM)
1	STS	75.0
1	PHKA2	72.0
1	SAT	65.2
2	NROB1	34.0
2	DMD	33.0
3	CYBB	2.8
3	MAOA	5.2
3	ARAF1	6.2
4	GATA1	1.9
4	SYP	1.5
5	SMCX	64.0
5	ALAS2	63.0
6	EFNB1	37.0
6	ATP7A	44.0
6	BTK	51.0
6	PLP1	56.0
6	COL4A5	62.5
7	AGTR2	12.5
7	PlAC1	16.0
7	F9	22.0
7	SOX3	24.5
7	IDH3G	29.5

Hannenhalli and Pevzner (1995a) developed a polynomial algorithm for computing the reversal distance between chromosomes, i.e., what the smallest number of inversions needed to transform one chromosome into another? 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 doubled markers use up the integers 1 to 14. To these we add a 0 at the front and a 15 at the end. Using commas to separate the ends of the markers we can write the two genomes as follows:

$$\begin{aligned} \text{mouse} \quad & 0, 8\,7, 5\,6, 13\,14, 4\,3, 11\,12, 10\,9, 2\,1, 15 \\ \text{human} \quad & 0, 1\,2, 3\,4, 5\,6, 7\,8, 9\,10, 11\,12, 13\,14, 15 \end{aligned}$$

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. In the picture we write the vertices in their order in the mouse genome. Commas in the mouse order become thick lines (black edges), while those in the human genome are thin lines (gray edges). Each vertex has one black and one gray edge so its connected components are easy to find: start with a vertex and follow the connections in either direction until you come back to where you start. In this case all the vertices are part of one cycle: 0, 8, 9, 2, 3, 11, 10, 12, 13, 6, 7, 5, 4, 14, 15, 1, 0.



To compute a lower bound for the distance now we take the number of commas seen when we write out one genome. In this example that is the number of segments +1 or 8, then we subtract the number of connected components in the breakpoint graph. In this example that is 1, so the result is 7. This is a lower bound on the distance since any inversion 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. 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. (See Hannenhalli and Pevzner 1995a.) In symbols if π is the signed permutation that represents the relative order and orientation of segments in the two genomes

$$d(\pi) = n + 1 - c(\pi) + h(\pi) + f(\pi)$$

where $d(\pi)$ is the distance from the identity, n is the number of markers, $c(\pi)$ is the number of components in the breakpoint graph, $h(\pi)$ is the number of hurdles, and $f(\pi)$ is the indicator of the event π is a fortress of hurdles.

Fortunately the complexities associated with hurdles 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 was equal to the lower bound. We can verify that 7 is the minimum distance by constructing a sequence of 7 moves that transforms the mouse X chromosome into the human order. There are 9672 solutions so we leave it as an exercise for the reader to find one. Here are some hints: (i) To do this it suffices to at each step choose an inversion that increases the number of cycles by 1. (ii) This never occurs if the two chosen black edges are in different cycles. (iii) If the two black edges are in the same cycle and are $a - b$ and $c - d$ as we read from left to right, this will occur unless in the cycle minus these two edges a is connected to d and b to c, in which case the number of cycles will not change. For example in the graph above an inversion that breaks black edges 7-5 and 6-13 or 7-5 and 14-4 will increase the number of cycles but the one that breaks 7-5 and 3-11 will not. See Section 5.2 of Durrett (2002) or Chapter 10 of Pevzner (2000) for more details.

Ranz, Segarra, and Ruiz (1997) did a comparative study of chromosome 2 of *Drosophila repleta* and chromosome arm 3R of *D. melanogaster.* If we number the 26 genes that they studied according to their order on the *D. repleta* chromosome then their order on *D. melanogaster* is given by

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

where we have used italics to indicate adjacencies that have been preserved. Since the divergence of these two species, this chromosome region has been subjected to many inversions that reverse a segment of the chromosome. Our first questions is: How many such reversals have occurred? To answer this question we need to formulate and analyze a model.

n-reversal chain. Consider n markers on a chromosome, which we label with $1, 2, \ldots n$ and can be in any of the $n!$ possible orders. To these markers we add two others: one called 0 at the beginning and one called $n+1$ at the end. Finally, for convenience of description, we connect adjacent markers by edges. For example, when $n = 7$ the state of the chromosome might be

$$0 - 5 - 3 - 4 - 1 - 7 - 2 - 6 - 8$$

Note that we no longer keep track of the orientation of the segments.

In our biological applications the probability of an inversion in a given generation is small so, in contrast to the usual card shuffling problems, we will formulate the dynamics in continuous time. The labels 0 and $n+1$ never move. To shuffle the others, at times of a rate one Poisson process we pick two of the $n+1$ edges at random and invert the order of the markers in between. For example, if we pick the edges $5-3$ and $7-2$ the result is

$$0 - 5 - 7 - 1 - 4 - 3 - 2 - 6 - 8$$

If we pick $3-4$ and $4-1$ in the first arrangement there is no visible change. However, allowing this move will simplify the mathematical analysis and only amounts to a small time change of the dynamics in which one picks two markers $1 \le i < j \le n$ at random and reverses the segment with those endpoints.

It is clear that if the chromosome is shuffled repeatedly then in the limit all of the $n!$ orders for the interior markers will have equal probability. The first question is how long does it take for the marker order to be randomized.

