

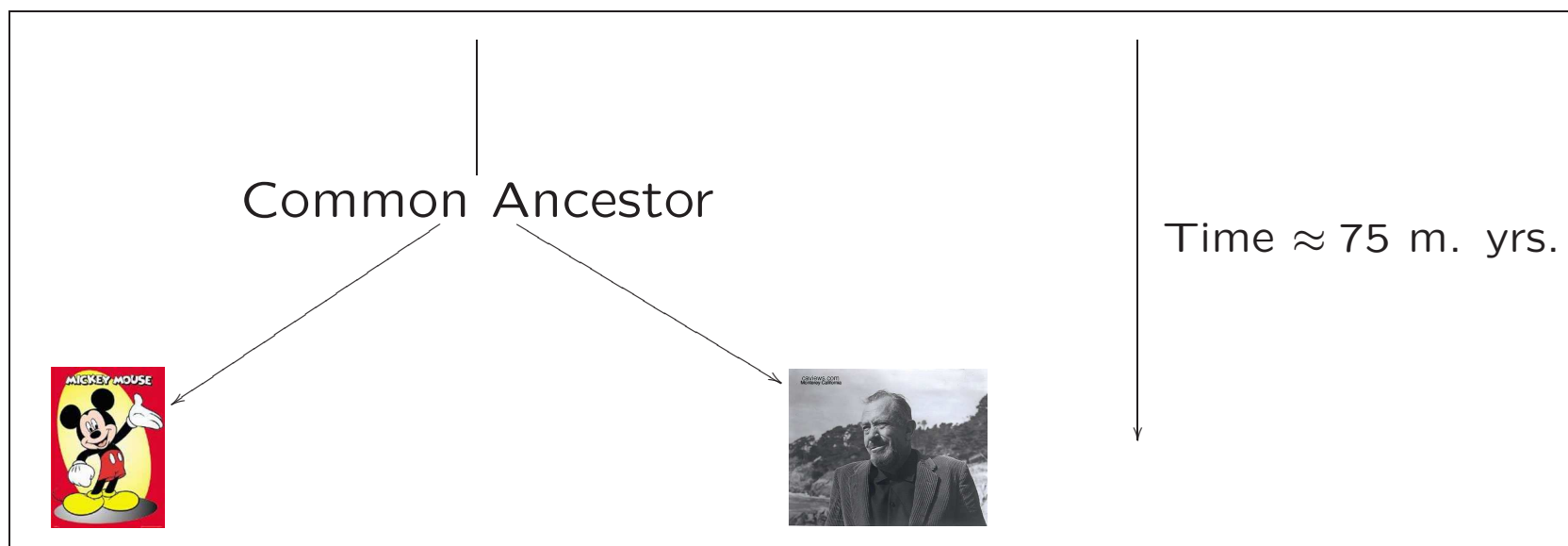
The small-scale effect
for large-scale mutations

