HAL
open science

Zombi: A phylogenetic simulator of trees, genomes and sequences that accounts for dead lineages

Adrian A Davin, Theo Tricou, Eric Tannier, Damien M. de Vienne, Gergely J Szöllosi

To cite this version:

Adrian A Davin, Theo Tricou, Eric Tannier, Damien M. de Vienne, Gergely J Szöllosi. Zombi: A phylogenetic simulator of trees, genomes and sequences that accounts for dead lineages. Bioinformatics, 2019, 10.1093/bioinformatics/btz710 . hal-02302010

HAL Id: hal-02302010

https://hal.science/hal-02302010v1

Submitted on 22 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bioinformatics: Application note

Zombi: A phylogenetic simulator of trees, genomes and sequences that accounts for dead lineages

Adrián A. Davín, [1, 2*], Théo Tricou[3], Eric Tannier[3,4], Damien M. de Vienne[3†] and Gergely J. Szöllősi[1,2,5†]

[1]MTA-ELTE Lendület Evolutionary Genomics Research Group, Budapest, Hungary

[2]Department of Biological Physics, Eötvös Loránd , Budapest, Hungary

[3]Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558, F-69622 Villeurbanne, France

[4]INRIA Grenoble Rhône-Alpes, F-38334, France

[5]Evolutionary Systems Research Group, Centre for Ecological Research, Hungarian Academy of Sciences, H-8237 Tihany, Hungary

*** Corresponding author:** Email aaredav@gmail.com

[†] **Equal contribution**

Abstract

Summary: Here we present **Zombi**, a tool to simulate the evolution of species, genomes and sequences *in silico*, that considers for the first time the evolution of genomes in extinct lineages. It also incorporates various features that have not to date been combined in a single simulator, such as the possibility of generating species trees with a pre-defined variation of speciation and extinction rates through time, simulating explicitly intergenic sequences of variable length and outputting gene tree - species tree reconciliations.

Availability and implementation: Source code and manual are freely available in https://github.com/AADavin/ZOMBI/

Contact: aaredav@gmail.com

Supplementary information: Supplementary materials can be found at Bionformatics online

1. Introduction

Reconstructing the pattern of horizontal gene transfers between species can help us date the origin of different taxa (Davín et al. 2018; Wolfe and Fournier 2018), understand the spread of genes of clinical importance (Lerminiaux and Cameron 2019) and resolve difficult phylogenetic questions (Abby et al. 2012). In the last decades, a large number of simulators have been developed to model a wide range of evolutionary scenarios (Dalquen et al. 2011; Mallo, De Oliveira Martins, and Posada 2016; Beiko and Charlebois 2007; Sjöstrand et al. 2013; Carvajal-Rodríguez 2008; Kundu and Bansal 2019) but none so far have considered the existence of extinct lineages and the horizontal transmission of genes (by lateral gene transfers) involving species that are not represented in the phylogeny (see: (Szöllősi et al.

2013; Fournier, Huang, and Gogarten 2009; Zhaxybayeva and Peter Gogarten 2004)). Zombi simulates explicitly the genome evolution taking place in these *extinct* lineages, which is expected to have an impact in *extant* lineages by means of Lateral Gene Transfers (Szöllősi et al. 2013). By not considering extinct lineages, other simulators make the implicit assumption that the transfer donor always leaves a surviving descendant among sampled species, while we know that this is most often not true (Szöllősi et al. 2013). Making this assumption may potentially hamper our ability to simulate realistic scenarios of evolution. In addition to considering evolution along extinct lineages, Zombi includes several features hitherto not found together in any other simulator (Table S1).

2. Basic features of Zombi

Zombi is a multilevel simulator, where a species tree is first simulated, then genomes evolve along the branches of this species tree, and finally, sequences are generated for each genome. These three steps, depicted in Figure 1 and detailed hereafter, are controlled by three main "modes", named **T**, **G**, and **S,** for species **T**ree, **G**enome and **S**equence, respectively.

The **T** mode simulates a species tree under the birth-death model (Kendall 1948), using the Gillespie algorithm with exponential waiting times (Gillespie 1977). The species tree generated is called the "Complete Species Tree" (CST) and it contains all lineages that have ever existed in the simulation. This tree is subsequently pruned to obtain the "Extant Species Tree" (EST), by removing all the lineages that did not survive until the end of the simulation (Figure 1A).

