Inferring fitness costs from within patient frequencies in HIV

Marion Hartl*†, Kristof Theys *†, Alison Feder, Maoz Gelbart, Adi Stern Pleuni S. Pennings

March 2016

1 Introduction

Mutations are crucial for adaptive processes, including the evolution of drug resistance and immune escape, yet, most mutations are costly. It is important to know how costly mutations are because these costs influence the probability of evolution from standing genetic variation and they they determine the effects of background selection and thus optimal recombination rates. A detailed knowledge of costs of mutations will also help us to discover new functional elements in genomes.

Traditionally, fitness effects are assessed either in in vitro systems (cell culture 1,2) or in a phylogenetic framework 3–5. Here, we use a novel approach to determine the fitness effects of mutations in HIV. HIV has unique properties that allow us to study fitness effects in vivo: it is fast evolving and leads to persistent infections. This means that genetic diversity accumulates quickly and independently in every host, and samples from different patients can thus be treated as independent replicate populations. With many replicate populations it is possible to use frequencies of individual mutations (averaged across populations) to estimate their fitness cost, something that is not possible when only one or a few populations are available.

Deleterious mutations occur in populations naturally and are purged from the population by selection 12. In infinitely large populations, the opposing forces of mutation and selection cause mutations to be present at a constant frequency in the population equal to u/s (where u is the mutation rate from wild-type to the mutant and s is the negative fitness effect of the mutation). This is called mutation-selection balance. In natural populations of finite size, the frequency of mutations is not constant, but fluctuates around the expected frequency of u/s, because of the stochastic nature of mutation and replication 12. Due to these stochastic fluctuations, it is impossible to accurately infer the selection acting on individual mutations from a single observation of a single population. In HIV however, it is possible to sample many independent populations (each patient harbors an independent HIV population 13), something that is impossible for most other evolutionary model systems. The mean frequency of mutations across populations will approach u/s (technically, because the fluctuations of mutation frequencies form an ergodic process 14), which allows us to accurately estimate the selection coefficients of individual mutations. With sufficient sequencing data, we can thus estimate the fitness cost of every point mutation at every position in the HIV-1 genome. In the current study, we focus on transition mutations (A¡¿G and C¡¿T) in the Pol gene because sufficient data are available for these mutations (transition mutations are much more common in HIV than transversions). It is expected that positive and balancing selection are not very important in Pol, so that most mutations should be deleterious, which simplifies our analysis.

We find that estimated costs of individual mutations vary widely between the sites in Pol. Some of this variation can be explained by whether a mutation causes an amino-acid change or not (synonymous mutations are less costly than non-synonymous mutations), and whether the resulting amino-acid change is drastic or not (costs are lower if a mutation leads to a change to a similar amino acid). In addition, we find that the identity of the ancestral nucleotide has a large effect, positions that are A or T in the ancestral state have less costly mutations (when they change to G and C respectively), whereas positions that are C or G ancestrally have more costly mutations (when they change to T and A respecitively).

2 Data

Most of our analysis is done on the Bacheler et al (2000) dataset. In addition, we look at the Lehman et al data and the Zanini et al data.

Description of the data / filtering. The Bacheler et al data is cloned and sequenced (median 18 sequences per patient, 171 patients), so it is of high quality. For each patient, we lump together all sequences, even though they come from different time points. We don't do much filtering, except that 1. we remove sites for which more than 10 percent of the patients have a majority of a non-WT nucleotide at the first time point and 2. we remove sites from specific sequences from specific patients if there is a mutation in one or both of the other two sites that make up a triplet. This last filtering is done because if there

*These authors contributed equally and are listed in reverse alphabetical order
†Department of Biology, San Francisco State University, San Francisco, CA 94132

is a mutation elsewhere in the triplet, it is no longer clear whether a given mutation should be counted as synonymous or non-synonymous.

The Lehman data is 454 sequences, much deeper, but higher error rate, there are multiple time points for most patients, but we only use one time point per patient (most of the time this is one month after seroconversion). The sequences span around 600 sites in RT. Because the Lehman data are different HIV subtypes (mostly C and A), we only consider sites that are conserved between A, B and C.