Nathanaël Berestycki

University of British Columbia
partly joint with Rick Durrett

Workshop on Mathematical Population Genetics

Of mice and men

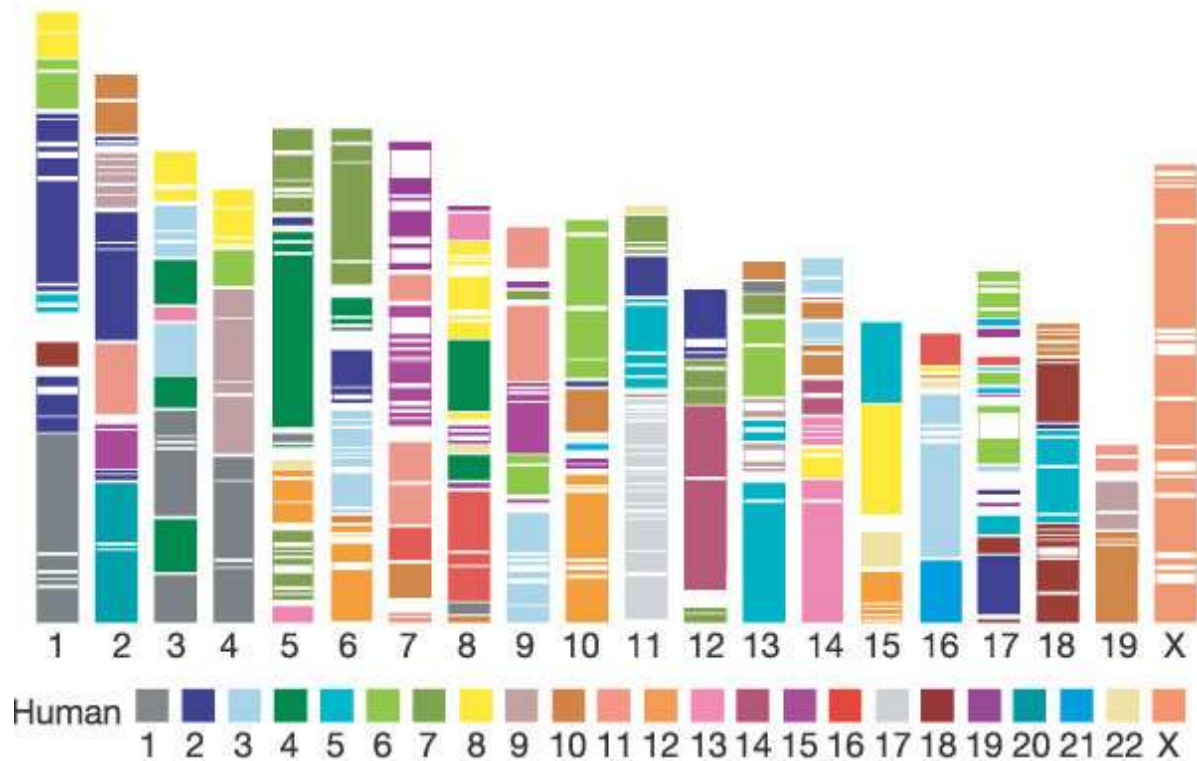


Question: What is the evolutionary distance between the mouse and the human?

► More precisely:

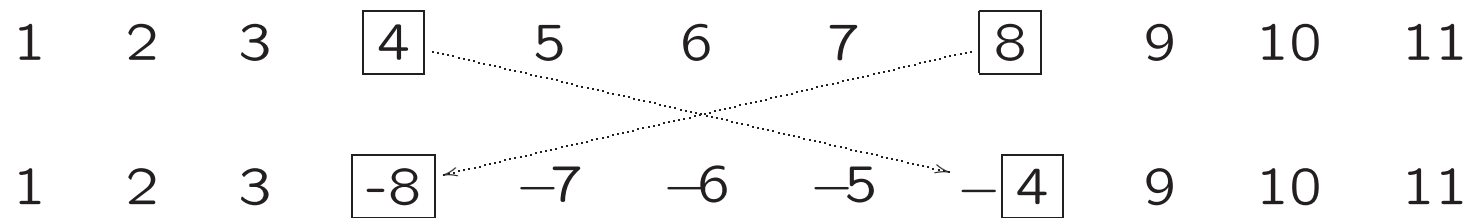
- Genomes are shuffled through large-scale mutations. (Inversions, translocations, fissions, fusions)
- Estimate the number of genome rearrangements that have occurred on the evolutionary path.

► Typically, this questions has been approached by **parsimony methods**: find a “greedy scenario” that transforms one genome into another using the least number of moves.



Comparative Human-Mouse Genome Map (Nature (2005))

► Focus on **Reversals**: (aka inversions) Large-scale mutation on a chromosome.



► If we use X chromosome or mitochondrial DNA for comparison, then reversals are the only large-scale mutations to be considered.

► For human-mouse X chromosome:

1 -7 6 -10 9 -8 2 -11 -3 5 4

► Hannehalli-Pevzner (95): Polynomial algorithm to compute parsimony distance for reversals. Uses the “breakpoint graph” and builds on older ideas of Watterson et al. (Ewens, Hall, Morgan)(1982), Nadeau-Taylor (1984), Bafna-Pevzner (1993, 1997).

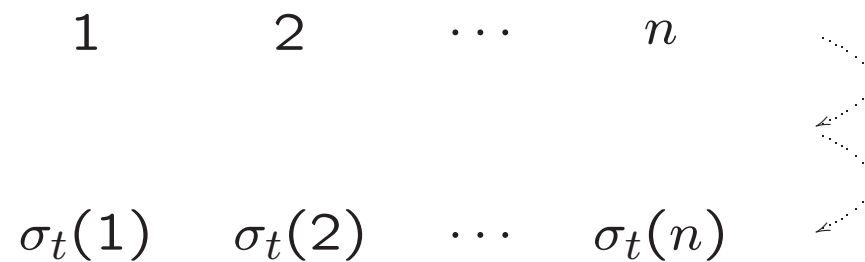
► Second data set from Ranz, Casals, and Ruiz (2001). They located 79 genes on chromosome 2 of *D. repleta* and on chromosome arm 3R of *D. melanogaster* and compared their order.

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	

In this case the distance is 54.

► **Question:** Given the large number of rearrangements, how reliable is the parsimony estimate?