Theorem 1. *Consider the state of the system at time $t = cn \ln n$ starting with all markers in order. If $c < 1/2$ then the total variation distance to the uniform distribution ν goes to 1 as $n \to \infty$. If $c > 2$ then the distance goes to 0.*

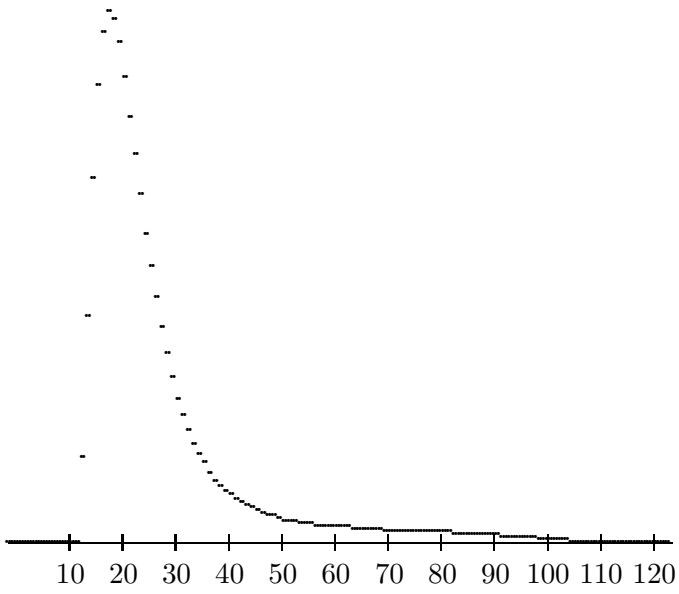
Lower bound. To prove the first half of the result, we define an edge to be *conserved* if the markers at its two ends differ by exactly 1. It is easy to see that the expected number of conserved edges in equilibrium is 2. Suppose now that $t(\epsilon) = (1-\epsilon)(n+1)\ln(n+1)/2$. We say that an edge is *undisturbed* if it has not been involved in a reversal before time t. Let U be the total number of undisturbed edges at time $t(\epsilon)$. A simple computation shows that $EU = (n+1)^{\epsilon}$ and $\mathrm{var}(U)/EU \to 1$ as $n \to \infty$. Letting A_ϵ be the event that there are at most $EU/2$ conserved edges and using Chebyshev's inequality it follows that for large n the total variation distance

$$\|p_{t(\epsilon)} - \nu\|_{TV} \ge |p_{t(\epsilon)}(A_\epsilon) - \nu(A_\epsilon)| \ge 1 - 9/EU \quad \text{if } n \text{ is large}$$

Upper bound. To prove a result in the other direction, we will use techniques that allow the comparison of convergence rates of two reversible Markov chains. The key observation is that the transposition which exchanges markers i and j can be obtained by reversing the segment with end points i and j and then reversing the interior segment obtained by dropping the end points i and j. Using this with Theorem 1 on page 2138 of Diaconis and Saloff-Coste (1993), one can show that reversal chain converges at most $4 = 2 \cdot 2$ times as slowly as the transposition chain. The desired conclusion then follows from a theorem of Diaconis and Shahshahani (1981) for random transpositions, see also Diaconis (1988). □

While Theorem 1 may be interesting, its conclusion does not tell us much about the number of inversions that occurred in our data set. To begin to investigate this question, we note that there are 6 conserved segments. This means that at least $27 - 6 = 21$ edges have been disturbed, so at least 11 reversals have occurred. Biologists often use this easily computed estimate, which is called the *breakpoint distance*. However, this lower bound is usually not sharp. In this example it can be shown that at least 14 reversals are needed to put the markers in order.

The maximum parsimony solution is 14 but there is no guarantee that nature took the shortest path between the two genomes. York, Durrett, and Nielsen (2002) have introduced a Bayesian approach to the problem of inferring the history of inversions separating two chromosomes. We refer the reader to their paper for the details of this Markov chain Monte Carlo method and only show a picture of the posterior distribution of the number of inversions for this data set. See Figure 1. This density assigns a small probability to the shortest path and has a mode at 19.



An alternative and simpler approach to our question comes from considering $\phi(\eta) =$ the number of conserved edges minus 2. Subtracting 2 makes ϕ is orthogonal to the constant eigenfunction. A simple calculation shows that ϕ is an eigenfunction of the chain with eigenvalue $(n-1)/(n+1)$. In our case $n = 26$ and $\phi = 4$ so solving

$$27\left(\frac{25}{27}\right)^m = 4 \quad \text{gives} \quad m = \frac{\ln(4/27)}{\ln(25/27)} = 24.8$$

gives a moment estimate of the number of inversions which seems consistent with the distribution in Figure 1.

