Supporting Information for

Repeat sequences limit the effectiveness of lateral gene transfer and favoured the evolution of meiotic sex in early eukaryotes

Marco Colnaghi\textsuperscript{1,2}, Nick Lane\textsuperscript{1,2} and Andrew Pomiankowski\textsuperscript{1,2*}

Affiliations:

\textsuperscript{1}CoMPELEX and \textsuperscript{2}Department of Genetics, Evolution and Environment

University College London, Gower Street, London WC1E 6BT, UK

* Corresponding author: Andrew Pomiankowski (ucbhpom@ucl.ac.uk),

+44 20 76797697

This PDF file includes:

Supporting text

SI references

Figures S1 to S4
Supporting Information Text

Supplementary Materials and Methods

At the beginning of each simulation, every individual possesses a genome composed of $g$ unique protein-coding genes, each of which can exist in either a wildtype or deleterious mutant state. The genome is assumed to be circular (locus $g$ is contiguous with locus 1). The genome is interspersed at random intervals with a generic repeat sequence, at an initial density $\rho$ per protein-coding gene. The initial positions of the repeats are randomly sampled from a uniform distribution (i.e., all loci regions have the same probability of harbouring a repeat sequence at the beginning of the simulation). For simplicity, protein-coding genes and repeats are treated as unitary entities, and we neglect sequence variability within these genes.

The new generation is obtained by sampling $N$ individuals, with replacement, from the old population (Figure S1). The probability of reproduction is proportional to the individual fitness. Following previous theoretical studies (1-4), we assume no epistatic interactions and measure fitness as a multiplicative function

$$w_i(t) = (1 - s)^{g - n_i(t)},$$

(1)

where $n_i(t)$ is the number of different functional protein-coding genes possessed by an individual $i$ at time $t$ ($g - n_i(t)$ is the number of genes that have been lost, either because of mutations or deletions). To avoid unnecessary complexity, we assume that
gene duplication has a negligible effect on fitness, and only consider whether there is at least one functional copy of each protein-coding gene. Once a new generation is formed, the old generation dies, and its DNA forms the genetic pool from which the new generation acquires environmental DNA (eDNA) for recombination. We assume that eDNA strands are only stable for one generation before decaying irreversibly. Individuals of the new generation undergo LGT with a probability $\lambda$, which is taken to be a constant throughout each simulation (i.e. $\lambda$ is independent of genome size and content or the amount of available eDNA, which may change during a simulation run).

For each individual that undergoes LGT, a sequence of eDNA of length $L$ is randomly sampled from the eDNA pool. The requisite for successful recombination is homology between the terminal loci of the eDNA sequence and the host genome. Homology can be either to a protein-coding gene or a repeat sequence. These dynamics follow experimental evidence that recombination of nonhomologous DNA can take place fairly easily in the presence of homologous flanking sequences, but not in their absence (5, 6); integration of homologous DNA is estimated to be very considerably more likely than strongly divergent “foreign” DNA (5, 6).

After a sequence is sampled from the eDNA pool, one of its terminal loci is randomly selected and matched with a homologous locus $x^*$ in the host genome. If multiple homologous loci are present, one site is selected at random. The other terminal locus of the eDNA is then matched to a homologous locus $x_k$ in the host genome. If there is a
single homologous locus $x_k$ in the recipient genome, the recombination probability is
generated according to a Gaussian distribution,

$$p(d_k) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{d_k-L}{\sigma})^2}$$  \hspace{1cm} (2)

where $d_k$ is the distance between $x^*$ and $x_k$, and the standard deviation is $\sigma = \sqrt{L}$
(where $L$ is the length of the eDNA). Note that no recombination occurs with probability
$1 - p(d_k)$. This function penalizes the probability of exchange over large distances
(e.g., it is more likely that an eDNA sequence spanning 2 loci would recombine over a
length spanning 2 loci than 20 loci). This is in agreement with experimental data on
recombination length in bacteria, showing the rarity of large recombination events and
deletions (7, 8). It also reflects the assumption that longer eDNA sequences have greater
variance in the distribution of recombination probabilities. If there are multiple
homologous loci $x_k$ in the recipient genome, one of the $x_k$ loci is selected to be the
other terminal region of recombination given weights proportional to $p(d_k)$. This favors
loci with a small mismatch between $d_k$ and $L$. We then as before apply the
recombination probability $p(d_k)$ at this chosen site. If recombination occurs, all the
elements between $x^*$ and $x_k$ (included) are substituted by the recombining eDNA
sequence. If there is no match to either $x^*$ or $x_k$, recombination with the eDNA
sequence is not possible and no genetic exchange takes place.
To model fully homologous recombination, we select two homologous loci $x^*$ and $x_k$ as above. If the eDNA and genomic sequences contain exactly the same genes in the same order (either as wild-type or mutant alleles) recombination successfully takes place (we assume that point mutations do not significantly affect the probability of homologous recombination). The sequence between $x^*$ and $x_k$ is excised and replaced by the eDNA sequence. Otherwise, no genetic exchange takes place.

