Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genetic and population analysis

SeDuS: Segmental Duplication Simulator

Diego A. Hartasánchez[1], Marina Brasó-Vives[1], Juanma Fuentes-Díaz[1], Oriol Vallès-Codina[1], Arcadi Navarro[1,2,3,4,*]

[1]Institute of Evolutionary Biology (Universitat Pompeu Fabra – CSIC), PRBB, 08003, Barcelona. [2]National Institute for Bioinformatics, 08003, Barcelona. [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010, Barcelona. [4]Centre for Genomic Regulation (CRG), 08003, Barcelona.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: SeDuS is the first flexible and user-friendly forward-in-time simulator of patterns of molecular evolution within segmental duplications undergoing interlocus gene conversion and crossover. SeDuS introduces known features of interlocus gene conversion such as biased directionality and dependence on local sequence identity. Additionally, it includes aspects such as different selective pressures acting upon copy number and flexible crossover distributions. A graphical user interface allows fast fine-tuning of relevant parameters and straightforward real-time analysis of the evolution of duplicates.

Availability and implementation: SeDuS is implemented in C++ and can be run via command line or through a GUI developed using Qt C++. Source code and binary executables for Linux, OS X and Windows are freely available at www.biologiaevolutiva.org/sedus/

Contact: sedus@upf.edu, arcadi.navarro@upf.edu

Supplementary information: A tutorial with a detailed description of implementation, parameters and output files is available online.

1 Introduction

The evolution of duplicated regions of the genome has attracted the attention of evolutionary biologists since Susumu Ohno proposed that they are a fundamental source of novel genes and functions (Ohno, 1970). Duplicated regions are a pervasive characteristic of eukaryotic genomes, and can span up to hundreds of kilobases encompassing several genes. Such is the case of segmental duplications (>1kb, >90% similarity), which are known to originate copy number variation and chromosomal rearrangements and to underlie the susceptibility to many diseases (Iskow et al., 2012).

Duplicated regions have a distinctive feature that crucially affects their evolution: they exchange genetic information through a type of gene conversion referred to as ectopic, non-allelic or interlocus gene conversion (IGC) (Ohta, 1982), which differs from usual allelic gene conversion in that it happens between paralog genomic regions. IGC is a major driver of the concerted evolution of duplicates, which complicates the application of conventional population-genetic interpretations, such as the molecular clock, to these regions (Teshima and Innan, 2004).

Simulating duplicated sequences under the coalescent has provided important insights into their neutral molecular evolution (Thornton, 2007). However, due to computing-time limitations (Yang et al., 2014), coalescent simulators can only explore a restricted range of parameters, particularly regarding recombination. Moreover, they preclude simulating IGC rate dependence on sequence similarity.

Here, we present a forward-in-time simulator of the molecular evolution of segmental duplications designed to explore their patterns of concerted evolution under a wide range of parameters. We have named this software SeDuS (Segmental Duplication Simulator). SeDuS is an improved and extended version of the in-house scripts used in previous work from our group (Hartasánchez et al., 2014). On top of command-line execution, SeDuS comes with an independent, user-friendly graphical user interface (GUI) allowing control over the most important parameters and direct visualization of simulation results. Thus, SeDuS has not only research applicability but can also be a great tool for educators.

To our knowledge, SeDuS is the first user-friendly, forward-in-time population genetics simulator specifically aimed at addressing the evolution of segmental duplications while giving full consideration to IGC. Our algorithm has a modular architecture allowing the user to easily

modify specific functions or to incorporate new features. SeDuS is under constant development and updates will be presented accordingly.

2 Design and implementation

The core of SeDuS is built on the C++ code published in Hartasánchez et al. (2014). Here, we briefly describe the underlying structure of the software and then expand on its novel biology-oriented additions and technical improvements.

SeDuS is a forward-in-time simulator of a Wright-Fisher diploid population evolving under neutrality. Each individual is represented by a single pair of homologous chromosomes, and each chromosome is initially composed of two blocks (*original* and *single-copy*) of equal length L. During a burn-in phase, each chromosome undergoes mutation and crossover. At a given point in time, a duplication event takes place in which the original block on a randomly chosen chromosome is copied to the right of the single-copy block, either on the same chromosome or on its homologous (Figure 1a). The duplication is conditioned to fixation following a neutral or selective trajectory. The original and *duplicated* blocks exchange information via IGC, which occurs at rate C in all chromosomes carrying the duplication (Figure 1b). For further details on the structure of SeDuS' please refer to the SeDuS tutorial.