Ranz, Ruiz, and Casals (2001) enriched the comparative map so that 79 markers can be located in both species. Again numbering the markers on the *D. repleta* chromosome by their order on *D. melanogaster* we have:

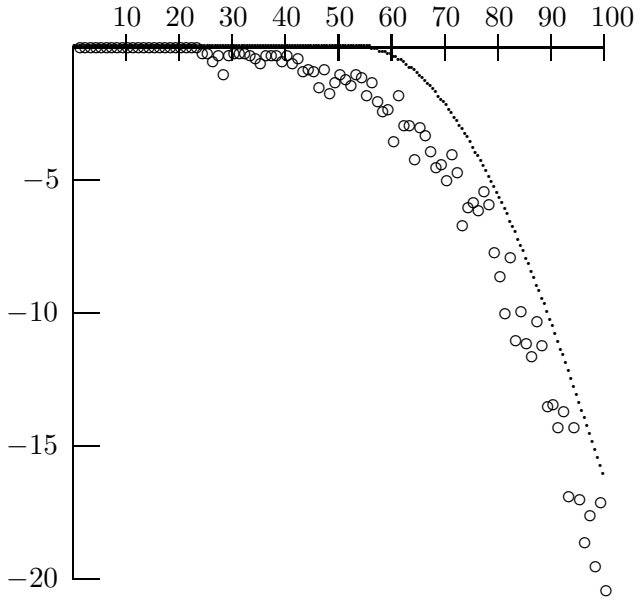
36	*37*	17	40	*16*	*15*	*14*	63	*10*	*9*	55	28
13	51	22	79	39	70	66	*5*	*6*	*7*	35	64
33	*32*	*60*	*61*	18	65	62	12	1	11	23	20
4	52	68	29	48	3	21	53	8	43	72	*58*
57	*56*	19	49	34	59	30	77	31	67	44	2
27	38	50	*26*	*25*	76	69	41	24	75	71	78
73	47	54	45	74	42	46					

The number of conserved adjacencies (again indicated with italics) is 11 so our moment estimate is

$$m = \frac{\ln(9/80)}{\ln(78/80)} = 86.3$$

This agrees with the Bayesian analysis in York, Durrett, and Nielsen (2002) where the mean of the posterior distribution is 92.61 and the mode is at 87. However these two numbers differ drastically from the parsimony analyses. The breakpoint distance is $(80 - 11)/2 = 35$, while the parsimony distance is 54. This lies outside the 95% credible interval of $[65, 120]$ that comes from the Bayesian estimate. Indeed the posterior probability of 54 is so small that this value that it was never seen in the 258 million MCMC updates in the simulation run.

2. Distances. In the last two examples we saw that the breakpoint distance was likely to be an underestimate of the true distance. This brings up the question: when is the parsimony estimate reliable? Bourque and Pevzner (2002) have approached this question by taking 100 markers in order performing k randomly chosen inversions, computing D_k the minimum number of inversions needed to return to the identity and then plotting the average value of $D_k - k \leq 0$ (the circles in the graph below). They concluded based on this and other simulations that the parsimony distance based on n markers was a good as long as the number of inversions was at most $0.4n$. The smooth curve, which we will describe below, gives the limiting behavior of $(D_{cn} - cn)/n$.



For simplicity we will consider the analogous problem for random transpositions. In that case the distance from the identity can be easily computed: it is the number of markers n minus the number of cycles in the permutation. For an example, consider the following permutation of 14 objects written in its cyclic decomposition:

$$(1\,7\,4)\ (2)\ (3\,12)\ (5\,13\,9\,11\,6)\ (8\,10\,14)$$

which indicates that $1 \to 7$, $7 \to 4$, $4 \to 1$, $2 \to 2$, $3 \to 12$, $12 \to 3$, etc. There are 5 cycles so the distance from the identity is 9. If the we perform a transposition that includes markers from two different cycles (e.g., 7 and 9) the two cycles merge into 1, while if we pick two in the same cycle (e.g., 13 and 11) it splits into two.

The situation is similar but slightly more complicated for reversals. There if we ignore the complexity of hurdles, the distance is $n + 1$ minus the number of components in the breakpoint graph. A reversal that involves edges in two different components merges them into 1 but, as we saw in the human mouse comparison, a reversal that involves two edges of the same cycle may or may not increase the number of cycles. To have a cleaner mathematical problem, we will consider the biologically less relevant case of random transpositions, and ask a question that in terms of the rate 1 continuous time random walk on the permutation group is: how far from the identity are we at time cn?

The first step in attacking this problem is to notice that by our description above the cycle structure evolves according to a coagulation-fragmentation process. Suppose that for the moment we ignore fragmentation and draw an edge from i to j whenever we transpose i and j. In this case the cycles are the components of the resulting random graph. There are $n(n-1)/2$ potential edges, so results of Erdös and Renyi imply that when $c < 1/2$ there are no very large components and we can ignore fragmentations. In this phase the distance will typically increase by 1 on each step or in the notation of Bourque and Pevzner, $D_k - k \approx 0$.

When $c > 1/2$ a giant component emerges in the percolation model and its behavior is much different from the large cycles in the permutation which experience a number of fragmentations and coagulations along the way to establishing a well known equilibrium distribution of sizes in a random permutation. As Hansen (1990) has shown (see also Donnelly, Kurtz, and Tavaré 1991) if $K_n(u)$ is the number of cycles of size $\le n^u$ then

$$\frac{K_n(u) - u \log n}{\sqrt{\log n}} \Rightarrow B_u \quad \text{a Brownian motion}$$

The dynamics of the large components are quite complicated but (i) there can never be more than $\sqrt{n}$ of size $\sqrt{n}$ or larger and (ii) an easy argument shows that the number of fragmentations occurring to clusters of size $\le \sqrt{n}$ is $O(\sqrt{n})$. These two observations plus results from the theory of random graphs (see Theorem 12 in Section V.2 of Bollobás 1985) imply

Theorem 2.1. *The number of cycles at time cn is* $g(2c)n + O(\sqrt{n})$ *where*

$$g(b) = \sum_{k=1}^{\infty} \frac{1}{k} p_k(b) \quad \textit{and} \quad p_k(b) = \frac{1}{b} \frac{k^{k-1}}{k!} (be^{-b})^k$$

Using Stirling's formula $k! \sim k^k e^{-k} \sqrt{2\pi k}$ it is easy to see that g' is continuous but $g''(1)$ does not exist. It is somewhat remarkable that $g(b) = 1 - b/2$ for $b < 1$. Thus there is a phase transition

in the behavior of the distance of the random transposition random walk from the identity at time $n/2$ from linear with time to sublinear.