The Zanini data is Illumina sequenced, multiple time points per patient. Time points are usually a few months or more apart and are treated as independent samples. Though the Zanini data are genome wide, we only consider the sites for which we also have data from the Bacheler data. Because the Zanini data are different HIV subtypes (B and C), we only consider sites that are conserved between B and C.

For all three datasets, we only consider transition mutations because they are more common in HIV than transversion mutations. For example, for a site that is A in the ancestral state, the frequency of a transition mutation is calculated for each patient (and each time point in the case of the Zanini data) as the number of sequences that carry a G divided by the number of sequences that carry a G or an A. These frequencies are then averaged over all patients (and time points in the case of the Zanini data). Sequences with a C or a T are thus not considered if the ancestral state is A. Selection coefficients are estimated by dividing the nucleotide-specific mutation rate by the average frequency, bootstrapping is done over sites (not patients). For the Generalized Linear Model, actual counts are considered as opposed to frequencies.

3 Result 1. Mean frequencies of non-synonymous site mutations are lower than mean frequencies of synonymous site mutations

Because we only consider transition mutations, there is exactly one possible mutation per site. We split all sites/mutations in three categories: synonymous (these do not change the amino acid), non-synonymous (these do change the amino acid to another amino acid) and stop-creating mutations (these mutations create a premature stop codon).

We find that the three categories of mutations show clearly different distributions of frequencies in the Bacheler dataset (Kolmogorov-Smirnov test, p-value ¡ 2.2e-16 for stop vs. non-synonymous sites and for non-synonymous vs. synonymous sites), see figure 1. This result is not surprising, but it is reassuring that the three categories of sites show clearly different frequencies.

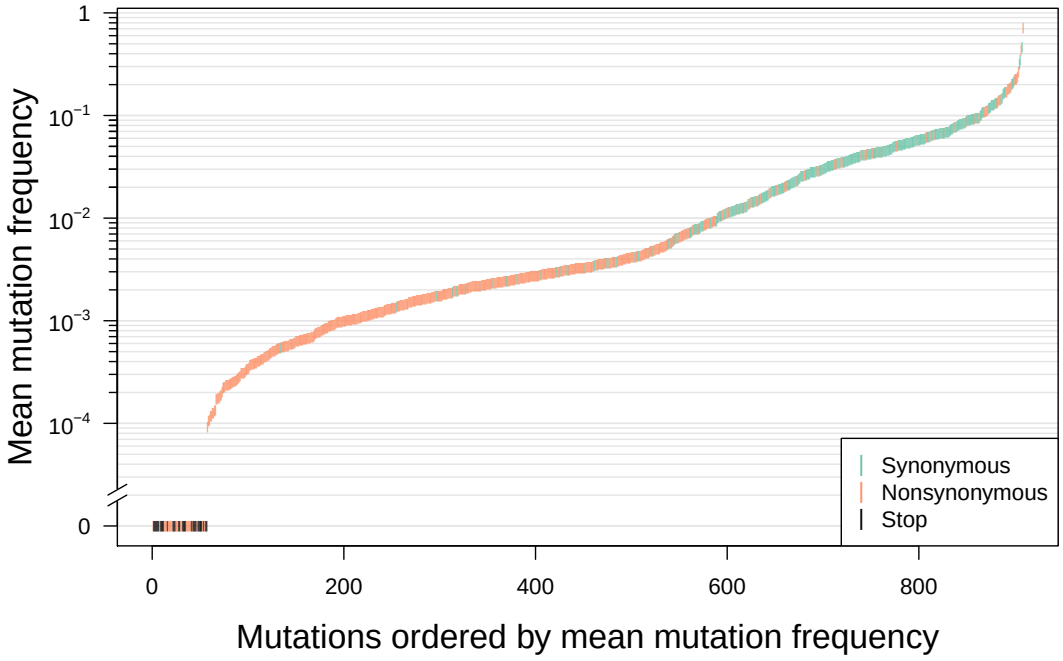

Figure 1: Mutation frequency for all sites from the Bacheler data, ordered by mutation frequency. Mutations that cause a premature stop codon are shown in black, mutations that cause an amino acid change (non-synonymous) are shown in red and mutations that do not change the amino acid (synonymous) are shown in green.

