



This is the accepted manuscript made available via CHORUS. The article has been published as:

# Stochastic Predator-Prey Dynamics of Transposons in the Human Genome

Chi Xue and Nigel Goldenfeld

Phys. Rev. Lett. **117**, 208101 — Published 10 November 2016

DOI: [10.1103/PhysRevLett.117.208101](https://doi.org/10.1103/PhysRevLett.117.208101)

# Stochastic predator-prey dynamics of transposons in the human genome

Chi Xue and Nigel Goldenfeld

*Department of Physics, Center for the Physics of Living Cells and  
Institute for Universal Biology, University of Illinois at Urbana-Champaign  
Loomis Laboratory of Physics, 1110 West Green Street, Urbana, Illinois 61801-3080, USA and  
Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign  
1206 West Gregory Drive, Urbana, Illinois 61801, USA*

(Dated: October 6, 2016)

Transposable elements, or transposons, are DNA sequences that can jump from site to site in the genome during the life cycle of a cell, usually encoding the very enzymes which perform their excision. However, some transposons are parasitic, relying on the enzymes produced by the regular transposons. In this case, we show that a stochastic model, which takes into account the small copy numbers of the transposons in a cell, predicts noise-induced predator-prey oscillations with a characteristic time scale that is much longer than the cell replication time, indicating that the state of the predator-prey oscillator is stored in the genome and transmitted to successive generations. Our work demonstrates the important role of number fluctuations in the expression of mobile genetic elements, and shows explicitly how ecological concepts can be applied to the dynamics and fluctuations of living genomes.

PACS numbers: 87.23.Cc, 87.10.Mn, 87.23.Kg

Transposable elements (TE) [1, 2] or transposons are DNA sequences that can migrate from site to site in a host genome. These mobile genetic elements are found in all three domains of life, but especially compose a significant fraction of eukaryotic genomes, for example, occupying 45% of the human genomic sequence [3]. Transposons are regarded as a major driver of adaptation and evolution [4], since they can induce both beneficial and deleterious transformations in the host genome, by inserting into encoding or regulation sequences, or causing misaligned pairing and unequal crossovers of chromosomes. In most cases, the modifications are disadvantageous to the host, for example causing hemophilia A in humans [5]. The activity of TEs has historically been observed through detailed population level assays, but recent measurements have demonstrated their activity in real time in living cells, using sophisticated fluorescence techniques [6], quantifying in detail how the stochastic processes of excision are not purely random, but reflect a cell's environment and genetic history. The interplay between transposon dynamics and replication, the cell's genotype and phenotype and the interactions with the environment are all reminiscent of population dynamics of organisms within an ecosystem, and this perspective is one that we explore and quantify here.

The dynamics of TEs are complex, but can be conveniently separated into two types of edit operation on the host genome: copy-and-paste, and cut-and-paste [7]. DNA transposons cut themselves out of the original site on the genome and later reintegrate at another site, thus performing a “cut-and-paste” operation which leaves the genome size invariant. Retrotransposons first transcribe into mRNA intermediates and then retrotranscribe to a new site on the genome sequence. This “copy-and-paste” dynamics leads to the growth of the genome

size. Some transposons (autonomous) encode the very enzymes which perform their excision, while others are parasitic (non-autonomous), relying on the enzymes produced by the regular TEs.

Several theoretical approaches have been proposed to study the dynamics of transposons. Population genetics models [8–13] were first developed to describe the equilibrium distribution of transposons in a population. Recent development views the genome as an ecosystem, with genetic elements of different types playing the role of individuals from different species [14–19]. In the case of non-autonomous transposons, a mean-field predator-prey type model describes their parasitic relationship with an autonomous transposon [16, 17]. However, these models do not account for the molecular level interactions between transposable elements and the dynamic behavior turns out to be sensitively dependent on these details. Furthermore, in a cell the copy number fluctuations are large, since the number of active (expressed) transposons is usually of order ten to a hundred [20]. Thus, the next generation of transposon models needs to take into account molecular details and stochasticity.