Sketch of Proof. We begin with Cayley's result that there are k^{k-2} trees with k labeled vertices. At time cn each edge is present with probability $\approx cn/\binom{n}{2} \approx 2c/n$ so the expected number of trees of size k is present is

$$\binom{n}{k} k^{k-2} \left(\frac{2c}{n}\right)^{k-1} \left(1-\frac{2c}{n}\right)^{k(n-k)+\binom{k}{2}-k+1}$$

since each of the $k-1$ edges need to be present and there can be no edges connecting the k point set to its complement or any other edges connecting the k points. For fixed k the above is

$$\approx n\frac{k^{k-2}}{k!}(2c)^{k-1}\left(1-\frac{2c}{n}\right)^{kn}$$

from which the result follows easily. □

Corollary. We have written the conclusion in the form given above so that $p_k(b)$ is the probability that 1 belongs to a component of size k in an Erdös-Renyi graph with edge occupancy probability b/n. A corollary of this result is that $p_k(b)$ gives the distribution of the total progeny in a Poisson branching process with mean b. That result is due to Borel (1942) for the equivalent problem of the total number of units served in the first busy period of a queue with Poisson arrivals and constant service times (the customers that arrive during a person's service time are their offspring). For a proof see Tanner (1961).

Having found laws of large numbers for the distance, it is natural to ask about fluctuations. This project is being carried out with Nathaniel Berestycki, a visiting student from Paris, who will eventually get a dual degree from Cornell and Paris. The subcritical regime is easy. Let F_t be the number of fragmentations at time t in a system in which transpositions occur at rate 1. The continuous time setting is more convenient since it leads to a random graph with independent edges. If N_t is the number of transpositions at time t then $D_t - N_t = -2F_t$ so we study the latter quantity.

Theorem 2.2. *Suppose $0 < c < 1$. As $n \to \infty$, $F_{cn/2}$ converges in distribution to a Poisson random variable with mean $(-\ln(1-c)-c)/2$.*

Sketch of Proof. Let $f_k(t)$ be the fraction of vertices at time t that are in components of size k. The rate at which fragmentations occur at time t is $\sum_k f_k(t)(k-1)/n$. If we look in the subcritical random graph at the component containing 1 then results from the theory of random graphs imply that with high probability it is a tree and if $t = bn/2$ then the number of individuals at distance k from 1 has roughly the same distribution as the kth generation of a branching process in which each individual has a Poisson mean b number of descendants. Thus the expected rate at which fragmentations occur is $1/(1-b)-1$. Integrating from 0 to c and recalling $t = bn/2$ gives the expected value. To prove convergence to the Poisson we show that the point process of fragmentations has in the limit a deterministic compensator and then we appeal to results in Jacod and Shirayev (1987). □

Theorem 2.3. *Let $c_n(r) = 1 - n^{-r/3}$ for $0 \le r \le 1$. As $n \to \infty$*

$$X_n(r) = \frac{F_{c_n(r)n/2} - (r/6)\log n}{(1/6)\log n)^{1/2}}$$

converges to a standard Brownian motion.

Sketch of Proof. The first thing to do is to explain why we stop at time $1 - n^{-1/3}$. This is the boundary between the subcritical phase where the arguments in the proof of Theorem 2.1 work and the critical phase where the largest component is of order $n^{2/3}$. See Aldous (1997) for a very interesting account of cluster growth in the random graph in the critical regime that relates it to a multiplicative coalescent. Expected value estimates (see Luczak, Pittel, and Wierman 1994) imply that the number of fragmentations in $[1 - n^{-1/3}, 1]$ is $O(1)$ and hence can be ignored. Our time change in the subcritical regime makes $X_n(r)$ asymptotically a martingale with variance process r and hence a Brownian motion. □

Conjecture 2.4. *Let $\sigma(c)$ be the standard deviation of the random variable that is equal to $1/k$ with probability $p_k(c)$. If $c > 1$ then $(D(cn/2) - (1 - g(c))n)/\sigma(c)n^{1/2}$ converges to a standard normal distribution.*

Why is this true? The conjecture is based on our belief that the number of cycles in the random permutation is to within $o(\sqrt{n})$ the number of cycles $\Gamma(c)$ in the random graph that comes from ignoring fragmentations. The proof is made difficult by the fact that the number of fragmentations is not small enough to be ignored but one must use the fact that pieces that break off the giant component are also reabsorbed by it. Once this is established, the rest is easy. $\Gamma(c) = \sum_i 1/|\mathcal{C}_i|$ where $|\mathcal{C}_i|$ is the size of the component containing i. The variables $-1/|\mathcal{C}_i|$ are increasing functions of the i.i.d. variables that define the random graph, and are asymptotically uncorrelated so the desired conclusion from a result of Newman and Wright (1981). □

3. Genomic Distance. In general genomes evolve not only by inversions within chromosomes but due to translocations between chromosomes, and fissions and fusions that change the number of chromosomes. To reduce the number of events considered from four to two, we note that a translocation splits two chromosomes (into say $a - b$ and $c - d$) and then recombines the pieces (to make $a - d$ and $b - c$ say). A fission is the special case in which the segments c and d are empty, a fusion when b and c are. To illustrate the problem we will consider part of the data of Doganlar et al. (2002) who constructed a comparative genetic linkage map of eggplant with 233 markers based on tomato cDNA, genomic DNA and ESTs. 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 6 singleton segments (numbers 3, 4, 12, 14, 15, 16) so one can easily consider all $2^6 = 64$ possibilities. The next table shows the two genomes with an assignment of signs to the singleton markers that minimizes the distance.