4 Result 2. Nucleotide identity, type of amino-acid change and SHAPE parameter predict selection coefficients of mutations

We used a Generalized Linear Model to determine which characteristics of sites can explain the observed frequencies of mutations. We then use estimated mutation rates from other studies (Abrams 2010, Zanini 2016) to translate frequencies into selection coefficients.

We find that nucleotide identity has a surprisingly large effect on the observed frequencies. Specifically, non-synonymous mutations at sites that are A or T ancestrally, are seen at higher frequencies (indicating lower selection coefficients) whereas non-synonymous mutations at sites that are C or G ancestrally, are seen at lower frequencies (indicating higher selection coefficients). For synonymous sites, the result is not as clear, although there appears to be an effect of G when compared to the other three nucleotides (see figure 2).

It is unclear to us what causes this nucleotide identity effect. In principle, this effect could be caused by spurious mutation rate estimates, but we find the same results if we use mutation rates from the Abrams paper or the Zanini 2016 paper. Also, the difference in mean selection coefficient between C sites and (A, T) sites is almost an order of magnitude. It seems unlikely that current mutation rate estimates are so far off.

We also find that the type of amino acid change caused by the mutation in non-synonymous mutations has a strong effect on observed frequencies. Amino acids were classified into five groups: positively charged, negatively charged, uncharged, hydrophobic and special cases (Cysteine, Glycine and Proline). Mutations causing drastic amino acid changes, so that the mutant amino acid belonged to a different group than the original amino acid were associated with lower frequencies, and thus higher selection coefficients.

Finally, we find that mutations at sites with lower SHAPE parameters tend to have lower mutation frequencies, thus higher selection coefficients. SHAPE contains information about the RNA folding of the position: positions that has lower shape score tend to create secondary RNA structures by being base-paired, which means that a mutation at such a site may be costly, even if the amino acid is not changed or not drastically changed.

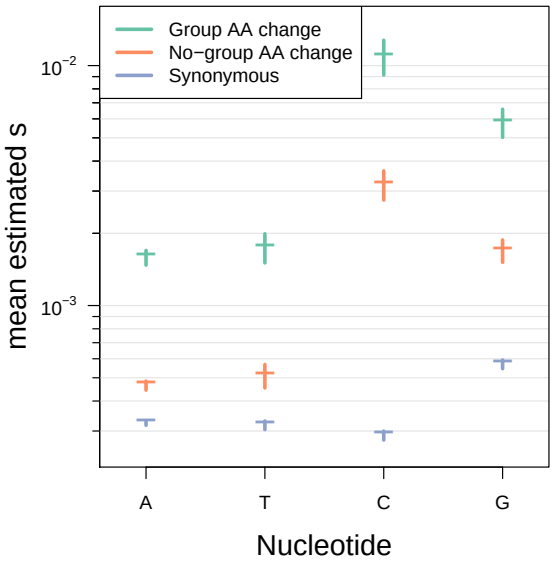

Figure 2: DFE for syn sites Bacheler.