After LGT, each individual acquires $m$ new deleterious mutations, where $m$ is a random integer drawn from a Poisson distribution with mean $U$. The genome wide mutation rate is given by $U = \mu g'$, where $g'$ is the number of wildtype protein-coding genes, as we assume that mutated genes cannot mutate again. The position of the particular locus or loci in the genome that mutates is then randomly determined. For simplicity, the possibility of back mutation in protein-coding genes is neglected.
SI References

1. J. Haigh, The accumulation of deleterious genes in a population—Muller’s ratchet. *Theor Popul Biol* **14**, 251-267 (1978).

2. I. Gordo, B. Charlesworth, On the speed of Muller’s ratchet. *Genetics* **156**, 2137-2140 (2000).

3. N. Takeuchi, K. Kaneko, E. V. Koonin, Horizontal gene transfer can rescue prokaryotes from Muller’s ratchet: benefit of DNA from dead cells and population subdivision. *G3* **4**, 325-339 (2014).

4. M. Colnaghi, N. Lane, A. Pomiankowski, Genome expansion in early eukaryotes drove the transition from lateral gene transfer to meiotic sex. *Elife* **9**, e58873 (2020).

5. J. de Vries, W. Wackernagel, Integration of foreign DNA during natural transformation of Acinetobacter sp. by homology-facilitated illegitimate recombination. *Proc Natl Acad Sci U S A* **99**, 2094-2099 (2002).

6. N. Hülter, W. Wackernagel, Double illegitimate recombination events integrate DNA segments through two different mechanisms during natural transformation of Acinetobacter baylyi. *Mol Microbiol* **67**, 984-995 (2008).

7. N. J. Croucher, S. R. Harris, L. Barquist, J. Parkhill, S. D. Bentley, A high-resolution view of genome-wide pneumococcal transformation. *PLoS Pathog* **8**, e1002745 (2012).

8. J. C. Mell, J. Y. Lee, M. Firme, S. Sinha, R. J. Redfield, Extensive cotransformation of natural variation into chromosomes of naturally competent Haemophilus influenzae. *G3* **4**, 717-731 (2014).
Fig. S1: Illustration of the model dynamics.

(A) Each generation, an individual has a probability $\lambda$ of acquiring a fragment of eDNA of length $L$ from the environment and recombining it. Following LGT, mutations are randomly introduced at a rate $\mu$ per locus. The new generation is then generated by
sampling with replacement from the old generation in proportion to individual reproductive fitness \( w_m \). The old generation dies and its DNA is released into the environment and constitutes the eDNA pool for the new generation. (B) The wildtype genome is represented as a circular series of genes indicated by different colours. (C) The wildtype genome is subject to mutation pressure, resulting in the accumulation of deleterious alleles through Muller's ratchet. (D) In a repeat-free population (\( \rho = 0 \)), LGT (\( \lambda = 0.1 \)) allows homologous recombination, increasing genetic variation and favouring the elimination of deleterious mutations. (E) In the presence of repeats (\( \rho = 0.1 \)), the possibility of ectopic recombination limits the benefits of LGT, causing gene deletions and duplications. Other simulation parameters: \( t = 5,000, g = 100, N = 2,500, L = 5, U = 0.003 \).
Fig. S2. Genome size and recombination length.

At low initial repeat density ($\rho = 0.01$), (A) increases in recombination length ($L$) limit mutation accumulation as genome size increase, (B) without increasing the rate of deletion. But in repeat-rich genomes ($\rho = 0.3$), (C) higher $L$ provides virtually no benefits, (D) while introducing a large number of new deletions. Error bars show the standard deviation over 100 independent simulations. Gene loss rate was calculated over $t_{max} = 5,000$ generations. Other parameters: $N = 2,500$, $\mu = 10^{-5}$ and $\lambda = 0.1$. 
Fig. S3. Time series of key model variables.

Illustration of the change in key variables during the course of a standard simulation ($t_{max} = 5,000$ generations) when there is a high repeat density ($\rho = 0.3$). Change in (A) mean genome size and (B) mean repeat numbers are slight. (C) The presence of repeats allows ectopic recombination, resulting in the mean duplication content (green line) and total gene loss (the number of wildtype genes that have been lost; yellow line) rising through time. In addition, the mean mutation load (calculated as the sum of all mutated genes, including duplicated ones; purple line) increases through time. All these processes occur with considerable variation and some reversals. Other parameters: $N = 2,500$, $g = 300$, $\mu = 10^{-4}$ and $\lambda = 0.1$. 
Fig. S4. Distribution of key model variables.

Illustration of the final distribution at the end of a standard simulation ($t_{\text{max}} = 5,000$ generations) using the same parameter values in Figure S3. Population frequency distribution of (A) genome size, (B) repeat content, (C) total gene loss, (D) mutation load and (E) duplication content. Other parameters: $N = 2,500$, $g = 300$, $\mu = 10^{-4}$, $\rho = 0.3$, and $\lambda = 0.1$. 