Zombi: A simulator of species, genes and genomes that accounts for extinct lineages.

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

[1]Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558, F-69622 Villeurbanne, France
[2]MTA-ELTE Lendület Evolutionary Genomics Research Group, Budapest, Hungary
[3]Department of Biological Physics, Eötvös Loránd , Budapest, Hungary
[4]INRIA Grenoble Rhône-Alpes, F-38334, France

† Corresponding author: Email aaredav@gmail.com
*** Equal contribution**

Abstract

Most living organisms that ever existed on Earth have left no descendants. Because introgressions and lateral gene transfers are frequent, some of these extinct lineages have impacted the evolution of extant species and their ancestors. As a consequence, ignoring extinct lineages in evolutionary studies can lead to spurious conclusions. Here we present Zombi, a platform to simulate the evolution of species, genes and genomes taking extinct lineages into account. We demonstrate its utility by testing a statistical inference method used to detect introgression and show that ignoring the presence of extinct lineages yields inconsistent results.

Key words: Simulator, Lateral Gene Transfer, introgression, gene tree - species tree, extinct lineages

New Approaches

Zombi is a platform designed to simulate evolutionary processes at different levels ranging from the species tree to the genomes and its constituent genes and their sequences, that evolve along it . Zombi considers explicitly the evolution of genomes in extinct lineages and thus the possibility for lateral gene transfers and introgressions to occur between species that are not represented in the extant phylogeny. Zombi provides the user with a clear and detailed output of the complete evolutionary process simulated. Zombi is written in Python 3.6 using the ETE 3 toolkit (Huerta-Cepas, Serra, & Bork, 2016) and the Pyvolve package (Spielman & Wilke, 2015). It is freely available at https://github.com/AADavin/ZOMBI.

Introduction

Platforms to simulate evolution *in silico* are frequently used to test statistical inference methods, providing a controlled scenario in which complete information on the patterns and processes of evolution is available. In the last decades, a large number of simulators have been developed to model a wide range of complex evolutionary scenarios (Carvajal-Rodríguez, 2008, 2010; GSR, n.d.), but none so far have considered the existence of extinct lineages and the horizontal transmission of genes (lateral gene transfer or introgressions) involving species that are not represented in the phylogeny (Fournier, Huang, & Gogarten, 2009; Szöllősi, Tannier, Lartillot, & Daubin, 2013; Zhaxybayeva & Peter Gogarten, 2004). Here we present Zombi, a platform to simulate the evolution of genomes *in silico*, that considers that not all species in the simulation have survived until the present time and the horizontal transmission of genes among all lineages in the phylogeny, including extinct lineages. We demonstrate that the presence of extinct or unknown species can be decisive for scientific conclusions involving non vertical inheritance by revisiting a particular method of detection of introgression, which becomes misleading when dead lineages are ignored.

The Zombi simulator

Zombi has three main modes that are used to simulate different entities: species Trees (**T**), Genomes (**G**) and Sequences (**S**).

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”, by removing all the lineages that did not survive until the end of the simulation (Figure 1a).

The **G** mode simulates the evolution of genomes along the branches of the CST (Figure 1b). Genomes are circular and comprised of 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 representing the direction of the coding strand. All genes in the original genome are considered to have an evolutionary origin independent of each other, belonging to different gene families. The genomes evolve along the branches of the CST by undergoing six possible genome-level events: duplications, losses, inversions, translocations,transfers and originations, which are described below. These events affect a variable number of contiguous genes that is sampled from a geometric distribution with parameter p (specific for each kind of event). Duplications insert the new duplicated segment next to the ancient segment. Inversions revert the order of the genes in the segment by changing the current orientation of each gene. Translocations change the position of a gene or a given group of genes to a new position determined by an uniform distribution. Losses remove genes from their present genome. Transfers events place a copy of a gene from the donor lineage in the

recipient lineage. The recipient genome is determined by randomly sampling from all the alive lineages contemporary to the donor lineage. If the gene or group of genes transferred find an homologous equivalent in the recipient genome, a replacement transfer can take place with a probability determined by the user. New gene families can also appear by means of origination, an event in which a new gene is inserted in a random position of the genome determined by an uniform distribution. At each speciation event, two identical copies of the genome are created and start evolving independently in the descending branches.