The purpose of this Letter is to develop a minimal individual-level model based on the specific interaction mechanism between a pair of autonomous-non-autonomous transposons. We begin with a model of the interactions between the TEs, and then use techniques from statistical mechanics to derive stochastic differential equations [21]. Our model predicts that number fluctuations generate persistent, noisy oscillations in the populations of the TEs, with a characteristic time scale that is much longer than the cell replication time, indicating that the state of the predator-prey oscillator is stored in the genome and transmitted to successive generations. Our work builds upon recent results that have shown how

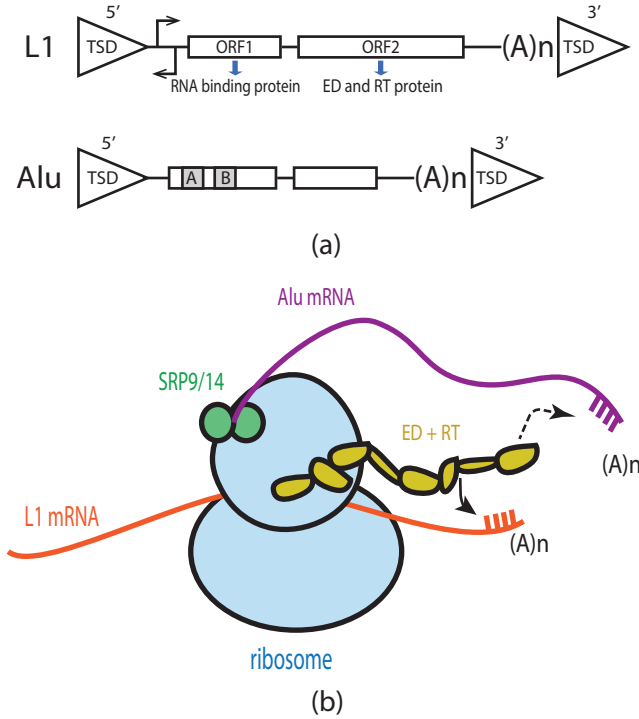


FIG. 1. (Color online.) (a) shows the structure of L1 and Alu elements. L1 has a pol II promoter (the right-pointing arrow) and an antisense promoter (the left-pointing arrow) followed by two open reading frames (ORFs) encoding a RNA binding protein and a protein that consists of an endonuclease (ED) and a reverse transcriptase (RT), respectively. Alu is composed of two non-coding monomers, with the left one bearing A and B boxes (the shaded area in the figure) as the pol III promoter. L1 and Alu elements share similar poly-A tails and are both flanked by target site duplicates (TSDs). (b) shows the *cis* and *trans* effects of L1 elements. When an ED+RT protein is translated at the ribosome, it *cis*-preferentially binds with the L1 mRNA that codes it, indicated by the solid arrow. An Alu mRNA can bind with two signal recognition proteins SRP9 and SPR14, and then attach to the ribosome. The nascent ED+RT protein then can *trans*-bind to the Alu mRNA, which has a similar poly-A tail (indicated by the dashed arrow), presumably with a similar probability to binding to L1 mRNA.

demographic stochasticity in ecosystems, where population size is integer-valued and locally finite, can lead to minimal models of persistent population cycles [21] or spatial patterns [22–26] without extra assumptions about the details of predation.

*Detailed model for transposon dynamics:-* Retrotransposons consist of two subgroups: LTR-transposons that have a long terminal repeat (LTR) structure, and non-LTR transposons that do not [7]. There are two types of non-LTR elements that show especially interesting interaction: the autonomous long interspersed nuclear elements (LINEs), and the non-autonomous short interspersed nuclear elements (SINEs) [27]. In the human

genome, the only active LINEs are LINE1 (L1) elements, which take up 17% of the entire genome [3]. They help SINEs, such as Alu elements, to transpose by providing critical enzymes used in the copy-and-paste dynamics [28]. We take L1 and Alu elements in the human genome as an example of a LINE-SINE pair and build a model of their interaction.