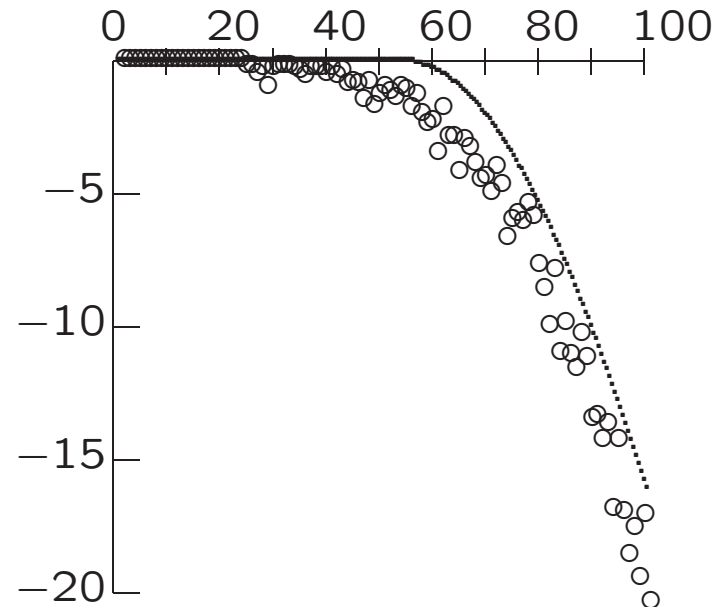
► Bourque-Pevzner (02): simulations. Start with the genes in order:



and perform **random reversals**.

For which range of t does $d(t) \approx t$?

- ▶ Simulations of Bourque - Pevzner (2002)
- ▶ Compared to Theorem 1 of B. - Durrett (2006)



- ▶ Concl: approx good as long as $t \leq 0.4n$.

Results B. and Durrett (2006)

- ▶ Simplify problem: instead perform **random transpositions**.

1 2 3 4 5 6 7 8 9 10 11



1 2 3 9 5 6 7 8 4 10 11

$\sigma(t)$ = random transposition random walk.

► Mathematical description of $\sigma(t)$:

– \mathcal{S}_n symmetric group on n labelled objects.

– Start σ at $\sigma(0) = I$ the identity.

– In cont. time (rate 1), multiply $\sigma(t-)$ by random uniform transposition.

► Another way to think about $\sigma(t)$.

– Turn \mathcal{S}_n into a Cayley graph G_n generated by $S = \{\text{transp.}\}$

– Then $\sigma(t) = \text{SRW}$ in cont. time on G_n .

► Let $D(t)$ = parsimony distance between $\sigma(t)$ and starting point.

$D(t)$ has a phase transition at $t = n/2$

Theorem B. and Durrett (2006)

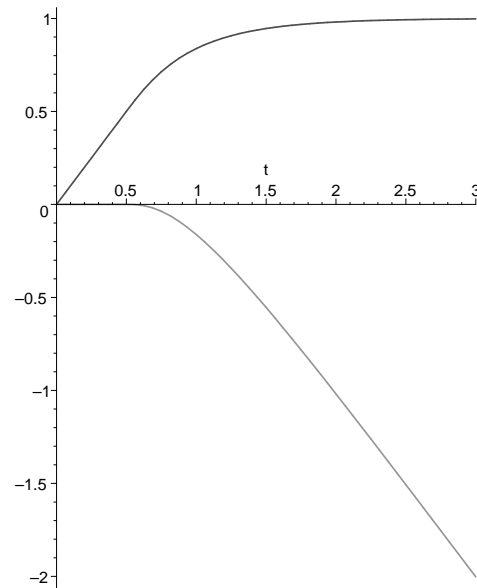
$$\frac{1}{n}D(tn) \rightarrow_p f(t)$$

where

$$f(t) = \begin{cases} t & \text{for } t \leq 1/2 \\ 1 - \sum_{k=1}^{\infty} \frac{1}{2t} \frac{k^{k-2}}{k!} (2te^{-2t})^k < t & \text{for } t > 1/2 \end{cases}$$

► Linear \longrightarrow Sublinear at $t = 1/2$.

► Non-smooth at $t = 1/2$ (no second derivative by Striling's formula)



► Using H-P algorithm, can be proved also for reversals.

- ▶ Proof is based on comparison with Erdős-Renyi random graphs.
- ▶ A crucial point is the existence of a formula for the distance: for any permutation σ ,

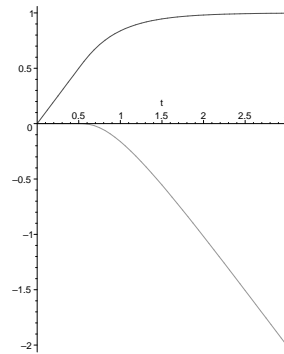
$$d_{\text{transp}}(\sigma) = n - \#\text{cycles}(\sigma)$$

Under random transpositions, cycles perform coalescence - fragmentation chain. The associated process of pure coalescence is precisely Erdős-Renyi random graphs.