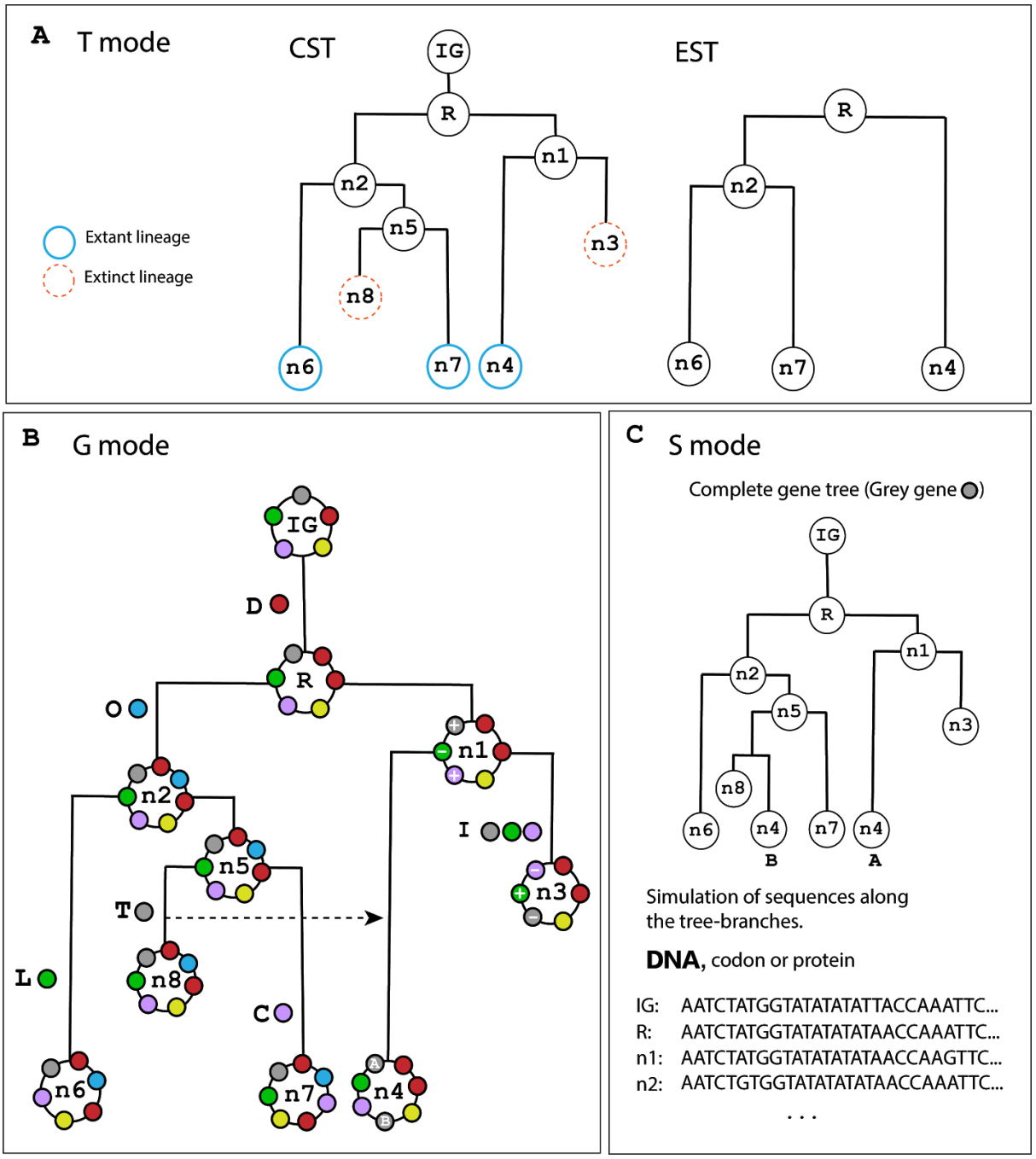

Fig 1. Overview of the three steps of the Zombi simulator. A: In T mode, Zombi simulates a species tree (the Complete Species Tree, CST) using a birth-death process and outputs the pruned version of it by removing extinct lineages (to obtain the Extant Species Tree, EST). In this example, lineages n3 and n8 go extinct before the simulation ends. **B**: in **G** mode, a circular genome evolves along the branches of the CST obtained with the **T** mode by

Duplications (D), Originations (O), Inversions (I), Translocations (C), Losses (L) and Transfers (T) of genes. Here, the initial genome (IG) is composed of five genes (small coloured circles), and various events affecting different genes and their impact on the genome structure are indicated along the branches. The inversion events not only modify the positions of the genes but also change their orientation. **C:** in **S** mode, Zombi can be used to simulate codon, nucleotides and amino acids along the branches of the gene family trees. Here, the gene tree of the grey coloured gene family from **B** has been depicted.