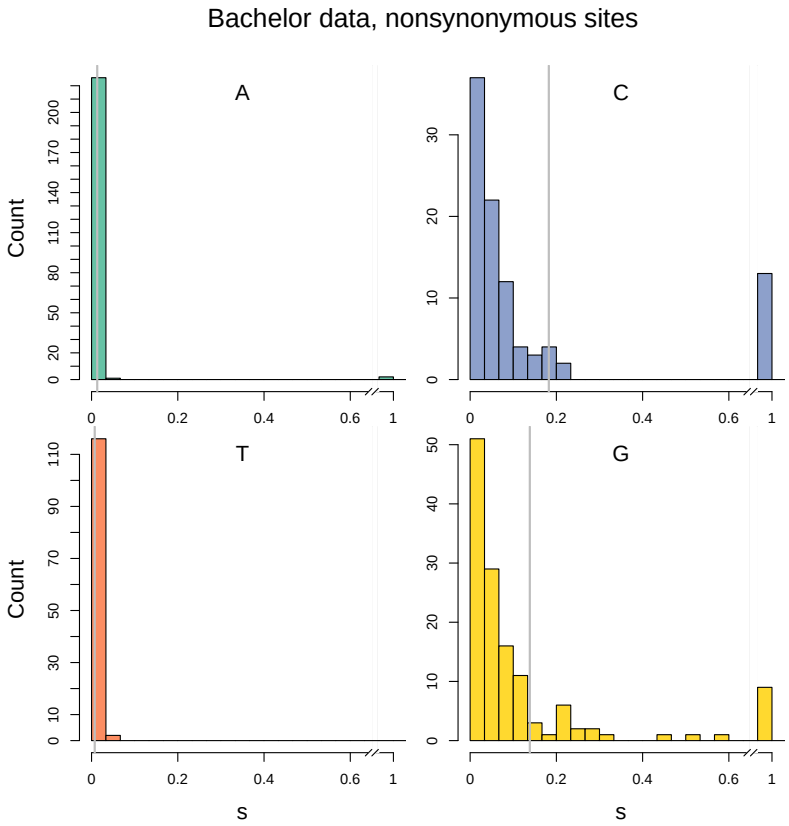

Figure 3: Estimated selection coefficients from GLM and mutation rates from Abrams et al 2010.

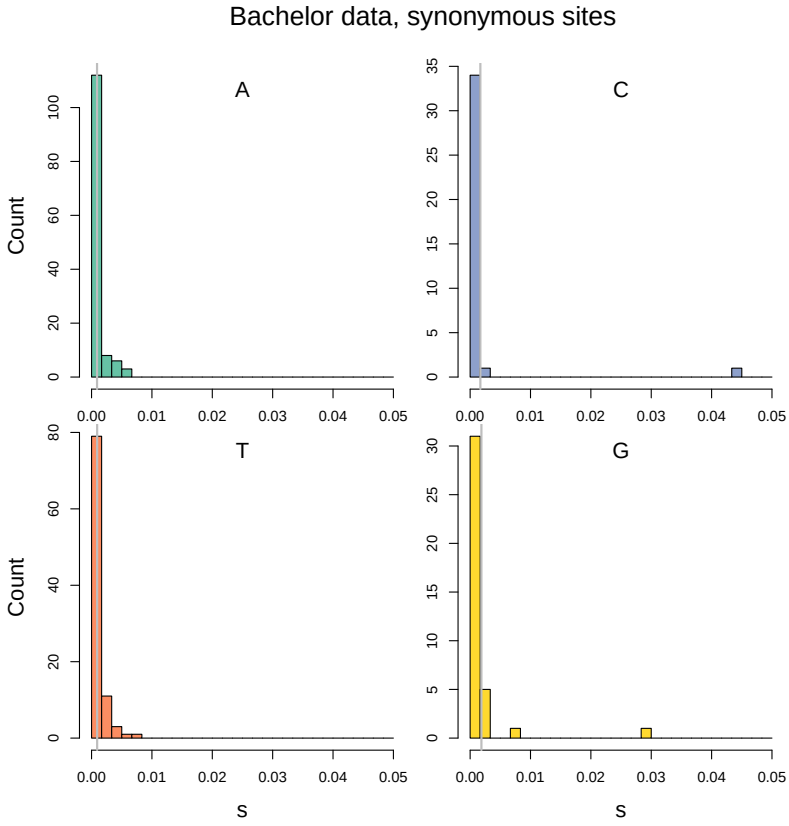

Figure 4: DFE for non-syn sites Bacheler.