- ▶ It is easy to count clusters in Erdős-Renyi random graphs. This gives the right answer for the number of components in coalescence-fragmentation.

- ▶ [More precisely, the transposition (i, j) yields a fragmentation if i and j are in the same cycle of the permutation, and a coagulation if i and j are in different cycles.]
- ▶ Hence the coalescence structure is obtained by putting an edge between i and j whenever they are transposed. This defines a coupling with a random graph process which has the same law as an Erdős-Renyi random graph process.
- ▶ Complexity of the components tells us roughly how much fragmentation there has been.

► Coming back to D. Repleta and D. Meganolester: $n = 79$ and data of Ruiz et al. gives $d = 54$, so we must find t such that $f(t) = 54/79 = 0.68$.

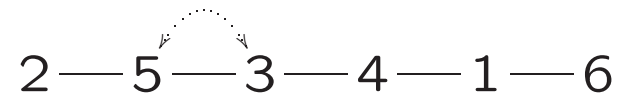


► Using this result, this means $t \approx 0.85$, i.e. roughly 65-70 rearrangements.

However:

- ▶ To derive this estimate we have implicitly assumed that all inversions are equally likely. However it seems intuitive that long-range inversions are less likely than short-range.
- ▶ According to Kent et al. (2001) median length of an inversion is less than 1Kb.
- ▶ What happens if we restrict to reversals where markers are no further than L units apart? (L -reversals, L -transpositions?)

- ▶ Simplest case: $L = 1$, i.e. random *adjacent* transpositions.



- ▶ Using only adjacent transpositions, σ has distance

$$d_{\text{adj}}(\sigma) = \#\{i < j : \sigma(i) > \sigma(j)\}$$

- ▶ In effect, many computer scientists study this model. (Eriksen (2004), Eriksson et al. (2000)).

- ▶ Eriksson et al. (2000), Eriksen (2004). After k steps

$$E(D_k^n) = \sum_{r=0}^k \frac{(-1)^r}{n^r} \left[\binom{k}{r+1} 2^r C_r + 4d_r \binom{k}{r} \right]$$

C_r = Catalan numbers, d_r “less famous” ... But hard to extract useful asymptotics.

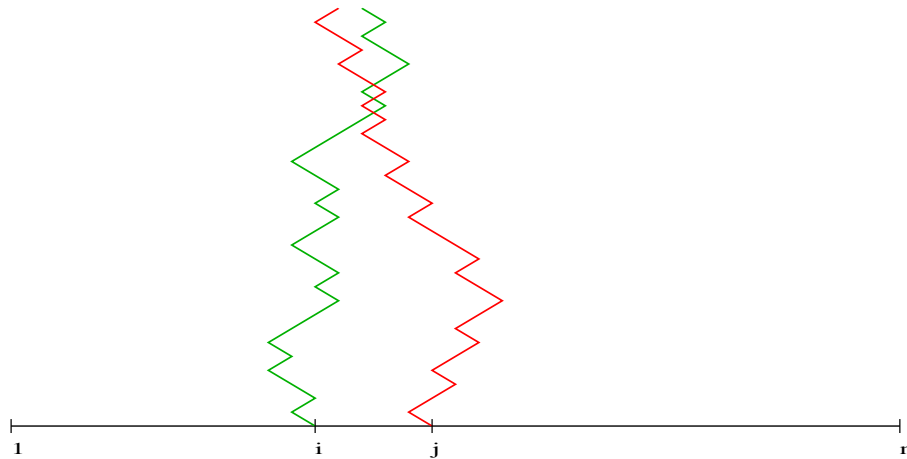
Some results

- ▶ After time t

$$E(D_t^n) = \sum_{i < j} P(\sigma_t(i) > \sigma_t(j))$$

- ▶ Because transpositions are adjacent for fixed i , $\sigma_t(i)$ is Simple Random Walk on $\{1, \dots, n\}$. (with refl. boundaries).

► In fact pairs of trajectories perform **simple exclusion** on $\{1, \dots, n\}$.



We obtain three different formulae for three different regimes.
Call D_t^n the distance at time t .

► 1. Small times. After $O(n)$ steps:

Theorem. B. and Durrett (2006) Let $t > 0$.

$$\frac{1}{n}E(D_{nt}^n) \rightarrow f(t)$$

as $n \rightarrow \infty$ for an explicit function $f(t)$.

► $f(t)$ smooth and behavior at infinity is **diffusive**

$$\lim_{t \rightarrow \infty} \frac{f(t)}{\sqrt{t}} = \sqrt{2/\pi}$$

► 2. Large times. After $O(n^3)$ steps:

Theorem. B. and Durrett (2006) Let $t > 0$. Then

$$\frac{1}{n^2} D_{n^3 t}^n \rightarrow_p P[B_1(2t) > B_2(2t)]$$

B_1, B_2 are two independent reflecting BM started uniformly on $0 < x < y < 1$.