When a protein is produced at a ribosome coded by an L1 mRNA, it tends to bind with that particular mRNA, presumably by recognizing its polyadenine (poly-A) tail [29], and later retrotranscribes it into the genome. This is known as the *cis*-preference of L1 elements [30]. However, if an Alu mRNA attaches to the same ribosome, then it can bind with the nascent protein by faking the L1 mRNA poly-A tail [31]. In this way, Alu elements steals the transposition machinery designed by L1 elements [32, 33]. This is known as the *trans*-effect of L1 elements [30]. The mechanism is sketched in Fig. 1(b).

*Minimal model for transposon dynamics:-* Based on the above detailed interaction, an individual level minimal model can be made, discarding all details about how proteins are made and how complexes are formed. The individual reactions are shown in Eq. (1), where  $L$  stands for an active LINE,  $S$  for an active SINE, and  $R_L$  for the complex of the ribosome, LINE mRNA and nascent protein. Deactivated transposons do not participate in the transposition events and thus are excluded from the model.

An  $L$  element encodes the complex  $R_L$  at the rate  $b_R$ . The complex  $R_L$  retro-transposes to produce a new  $L$  element at the rate  $b_L$ , if there is no interruption.  $S$  element hijacks the complex  $R_L$  to duplicate itself at the rate  $b_S/V$ , where  $V$  is the system size. The complex  $R_L$  decays at the rate  $d_R$ .  $L$  and  $S$  elements are deactivated, at the rates  $d_L$  and  $d_S$ , respectively.  $\emptyset$  stands for null. The reactions for this minimal model, with the corresponding forward rates are as follows:



We assume the system is well mixed because mixing of reactants is faster, happening constantly within the cell lifetime, than the reactions.

We first use the Gillespie algorithm [34] to simulate the above reactions. Copy number vs time curves are plotted in Fig. 2. As shown in the figure, LINE and SINE copy numbers fluctuate around constant values, in the form of quasi-cycles. The circular envelope of the trajectory on the  $L$ - $S$  plane indicates a phase difference of roughly  $\pi/2$ , with SINE lagging LINE, supporting the identification of

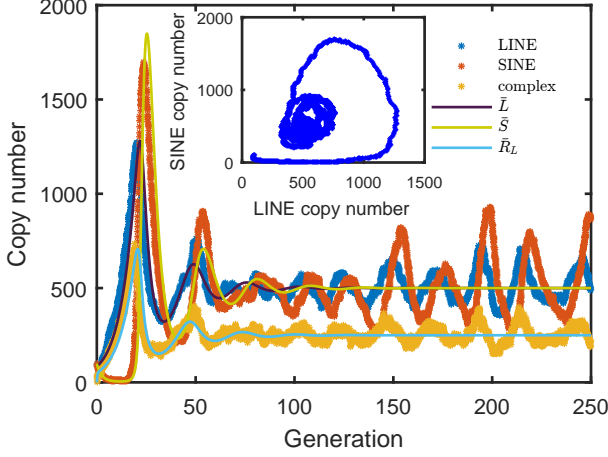


FIG. 2. (Color online.) Results of a typical stochastic simulation with illustrative parameters  $b_R = 2$ ,  $b_L = 1$ ,  $b_S = 1$ ,  $d_R = 2$ ,  $d_L = 0.5$ ,  $d_S = 0.5$ , and the system size  $V = 500$ . The main figure shows the copy numbers of active LINEs, SINEs and ribosome/L-mRNA/protein complexes as a function of time, in the unit of a cell generation. Copy numbers fluctuate around constant values, demonstrating quasi-cycles with period  $\sim 25$  generations. Solid lines are obtained by evolving the deterministic equations and show oscillatory decay toward steady values. Demographic noise induces quasi-cycles by constantly stimulating the deterministic oscillation mode. The inset shows the trajectory on the  $L$ - $S$  plane. The circular envelope indicates a phase difference of roughly  $\pi/2$ .