5 Result 3. Parameters for Gamma distribution for distribution of fitness effects

		Abram 2010		Zanini 2016		Proportion
	sites	shape	scale	shape	scale	very deleterious
all	909	0.358	0.054	0.407	0.017	0.063
		(0.338, 0.382)	(0.044, 0.066)	(0.384, 0.435)	(0.014, 0.021)	(0.047, 0.079)
PR	288	0.352	0.059	0.402	0.019	0.062
		(0.321, 0.393)	(0.04, 0.082)	(0.365, 0.454)	(0.013, 0.024)	(0.035, 0.09)
RT	621	0.36	0.052	0.409	0.017	0.063
		(0.336, 0.392)	(0.039, 0.067)	(0.381, 0.444)	(0.013, 0.021)	(0.045, 0.084)
a	358	0.751	0.004	0.751	0.002	0.006
		(0.681, 0.848)	(0.003, 0.005)	(0.678, 0.845)	(0.002, 0.002)	(0, 0.014)
t	213	0.652	0.007	0.652	0.005	0
		(0.579, 0.748)	(0.005, 0.009)	(0.582, 0.755)	(0.004, 0.006)	(0, 0)
c	150	0.422	0.094	0.422	0.043	0.2
		(0.364, 0.511)	(0.072, 0.117)	(0.366, 0.51)	(0.033, 0.053)	(0.14, 0.267)
g	188	0.434	0.136	0.434	0.037	0.133
		(0.382, 0.517)	(0.099, 0.176)	(0.382, 0.511)	(0.028, 0.048)	(0.085, 0.181)

Table 1: **Gamma distribution parameters fit to observed DFE.** Gamma distribution parameters shape and scale are estimated from all loci with an observed minor transition using the Subplex algorithm implemented in the *nloptr* package. The proportions of loci with observed minor allele frequency of 0 are recorded as very deleterious. Bootstrapping is performed across sites.

5.1 The relationship between mean frequency and worldwide conservation of the site

Do our selection coefficient estimates predict conservation on a phylogenetic scale?

5.2 No effect of spatial location within POL

Not sure whether to keep this or not. Marion randomization test. Result: it doesn't look like there is an effect. Marion will redo test with 1000 runs and get p-value.

6 Conclusion

I think we want to conclude that frequencies can help us find out the cost of individual mutations. Unsurprising effects of syn non syn, and amino acid change, shape parameter. Surprising effects of nt identity.

Adi, you asked about: "I would have actually expected getting a much larger percentage of lethal or near lethal (which would be round 1E-05 or 1E-04). For most RNA viruses 10-40 perc of mutations are lethal (see here for example). " We don't know much about how the data was filtered prior to it being uploaded to genbank, so it could be that some of the low freq. muts are actually sequencing mistakes in which case the mut would possibly be lethal. This should def. go in the discussion!!

7 Supplementary figures

Figure 5: Mutation frequency for all available sites from the Lehman data, ordered by mutation frequency. Mutations that cause a premature stop codon are shown in black, mutations that cause an amino acid change (non-synonymous) are shown in red and mutations that do not change the amino acid (synonymous) are shown in green.

Figure 6: Mutation frequency for 984 Pol sites from the Zanini data, ordered by mutation frequency. Mutations that cause a premature stop codon are shown in black, mutations that cause an amino acid change (non-synonymous) are shown in red and mutations that do not change the amino acid (synonymous) are shown in green.