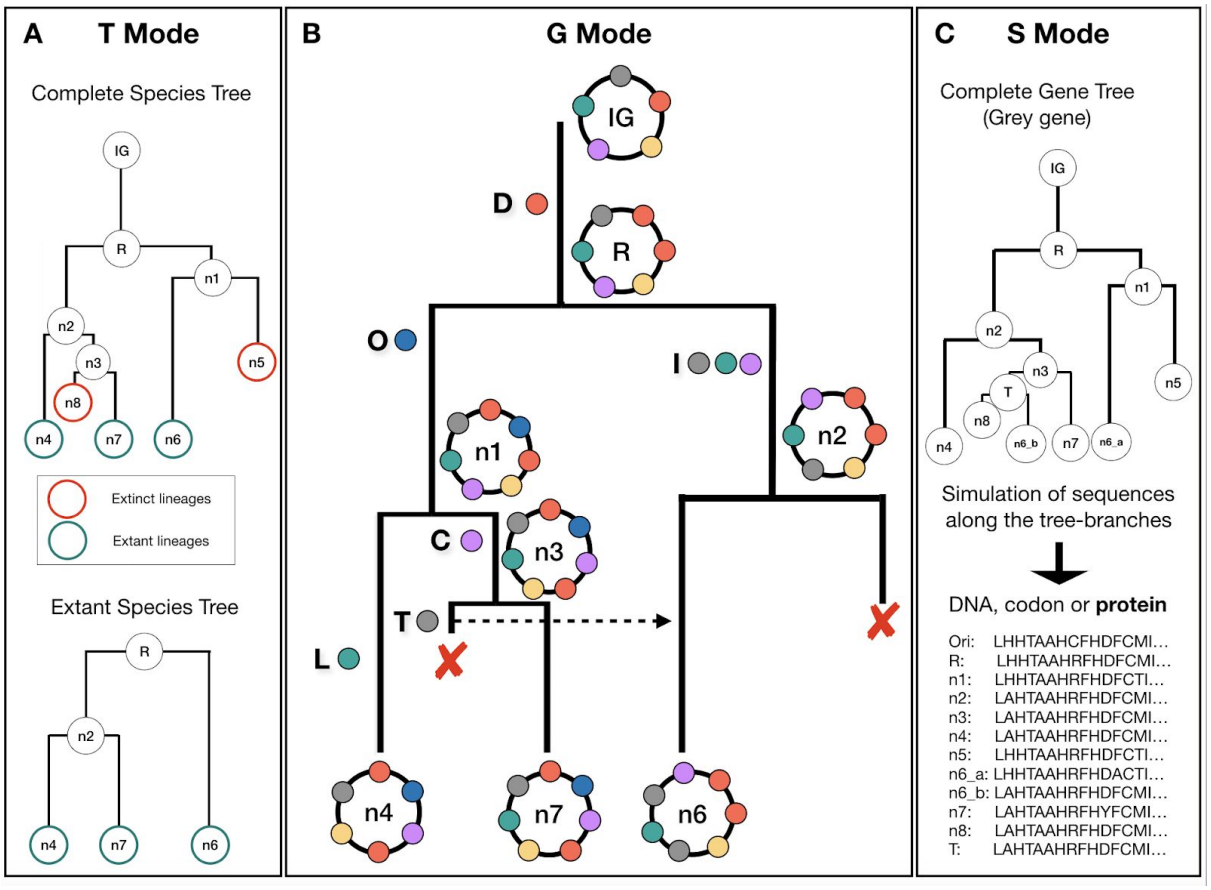

Figure 1. Overview of the three Zombi modes. A: In T mode, Zombi simulates a Species Tree (called Complete Species Tree, CST) using a birth-death process and outputs the pruned version of it by removing extinct lineages. In this example, lineages n5 and n8 go extinct before the simulation ends. **B**: in G mode, genomes evolve along the branches of the CST obtained in **A**. Genomes are circular and composed of a variable number of genes (represented by small coloured circles). The simulation starts with an initial genome (IG) comprised of 5 genes at the beginning of time. Different events (D Duplication, O Origination, I Inversion, C Translocation, L Loss, T Transfer) modify the genome composition. The genes affected are represented next to the letter indicating the event . For example, a duplication takes place between the IG and R (Root) affecting the red gene. A transfer event takes place between the branch leading to n8 (that goes extinct) and the branch leading to n6, affecting the green gene. The inversion event affects three contiguous genes (green, blue and purple). **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 outputs the structure of all the genomes at each node of the species tree (including the position in the genome of each gene and its orientation), a log of all the events

that occurred (per gene family and per branch), and the individual gene trees for each gene family, with and without the lineages that did not survive until the end of the simulation. Reconciled gene trees with the species tree can also be output in the RecPhyloXML reconciliation standard (Duchemin et al., 2018).