The **G** mode simulates the evolution of genomes along the branches of the CST (Figure 1B). Genomes are circular and contain a variable number of genes ordered along the chromosome. The simulation starts with a single genome, with an initial number of genes determined by the user. Each gene has an orientation (+ or -) that is determined randomly and represents the direction of the gene in the coding strand. The genomes evolve along the branches of the CST by undergoing six possible genome-level events: duplications, losses, inversions, translocations, transfers and originations. These events affect a variable number of contiguous genes, sampled from a geometric distribution with parameter p (specific to each type of event). At each speciation node of the CST, two identical copies of the genome are created and start evolving independently in the descending branches.

The **S** mode, finally, simulates gene sequences (at either the codon, nucleotide or protein level) along the gene family trees (Figure 1C). The user can modify the scaling of the tree to better control the number of substitutions that take place per unit of time, and thus simulate fast or slow-evolving genes.

3. Advanced features

In addition to the basic features presented above, "advanced" modes of Zombi (listed in Table S2) can be used to obtain richer and more realistic evolutionary scenarios. For example, it is possible to use a species tree input by the user, to generate species trees with variable extinction and speciation rates, or to control the number of living lineages at each unit of time (Figure S1). At the genome level, Zombi can simulate branch-specific rates for each possible evolutionary event. This allows the user to simulate very specific scenarios such as highways of transfers between a given pair of lineages or events of massive duplications in certain branches of the species tree. Zombi can also simulate genomes accounting for intergenic regions of variable length (drawn from a flat Dirichlet distribution (Biller et al. 2016). At the sequence level, finally, the user can fine-tune the substitution rates to make them branch specific.

Zombi provides the user with a clear and detailed output of the complete evolutionary process simulated, including a comprehensive log-file consisting of all the events taking place during the simulation, the reconciled gene trees with the species tree in the RecPhyloXML reconciliation standard (Duchemin et al. 2018) and the complete genomes at every node in the CST.

4. Performance and validation

Simulations with Zombi are fast: with a starting genome of 500 genes and a species tree of 2000 taxa (extinct + extant), it takes around 1 minute on a 3.4Ghz laptop to simulate all the genomes (Figure S2).

We validated that the distribution of waiting times between successive events was following an exponential distribution (Figure S3 and S4) and that the distribution of intergene sizes at

equilibrium was following a flat Dirichlet distribution, as expected from Biller et al. 2016 (Figure S5). We also checked by hand the validity of many simple scenarios to detect possible inconsistencies in the algorithm.

5. Implementation

Zombi is implemented in Python 3.6. It relies on the ETE 3 toolkit (Huerta-Cepas, Serra, and Bork 2016) and the Pyvolve package (Spielman and Wilke 2015). It is freely available at https://github.com/AADavin/ZOMBI along with detailed documentation in a wiki page.

Acknowledgments:

A.A.D. and G.J.Sz. received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under grant agreement no. 714774., in addition, G.J.Sz. was supported by the grant GINOP-2.3.2.-15-2016-00057. T.T and D.M.d.V received funding from grant ANR-18-CE02-0007-01 ("STHORIZ"). We thank Vincent Daubin, Wandrille Duchemin, Nicolas Lartillot and Thibault Latrille for insightful discussions during the preparation of this manuscript.

References

Abby, S. S., E. Tannier, M. Gouy, and V. Daubin. 2012. "Lateral Gene Transfer as a Support for the Tree of Life." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1116871109.

Beiko, Robert G., and Robert L. Charlebois. 2007. "A Simulation Test Bed for Hypotheses of Genome Evolution." *Bioinformatics* 23 (7): 825–31.

Biller, Priscila, Laurent Guéguen, Carole Knibbe, and Eric Tannier. 2016. "Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation." *Genome Biology and Evolution* 8 (5): 1427–39.

Carvajal-Rodríguez, Antonio. 2008. "Simulation of Genomes: A Review." *Current Genomics* 9 (3): 155–59.

Dalquen, Daniel A., Maria Anisimova, Gaston H. Gonnet, and Christophe Dessimoz. 2011.