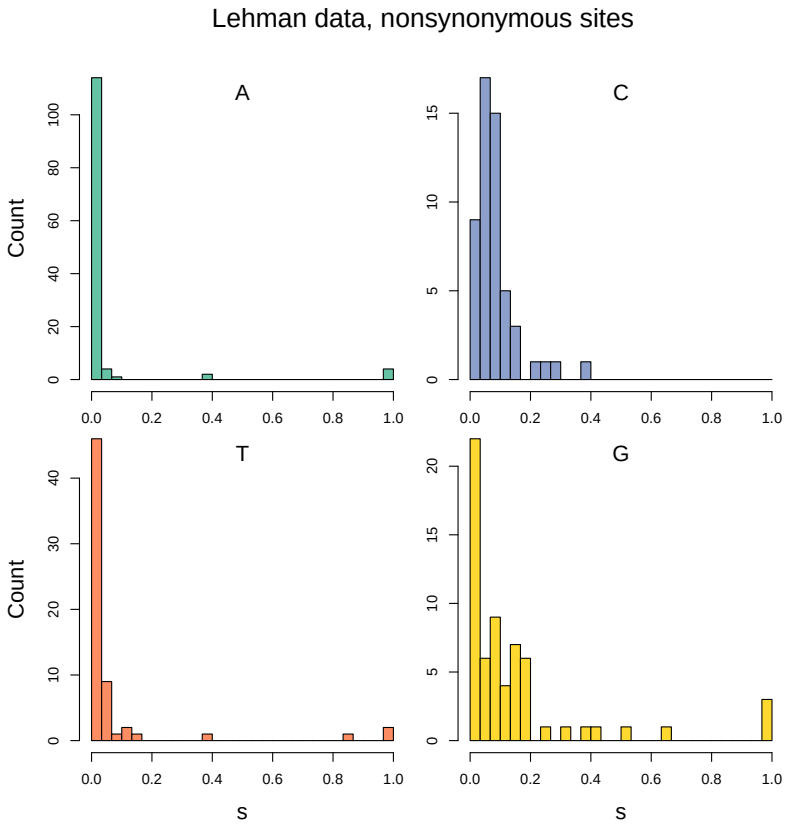

Figure 7: DFE for non-syn sites Lehman.

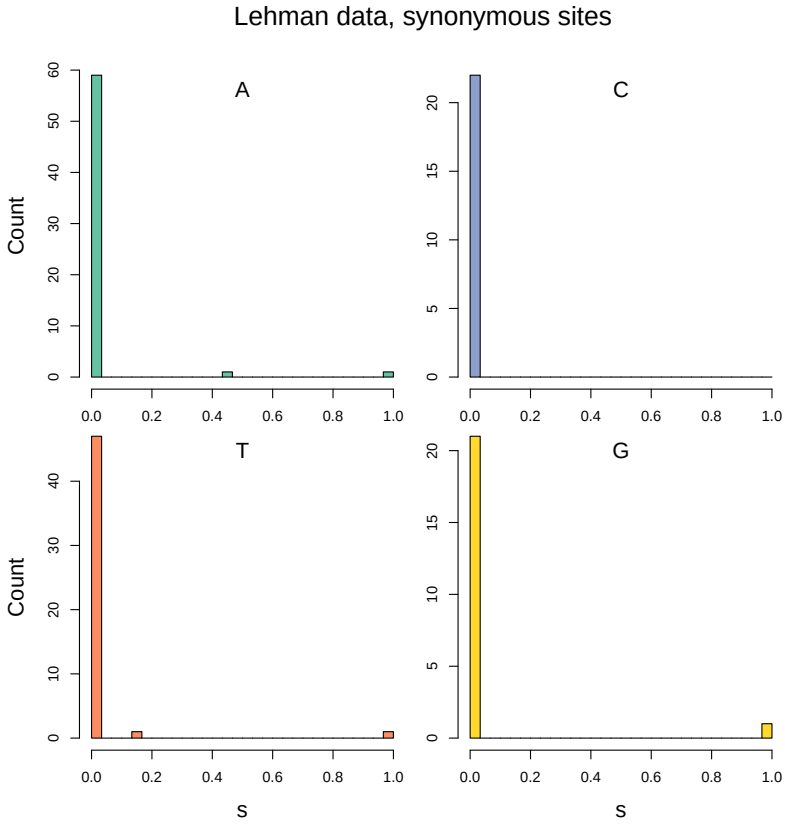

Figure 8: DFE for syn sites Lehman.

