Explanation and further reading for the methods in `wgd`

Arthur Zwaenepoel

Introduction

`wgd` is a python package and command line suite for analyzing ancient whole genome duplications (WGDs) in genomic data sets. The main focus of `wgd` is on the construction of whole paranome (the collection of all duplicate genes in a genome) |Ks| distributions, but it also provides tools for the inference of intragenomic synteny nd anchor points as well as downstream analyses and visualizations of whole paranome |Ks| distributions. On this page we provide some background information about the main functionality and methods, namely the inference of these so-called |Ks| distributions.

What is a |Ks| distribution?

|Ks|, also called dS, denotes the synonymous distance between two protein-coding DNA sequences (CDS), where synonymous distance means the expected number of synonymous substitutions per synonymous site. A |Ks| estimate for any two coding sequences can be acquired in two main ways, (1) by means of heuristic counting methods such as the method of Nei & Gojobori (1986) or (2) by using Markov models of codon substitution with Maximum-likelihood (ML) estimation algorithms. The latter methods are most commonly used nowadays, and trace back to the pioneering work of Goldman & Yang (1994) and Muse & Gaut (1994). For more information on the estimation of pairwise synonymous distances, we refer to the wonderful textbooks on molecular evolution authored by professor Ziheng Yang (e.g. Chapter 2 in Yang 2006).

Since synonymous substitutions are presumed to have largely negligible effects on fitness, they are commonly regarded as neutral. Therefore we expect the synonymous distance between two protein coding sequences to increase with time in a stochasti clock like fashion. It is therefore assumed that |Ks| can serve as a proxy for time, and that a larger |Ks| value indicates an older divergence time between two sequences.

A whole paranome |Ks| distribution is nothing more than the distribution of all |Ks| estimates for all inferred duplication events in the genome of some species.

It therefore serves as a proxy for the distribution of the divergence times of all duplication events that have left a trace in the genome of interest. Under the assumption of a constant gene duplication rate and a constant rate of loss of duplicated copies, such a distribution is expected to show an exponential decay shape (Lynch & Conery 2000, Lynch 2007), which is the result of a quasi-equilibrium linear birth-death process. A WGD event entails the duplication of all gene copies in the genome at some point in time (*i.e.* the polyploidization event) and subsequent massive loss or divergence of duplicated copies in a relatively short time frame (*i.e.* the rediploidization phase). This process of WGD followed by rediploidization is expected to leave a large number of duplication events with similar divergence times (and hence similar |Ks| values) that can be traced back in the extant genome of interest. As a result a WGD is expected to leave a peak signature in the |Ks| distribution that cannot be explained by constant rate small-scale duplication as in the model of Lynch & Conery.

As a side note, we note that if we assume synonymous substitution is a Poisson process, and the synonymous substitution rate is Gamma distributed across gene families, this WGD peak signature is expected to follow a negative binomial distribution in the ditribution of numbers of synonymous substititions (which is not the same as |Ks| distribution). In the |Ks| distribution, this would translate to a peak which has a distribution with positive skew, so which could e approximated by a Gamma or Log-Normal distribution. These considerations are important for mixture modeling.

An *ad hoc* illustration of such a hypothetical |Ks| distribution is shown below. This figure shows the hallmarks of duplicate gene demography, namely the exponential decay signature of the continuous birth and death of small-scale duplications, the uniform background of fixed duplicates, and the 'secondary peak' signature of a WGD (here modeled as a log-normal distribution).

How is the |Ks| distribution computed in `wgd`?

How to interpret a |Ks| distribution?

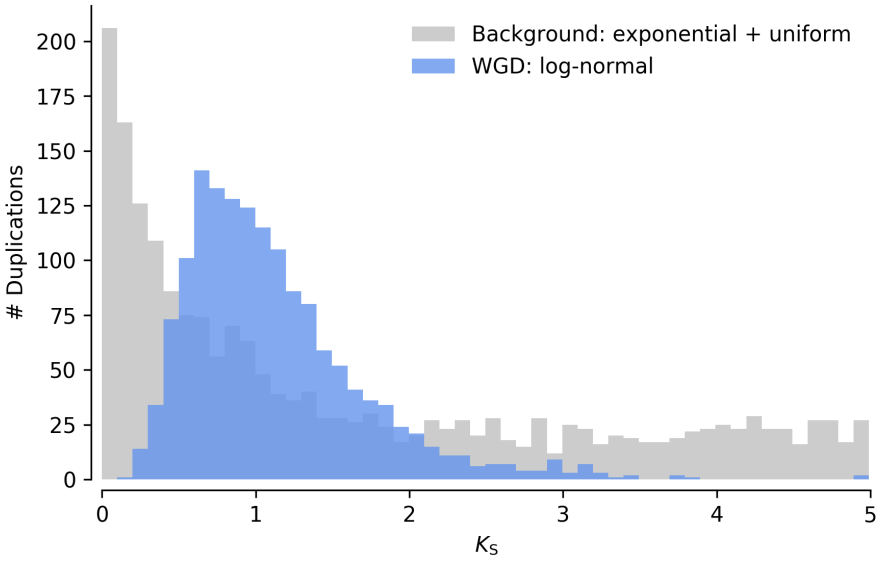

Figure 1: Illustration of the basic principles behind whole paranome |Ks| distributions.