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

As in the reversal distance problem, our first step is to double the markers. 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 next table shows the result of the first two preparatory steps. Commas indicate separations between two segments or between a segment and an end.

Eggplant
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
2000, 1 2 , 10 9 , 3 4 , 11 12 , 2001
2002, 41 42 , 44 43 , 40 39 , 15 16 , 2003
2004, 7 8 , 28 27 , 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

As before, 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: $1000 - 1 - 2000$, $1001 - 12 - 2001$, $1002 - 13 - 2006$, $1003 - 16 - 2003$, and $1005 - 20 - 2011$. These 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 remaining paths and cycles in the breakpoint graph are listed below.

1004 17 6 7 2004
1006 21 27 26 19 18 2005
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
2 3 9 8 28 29 5 4 11 10

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 $5 + 8 = 13$, 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 20 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. As before this is only a lower bound. For the genomic distance problem the full answer is quite complicated and involves 7 quantities associated with genome. (For more details see Hannehalli and Pevzner 1995b or Pevzner 2000.) At least in this example, nature is simpler than the mathematically worst possible case. It is easy to produce path of length 20 to show that the lower bound is achieved.

It is straightforward to generalize the methods of York, Durrett, and Nielsen (2002) to treat the genomic distance problem and this is being done in a continuing collaboration of these three authors. Not much is known about properties of the corresponding stochastic model, which can be thought of as shuffling multiple decks of cards. De, Ferguson, Sindi, and Durrett (2001), following up on work of Sankoff and Ferretti (1996), considered the evolution of chromosome sizes due to translocations. Suppose that a genome consists of N nucleotides total and k chromosome arms with lengths $N_1, \ldots N_k$ (where $N_i \geq 2$ and $\sum_i N_i = N$). We model chromosome arms rather than whole chromosomes to capture the biological constraint that after translocation each chromosome must have one centromere. There are $N - k$ spaces between nucleotides. Suppose we pick two spaces at random. If they are in the same arm we do nothing. Otherwise we perform the one possible translocation that results in one centromere on each chromosome.

It is easy to see that if we let x and y be two chromosome length vectors our chain has $p(x, y) = p(y, x)$ so a uniform distribution over all states is stationary. While this result is nice and simple, nature is unfortunately more complicated. Comparing with say the lengths of human chromosomes the short chromosomes in the model are too short and the long chromosomes are too long. One way to modify the chain is to note that recombination between homologous chromosomes is important to pair chromosomes for cell division and to introduce a fitness function

$$\phi(x) = \prod_{i=1}^{k} (1 - e^{-x_i})$$

which is the probability each homologous chromosome arm experiences recombination. We have no constant in the fitness function since we are now assuming lengths are measured in Morgans (=100 centiMorgans) and hence recombinations in each generation are a rate one Poisson process.

If we take inspiration from the Metropolis algorithm and define

$$q(x,y) = \begin{cases} p(x,y) & \text{if } \phi(y) \geq \phi(x) \\ p(x,y)\frac{\phi(y)}{\phi(x)} & \text{if } \phi(x) \geq \phi(y) \end{cases}$$

then $\phi(x)q(x,y) = \phi(y)q(y,x)$, so $\phi(x)$ is a reversible measure. The new model fits data from four species well (human, rat, pig, and yeast) but does not fit well for four others (mouse, sheep, wheat, and rice). We refer the reader to De et al. (2001) for more discussion of fitting the model to data and turn now to the statement of two mathematical problems.

Problem 3.1. *Consider the original version of the Markov chain without the fitness function. In terms of card shuffling we can pose the problem as follows. There are k decks of cards. We pick two spaces between cards and cut the decks at the chosen location. If both cuts are made in the same stack we do nothing. Otherwise we put the top of one deck on the bottom of the other. We know that the uniform distribution for the vector of deck sizes is stationary. The problem is to understand the rate of convergence to equilibrium.*

Problem 3.2. *Consider the previous chain but now (i) when the two spaces are chosen in the same deck we invert the order of the cards in between and (ii) if a translocation between two decks is proposed we do that move only with probability p. In general translocations occur an overall a rate 1/10 that of inversions. Since in a typical arrangement there are many more possible translocations than inversions p is quite small. In the case of inversion distance we had one statistic, the number of conserved adjacencies could be used to estimate the number of inversions. For the new problem, we would like to find two statistics that will allow us to estimate both the number of translocations and inversions.*

When distances between the markers are known in one genome, there is another method due to Nadeau and Taylor (1984) that can be used. We will describe their technique in some detail since it makes an interesting use of elementary results for Poisson processes. The basic data for the process is the set of lengths of conserved segments, i.e., two or more consecutive markers in one genome that are adjacent (possibly in the reverse order) in the other. The actual conserved segment in the genome is larger than the distance r between the two markers at the end of the conserved interval. Thinking about what happens when we put n points at random in the unit interval we estimate the length of the conserved segment containing these markers by $\hat{r} = r(n+1)/(n-1)$ where n is the number of makers in the segment.