"ALF—A Simulation Framework for Genome Evolution." *Molecular Biology and Evolution* 29 (4): 1115–23.
Davín, Adrián A., Eric Tannier, Tom A. Williams, Bastien Boussau, Vincent Daubin, and Gergely J. Szöllősi. 2018. "Gene Transfers Can Date the Tree of Life." *Nature Ecology & Evolution* 2 (5): 904–9.
Duchemin, Wandrille, Guillaume Gence, Anne-Muriel Arigon Chifolleau, Lars Arvestad, Mukul S. Bansal, Vincent Berry, Bastien Boussau, et al. 2018. "RecPhyloXML - a Format for Reconciled Gene Trees." *Bioinformatics* , May. https://doi.org/10.1093/bioinformatics/bty389.
Fournier, Gregory P., Jinling Huang, and J. Peter Gogarten. 2009. "Horizontal Gene Transfer from Extinct and Extant Lineages: Biological Innovation and the Coral of Life." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364 (1527): 2229–39.
Gillespie, Daniel T. 1977. "Exact Stochastic Simulation of Coupled Chemical Reactions." *The Journal of Physical Chemistry* 81 (25): 2340–61.
Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data." *Molecular Biology and Evolution* 33 (6): 1635–38.
Kendall, David G. 1948. "On the Generalized 'Birth-and-Death' Process." *Annals of Mathematical Statistics* 19 (1): 1–15.
Kundu, Soumya, and Mukul S. Bansal. 2019. "SaGePhy: An Improved Phylogenetic Simulation Framework for Gene and Subgene Evolution." *Bioinformatics* , February. https://doi.org/10.1093/bioinformatics/btz081.
Lerminiaux, Nicole A., and Andrew D. S. Cameron. 2019. "Horizontal Transfer of Antibiotic Resistance Genes in Clinical Environments." *Canadian Journal of Microbiology* 65 (1): 34–44.
Mallo, Diego, Leonardo De Oliveira Martins, and David Posada. 2016. "SimPhy: Phylogenomic Simulation of Gene, Locus, and Species Trees." *Systematic Biology* 65 (2): 334–44.
Sjöstrand, Joel, Lars Arvestad, Jens Lagergren, and Bengt Sennblad. 2013. "GenPhyloData: Realistic Simulation of Gene Family Evolution." *BMC Bioinformatics* 14 (June): 209.
Spielman, Stephanie J., and Claus O. Wilke. 2015. "Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies." *PloS One* 10 (9): e0139047.
Szöllősi, Gergely J., Eric Tannier, Nicolas Lartillot, and Vincent Daubin. 2013. "Lateral Gene Transfer from the Dead." *Systematic Biology* 62 (3): 386–97.
Wolfe, Joanna M., and Gregory P. Fournier. 2018. "Horizontal Gene Transfer Constrains the Timing of Methanogen Evolution." *Nature Ecology & Evolution* 2 (5): 897–903.
Zhaxybayeva, Olga, and J. Peter Gogarten. 2004. "Cladogenesis, Coalescence and the Evolution of the Three Domains of Life." *Trends in Genetics: TIG* 20 (4): 182–87.

Name	Species Tree Level	Geno-me Level	Sequen-ce Level	Extinct lineages	Sam-pling	Intergenic regions	Reconciled trees	Gene fusion-fission	ILS
Zombi	•	•	•	•	•	•	•		
ALF	•	•	•			•		•	
SimPhy	•		•				•		•
EvolSimulator	•		•			•			
GenPhyloData	•		•						
SaGePhy	•		•					•	

Table S1: Comparison of the features available in the main evolution simulators. Zombi (this paper), ALF (Dalquen et al. 2011), SimPhy (Mallo, De Oliveira Martins, and Posada 2016), EvolSimulator (Beiko and Charlebois 2007), GenPhyloData (Sjöstrand et al. 2013) and SaGePhy (Kundu and Bansal 2019). The features presented are whether the tool is capable of simulating species trees (Species Tree level), genomes (Genome level, meaning that it considers the structure of the genome, i.e. the physical adjacencies of genes in a genome), sequences (Sequences level), the presence of extinct lineages (Extinct lineages), the possibility of sampling species integrated in the simulator and pruning gene trees according to the species sampled (Sampling), the simulation of intergenic regions (Intergenic regions), outputting reconciled trees (Reconciled trees), considering fusion and fission of genes (Fusion-fission of genes) and producing ILS-induced gene tree/species tree discrepancy (ILS).

Mode	Description
Species Tree	
T	Basic mode
Tb	Branch-wise extinction/speciation rates
Tp	Lineage profiling (controls the number of extant lineages per unit of time)
Ti	Input tree by the user
Genomes	
G	Basic mode
Gu	Branch-wise event rates defined by the user
Gf	Simulate full genomes, including intergenic regions
Sequences	
S	Basic mode
Su	Branch-wise substitution rates defined by the user
Sf	Simulate sequences in combination with the Gu mode

Table S2. Zombi modes. Zombi implements a total of 10 different modes assigned to three main categories (Species Tree, Genome and Sequence). The basic mode of each category is explained in the main text of this paper.