The **S** mode simulates gene sequences (at either the codon, nucleotide or protein level) (Figure 1c). For this, Zombi resorts to the Python library Pyvolve (Spielman & Wilke, 2015), that takes as an input a gene tree and simulate the evolution of genes along the branches of this tree. The user can control also the scaling of the tree, to have a better control on the number of substitutions that take place per unit of time.

Zombi includes additional functions to compute species phylogenies with variable speciation and extinction rates, sample a fraction of the total number number of extant lineages, create phylogenies with a fine control of the number of lineages per unit of time (Figure S1), compute genomes on species trees input by the user, compute genomes with branch-wise rates and model the heterogeneity in the substitution rate along different branches of the species tree and different gene families.

Performance

To measure the performance of Zombi we simulated datasets with varying number of species (extinct and extant) and the size of the genome (number of genes) at the beginning of the simulation (Figure S4). On a 3.4 GHz Intel Core i5 processor, Zombi can simulate the genomes of around 500 genes for 15,000 species in less than 10 minutes.

Use case and importance of modelling extinct lineages

Some of the existing methods to detect non-vertical inheritance (Durand, Patterson, Reich, & Slatkin, 2011; Fontaine et al., 2015; Pease & Hahn, 2014) do not consider the importance that extinct lineages can have. To illustrate a use of Zombi we tested the validity of one of these methods (Fontaine et al., 2015), used earlier to detect introgression-free genes, thought to be good markers for retrieving the phylogeny of extant species. The rationale behind the approach is depicted in Figure 2. For three species or clades (A, B and C), if introgression occurred *within* these three groups (Fig 2, cases 1 and 2), it will not only change the topology of the introgressed gene trees, but also decrease the branch lengths in the trees (T1 and T2). Consequently, under this assumption, trees with high T1 and T2 are expected to represent genes that were not introgressed and thus display a topology similar to the species tree topology. However, if introgression occurs *between* the three groups and extinct lineages (or unsampled ones, Figure 2, cases 3a and 3b), the expectations are opposite. In this case, introgression is expected to increase T1 and T2, and the genes that were not affected by introgression (and are thus topologically similar to the species tree) are those with the smaller T1 and T2 values.

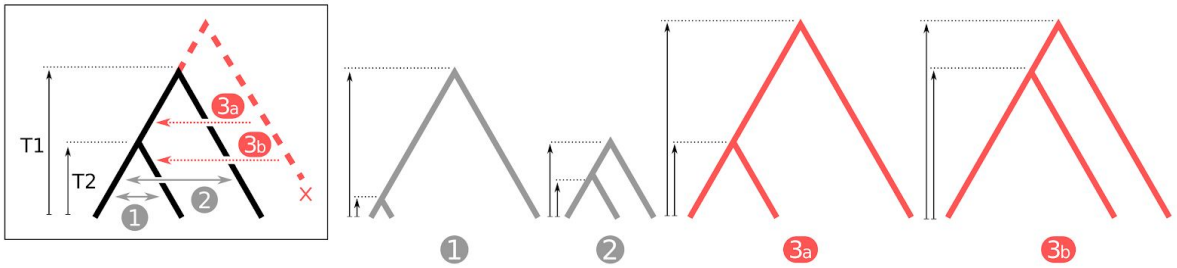

Figure 2. Illustration of the use of branch lengths (T1 and T2) to detect introgression. In the classical view introgression is only considered to take place among the species considered (cases 1 and 2) and results on average on a decrease of branch length in genes that have been introgressed. If we consider that an unsampled or extinct lineage is the source of the transfer (cases 3a and 3b), then the result of the introgression is an increase of branch length as compared to non-introgressed genomic regions. This can cause confusions in studies relying on this approach to tease apart introgressed and non-introgressed genomic regions.

To go further and verify these assumptions, we simulated with Zombi two species trees, one with extinct lineages and the other without them (but same tree of extant species), and simulated the evolution of genomes along their branches (Figure 3a 3b). We then chose three clades (A, B, C) and computed the mean branch length between all the genes supporting each of the three possible topologies for the three clades (Figure 3c and 3d), following the approach used in Fontaine et al (2015). We observed that when extinct lineages were not considered, the correct species tree was indeed supported by gene trees with the statistically highest T1 and T2 (Fig 3c). However, under the more realistic scenario where extinct lineages were considered, the correct species tree was the one supported by genes with the smallest T1 and T2 (Fig 3d). Thus detecting the correct species tree based on this test leads to opposite conclusions depending on the existence (or not) of extinct (or hidden) lineages. This experiment shows how Zombi can be used to test an evolutionary hypothesis in silico and illustrates the importance of considering lineages not directly represented in the phylogeny.