SINEs as predators on the LINEs.

*System size expansion:-* Let the copy number concentrations of active LINEs, SINEs and complexes be  $L$ ,  $S$  and  $R_L$  respectively. Then, with the system size being  $V$ ,  $VL$ ,  $VS$  and  $VR_L$  are equal to the actual copy numbers of the corresponding groups. The master equation about the probability  $\mathcal{P}(L, S, R_L)$  of the system being in the state  $(L, S, R_L)$  is written down as follows,

$$\begin{aligned} & \frac{d}{dt} \mathcal{P}(L, S, R_L) \\ = & V \left\{ (\mathcal{E}_{R_L}^- - 1) b_R L + (\mathcal{E}_{R_L}^+ \mathcal{E}_L^- - 1) b_L R_L \right. \\ & + (\mathcal{E}_{R_L}^+ \mathcal{E}_S^- - 1) b_S R_L S + (\mathcal{E}_{R_L}^+ - 1) d_R R_L \\ & \left. + (\mathcal{E}_L^+ - 1) d_L L + (\mathcal{E}_S^+ - 1) d_S S \right\} \mathcal{P}, \end{aligned} \quad (2)$$

with the raising and lowering operators given by

$$\mathcal{E}_X^\pm f(X) \equiv f\left(\frac{N_X \pm 1}{V}\right) \approx f(X) \pm \frac{1}{V} \partial_X f + \frac{1}{2V^2} \partial_X^2 f, \quad (3)$$

where  $f$  is an arbitrary function of the concentration  $X$ , and  $X$  stands for  $L$ ,  $S$  or  $R_L$ .

Substituting the expansions of operators into the master Eq. (2), and saving terms up to order  $O(V^{-1})$ , we obtain a non-linear Fokker-Planck equation. The corresponding Langevin equations about concentrations  $L$ ,  $S$  and  $R_L$  are nonlinear with multiplicative noises.

To obtain a set of linearized Langevin equations for concentration fluctuations, we perform the van Kampen's system size expansion, separating concentrations into deterministic part,  $\bar{L}$ ,  $\bar{S}$  and  $\bar{R}_L$ , and stochastic part,  $\xi$ ,  $\eta$  and  $\theta$ , as follows.

$$L = \bar{L} + \frac{\xi}{\sqrt{V}}, \quad S = \bar{S} + \frac{\eta}{\sqrt{V}}, \quad R_L = \bar{R}_L + \frac{\theta}{\sqrt{V}}. \quad (4)$$

Writing

$$\Pi(\xi, \eta, \theta) \equiv \mathcal{P}(L, S, R_L), \quad (5)$$

we find that

$$\frac{d}{dt} \mathcal{P} = \partial_t \Pi - \sqrt{V} \frac{d\bar{L}}{dt} \partial_\xi \Pi - \sqrt{V} \frac{d\bar{S}}{dt} \partial_\eta \Pi - \sqrt{V} \frac{d\bar{R}_L}{dt} \partial_\theta \Pi. \quad (6)$$

Substituting the system size expansion expressions Eq. (4) into the nonlinear Fokker-Planck equation and matching orders of  $V$ , we obtain to  $O(\sqrt{V})$

$$\frac{d\bar{L}}{dt} = b_L \bar{R}_L - d_L \bar{L}, \quad (7a)$$

$$\frac{d\bar{S}}{dt} = b_S \bar{S} \bar{R}_L - d_S \bar{S}, \quad (7b)$$

$$\frac{d\bar{R}_L}{dt} = b_R \bar{L} - b_L \bar{R}_L - b_S \bar{S} \bar{R}_L - d_R \bar{R}_L. \quad (7c)$$

These are the deterministic, or mean field, equations. The coexistence steady state, where  $\bar{L}$ ,  $\bar{S}$  and  $\bar{R}_L$  are all non-zero, is always exponentially stable, according to linear stability analysis. We have verified numerically that the imaginary part of the linear stability matrix eigenvalues provides a reasonable estimate for the angular frequency of quasi-cycles. Specifically, for the parameters used to generate Fig. 2 and Fig. 3, the eigenvalue imaginary part is equal to 0.2330, and agrees well with the Gillespie simulation value for the peak angular frequency, 0.23 generation $^{-1}$ , of the quasi-cycle power spectra shown in Fig. 3.