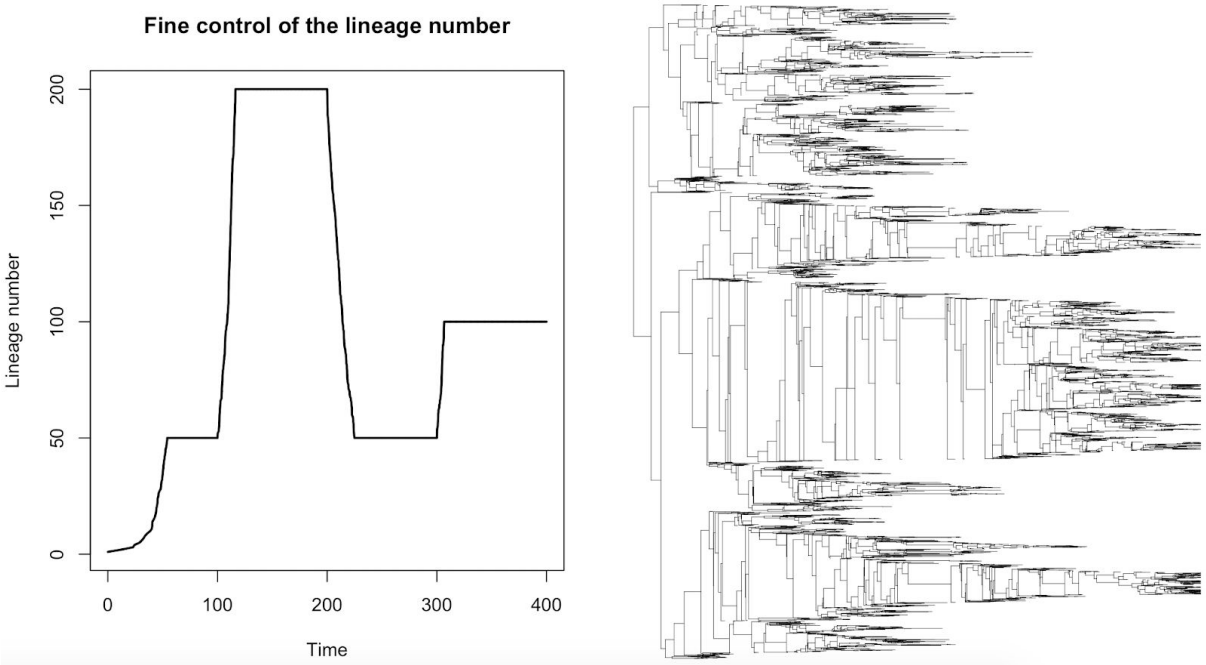

Figure S1. Fine control of the lineage number. Zombi can compute species tree using as input a list of times and the corresponding lineage number that should be attained by that time (in the example t 100 = 50; t 200 = 200; t 300 = 50; t 400 = 100). Zombi tries to attain the lineage number specified for each time interval using the speciation and extinction rates input by the user. At first, there is 1 living lineage and only speciations take place until the number of lineages = 50, number attained in this example when t~50. After that, and because time < 100, the number of lineages reaches an equilibrium in which there is a turnover of species controlled by a parameter also input by the user. Each time that a turnover event takes place two species are randomly sampled in the phylogeny. The first species undergoes a speciation and the second one dies, thus maintaining the total lineage number. The simulation continues until time = 400. In the right panel we can find the resulting species tree.