- ▶ 3. Intermediate regime: After $n \ll t \ll n^3$ steps:

Theorem. B. and Durrett (2006) Let $t = t(n)$ with $t/n \rightarrow \infty$ and $t/n^3 \rightarrow 0$. Then

$$\frac{1}{\sqrt{nt}} D_t^n \xrightarrow{p} \sqrt{\frac{2}{\pi}}$$

- ▶ Behavior contrasts sharply with random transpositions.
- ▶ Different regimes, but **no cut-off times** to separate the regimes. No phase transition, but instead hidden transitions.

A digression for mathematicians.

► Following W. Ewens, what happens when we close the torus and the transposition $(1\ n)$ is allowed?

Conjecture Only large time behavior is different. In this case

$$\frac{1}{n^2} E(D_{n^3 t}^n) \rightarrow \frac{1}{4\pi} E(|R_t|)$$

the range of a Brownian motion on the unit circle.

► Heuristics. On a shortest path to the identity, no pair of particles may be switched more than once. One can still write:

$$D(\sigma) = \sum_{i < j} \mathbf{1}\{\text{particles } i \text{ and } j \text{ are switched along the path}\}$$

- ▶ Hence need to estimate $P(i \text{ and } j \text{ must be switched.})$.
- ▶ Claim: this is $\approx (1/2)P(\text{particles } i \text{ and } j \text{ have hit})$.
- ▶ After a little algebra and symmetry of Brownian motion, we end up with $E(|R_t|)/4\pi$.

The small-scale effect: local transpositions.

- ▶ Fix a number L , and do random transpositions of markers spaced uniformly on $\{1, \dots, L\}$.
- ▶ Potentially, $L = L(n)$. $L = 1$ is adjacent transpositions, $L = n$ is full transpositions.
- ▶ For which values of L is there a phase transition? hidden transitions? What time is phase transition? (how does it depend on L ?)

Randomized local transpositions.

- ▶ In a way, the answer depends on whether $L < \infty$ or $L \rightarrow \infty$.
- ▶ Assume L is a random variable. I.e., pick L according to a given distribution, with $p_i = P(L = i)$, and assume that

$$\forall 1 \leq i \leq n, \quad p_i > 0$$

- ▶ Since all transpositions are possible, the formula for the distance $d = n - \#\text{cycles}$ is still available!
- ▶ Case of random transpositions: $p_i = 1/n$. Adjacent transpositions: $p_1 = 1$, and $p_i = 0$ otherwise.

► Result: the behavior depends on $\max p_i$.

Theorem. If $\max p_i \rightarrow 0$ then

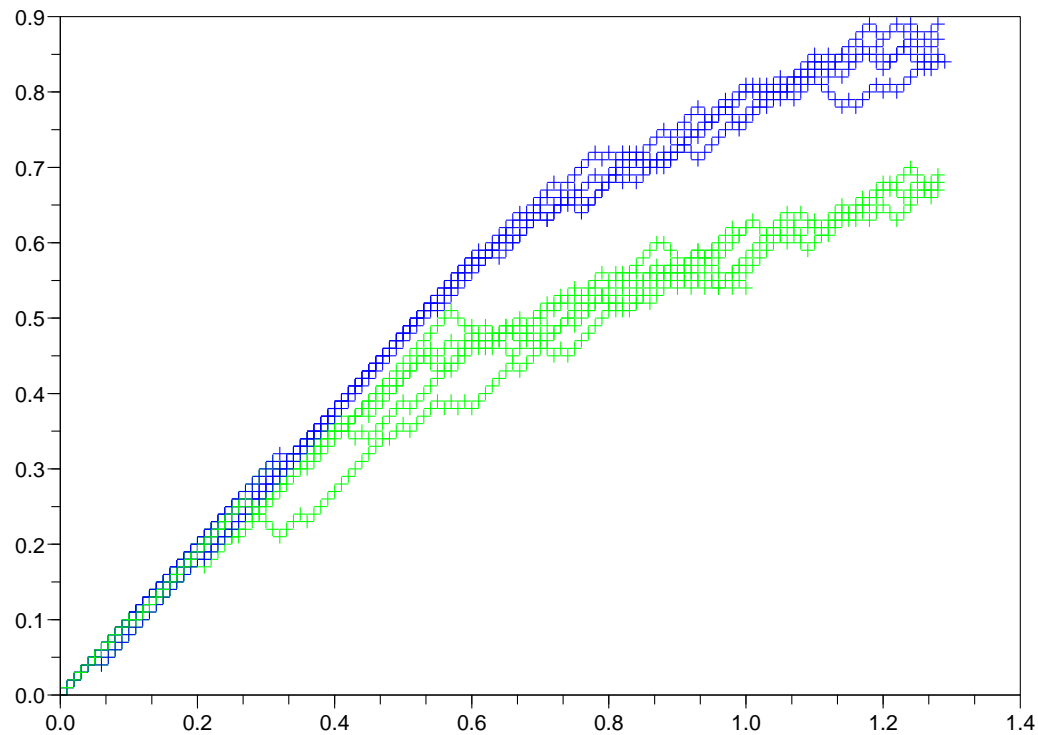
$$\frac{1}{n}E(D_{nt}) \rightarrow f(t)$$