By matching  $O(1)$  terms, we obtain the linearized Langevin equations for  $\xi$ ,  $\eta$  and  $\theta$  using Ito's Lemma [35]:

$$\frac{d\xi}{dt} = b_L \theta - d_L \xi + r(t), \quad (8a)$$

$$\frac{d\eta}{dt} = b_S \bar{R}_L \eta + b_S \bar{S} \theta - d_S \eta + s(t), \quad (8b)$$

$$\frac{d\theta}{dt} = b_R \xi - b_L \theta - b_S \bar{R}_L \eta - b_S \bar{S} \theta - d_R \theta + h(t). \quad (8c)$$

$r(t)$ ,  $s(t)$  and  $h(t)$  are noises in  $\xi$ ,  $\eta$  and  $\theta$ , respectively.

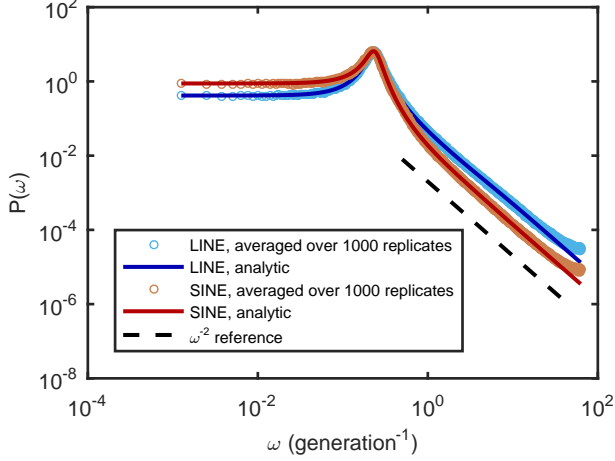


FIG. 3. (Color online.) Power spectra of the LINE and SINE concentration fluctuations. Circles stand for the power spectra obtained by averaging over 1000 replicates. Solid lines stand for the calculated spectra. The dash line is a reference function  $\sim \omega^{-2}$ . Parameters are  $b_R = 2$ ,  $b_L = 1$ ,  $b_S = 1$ ,  $d_R = 2$ ,  $d_L = 0.5$ ,  $d_S = 0.5$ ,  $V = 500$ . The peak angular frequency is equal to  $0.23 \text{ generation}^{-1}$ , corresponding to a period of 27 generations. The straight tail, in log-log scale, has a slope of  $-2$ , indicating a  $\omega^{-2}$  asymptotic behavior.

The correlations between these noises are given by

$$\langle h(t)h(t') \rangle = \delta(t - t')(b_R \bar{L} + b_L \bar{R}_L + b_S \bar{S} \bar{R}_L + d_R \bar{R}_L), \quad (9a)$$

$$\langle r(t)r(t') \rangle = \delta(t - t')(b_L \bar{R}_L + d_L \bar{L}), \quad (9b)$$

$$\langle s(t)s(t') \rangle = \delta(t - t')(b_S \bar{S} \bar{R}_L + d_S \bar{S}), \quad (9c)$$

$$\langle h(t)r(t') \rangle = \delta(t - t')(-b_L \bar{R}_L), \quad (9d)$$

$$\langle h(t)s(t') \rangle = \delta(t - t')(-b_S \bar{S} \bar{R}_L), \quad (9e)$$

$$\langle r(t)s(t') \rangle = 0. \quad (9f)$$

These Langevin equations describe the fluctuations of concentrations around the steady state values.