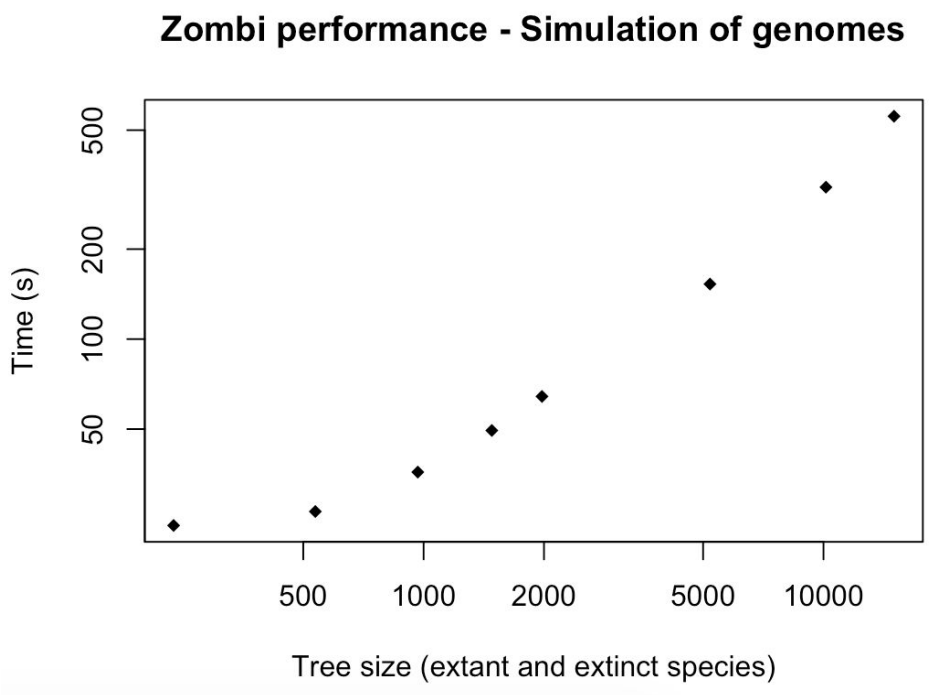

Figure S2. Computing time for different simulation in a computer with a 3,4 GHz Intel Core i5 processor. The rates used were Duplication rate: 0.2, Transfer rate: 0.2, Loss rate: 0.6, Origination rate:0.05, Inversion rate: 0.2, Translocation rate: 0.2. The initial genome was composed of 500 genes. All extension rates were set to 1. Species trees were obtained using by setting Speciation rate: 1 and Extinction rate: 0.5.

