# Comparison of different McDonald and Kreitman approaches using *Drosophila melanogaster* population data

Marta Coronado-Zamora

December 5, 2017

## 1 Introduction

Several tests have been developed to quantify the amount of selection in the genome using divergence and/or polymorphism data.

A limitation of the aforementioned studies is that $\omega$ is not able to disentangle adaptive from non-adaptive substitutions. In fact, non-synonymous substitutions can turn out to be adaptive, neutral or slightingly deleterious (strongly deleterious mutations are purged from the population and so they do not contribute to differences between species). In other words, a high $\omega$ can be the result of either relaxed selection (low conservation), a high adaptive substitution rate, or a combination of both. Since adaptive mutations tend to be fixed quickly (**????**), they will rarely be detected as polymorphic variants but only as a divergent site (that is once fixed). Thus, the adaptive substitution rate in the genome can be inferred when there is an excess of non-synonymous divergence relative to its non-synonymous polymorphism. Accordinly, in the McDonald-Kreitman test (MKT) (**?**) the divergence ratio ($d_N/d_S$) is normalized by the polymorphism ratio ($\pi_n/\pi_s$), which allows taking into account the constraint on non-synonymous sites and, thus, better detect adaptive substitutions (($d_N/d_S$)/($\pi_n/\pi_s$) ¿ 1). From that one can also estimate the proportion of fixed variants that are adaptive ($\alpha$).However, the estimation of $\alpha$ can be biased due to the segregation of slightly deleterious non-synonymous mutations (**?**). Given a stable population size, slightly deleterious mutations can produce an understimation of $\alpha$ because they tend to contribute more to polymorphism than to divergence. Because slightly deleterious substitutions tend to segregate at low frequency, its effect can be partially controlled by removing low-frequency polymorphisms from the analysis (**?**). This procedure still gives biased estimations especially when the REPASAR QUE VOLS DIR PER HIGH LEVEL OF ADAPTATION MARTA level of adaptation is very high (**?**). The integrative MKT (**?**) MARTA CITA was proposed as a better method for correcting the effect of slightly deleterious substitutions. Instead of simply removing low-frequency polymorphisms, the count of segregating sites in non-synonymous sites is separated into the number of neutral variants and the number of weakly deleterious variants, allowing evaluating independently adaptive and weakly deleterious selection (see Methods). The DFE-alpha method is another derivative of MKT, which estimates the Distribution of Fitness Effects

(DFE) of mutations from DNA sequence polymorphism data. The DFE-alpha method corrects for the segregation of slightly deleterious substitutions by first estimating the DFE at selected sites by a gamma distribution and then calculating how many non-adaptive substitutions are expected to become fixed given the inferred DFE from polymorphism data.

# 2 Data

Data from 13,753 genes from a North American population of *Drosophila melanogaster* (Raleigh, North Carolina). Outgroup: *Drosophila simulans*. DAF 20. compute m0, m, p0, p1 with ad-hoc python scripts.

## 2.1 Tests applied

MKT standard
FWW method
DGRP correction
MKT asymptotic
Data simulated with Slim:

# 3 Results

## 3.1 Estimation of alpha using the MK standard

Summary of the method

Out of 13,753 Drosophila melanogaster genes, 11,009 genes are analyzable with the MKT, because they have polymorphic (P0) and divergence sites (di). Alpha mean: -1.275062 (sd:3.96439). From them, only 748 genes are positive and significant (as determined with a Fischer exact test) see table? Figure created.Mean: 0.7874545, sd: 0.1394733.

## 3.2 Estimation of alpha using the FWW correction

Summary of the method

Cutoffs: 0, 0.025, 0.075, 0.125, 0.175, 0.225, 0.275, 0.325
the number of genes available in each cutoff change:
figure:

alpha increments as the cutoff is increased, but the number of alphas significant disminishes. We analyzed the 724 genes that are significant with FWW to compare it with DGRP. we can analyze X genes out of 724. Alpha is significant for X genes. Comparison (t-test between the two tables):

Analyzed the genes of a cutoff in another one.

## 3.3 Estimation of alpha using the DGRP correction

Summary of the method

We can analyze more genes. Contrary to before, Alpha dicreases.
Analyzed the genes of a cutoff in another one.

### 3.4 Estimation of genes with asymptotic (integrative)

Only 118 genes can be analyzed, the ones that doesn't have a 0 in any on their DAF categories.

Summary alfas.

Comparison of the 118 genes between Standard - Li (5%) - DGRP - Asymptotic

## References