*Persistent oscillations:-* The power spectra  $P_{\xi\xi}(\omega)$  and  $P_{\eta\eta}(\omega)$  can be calculated by manipulating the Fourier transform of Eq. (8) and the correlations Eq. (9). The result is a complicated fraction, of which the numerator is a fourth order polynomial of  $\omega$  and the denominator a sixth order polynomial of  $\omega$ . Asymptotically, the power spectra have a tail in the form of  $\omega^{-2}$ . Figure 3 shows a comparison between the power spectra obtained from simulation and the analytic calculation, which demonstrates a satisfactory agreement. This minimal model shows that the negative feedback of SINEs on LINE transposition rate results in a predator-prey like dynamics [21], with noise induced quasi-cycles.

*Estimation of parameters:-* For the human genome, transposition rates of L1 and Alu elements measured by the mutation accumulation method are of order 1 in  $O(10) \sim O(100)$  births [36–38]. The deactivation rates

have a lower limit set by the base pair point mutation rate, which is roughly  $10^{-8}$  per base pair per generation [39, 40]. These rates seem to be too slow to generate any experimentally detectable dynamical behaviors. However, this estimate only accounts for fixed mutations that are not lethal, and thus underestimates the actual mutation rates. In a recent experiment [6] on real-time transposition events in living bacterium cells, the actual transposition rate directly observed was  $10^3$  times higher than that obtained by the mutation accumulation method. Moreover, the point mutation rate can be raised by a factor of  $10^2$  by deactivating the base pair mismatch repair machinery [41]. Thus, for a single-cell experiment rather than a large population, the relevant estimate is:  $b_R = 2$ ,  $b_L = 1 \times 10^{-2}$ ,  $b_S = 1 \times 10^{-2}$ ,  $d_R = 1$ ,  $d_L = 1 \times 10^{-2}$ ,  $d_S = 1 \times 10^{-2}$ , with units being  $\text{generation}^{-1}$ . The resultant quasi-cycle period should be roughly  $1 \times 10^3$  generations. Such oscillations could potentially be observed by integration of the LINE/SINE elements into a host microbial cell, *E. coli* for example, and using novel reporter techniques [6, 42].

In conclusion, we have shown that the dynamics of transposons can fruitfully be analyzed using analogy to ecological models, equipped with tools from statistical physics. Our calculations predict the existence of potentially observable, persistent and noisy oscillations in the populations of active SINEs and LINES.

*Acknowledgments:-* We acknowledge helpful discussions with Oleg Simakov and Tom Kuhlman. This work was partially supported by the National Science Foundation through grant PHY-1430124 to the NSF Center for the Physics of Living Cells, and by the National Aeronautics and Space Administration Astrobiology Institute (NAI) under Cooperative Agreement No. NNA13AA91A issued through the Science Mission Directorate.

- 
- [1] B. McClintock, Proceedings of the National Academy of Sciences **36**, 344 (1950).
  - [2] B. McClintock, Genetics **38**, 579 (1953).
  - [3] International Human Genome Sequencing Consortium, Nature **409**, 860 (2001).
  - [4] H. H. Kazazian, Jr., Science **303**, 1626 (2004).
  - [5] B. A. Dombroski, S. L. Mathias, E. Nanthakumar, A. F. Scott, and H. H. Kazazian, Jr., Science **254**, 1805 (1991).
  - [6] N. H. Kim, G. Lee, N. A. Sherer, K. M. Martini, N. Goldenfeld, and T. E. Kuhlman, Proceedings of the National Academy of Sciences **113**, 7278 (2016).
  - [7] T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman, Nature Reviews Genetics **8**, 973 (2007).
  - [8] B. Charlesworth and D. Charlesworth, Genetical Research **42**, 1 (1983).
  - [9] B. Charlesworth, P. Sniegowski, and W. Stephan, Nature **371**, 215 (1994).