the same function. In particular the critical time is always $n/2$.

Theorem/Conj. If $\liminf_{n \rightarrow \infty} \max p_i > 0$, then

$$\frac{1}{2}u(t) \leq \liminf_{n \rightarrow \infty} E\left(\frac{D_{tn}}{n}\right) \leq \limsup_{n \rightarrow \infty} E\left(\frac{D_{tn}}{n}\right) \leq u(t)$$

where u is a strictly concave function.



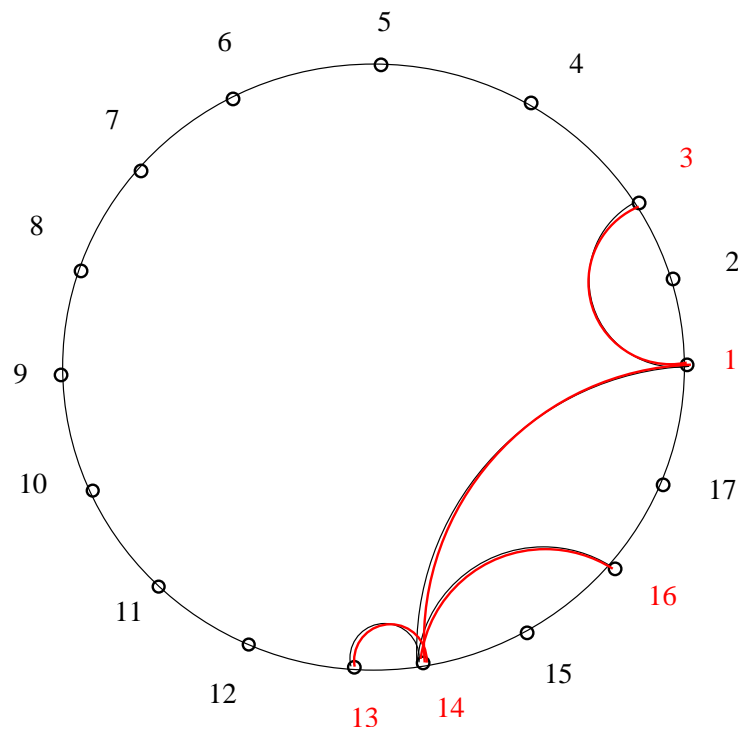
Simulations for $n = 100$. Blue=random transpositions. Green=local transpositions with $L = \text{Uniform on } \{1, \dots, 5\}$.

- ▶ Using these simulations we would conclude for *D. Repleta* and *D. Melanogaster* that ≈ 100 rearrangements were performed
- ▶ To be compared with distance of 54 and correction of 65 using full transpositions.
- ▶ (However 5 is arbitrary so this shouldn't be taken too seriously...)
- ▶ Still, locality clearly changes how we need to interpret the data!

Where do these results come from?

- ▶ We still use the coalescence-fragmentation approach.
- ▶ Linear behavior \leftrightarrow very little fragmentation.
Similarly, strictly convex behavior \leftrightarrow rate of frag. is > 0 and increasing.
- ▶ The coalescing structure can be understood using Branching Random Walks on the discrete torus.

► At time $\lambda n/2$, the component containing a given vertex converges to the range of a discrete BRW with offspring $\text{Poisson}(\lambda)$ and step distribution $(p_{\pm i}/2)$.



► To see how much fragmentation, we see if the BRW bumps onto itself.

► For a random walk on \mathbf{Z} ,

$$P(\text{return to 0 in } K \text{ steps}) \rightarrow 0 \iff \max p_i \rightarrow 0$$

► Case 1. When $\max p_i \rightarrow 0$ the range of BRW doesn't come back onto itself and this implies no fragmentation as long as the cluster stays finite.

► Combinatorially, $\#\text{cycles} = \sum_{i=1}^n \frac{1}{|\mathcal{C}_i|}$, so at a given time $\lambda n/2$

$$E(\#\text{cycles}) = nE(1/|\mathcal{C}_1|) \approx nE(1/Z_{BRW}(\lambda))$$

Indeed: when clusters are small the approximation is good, when they are large they don't contribute!

