COLLAPSED DUPLICATIONS? WHAT TO EXPECT AND WHAT TO LOOK FOR.

Diego A. Hartasánchez[1], Marina Brasó-Vives[1], Marc Pybus[1], and Arcadi Navarro[1,2,3,4]

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

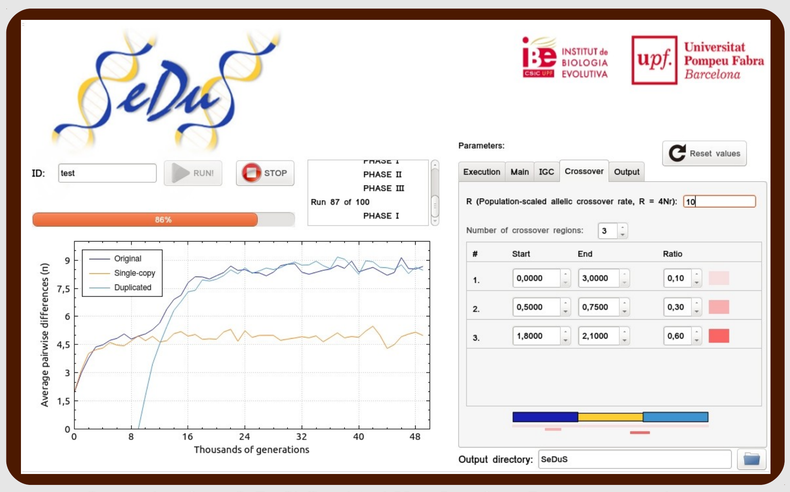
INTRODUCTION

Segmental duplications (SDs), defined as >1 kb regions of the genome with >90% similarity between copies, are an ubiquitous characteristic of eukaryotic genomes. Their evolution is known to be complex for several reasons: first, because SDs undergo interlocus gene conversion (IGC), a possible source of variation; second, reduced selective pressures may allow variants to increase in frequency more easily; and third, SDs are mediators of NAHR and formation of CNVs, which in turn are associated with susceptibility to disease.

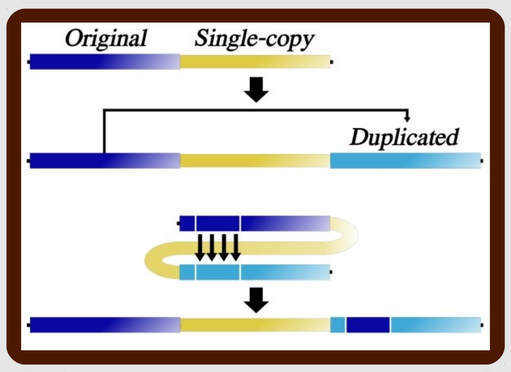
SD detection and characterization has been recognized as being of great importance. Ironically, most of the efforts dedicated to these tasks are aimed at eliminating SDs from genome-wide scans in order to avoid spurious signals coming from duplicated regions. Given that SDs are likely to be possible targets of natural selection, it would seem natural to look for SNPs under selection in duplications. However, to date there is no adequate test to detect selection in duplications precisely because none takes into account their complex evolution. Furthermore, the effect of applying neutrality tests to collapsed duplications is mostly ignored.

METHODS

We performed simulations of SDs undergoing concerted evolution with **SeDuS: Segmental Duplication Simulator[1]**. SeDuS is a forward-in-time simulator of SDs that allows for the exploration of a wide set of parameters. In particular, here we present results for a range of interlocus gene conversion and allelic crossover rates.



Screenshot from SeDuS' Graphical User Interface. Parameters can be easily modified and results can be visualized in real-time.



Interlocus gene conversion occurs with rate *C* between the original and duplicated blocks, driving the concerted evolution of segmental duplications.

Additionally to the results from SeDuS, we have run simulations with MSMS[2]. Simulated scenarios involve neutrality, a complete selective sweep, an incomplete selective sweep and a case of balancing selection. We simulate selective scenarios to show that duplications might be mistakenly taken as regions under selection if summary statistics are not properly chosen.

RESULTS

We have explored what type of signal would be obtained if traditional neutrality tests were applied to duplicated regions in the following two scenarios: if the duplication is known to exist, or if the duplication is collapsed onto one single region as is common when reference genomes are of poor quality.