- [10] C. H. Langley, J. F. Brookfield, and N. Kaplan, *Genetics* **104**, 457 (1983).
- [11] J. F. Brookfield and R. M. Badge, *Genetica* **100**, 281 (1997).
- [12] A. Le Rouzic and P. Capy, *Genetics* **174**, 785 (2006).
- [13] A. Le Rouzic and G. Deceliere, *Genetical Research* **85**, 171 (2005).
- [14] J. F. Brookfield, *Nature Reviews Genetics* **6**, 128 (2005).
- [15] S. Venner, C. Feschotte, and C. Biémont, *Trends in Genetics* **25**, 317 (2009).
- [16] A. Le Rouzic, S. Dupas, and P. Capy, *Gene* **390**, 214 (2007).
- [17] A. Le Rouzic, T. S. Boutin, and P. Capy, *Proceedings of the National Academy of Sciences* **104**, 19375 (2007).
- [18] F. Serra, V. Becher, and H. Dopazo, *PLoS ONE* **8**, e63915 (2013).
- [19] S. Linquist, K. Cottenie, T. A. Elliott, B. Saylor, S. C. Kremer, and T. R. Gregory, *Molecular Ecology* **24**, 3232 (2015).
- [20] B. Brouha, J. Schustak, R. M. Badge, S. Lutz-Prigge, A. H. Farley, J. V. Moran, and H. H. Kazazian, Jr., *Proceedings of the National Academy of Sciences* **100**, 5280 (2003).
- [21] A. J. McKane and T. J. Newman, *Physical Review Letters* **94**, 218102 (2005).
- [22] T. Butler and N. Goldenfeld, *Physical Review E* **80**, 030902 (2009).
- [23] T. Biancalani, D. Fanelli, and F. Di Patti, *Physical Review E* **81**, 046215 (2010).
- [24] U. C. Täuber, *Journal of Physics A: Mathematical and Theoretical* **45**, 405002 (2012).
- [25] B. Houchmandzadeh, *Journal of biosciences* **39**, 249 (2014).
- [26] H. Fort, *Entropy* **15**, 5237 (2013).
- [27] M. F. Singer, *Cell* **28**, 433 (1982).
- [28] R. E. Mills, E. A. Bennett, R. C. Iskow, and S. E. Devine, *Trends in Genetics* **23**, 183 (2007).
- [29] A. J. Doucet, J. E. Wilusz, T. Miyoshi, Y. Liu, and J. V. Moran, *Molecular Cell* **60**, 728 (2015).
- [30] W. Wei, N. Gilbert, S. L. Ooi, J. F. Lawler, E. M. Ostertag, H. H. Kazazian, Jr., J. D. Boeke, and J. V. Moran, *Molecular and Cellular Biology* **21**, 1429 (2001).
- [31] J. D. Boeke, *Nature Genetics* **16**, 6 (1997).
- [32] A. M. Weiner, *Current Opinion in Cell Biology* **14**, 343 (2002).
- [33] M. Dewannieux, C. Esnault, and T. Heidmann, *Nature Genetics* **35**, 41 (2003).
- [34] D. T. Gillespie, *Journal of Computational Physics* **22**, 403 (1976).
- [35] K. Itô, *Proceedings of the Imperial Academy* **20**, 519 (1944).
- [36] N. A. Rosenberg, A. G. Tsolaki, and M. M. Tanaka, *Theoretical Population Biology* **63**, 347 (2003).
- [37] C. R. L. Huang, K. H. Burns, and J. D. Boeke, *Annual Review of Genetics* **46**, 651 (2012).
- [38] R. Cordaux, D. J. Hedges, S. W. Herke, and M. A. Batzer, *Gene* **373**, 134 (2006).
- [39] M. W. Nachman and S. L. Crowell, *Genetics* **156**, 297 (2000).
- [40] J. C. Roach, G. Glusman, A. F. A. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas, *Science* **328**, 636 (2010).
- [41] M. Elez, A. W. Murray, L.-J. Bi, X.-E. Zhang, I. Matic, and M. Radman, *Current Biology* **20**, 1432 (2010).
- [42] T. E. Kuhlman, Private communication.