ksrates: positioning whole-genome duplications relative to speciation events using rate-adjusted mixed paralog–ortholog K_S distributions

Cecilia Sensalari[1,2], Steven Maere[1,2,*] and Rolf Lohaus[1,2,*]

[1]Dept. of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium
[2]VIB-UGent Center for Plant Systems Biology, 9052 Ghent, Belgium

∗ co-last authors

Abstract

Summary: To position ancient whole-genome duplication (WGD) events with respect to speciation events in a phylogeny, the K_S values of WGD paralog pairs in a species of interest are often compared with the K_S values of ortholog pairs between this species and other species. However, if the lineages involved exhibit different substitution rates, direct comparison of paralog and ortholog K_S estimates can be misleading and result in phylogenetic misinterpretation of WGD signatures. Here we present *ksrates*, a user-friendly command-line tool to compare paralog and ortholog K_S distributions derived from genomic or transcriptomic sequences. *ksrates* estimates differences in synonymous substitution rates among the lineages involved and generates an adjusted mixed plot of paralog and ortholog K_S distributions that allows to assess the relative phylogenetic positioning of presumed WGD and speciation events.

Availability and implementation: *ksrates* is open-source software implemented in Python 3 and as a Nextflow pipeline. The source code, Singularity and Docker containers, documentation and tutorial are available via https://github.com/VIB-PSB/ksrates.

Contact: steven.maere@ugent.vib.be, rolf.lohaus@ugent.vib.be

1 Introduction

The K_S value of a pair of homologous sequences, i.e. the estimated number of synonymous substitutions separating them per synonymous site, is used as a proxy for the time elapsed since the sequences diverged (Lynch and Conery, 2000; Blanc and Wolfe, 2004). K_S values of paralog pairs can be used to construct a relative age distribution of duplication events in a species, offering insight into the species' gene duplication history. Distinctive peaks in such paralog K_S distributions are often used to infer the presence of large-scale duplication events, such as whole-genome duplications (WGDs) (Blanc and Wolfe, 2004). Ortholog K_S distributions on the other hand are informative of the age of divergence between two species. A common practice to assess the temporal order of speciation and duplication events, e.g. to characterize the phylogenetic position of inferred WGDs, is to superimpose ortholog and paralog K_S distributions in mixed plots (Blanc and Wolfe, 2004; Zhang *et al.*, 2017; Li *et al.*, 2018).

However, the sequence of events inferred from such naive mixed K_S plots is not always reliable. K_S estimates for events of the same absolute age may differ depending on the synonymous

substitution rates in the lineages involved. A pair of duplicates in a fast-evolving species for instance will have a higher K_S value than a simultaneously duplicated pair in a more slowly evolving species. Hence, paralog K_S distributions of species with a different substitution rate history are not directly comparable, as they feature different timescales. Similarly, ortholog K_S values are influenced by the substitution rates in both of the species lineages being compared. An ortholog K_S distribution is thus not directly comparable to the paralog K_S distribution of either species or the ortholog K_S distributions of other species pairs, unless all species share the same substitution rate in the given time frame. Therefore, the straightforward superimposition of paralog and ortholog K_S distributions in a mixed plot can be misleading if the lineages involved evolved at different rates.

ksrates generates adjusted mixed plots of K_S distributions by rescaling ortholog K_S estimates of species divergence times to the paralog K_S scale of a focal species, producing shifts in the estimated K_S position of each speciation event proportional to the estimated substitution rate difference between the diverged lineage and that of the focal species. These substitution rate differences are estimated from the ortholog K_S data, based on species trios including the focal species, a diverged species and an outgroup, similar to relative rate test calculations (Sarich and Wilson, 1973; Graur, 2016). A use case is presented to show that *ksrates* generates adjusted mixed paralog–ortholog K_S plots from which the phylogenetic placement of hypothesized WGDs can be inferred more accurately than from naive mixed plots.