In the first case, we describe summary statistics that might be helpful to distinguish duplicated regions undergoing interlocus gene conversion. In the second case, we present the possible misinterpretations of summary statistics when applied to regions that are duplicated but unknown to be so.

We have applied a set of summary statistics to simulated data.

Diversity estimators:

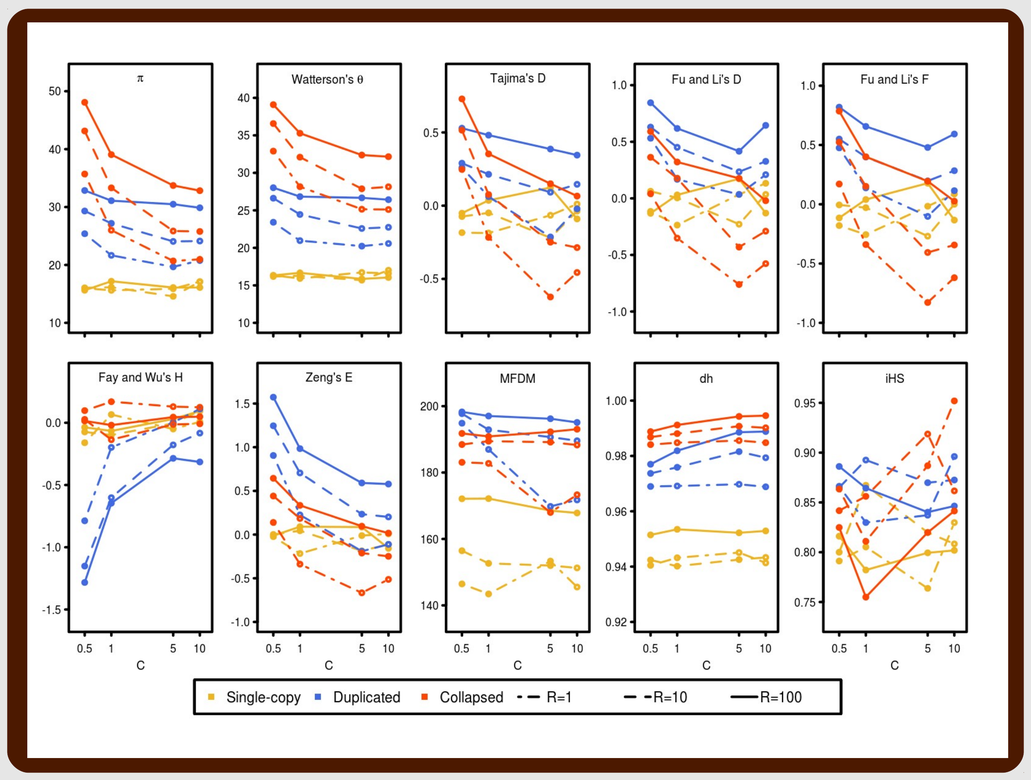
π (Nei and Li, 1979)
Watterson's θ (Watterson, 1975)

Neutrality statistics:

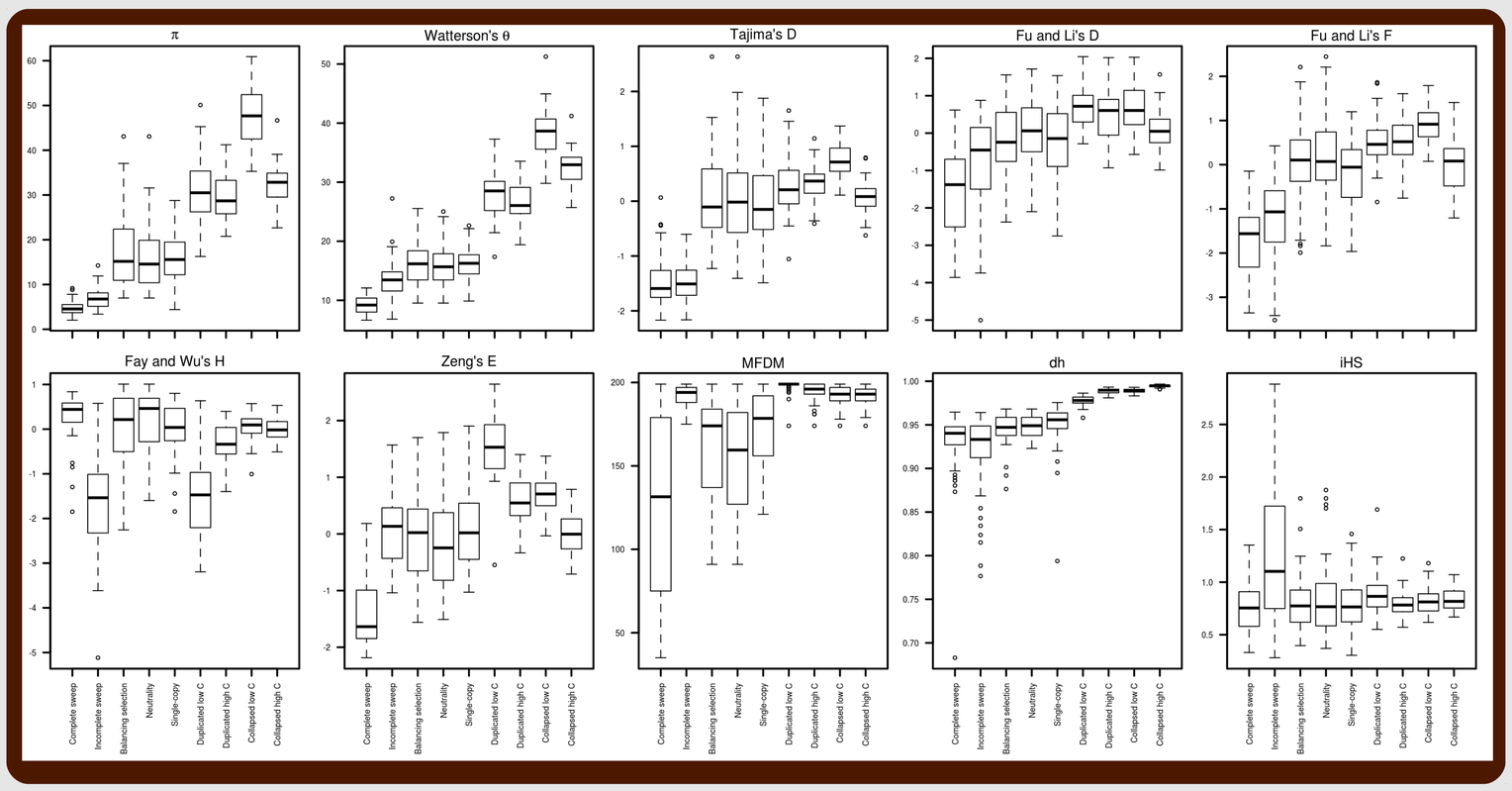
Tajima's D (Tajima, 1989)
Fu and Li's D (Fu and Li, 1993)
Fu and Li's F (Fu and Li, 1993)
Fay and Wu's H (Fay and Wu, 2000)
Zeng's E (Zeng *et al.*, 2006)
MFDM (Li, 2011)

Haplotype based:

dh (Nei, 1987)
iHS (Voight *et al.*, 2006)



Summary statistics for the Single-copy, Duplicated and Collapsed blocks at equilibrium under concerted evolution for a range of interlocus gene conversion rates (*C*) and allelic crossover rates (*R*).



Summary statistics for four evolutionary scenarios (complete sweep, incomplete sweep, balancing selection, and neutrality) and for the single-copy, duplicated and collapsed blocks. Duplicated and collapsed are shown for low and high IGC rates (*C*) with high allelic crossover rates (*R*=10).

CONCLUSIONS

We have modeled and analyzed the effect of concerted evolution of SDs on common statistical tests. We describe the type of signature imprinted by natural selection on duplicated regions when these are collapsed.

Our results show that unidentified duplications can render confounding results if collapsed when building genome assemblies, and on the other hand, that by collapsing duplications, one can actually extract relevant information about their evolution.

REFERENCES

1. Hartasánchez, et al. (Submitted, 2015).
2. Ewing and Hermisson, *Bioinformatics* **26**, 2010.