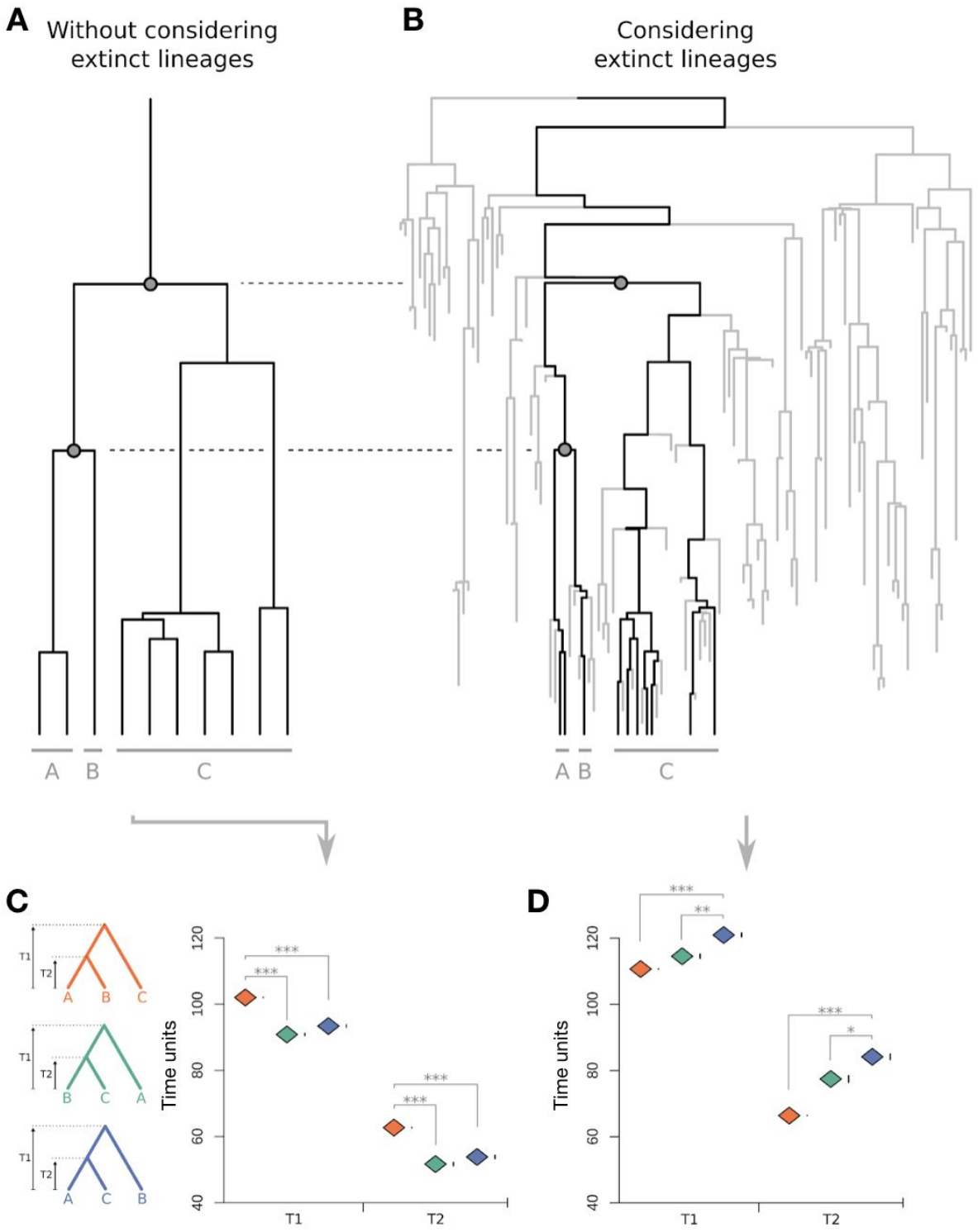

Figure 3: Use of Zombi to re-evaluate a method aiming at detecting introgression-free genes in phylogenomics**.** The evolution of genomes (3000 genes) was simulated on two trees, one with only extant species (**A**), the other with all species (the Complete Species Tree, **B**). Mean T1 and T2 were computed for all genes supporting each one of the three possible topologies (orange, green and blue trees).The true topology (orange) is the one supported by the trees with the highest mean T1 and T2 when considering only extant species (**C**), but when considering also extinct lineages (**D**), the gene trees with the highest T1 and T2 support instead an incorrect topology. One-tailed t-test p-values: * < 0.01; ** < 0.001; *** < 0.0001. Vertical bars on the right of diamonds represent standard error.

