Bioinformatics, 35(16), 2019, 2868–2869
doi: 10.1093/bioinformatics/bty1060
Advance Access Publication Date: 7 January 2019
Applications Note

OXFORD

Genetics and population analysis

polyDFEv2.0: testing for invariance of the distribution of fitness effects within and across species

Paula Tataru and Thomas Bataillon *

Bioinformatics Research Centre, Aarhus University, DK-8000 Aarhus, Denmark

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on July 17, 2018; revised on December 18, 2018; editorial decision on December 20, 2018; accepted on January 3, 2019

Abstract

Summary: Distribution of fitness effects (DFE) of mutations can be inferred from site frequency spectrum (SFS) data. There is mounting interest to determine whether distinct genomic regions and/or species share a common DFE, or whether evidence exists for differences among them. polyDFEv2.0 fits multiple SFS datasets at once and provides likelihood ratio tests for DFE invariance across datasets. Simulations show that testing for DFE invariance across genomic regions within a species requires models accounting for distinct sources of heterogeneity (chance and genuine difference in DFE) underlying differences in SFS data in these regions. Not accounting for this will result in the spurious detection of DFE differences.

Availability and Implementation: polyDFEv2.0 is implemented in C and is accompanied by a series of R functions that facilitate post-processing of the output. It is available as source code and compiled binaries under a GNU General Public License v3.0 from https://github.com/paula-tataru/polyDFE.

Contact: tbata@birc.au.dk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Levels of purifying and positive selection vary throughout the genome and a powerful way to study such variation is by inferring the distribution of fitness effects (DFE) for different genomic regions that are a priori expected to undergo distinct selective pressures (Gronau *et al.*, 2013; Racimo and Schraiber 2014). Similarly, there is renewed interest in understanding what genomic, demographic and ecological factors explain differences in genome-wide polymorphism patterns across species (Chen *et al.*, 2017; Ellegren and Galtier 2016; Huber *et al.*, 2017). The site frequency spectrum (SFS) contains information to infer the DFE of new mutations. Existing methods infer the DFE while accounting for demography and other sources of distortion in the SFS (Barton *et al.*, 2018; Galtier 2016; Kim *et al.*, 2017; Kousathanas and Keightley 2013; Schneider *et al.*, 2011; Tataru *et al.*, 2017). Yet, current methods assume that a mutation's fitness is drawn from a single common DFE and are therefore not well suited to determine whether distinct gene categories, genomic regions and/or species share invariant DFEs, or whether there are genuine differences between such categories.

Here, we present a new method, polyDFEv2.0, for testing invariance of DFEs across datasets, be it distinct genomic regions within species, or different species. Simulations demonstrate that the method guards against excessive type I error while retaining substantial power to detect differences across datasets.

2 Results and discussion

polyDFE implements a likelihood framework that allows fitting simultaneously DFE parameters, nuisance parameters that account for distortions in the SFS data induced by linkage and demography and errors when polarizing the SFS data (Tataru *et al.*, 2017). Here, we extend polyDFE for fitting multiple datasets. Any fitted parameter can be constrained to be shared (invariant) across datasets or fitted independently for each dataset (models M_1–M_4, Table 1).

Table 1. Models for detecting heterogeneity in DFEs across datasets fitted in polyDFEv2.0

Models	Model parameters fitted			
	DFE parameters	Demography	Mutation rate	Polarization error
M_1	Shared	Shared	Shared	Shared
M_2	Independent	Shared	Shared	Shared
M_3	Shared	Independent	Independent	Independent
M_4	Independent	Independent	Independent	Independent

Likelihood ratio tests (LRTs) are used to determine if the datasets have different DFEs.

When testing for invariance in DFEs across species, we use LRTs comparing the fit of M_3 and M_4, as we expect demography and scaled mutation rates to vary between species (Table 1). Simulations (Supplementary Material) show that comparing M_3 and M_4 provides a reliable test for the null hypothesis of DFE invariance: no inflation in type I error under H_0 and good power to detect differences in strength of purifying and positive selection (Supplementary Fig. S1).