Let D be the density of markers, i.e., the total number divided by the size of the genome. If the average length of conserved segments is L and we assume that their lengths are exponentially distributed then since we only detect segments with at least two markers the distribution of their lengths is

$$(1 - e^{-Dx} - Dxe^{-Dx})\frac{1}{L}e^{-x/L}$$

normalized to be a probability density. A little calculus shows that the mean of this distribution is $(L^2D + 3L)/(LD + 1)$.

Historically the first application of this technique was to a human-mouse comparative map with a total of 56 markers. Based on this limited amount of data they estimated that there were 178 ± 39 conserved segments. This estimate has held up remarkably well as the density of the comparative

map has increased. See Nadeau and Sankoff (1998). To illustrate this computation we will use a comparative map of the human and cattle autosomes (non-sex chromosomes) constructed by Band et al. (2000). Using resources on the NCBI home page we were able to determine the location in the human genome of 422 genes in the map. These defined 125 conserved segments of actual average length 7.188 Mb (megabases) giving rise to an adjusted average length of 14.501 Mb. Assuming 3200 Mb for the size of the human genome the marker density was $D = 1.32 \times 10^{-4}$ or one every 7.582 Mb. Setting $14.501 = (L^2D + 3L)/(LD + 1)$ and solving the quadratic equation for L gives an estimate $\hat{L} = 7.144$ Mb, which translates into approximately 448 segments. Subtracting 22 chromosomes we infer there were 424 breakpoints, which leads to an estimate of 212 inversions and translocations. As a check on the assumptions of the Nadeau and Taylor computation, we note that if markers and segment endpoints are distributed randomly then the number of markers in a conserved segment would have a geometric distribution. The next table compares the observed counts with what was expected

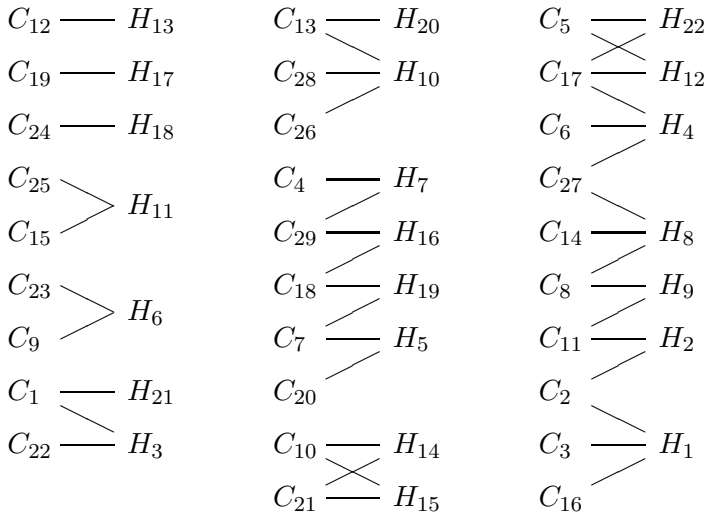
markers	observed	expected
0	–	222.9
1	85	108.1
2	76	52.5
3	29	25.4
4	10	12.3
5	5	6.0
6	3	2.9
7	1	1.4
8	1	0.7

To get an idea of the number of translocations that have occurred we will look at the human-cattle correspondence through the eyes of FISH (fluorescent in situ hybridization) data of Hayes (1995) and Chowdary et al. (1996). In this technique one takes individual human chromosomes labels them with fluorescent chemicals and then determines where they hybridize to cattle chromosomes. In the case of the human cattle comparison taking a consensus of the two painting experiments and the comparative map gives the following data relating the 29 cattle autosomes to the 22 human ones. The first number is the cattle chromosome number. Those that follow the period indicate where the segment appears in the human genome.

1. 21, 3, 21	11. 2, 9	21. 15, 14
2. 2, 1	12. 13	22. 3
3. 1	13. 10, 20	23. 6
4. 7	14. 8	24. 18
5. 12, 22, 12, 22	15. 11	25. 11
6. 4	16. 1	26. 10
7. 19, 5	17. 4, 12, 22	27. 4, 8
8. 8, 9	18. 16, 19	28. 10
9. 6	19. 17	29. 16, 7
10. 15, 14	20. 5	

To visualize the relationship between the genomes it is useful to draw the bipartite graph with vertices the chromosome numbers in the two genomes and an edge from C_i to H_j if there is part

of cattle chromosome i is homologous part of human chromosome j. We call this the *Oxford graph* since the adjacency matrix of this graph is what biologists would call an *Oxford grid*.



Parsimony analysis reveals that a minimum of 19 translocations is needed to rearrange the cattle genome to match the chromosomes of the human genome. Using the estimate of Nadeau and Taylor and subtracting 202 - 19 = 183 we get an estimate for the number of inversions consistent with the belief that inversions occur at roughly 10 times the rate of translocations. With Problem 3.2 in mind, we observe that in the human cattle comparison there are 212 conserved adjacenices out of 400 possible so the moment estimate is 127.9. This is much smaller than the other estimate of the total number of events, so we need two new statistics for Problem 3.2.