Materials and Methods

The species tree with dead lineages was obtained using the **Tp** mode with a speciation and extinction rates of 0.15 and 0.09 respectively. The lineage number was set to 20 species for the 140 first units of times then to 3 species for the last 10 units with a turnover set to 0.05, which produces a complete tree of 119 species of which only 10 are extant (Figure 3b).
For the second tree without extinct lineages the extant tree from the precedent simulation (Figure 3a) was used with the **Ti** mode, which allows the user to input an ultrametric tree as the CST. Genome with an initial size of 3000 genes were made to evolve in both trees using the **G** mode. In both case all main events rates except for transfers were set to 0. The transfer rate was set to 5 for the tree with extinction and 10 with no extinction. The probability of replacement transfers was set to 1. The extension parameter of the transfer events was set to 1, meaning that every time that a transfer event takes place it affects only a single gene. The choice of the three clades A, B and C was arbitrary. Gene families were grouped according to which topologies they supported: either the topology reflecting the species history: ((A,B),C) or one of the two topologies resulting from introgressions, ((A,C),B) and ((B,C),A). For each of the three topologies for both sets of genes simulation, mean T1 and T2 was computed.

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. We thank Vincent Daubin, Wandrille Duchemin, Nicolas Lartillot and Thibault Latrille for insightful discussions during the preparation of this manuscript.

References

Carvajal-Rodríguez, A. (2008). Simulation of genomes: a review. *Current Genomics*, *9*(3), 155–159.

Carvajal-Rodríguez, A. (2010). Simulation of genes and genomes forward in time. *Current Genomics*, *11*(1), 58–61.

Duchemin, W., Gence, G., Arigon Chifolleau, A.-M., Arvestad, L., Bansal, M. S., Berry, V., … Daubin, V. (2018). RecPhyloXML - a format for reconciled gene trees. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/bty389

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252.

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., … Besansky, N. J. (2015). Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, *347*(6217), 1258524.

Fournier, G. P., Huang, J., & Gogarten, J. P. (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), 2229–2239.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, *81*(25), 2340–2361.

GSR. (n.d.). Genetic Simulation Resources Home. Retrieved May 29, 2018, from https://popmodels.cancercontrol.cancer.gov/gsr/home/

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6),

1635–1638.

Kendall, D. G. (1948). On the Generalized “Birth-and-Death” Process. *Annals of Mathematical Statistics*, *19*(1), 1–15.

Pease, J. B., & Hahn, M. W. (2014). Detection and Polarization of Introgression in a Five-taxon Phylogeny. https://doi.org/10.1101/004689

Spielman, S. J., & Wilke, C. O. (2015). Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies. *PloS One*, *10*(9), e0139047.

Szöllősi, G. J., Tannier, E., Lartillot, N., & Daubin, V. (2013). Lateral Gene Transfer from the Dead. *Systematic Biology*, *62*(3), 386–397.

Zhaxybayeva, O., & Peter Gogarten, J. (2004). Cladogenesis, coalescence and the evolution of the three domains of life. *Trends in Genetics: TIG*, *20*(4), 182–187.