2 Methods

2.1 K_S distribution construction and mixture modeling

ksrates uses the *wgd* package (Zwaenepoel *et al.*, 2019) to detect paralog pairs and one-to-one ortholog pairs from genomic or transcriptomic sequence data and to calculate the associated K_S values, and then constructs K_S distributions from the raw K_S data (see Supplementary Methods). When genome structural annotation (GFF3 file) is provided as input for the focal species, the i-ADHoRe 3.0 package (Proost *et al.*, 2012) is run to identify anchor pairs, i.e. pairs of paralogs located on collinear duplicated segments remaining from large-scale duplication events such as WGDs. *ksrates* then uses mixture modeling on either the anchor-pair K_S distribution (if available) or the whole-paranome K_S distribution of the focal species to detect potential WGD signatures. When anchor-pair data is available, the anchor-pair K_S values are associated with putative WGDs based on lognormal mixture model clustering of the median K_S values of collinear segment pairs (see Supplementary Methods and Fig. 1A). Otherwise, an exponential-lognormal mixture model is fit to the whole-paranome K_S distribution, with an exponential component capturing the small-scale duplication background and lognormal components modeling putative WGD peaks. Optionally, it is also possible to fit more simple lognormal-only mixture models to the whole-paranome and anchor-pair K_S distributions. The reader should be aware that mixture modeling results should be interpreted cautiously because mixture models tend to overestimate the number of components present in the target K_S distribution and hence the number of WGDs (Tiley *et al.*, 2018).

2.2 Ortholog K_S adjustment and output

Phylogenetic relationships between species are provided as an input phylogenetic tree. Raw K_S estimates for the divergence times between the focal species and the other species in the tree are obtained by estimating the mode of the corresponding ortholog K_S distributions using bootstrapped kernel density estimation (see Supplementary Methods). For each divergence event, a substitution-rate-adjusted K_S estimate is then calculated by decomposing the raw K_S value into branch-specific contributions (with the help of an outgroup species, similar to the methodology used in relative rate testing (Sarich and Wilson, 1973; Graur, 2016)) and rescaling the contribution of the diverged species to the K_S-timescale of the focal species (see Supplementary Methods). When adjusted K_S estimates for a divergence event are computed using multiple outgroups, the mean of the adjusted K_S values is used as consensus. The adjusted divergence K_S estimates are then superimposed on the paralog K_S distribution of the focal species to produce an adjusted mixed K_S distribution plot