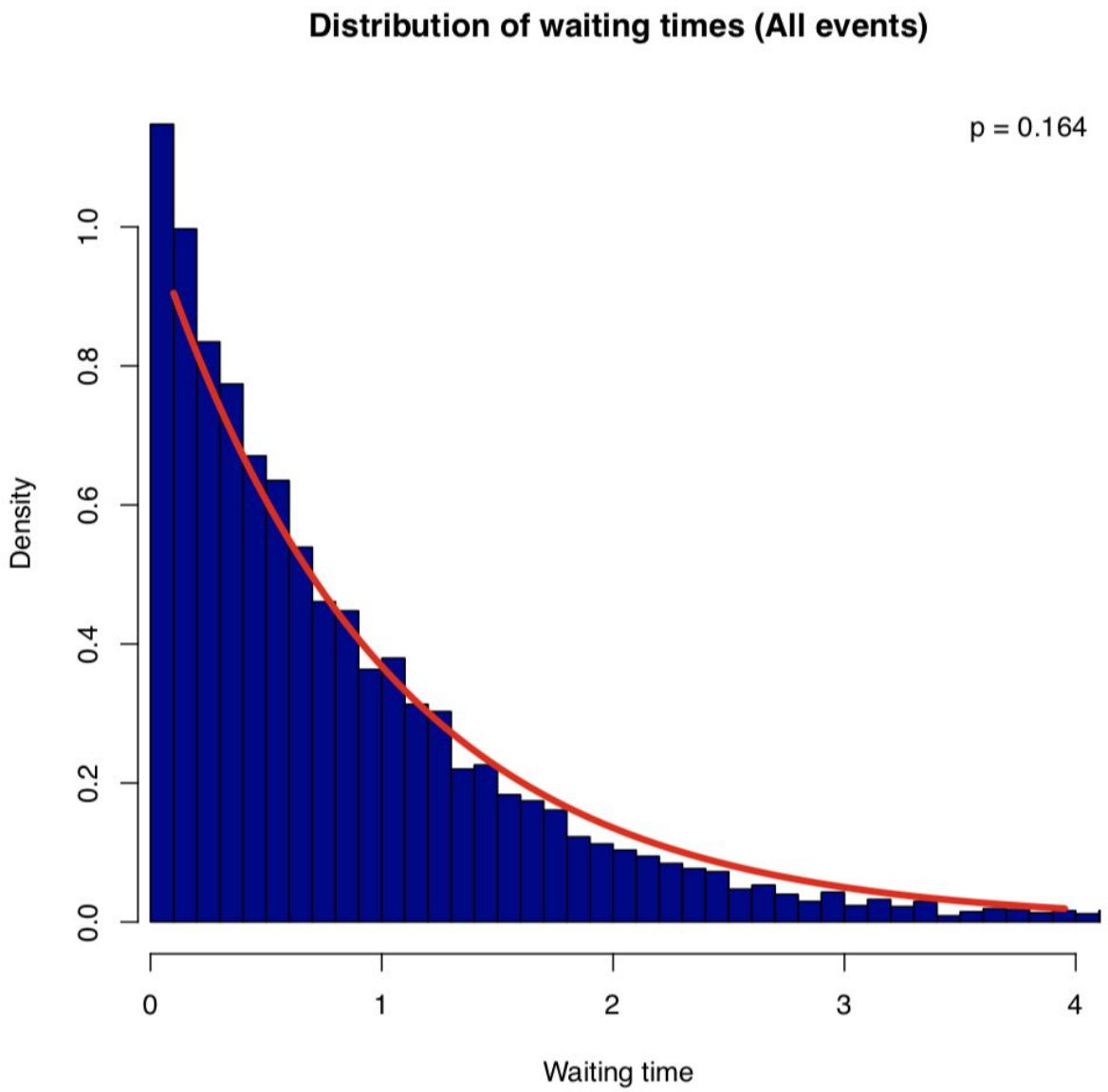

Figure S3. Comparison between the observed distribution of waiting times between consecutive events (duplications, transfers, losses, inversions, translocations and originations, blue bars) and the expected one (red line). The p-value corresponds to a KS test between the empirical distribution and an exponential distribution of the same rate than the same one used in the simulations.

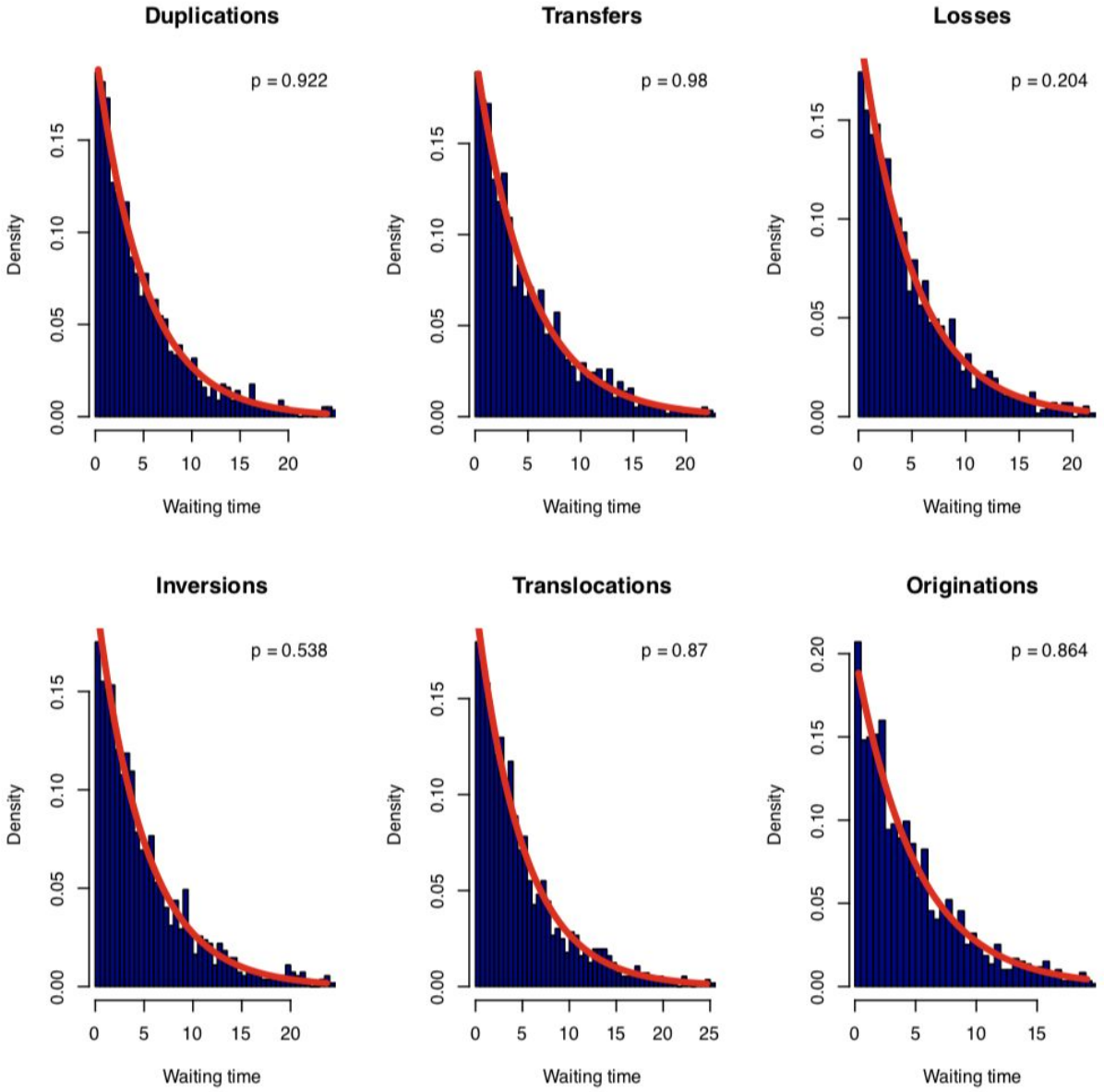

Figure S4. Comparison between the observed (blue bars) and the expected (red line) distribution of waiting times between consecutive events of each type. The p-value corresponds to a KS test between the empirical distribution and an exponential distribution of the same rate.