4. Nonuniformity of inversions. Define a *syntenic segment* to be a segment of chromosome where all of the markers come form the same chromosome in the other species, but not necessarily in the same order. A remarkable aspect of the cattle data is that although our estimates suggest that there have been roughly 19 translocations and 183 inversions, each chromosome consists of only a few syntenic segments. If inversions were uniformly distributed on the chromosome we would expect inversions that occur after a translocations would mingle the two segments.

A second piece of evidence that not all inversions are equally likely comes from the 79 marker *Drosophila* data. There the estimated number of reversals is large but there is still a strong correlation between the marker order in the two genomes. Spearman's rank correlation $\rho = 0.326$ which is significant at the $p = 0.001$ level. From the point of view of Theorem 1 this is not surprising: our lower bound on the mixing time predicts that $39.5 \ln 75 = 173$ reversals are needed to completely randomize the data. However, simulations in Durrett (2003) show that the rank correlation is randomized well before that time. In 10,000 runs the average rank correlation is only 0.0423 after 40 shuffles and only 4.3% of the runs had a rank correlation larger than 0.325.

To seek a biological explanation of the non-uniformity we note that the gene-to-gene pairing of homologous chromosomes implies that if one chromosome of the pair contains an inversion that the other does not, a loop will form in the region in which the gene order is inverted. (See e.g, page 367 of Hartl and Jones 2000.) If a recombination occurs in the inverted region then the recombined chromosomes will contain two copies of some regions and zero of others, which can have unpleasant consequences. A simple way to take this into account is

θ-reversal model. Inversions that reverse markers i to $i+j$ occur at rate $\theta^{j-1}/n(1-\theta)$.

The reasoning here is that the probability of no recombination decreases exponentially with the length of the segment reversed.

This model seems quite complicated to analyze, so we will make two simplifications. The first is that we will consider the markers $1, 2, \ldots n$ as lying on a circle and connect them by edges as before. For example, when $n = 7$ we have $1-2-3-4-5-6-7-1$. Departing slightly from our previous approach, we pick the location of the left marker of the segment to be inverted, i, uniformly over the set of possibilities, and then pick the location of the right marker to be $i+j$ with probability p_j, where the arithmetic is done modulo n. Some of the results in Durrett (2003) were obtained for this ***p*-reversal model**, but in most cases he considered the simpler

***L*-reversal model.** $p_j = 1/L$ for $1 \le j \le L$.

An easy extension of the lower bound in Theorem 1 shows that the amount of time for the p-reversal chain to reach equilibrium is at least $(n/2)\ln n$. To prove a second lower bound we will use an idea of Wilson (2001). The first step in the analysis to note that a single marker performs a symmetric random walk on the circle. To compute the jump distribution we note that the marker at n will be moved to i if the left endpoint of the inversion is at $n-k$ and the right is at $i+k$ where $k \ge 0$ and $n-k > i+k$. Summing we have the rate for jumps by $+i$ is q_i/n where

$$q_i = \sum_{k \ge 0} p_{i+2k}$$

The condition $n-k > i+k$ does not appear in the sum since we suppose $p_m = 0$ for $m \ge n$. Symmetry implies that $q_{-i} = q_i$. When p_i is uniform on $1, 2, \ldots L$ this has almost a triangular distribution:

$$q_i = (1 + [(L-i)/2])/L \quad \text{for } 1 \le i \le L$$

where $[z]$ denotes the integer part of z. When p is geometric and $n(1-\theta)$ is large so we can ignore truncation of the infinite series

$$q_i = \theta^{i-1}(1-\theta)/(1-\theta^2)$$

Theorem 4.1. *If the distribution p_i is fixed and $n \to \infty$ then convergence to equilibrium takes at least time*

$$\frac{1}{2} \cdot \frac{n^3}{4\pi^2 \sum_i i^2 q_i} \ln n$$

In the L-reversal chain if $L \to \infty$ and $(\ln L)/(\ln n) \to a \in [0,1)$ then convergence to equilibrium takes at least time

$$\frac{(1-a)}{2} \cdot \frac{6}{\pi^2} \cdot \frac{n^3}{L^3} \ln n$$

The key to the proof is the fact that $f(x) = \sin(2\pi x/n)$ is an eigenfunction for any symmetric random walk on the circle, which for the single particle walk has eigenvalue

$$-\lambda \equiv \sum_{i=1}^{n} \frac{2q_i}{n}[\cos(2\pi i/n) - 1] \sim -\frac{4\pi^2}{n^3}\sum_{i=1}^{n} i^2 q_i \quad \text{as } n \to \infty$$

Let $\eta_t(i)$ be the marker at position i at time t, and $X_t^j = \eta_t^{-1}(j)$ be the location of marker j at time t. Let

$$g(m) = \operatorname{sgn}\left(\frac{n+1}{2} - m\right) = \begin{cases} 1 & m < (n+1)/2 \\ 0 & m = (n+1)/2 \\ -1 & m > n+1)/2 \end{cases}$$

and let

$$\Phi(\eta_t) = \sum_i g(\eta_t(i)) \sin(2\pi i/n) = \sum_j g(j) \sin(\pi X_t^j/n)$$

Letting $\Phi_t = \Phi(\eta_t)$, it follows from the calculation above that

$$\frac{d}{dt} E\Phi_t = -\lambda E\Phi_t$$

When the markers start in order $\Phi_0 \approx 2n/\pi$. Since in equilibrium $\operatorname{var}(\Phi_\infty) = O(n)$, this suggests that the chain cannot be in equilibrium until $E\Phi_t = O(\sqrt{n})$ which takes about $(\ln n)/2\lambda$ units of time. To complete this outline we have to show that $\operatorname{var}(\Phi_t) \le Cn$. This can be done by using ideas from Wilson (2001). See Durrett (2003) for more details. We get a result with a worse constant for the L-reversal since we can only show that $\operatorname{var}(\Phi_t) \le CnL$.

