Peer Community In
Evolutionary Biology

A new statistical tool to identify the determinant of parallel evolution

Stéphanie Bedhomme

CEFE, CNRS -- Montpellier, France
Correspondence to Stéphanie Bedhomme (stephanie.bedhomme@cefe.cnrs.fr)
doi: 10.24072/pci.evolbiol.100045

Published: 31st Jan. 2018

Cite as: Bedhomme S. 2018. A new statistical tool to identify the determinant of parallel evolution. Peer Community in Evolutionary Biology, 100045 doi: 10.24072/pci.evolbiol.100045

A recommendation – based on reviews by an anonymous reviewer and Bastien Boussau – of

Bailey SF, Guo Q and Bataillon T. 2018. Identifying drivers of parallel evolution: A regression model approach. bioRxiv, 118695, ver. 4 peer-reeviewed by PCI Evol Biol. doi: 10.1101/118695

In experimental evolution followed by whole genome resequencing, parallel evolution, defined as the increase in frequency of identical changes in independent populations adapting to the same environment, is often considered as the product of similar selection pressures and the parallel changes are interpreted as adaptive. However, theory predicts that heterogeneity both in mutation rate and selection intensity across the genome can trigger patterns of parallel evolution. It is thus important to evaluate and quantify the contribution of both mutation and selection in determining parallel evolution to interpret more accurately experimental evolution genomic data and also potentially improve our capacity to predict the genes that will respond to selection. In their manuscript, Bailey, Guo and Bataillon [1] derive a framework of statistical models to partition the role of mutation and selection in determining patterns of parallel evolution at the gene level. The rationale is to use the synonymous mutations dataset as a baseline to characterize the mutation rate heterogeneity, assuming a negligible impact of selection on synonymous mutations and then analyse the non-synonymous dataset to identify additional source(s) of heterogeneity, by examining the proportion of the variation explained by a number of genomic variables. This framework is applied to a published data set of resequencing of 40 Saccharomyces cerevisiae populations adapting to a laboratory environment [2]. The model explaining at best the synonymous mutations dataset is one of homogeneous mutation rate along the genome with a significant positive effect of

gene length, likely reflecting variation in the size of the mutational target. For the non-synonymous mutations dataset, introducing heterogeneity between sites for the probability of a change to increase in frequency is improving the model fit and this heterogeneity can be partially explained by differences in gene length, recombination rate and number of functional protein domains. The application of the framework to an experimental data set illustrates its capacity to disentangle the role of mutation and selection and to identify genomic variables explaining heterogeneity in parallel evolution probability but also points to potential limits, cautiously discussed by the authors: first, the number of mutations in the dataset analysed needs to be sufficient, in particular to establish the baseline on the synonymous dataset. Here, despite a high replication (40 populations evolved in the exact same conditions), the total number of synonymous mutations that could be analysed was not very high and there was only one case of a gene with synonymous mutation in two independent populations. Second, although the models are able to identify factors affecting the mutation counts, the proportion of the variation explained is quite low. The consequence is that the models correctly predicts the mutation count distribution but the objective of predicting on which genes the response to selection will occur still seems quite far away. The framework developed in this manuscript [1] clearly represents a very useful tool for the analysis of large "evolve and resequence" data sets and to gain a better understanding of the determinants of parallel evolution in general. The extension of its application to mutations others than SNPs would provide the possibility to get a more complete picture of the differences in contributions of mutation and selection intensity heterogeneities depending on the mutation types.

References

[1] Bailey SF, Guo Q and Bataillon T. 2018. Identifying drivers of parallel evolution: A regression model approach. BioRxiv 118695, ver. 4 peer-reviewed by PCI Evol Biol. doi: http://doi.org/10.1101/118695

[2] Lang GI, Rice DP, Hickman, MJ, Sodergren E, Weinstock GM, Botstein D, and Desai MM. 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature 500: 571–574. doi: http://doi.org/10.1038/nature12344

Appendix

Reviews by an anonymous reviewer and Bastien Boussau: http://dx.doi.org/10.24072/pci.evolbiol.100045