► For instance in case 1, we conclude $E(D_{\lambda n/2}) \approx n(1 - E(1/Z_\lambda))$ where Z_λ is the total progeny of a $PGW(\lambda)$. By known results (e.g. B. and Durrett) this is exactly $f(2\lambda)$.

► In case 2, when $\max p_i$ remains > 0 , the BRW bumps onto itself all the time. Each step is a fragmentation with probability at least $\delta > 0$ for some $\delta > 0$.

► So the rate of frag. is $> \delta$ but is also increasing in fact (harder to prove...)

► Get an upper-bound on the distance by looking at BRW estimates \rightarrow strictly convex function.

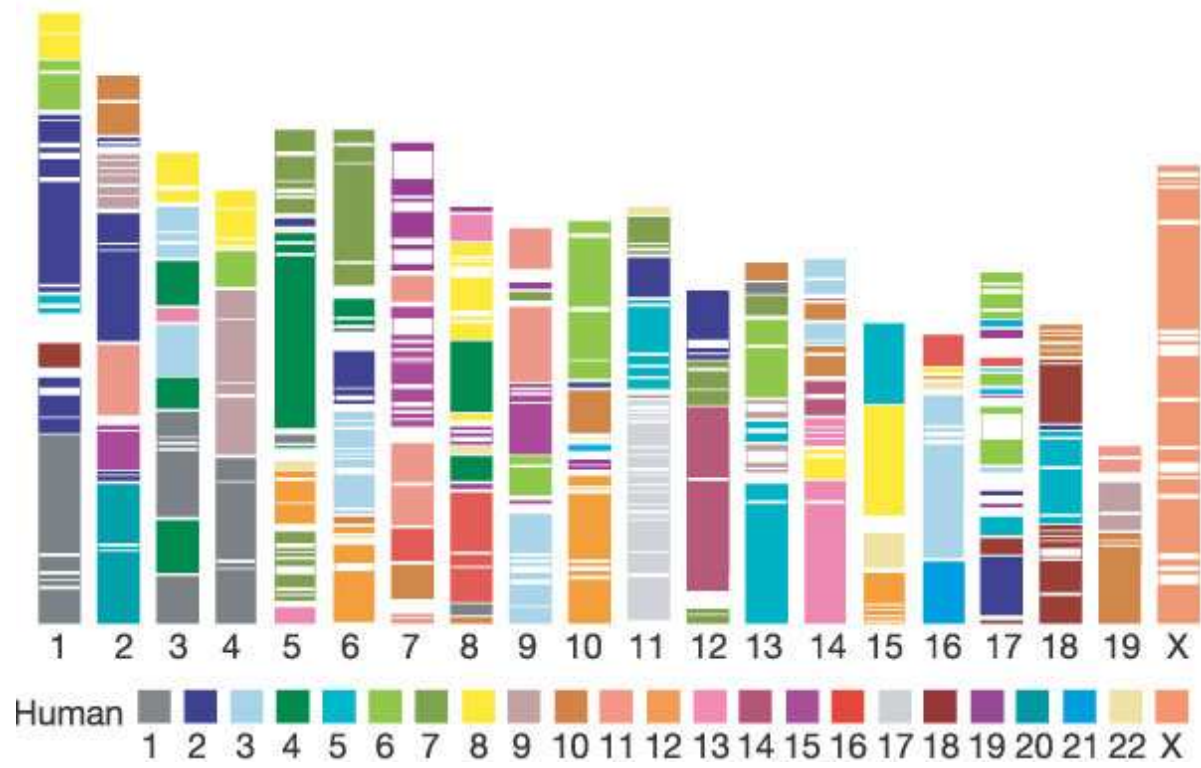
► One needs to look more carefully. By looking at reasonable p_i one should be able to prove more precise results. In particular, we **conjecture** that

$$\lim_{n \rightarrow \infty} \frac{D_{tn}}{n} = \rho(t)u(t)$$

for some $1/2 < \rho(t) < 1$.

Conclusions

1. Importance of estimates for the lengths of inversions. This will allow to make simulations and interpret data correctly.
2. Emphasize stochastic methods for estimating genome rearrangements when classical estimators become fuzzy.
3. In particular, methods based on random graphs and coalescence.
4. Much work is still needed to understand this picture:



Nature (2005)

... but even more work is needed for...