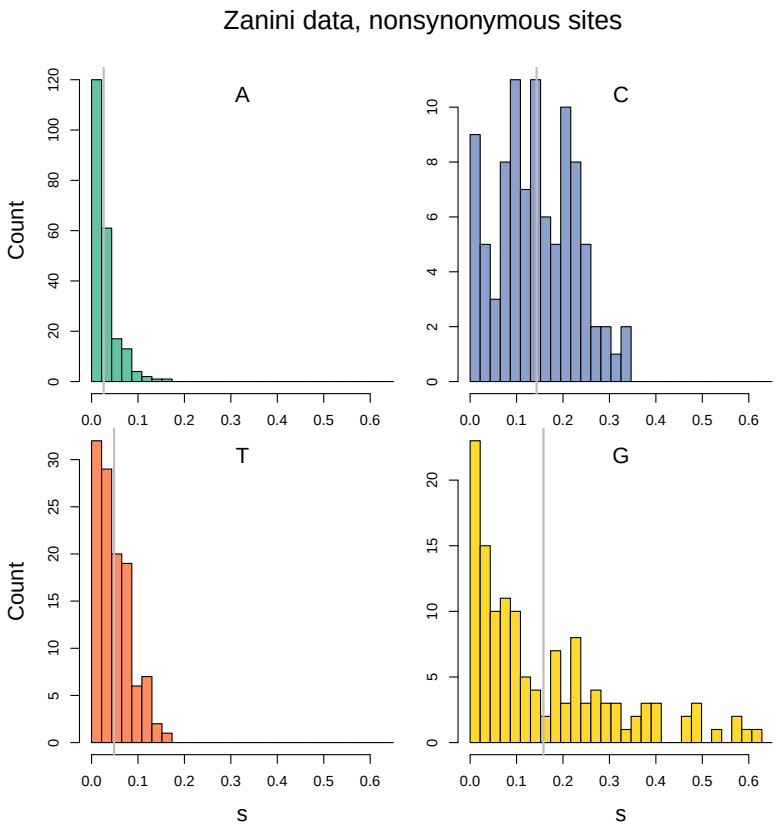

Figure 9: DFE for non-syn sites Zanini.

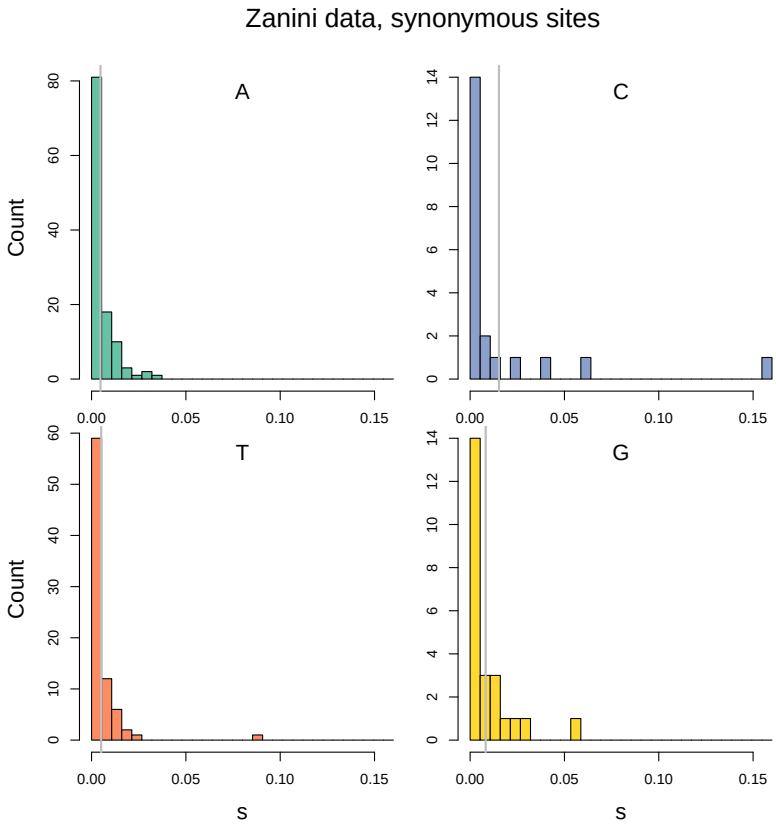

Figure 10: DFE for syn sites Zanini.