Supplementary Material

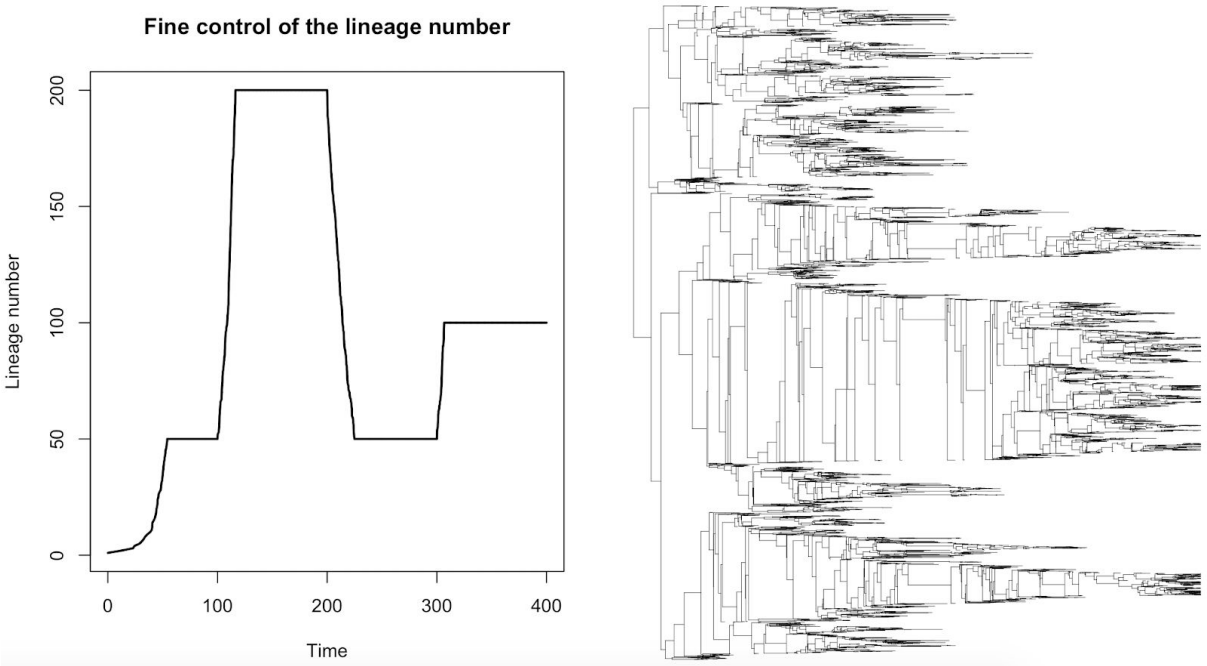

Figure S1. Fine control of the lineage number. Zombi can compute species tree using as input a list of times and the corresponding lineage number that should be attained by that time (in the example t 100 = 50; t 200 = 200; t 300 = 50; t 400 = 100). Zombi tries to attain the lineage number specified for each time interval given time with the given 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 speciates and the second one dies, keeping this way the total lineage number. The simulation continues until time = 400. In the right panel we can find the resulting species tree.

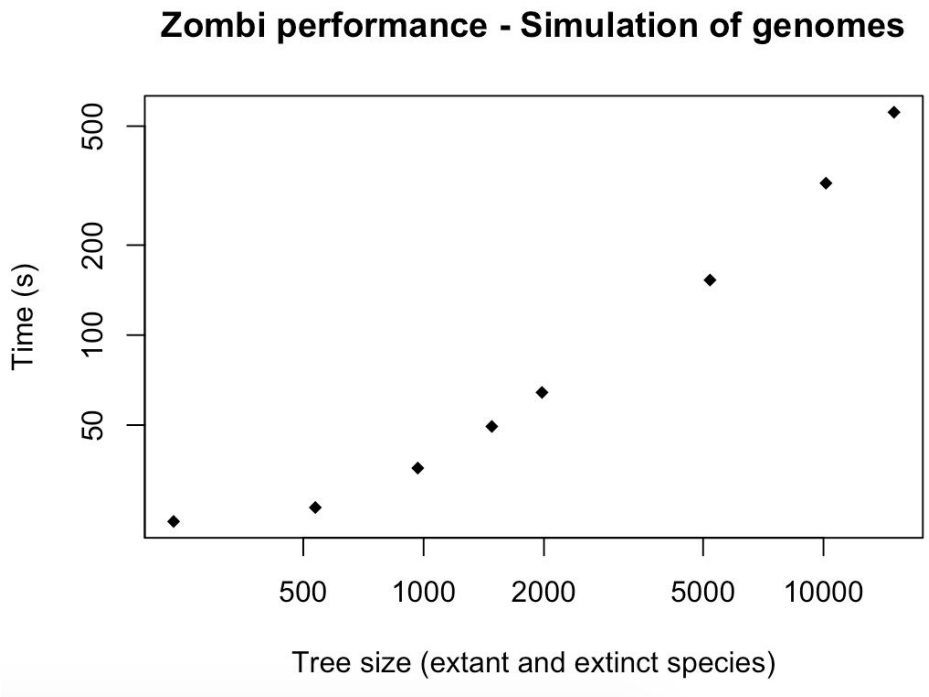

Figure S2. Computing time for different simulation in a computer with a 3,4 GHz Intel Core i5 processor. The rates used were Duplication rate: 0.2, Transfer rate: 0.2, Loss rate: 0.6, Origination rate:0.05, Inversion rate: 0.2, Translocation rate: 0.2. The initial genome was composed of 500 genes. All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