Problem 4.1. *When $L = n^a$ with $a > 2/3$ the lower bound in Theorem 4.1 is for large n worse than the $(n/2)\ln n$ bound we get by looking at conserved adjacencies. We believe that the maximum of these two lower bounds gives the right order of magnitude for the convergence time but we have not been able to prove matching upper bounds.*

Using Wilson's idea we can introduce a statistic for the shuffling a linear (not circular chromosome)

$$\sum_{i=1}^{79} g(\eta_t(i)) \cos(\pi(i - 0.5)/n)$$

Here we have replaced the $\sin(2\pi i/n)$ by something that is an eigenfunction for the nearest neighbor random walk on $1, 2, \ldots n$ with reflecting boundary conditions. Durrett (2003) gives results of 10,000 simulations of the 23-reversal chain acting on 79 markers. Plotting the logarithm of the rank correlation, Wilson's statistic and the number of conserved segments -2, and scaling these statistics to have maximum value 1, the three curves are almost straight lines. The value at time 85 for the (conserved segments - 2)/77 is .1224 which is a little larger than the value of .1168 for the data. The value of 23 was chosen so that the value of the rank correlation 0.312 matched the 0.326 of the data as closely as possible. A simulation of the 22-reversal chain gave a value of 0.355 at that time.

Problem 4.2. *Do these quantities decay exponentially for the p-reversal model? or at least for the special cases of the L- or θ-reversal models?*

A second way to approach the problem of estimating inversion track lengths is to see how estimates of the number of reversals depend on the density of markers in the map. If n markers (blue balls) are randomly distributed and we pick two inversion end points (red balls) at random then the relative positions of the $n+2$ balls are all equally likely. The inversion will not be detected by the set of markers if there are 0 or 1 blue balls between the two red ones an event of probability

$$\frac{n+1+n}{\binom{n+2}{2}} = \frac{4n+2}{(n+2)(n+1)} \approx \frac{4}{n+2}$$

This means that the 26 markers in the first *Drosophila* data set should have missed only 1/7 of the inversions in sharp contrast to the fact that our estimate jumped from 24.8 with 26 markers to 86.3 with 79.

Suppose now that markers are distributed according to a Poisson process with mean spacing M while inversion track lengths have an exponential distribution with mean L. If we place one inversion end point at random on the chromosome and then move to the right to locate the second one then the probability a marker comes before the other inversion endpoint is

$$\frac{1/M}{1/M+1/L} = \frac{L}{L+M}$$

so the fraction detected is $L^2/(L+M)^2$. If we take 30Mb as an estimate for the size of the chromosome arm studied, we see that the marker spacings in the two studies are: $M_1 = 30/27 = 1.11$ Mb and $M_2 = 30/80 = .375$ Mb respectively. Taking ratios we can estimate L by

$$\frac{86.3}{24.8} = \frac{(L+1.1)^2}{(L+0.375)^2}$$

Taking square roots of each side and solving we have $1.865L+0.375 = L+1.1$ or $L = 0.725/0.765 = 0.948$ Mb. If this is accurate then the larger data set only detects

$$\left(\frac{0.948}{1.273}\right)^2 = 0.554$$

or 55.4% of the inversions that have occurred.

While two data points are enough to make estimates, it would be nice to have more extensive data on which to test our model. This is the topic of an undergraduate research project being conducted by David Russell. We are of course not able to refine an existing comparative map but we can take a detailed one: the comparison of human and mouse X chromosome and extract randomly chosen subsets of various sizes to examine the dependence of the estimate on the marker density.

REFERENCES

Aldous, D. (1997) Brownian excursions, critical random graphs, and the multiplicative coalescent. *Annals of Probability.* **25**, 812–854

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

Bollobás, B. (1985) *Random Graphs.* Academic Press, New York

Borel, E. (1942) Sur l'empoi du théorème de Bernoulli pour faciliter le calcul d'un infinité de coefficients. Application au problem de l'attente á un guichet. *C. R. Aacd. Sci. Paris.* **214**, 452–456

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

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

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. *J. Appl. Prob.* **38**, 324–334

Diaconis, P. (1988) *Group Representations in Probability and Statistics.* Institute of Mathematical Statistics Lecture Notes, Volume 11

Diaconis, P., and Saloff-Coste, L. (1993) Comparison techniques for random walks on finite groups. *Annals of Probability.* **21**, 2131–2156

Diaconis, P., and Shahshahani, M. (1981) Generating a random permutation with random transpositions. *Z. fur. Wahr.* **57**, 159–179

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

Donnelly, P., Kurtz, T.G., and Tavaré, S. (1991) On the functional central limit theorem for the Ewens sampling formula. *Ann. Appl. Prob.* **1**, 539–545

Durrett, R. (2002) *Probability Models for DNA Sequence Evolution.* Springer-Verlag, New York

Durrett, R. (2003) Shuffling chromosomes. *J. Theoretical Prob.* to appear

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

Hansen, J.C. (1990) A functional central limit theorem for the Ewens sampling formula. *J. Appl. Prob.* **27**, 28–43