When analyzing SFS datasets from distinct regions or gene ontologies within one species, we investigated the type I error and power of two LRTs: M_1 versus M_2 and M_3 versus M_4. Comparing models M_1 and M_2 amounts to assume that all genomic regions share the same nuisance parameters. Simulations show that M_1 might not be a proper null model for DFE invariance within species and accordingly that the χ^2 approximation for this LRT might not apply even with large datasets (Supplementary Figs S1 and S2A). This is because even under DFE invariance, differences in coalescent histories between genomic regions and single nucleotide polymorphism polarization error rates will generate differences in the observed SFS data across regions. It is a challenge to devise inference models that do not spuriously attribute the differences observed in the SFS counts as underlying differences in the DFEs. Modeling accurately the differences in coalescent histories and the interaction of selection and demography along the genome remains difficult (Hartfield *et al.*, 2017; Li *et al.*, 2012) and the object of much current research. Here, the models fitted in the polyDFE framework account for this via the use of distinct nuisance parameters for each SFS dataset.

Accordingly, even when fitting different SFS datasets within a single species, we recommend fitting models with nuisance parameters for each SFS. Using models that account for both differences in DFE and nuisance parameters (M_3 and M_4) allows to control for type I error while retaining substantial power to detect differences in DFE (Supplementary Fig. S1). An alternative that is computationally more demanding is to rely on simulations to obtain the empirical distribution of the LRT under the null hypothesis of invariance instead of relying on the χ^2 approximations (see Supplementary Material for details).

We re-analyzed a chimpanzee dataset (Bataillon *et al.*, 2015) and tested for DFEs invariance across two subspecies (central versus eastern chimpanzee) and between autosomal versus X-linked regions (see Supplementary Material for details). We found autosomal DFEs to be different between subspecies. Central chimpanzees exhibit a higher effective size resulting in overall a few more strongly deleterious mutations (Supplementary Fig. S3A). No significant differences are found when the comparison involved X-linked regions (Supplementary Fig. S3B and C). While genuine differences in DFEs probably exist, simulations show that the amount of data available on the X chromosome (0.89 Mb of X-linked sites versus 20 Mb of autosomes) entails substantial loss of power. The amount of data available can limit drastically the statistical power (Supplementary Fig. S4): the observed differences in the inferred DFEs are very modest among autosomes but they are detected as significantly different while the more substantial differences between X and autosomes are not.

3 Conclusion

polyDFEv2.0 provides tests for DFE invariance. Flexibility in the fitted models also enables testing more specific hypothesis regarding the nature of the differences across DFEs: e.g. did the shape of the DFE change and/or the proportion of beneficial mutations change across DFEs? Using models that are flexible enough to account jointly for differences in nuisance parameters across categories/species is of paramount importance to make reliable tests for DFE invariance across and within species. Ultimately, amounts of single nucleotide polymorphisms within each compared SFS drastically condition the power of our method to detect heterogeneity. We recommend prudence when dividing data among too many separate SFSs to be fitted using different DFEs.

Acknowledgements

We thank D. Castellano, M. Hartfield and E. Lucotte and three reviewers (J. Schraiber, A. Eyre-Walker and I. Gronau) for comments on the manuscript and trying out the method.

Funding

This work was supported by the European Research Council [FP7/20072013, ERC grant number 311341].

Conflict of Interest: none declared.

References

Barton,H.J. *et al.* (2018) New methods for inferring the distribution of fitness effects for INDELs and SNPs. *Mol. Biol. Evol.*, **35**, 1536–1546.

Bataillon,T. *et al.* (2015) Inference of purifying and positive selection in three subspecies of chimpanzees (*Pan troglodytes*) from exome sequencing. *Genome Biol. Evol.*, **7**, 1122–1132.

Chen,J. *et al.* (2017) Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. *Mol. Biol. Evol.*, **34**, 1417–1428.

Ellegren,H. and Galtier,N. (2016) Determinants of genetic diversity. *Nat. Rev. Genet.*, **17**, 422–433.

Galtier,N. (2016) Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genet.*, **12**, e1005774.

Gronau,I. *et al.* (2013) Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Mol. Biol. Evol.*, **30**, 1159–1171.

Hartfield,M. *et al.* (2017) The evolutionary interplay between adaptation and self-fertilization. *Trends Genet.*, **33**, 420–431.

Huber,C.D. *et al.* (2017) Determining the factors driving selective effects of new nonsynonymous mutations. *PNAS*, **114**, 4465–4470.

Kim,B.Y. *et al.* (2017) Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics*, **206**, 345–361.

Kousathanas,A. and Keightley,P.D. (2013) A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics*, **193**, 1197–1208.

Li,J. *et al.* (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Mol. Ecol.*, **21**, 28–44.

Racimo,F. and Schraiber,J.G. (2014) Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. *PLoS Genet.*, **10**, e1004697.

Schneider,A. *et al.* (2011) A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics*, **189**, 1427–1437.

Tataru,P. *et al.* (2017) Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics*, **207**, 1103–1119.