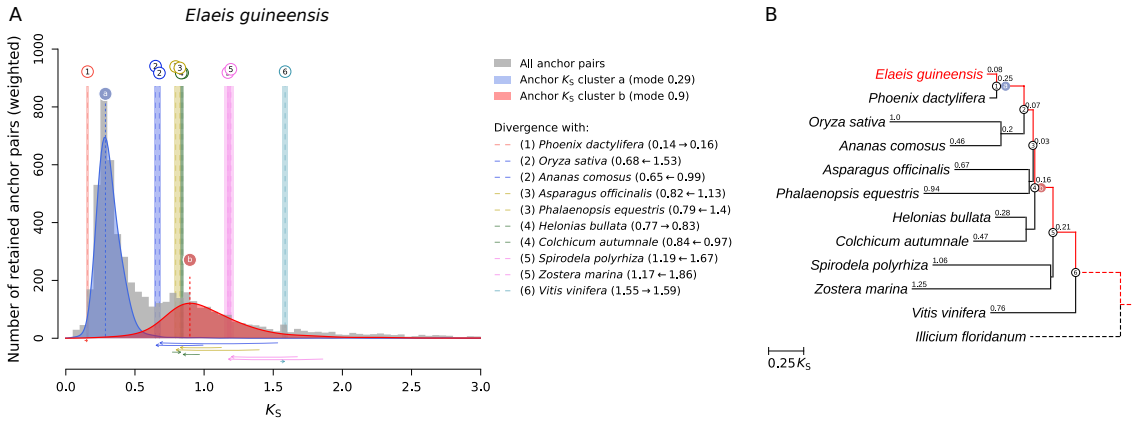

Figure 1. Use case analyzing WGD signatures in monocot plants with oil palm (*Elaeis guineensis*) as the focal species. (**A**) Substitution-rate-adjusted mixed paralog–ortholog K_S plot for oil palm as produced by *ksrates*. The paralog K_S distribution for oil palm anchor pairs is shown in gray, with two putative WGD components inferred from clustering of the median K_S values of collinear segment pairs (see Supplementary Methods) indicated in blue and red. The vertical dashed lines labeled 'a' and 'b' indicate the modes of these components, which are taken as K_S-based WGD age estimates. Rate-adjusted mode estimates of one-to-one ortholog K_S distributions between oil palm and other species, representing speciation events, are drawn as numbered vertical long-dashed lines. Colored boxes around the ortholog K_S mode estimates range from one standard deviation (sd) below to one sd above the mean mode estimate. Lines representing the same speciation event in the phylogeny (numbered circles in (**B**)) share color and numbering. The speciation-line shifts produced by *ksrates*' substitution rate adjustments are indicated by horizontal arrows at the bottom of the plot, and the corresponding shifts in K_S values are given in the legend on the right of the panel. (**B**) Phylogram generated by *ksrates* from the input phylogenetic tree, with branch lengths set to the K_S distances estimated from ortholog K_S distributions. Short branches are not labeled. Lengths of dashed branches cannot be computed. The evolutionary lineage of oil palm to the root of the tree has been highlighted in red. Numbered circles were added manually to indicate the speciation events along this lineage (numbering as in (**A**)). The blue and red filled circles labeled 'a' and 'b' were also added manually to indicate the K_S age estimates of the WGDs inferred in (**A**).

(Fig. 1A). The branch-specific K_S values estimated by *ksrates* are also used to obtain a phylogram for the input phylogeny with branch lengths in K_S units (Fig. 1B), providing a visual representation of the relative substitution rate heterogeneity in the dataset.

3 Use case

Shown in Fig. 1A is a *ksrates* mixed plot of the anchor-pair paralog K_S distribution for the monocot plant *Elaeis guineensis* (oil palm) and substitution-rate-adjusted ortholog K_S peak estimates representing divergence events with other flowering plant species (see Fig. 1B). The two peaks detected in the anchor pair K_S distribution are caused by a known palm-specific WGD (Singh *et al.*, 2013) (blue peak) and a much older WGD (τ, red peak) thought to be shared by all monocots except a few early-diverging clades such as the Alismatales (here represented by *Spirodela polyrhiza* and *Zostera marina*) (Ming *et al.*, 2015). Because palms have a much lower substitution rate than other monocots such as grasses or Asparagales (Barrett *et al.*, 2015, also Fig. 1B), a naive mixed paralog–ortholog K_S plot (Supplementary Fig. 1) however suggests that the older τ WGD would be a second palm-specific WGD event. Moreover, the naive mixed plot contains widely different K_S estimates for the same divergence events and produces an order of divergences inconsistent with the known phylogeny of flowering plants (Byng *et al.*, 2013) (Supplementary Fig. 1). Substitution

rate adjustment with *ksrates* shifts the ortholog K_S estimates of most divergence events leftward towards younger K_S values (see arrows at the bottom of Fig. 1A) and groups estimates of the same event together, resulting in a correct ordering of divergence events and a correct phylogenetic positioning of the two WGDs in the oil palm lineage.

Data availability

The data underlying this article are available in Zenodo, at https://dx.doi.org/10.5281/zenodo.4717851.

Author contributions

R.L. conceived the K_S rate-adjustment methodology. R.L. and S.M. designed tool functionality. C.S. and R.L. implemented the tool. R.L. and S.M. supervised the study. All authors contributed to writing the manuscript.

Acknowledgements

We thank Bert Droesbeke for technical assistance with the Singularity and Docker containers.

Funding

Research in the lab of S.M. is supported by VIB and Ghent University.

Conflict of Interest: none

References

Barrett,C.F. *et al.* (2015) Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. *New Phytol.*, **209**, 855–870.

Blanc,G. and Wolfe,K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, **16**, 1667–1678.

Byng,J.W. *et al.* (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.*, **181**, 1–20.

Graur,D. (2016) *Molecular and Genome Evolution.* Sinauer Associates, Sunderland (Massachusetts).

Li,F.W. *et al.* (2018) Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature Plants*, **4**, 460–472.

Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Ming,R. *et al.* (2015) The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.*, **47**, 1435–1442.

Proost,S. *et al.* (2012) i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.*, **40**, e11.

Sarich,V.M. and Wilson,A.C. (1973) Generation time and genomic evolution in primates. *Science*, **179**, 1144–1147.

Singh,R. *et al.* (2013) Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature*, **500**, 335–339.

Tiley,G.P. *et al.* (2018) Assessing the performance of *Ks* plots for detecting ancient whole genome duplications. *Genome. Biol. Evol.*, **10**, 2882–2898.

Zhang,G.Q. *et al.* (2017) The Apostasia genome and the evolution of orchids. *Nature*, **549**, 379–383.

Zwaenepoel,A. and Van de Peer,Y. (2019) wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics*, **35**, 2153–2155.