SeDuS includes a series of new features regarding distinct aspects of IGC. Even though IGC between duplicates has been known for a long time, the molecular mechanisms underlying IGC remain relatively obscure (Hastings, 2010) and largely unexplored via simulations. One major reason for the latter is that some of the characteristics of IGC violate the basic assumptions of the coalescent model (Thornton, 2007). For instance, the rate of IGC depends on local sequence similarity, with research indicating that for an IGC event to occur, a tract of 100% identity between duplicates, called a minimal efficient processing segment (MEPS) must be present near the IGC initiation site (Shen and Huang, 1986). This phenomenon makes the probability of an IGC event between two particular sequences dependent on their level of divergence, which makes it impossible to separate, as the coalescent does, the processes of mutation and genealogy building. In contrast, SeDuS easily simulates MEPS. Additionally, SeDuS incorporates biased directionality in IGC by establishing different probabilities for the duplicated block to act as donor or acceptor of IGC events. Moreover, IGC can occur between paralogs in the same chromosome or in homologous chromosomes with user-defined probabilities.

Another novelty is that SeDuS can simulate both neutral and non-neutral fixation-conditioned trajectories of the duplication. For example, fast fixation events, characteristic of the presence of a duplication being positively selected (or slightly deleterious), can be simulated in SeDuS by forcing the duplication to reach fixation in a given number of generations through a linear trajectory (Teshima and Innan, 2012).

Previous work has showed that crossover hotspots overlapping duplicated regions might generate important deviations from neutral expectations (Hartasánchez et al., 2014), highlighting the importance of incorporating specific recombination landscapes when simulating concerted evolution between duplicates. To allow simulating such scenarios, SeDuS allows meiotic crossover to occur at rate R at user-defined regions that might include several hotspots of any specified intensity (up to five regions in the GUI and an unlimited number in the command-line version). Regions can overlap, allowing the user to easily simulate, for instance, a crossover hotspot over a background crossover rate.

In terms of technical improvements, SeDuS has efficient memory management, a structure that enables parallelization of simulation runs, and shorter execution times. On a typical desktop computer, the simulation of a population of size N=1,000 with a duplicated region of 10 kb evolving under concerted evolution for 10,000 generations takes ~2 seconds if executed via command line.

Another major feature of SeDuS is its GUI (implemented in Qt C++), which provides real-time feedback and allows the visualization of variation measures, such as the average number of pairwise differences within each block. The GUI is user-friendly and can be used for quick explorations of the molecular evolution of segmental duplications with both research and educational purposes.

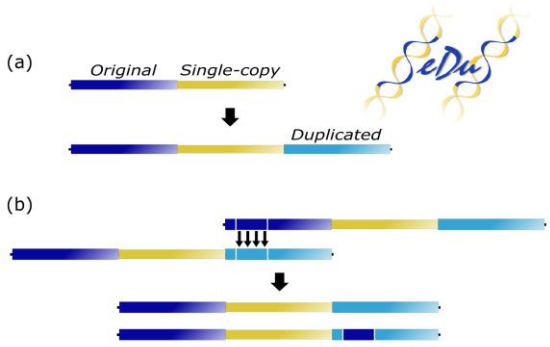

Fig. 1. (a) Unique duplication event. **(b)** Interlocus gene conversion occurs with rate C between the original and duplicated blocks in homologous chromosomes or in the same chromosome (not represented) driving the concerted evolution of segmental duplications.

Acknowledgements

We thank David A. Hughes for helpful comments and Txema Heredia for technical assistance throughout the development of this software.

Funding

This work has been supported by the Spanish National Institute of Bioinformatics, a platform of the Instituto de Salud Carlos III (PT13/0001/0026), and the Spanish Government, Grant BFU2012-38236 to A.N.; by grants to D.A.H. from Conacyt and CSIC (JAE Predoc); by the Fondo Europeo de Desarrollo Regional (FEDER) and the Fondo Social Europeo (FSE); and by a grant to M.B-V. from AGAUR (FI – DGR 2015).

Conflict of Interest: none declared.

References

Hartasánchez,D.A. et al. (2014) Interplay of interlocus gene conversion and crossover in segmental duplications under a neutral scenario. *G3 (Bethesda)*, **4**, 1479-1489.

Hastings,P.J. (2010) Mechanisms of ectopic gene conversion. *Genes*, **1**, 427-439.

Iskow,R.C. et al. (2012) Exploring the role of copy number variants in human adaptation. *Trends Genet.*, **28**, 245-257.

Ohno,S. (1970) *Evolution by Gene Duplication*. Springer, New York.

Ohta,T. (1982) Allelic and nonallelic homology of a supergene family. *Proc. Natl. Acad. Sci. USA*, **79**, 3251-3254.

Shen,P. and Huang,H.V. (1986) Homologous recombination Escherichia coli: dependence on substrate length and homology. *Genetics*, **112**, 441-457.

Teshima,K.M. and Innan,H. (2004) The effect of gene conversion on the divergence between duplicated genes. *Genetics*, **166**, 1553-1560.

Teshima,K.M. and Innan,H. (2012) The coalescent with selection on copy number variants. *Genetics*, **190**, 1077-1086.

Thornton,K.R. (2007) The neutral coalescent process for recent gene duplications and copy-number variants. *Genetics*, **177**, 987-1000.

Yang,T. et al. (2014) Critical assessment of coalescent simulators in modeling recombination hotspots in genomic sequences. *BMC Bioinformatics*, **15**, 3.