(data from Bourque Pevzner and Tesler (2004), see also Durrett and Interian)

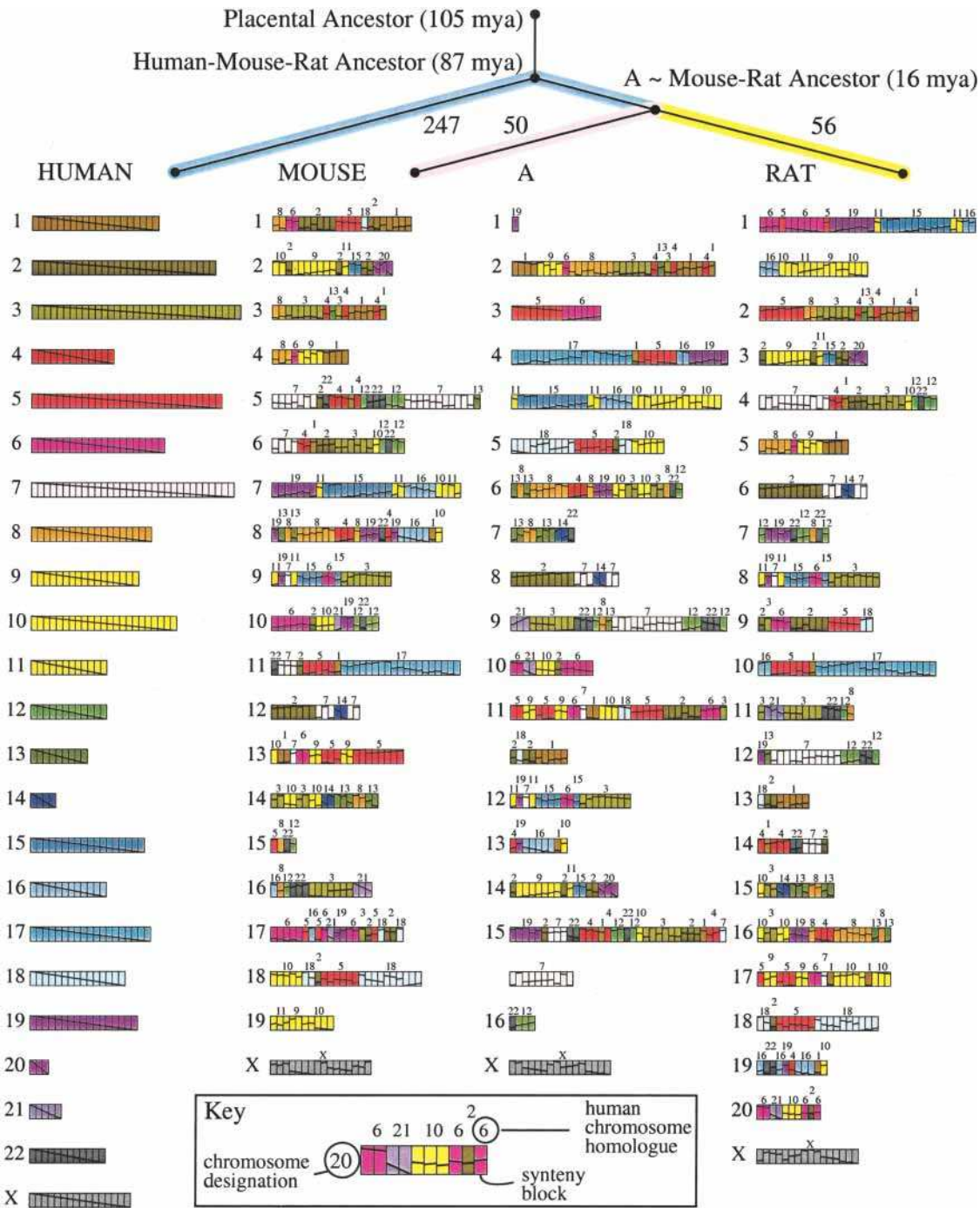


Figure 3 Ancestral murid rodent genome (A) and evolutionary tree computed by MGR, using mouse and rat with human as an outgroup. Each genome is represented as an arrangement of 391 syntenic blocks (longer than 300 kb) as computed by GRIMM-Syteny. The syntenic blocks are each represented as one unit, regardless of their length in nucleotides. Chromosomes with too many blocks are split into two lines. Each human chromosome is assigned a unique color, and a diagonal line is drawn through the whole chromosome. In other genomes, this diagonal line indicates the relative order and orientation of the rearranged blocks. The phylogram at the top of the figure indicates the number of rearrangements required to convert each genome (human, mouse, rat) into A, as computed by MGR. The estimated dates of divergence are from Springer et al. (2003).

chromosome 16] or in an ancestor. The reconstruction suggests that the ancestral murid rodent genome retained many previously postulated chromosome associations of the placental ancestor like 3/21, 4/8, 12/22, 16/19 (Murphy et al. 2003; Stanyon

et al. 2003). Because human is so distant from the putative murid rodent ancestor, features of this ancestor can also be studied by looking at how mouse chromosomes are conserved or perturbed in the rat genome and in the ancestor recovered. Some mouse