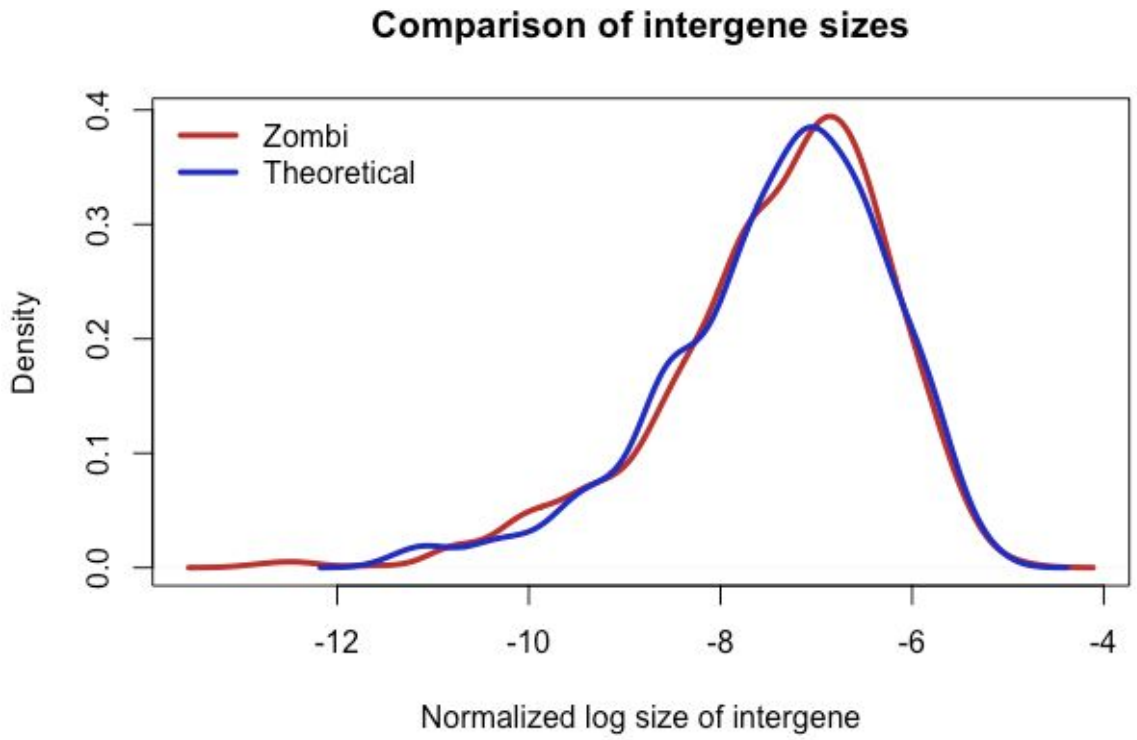

Figure S5: Validation of the mode Gf. To validate the mode **Gf** we simulated a genome with 1000 genes, whose intergene lengths had a constant size of 10000 nucleotides (instead of the Dirichlet as the default option). Then, we made it evolve under many inversion events (~10^6), to see whether at the equilibrium the intergene sizes followed a flat Dirichlet distribution, as expected (see Biller et al. 2016). We compared the obtained values with a randomly generated flat Dirichlet distribution using a KS test and obtained no significant difference (K-S test; p-value = 0.876)

References

Beiko, Robert G., and Robert L. Charlebois. 2007. "A Simulation Test Bed for Hypotheses of Genome Evolution." *Bioinformatics* 23 (7): 825–31.

Dalquen, Daniel A., Maria Anisimova, Gaston H. Gonnet, and Christophe Dessimoz. 2011. "ALF—A Simulation Framework for Genome Evolution." *Molecular Biology and Evolution* 29 (4): 1115–23.

Kundu, Soumya, and Mukul S. Bansal. 2019. "SaGePhy: An Improved Phylogenetic Simulation Framework for Gene and Subgene Evolution." *Bioinformatics* , February. https://doi.org/10.1093/bioinformatics/btz081.

Mallo, Diego, Leonardo De Oliveira Martins, and David Posada. 2016. "SimPhy: Phylogenomic Simulation of Gene, Locus, and Species Trees." *Systematic Biology* 65 (2): 334–44.

Sjöstrand, Joel, Lars Arvestad, Jens Lagergren, and Bengt Sennblad. 2013. "GenPhyloData: Realistic Simulation of Gene Family Evolution." *BMC Bioinformatics* 14 (June): 209.