OPEN ACCESS Freely available online

PLoS one

Statistical Inference of Selection and Divergence from a Time-Dependent Poisson Random Field Model

Amei Amei[1]*, Stanley Sawyer[2]

1 Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, Nevada, United States of America, **2** Department of Mathematics, Washington University in St.Louis, St.Louis, Missouri, United States of America

Abstract

We apply a recently developed time-dependent Poisson random field model to aligned DNA sequences from two related biological species to estimate selection coefficients and divergence time. We use Markov chain Monte Carlo methods to estimate species divergence time and selection coefficients for each locus. The model assumes that the selective effects of non-synonymous mutations are normally distributed across genetic loci but constant within loci, and synonymous mutations are selectively neutral. In contrast with previous models, we do not assume that the individual species are at population equilibrium after divergence. Using a data set of 91 genes in two *Drosophila* species, *D. melanogaster* and *D. simulans*, we estimate the species divergence time $t_{\mathrm{div}} = 2.61N_e$ (or 1.68 million years, assuming the haploid effective population size $N_e = 6.45 \times 10^5$ years) and a mean selection coefficient per generation $\mu_\gamma = 1.98/N_e$. Although the average selection coefficient is positive, the magnitude of the selection is quite small. Results from numerical simulations are also presented as an accuracy check for the time-dependent model.

Citation: Amei A, Sawyer S (2012) Statistical Inference of Selection and Divergence from a Time-Dependent Poisson Random Field Model. PLoS ONE 7(4): e34413. doi:10.1371/journal.pone.0034413

Editor: Attila Szolnoki, Hungarian Academy of Sciences, Hungary

Received September 21, 2011; **Accepted** February 27, 2012; **Published** April 3, 2012

Funding: The authors have no funding or support to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: amei.amei@unlv.edu

Introduction

Mutation, selection, and genetic drift are important forces that shape pattern of genetic polymorphism within and between species [1]. McDonald and Kreitman first used a 2×2 contingency table to test differences in selection between silent and amino acid replacement sites [2]. Data from the *Adh* gene encoding alcohol dehydrogenase in *Drosophila* suggested that adaptive fixation of selectively advantageous mutations was the cause of a statistically significant number of excess replacement substitutions. A quantitative theory for the amount of selection between two recently diverged species was developed by Sawyer and Hartl [3], who developed a sampling theory in which the set of frequencies of mutant sites is modeled as a Poisson random field (PRF). This theory was applied to the sample configurations of nucleotides in the *Adh* gene in *Drosophila* and led to maximum likelihood estimates of silent and replacement mutation rates, an average selection coefficient, and the species divergence time. Bayesian methods have proven useful for data sets with multiple genetic loci. Bustamante et al. [4] introduced a hierarchical Bayesian fixed effects model in which selective intensities of new replacement mutations are constant within genetic loci, but are normally distributed across genes. Application of this model using Markov chain Monte Carlo (MCMC) simulations yielded evidence for predominantly beneficial gene substitutions in *Drosophila* but detrimental substitutions in the mustard weed *Arabidopsis*. Sawyer et al. [5,6] extended this model to a Bayesian random effects model in which selective effects of non-synonymous mutations are normally distributed within genes, while, as in Bustamante et al. [4], within-locus means are normally distributed across genetic loci. Abel [7] considered similar models with more heavy-tailed distributions within loci (specifically, Laplace and chi-square distributions) and found similar numerical results.

Although the PRF model of Sawyer and Hartl [3,5,6] provides robust estimates [7–12] for mutation and selection parameters, numerical simulations have shown that estimate of species divergence time is somewhat biased, particularly for small divergence time [7]. This is due to the model assumption that the two species are individually at mutation-selection-drift equilibrium after divergence. Recently, we have derived a "time-dependent" PRF model that removes this equilibrium assumption [13] (see the next section for details). Williamson et al. [12] proposed a time-inhomogeneous PRF model to make inference about constant selection and population growth simultaneously. They applied the model to site frequency spectrum of 301 human genes and showed a strong evidence for recent population growth. Later, Boyko et al. [9] extended the site-frequency spectrum based PRF approach to allow for simultaneous inference of demography and a distribution of fitness effects among newly arising mutations. The application of their method to a Single Nucleotide Polymorphism (SNP) data of 20 European Americans and 15 African Americans showed evidence of an ancient population expansion in the sample of African population and a relatively recent bottleneck in the sample of European American population. Given the estimates of demographic parameters, they made inference of the distribution of the selection effects. Both studies are based on maximum likelihood methods and only applied to a single population. In order to make inference about both selective effects and species divergence time, we developed a hierarchical Bayesian framework for sample configuration formulas derived

from the time-dependent PRF model that contrasts the number of silent and replacement polymorphisms within species with that of fixed differences between species. We applied the model to 91 genes in African populations of *D. melanogaster* and *D. simulans* (Pröschel et al. [14]) and find that a large proportion of newly arising amino acid replacement mutations observed as polymorphisms are subject to weak positive selection. The model is first tested on a set of simulated data. Estimates of mutation and selection parameters are reasonably accurate. In particular, the point estimate of species divergence time matches the true parameter value almost perfectly for each simulated data. This shows the power of the time-dependent model in estimating species divergence time.

Methods

A 2×2 contingency table consisting of the number of fixed differences and polymorphisms at silent and replacement sites is called a MacDonald-Kreitman table and also a DPRS table. Assuming time equilibrium and independence among nucleotide sites, or equivalently linkage equilibrium, the four entries in a DPRS table can be regarded as independent Poisson random variables whose expected values can be derived from the fixation flux and limiting distribution of polymorphic nucleotide substitutions [3]. In the time-dependent case, we define two types of polymorphisms [3]. A site is a *legacy polymorphism* if the ancestors of the sequences in the DNA alignment were polymorphic at the time of divergence. The site is a *new polymorphism* if the polymorphism is caused by one or more mutations since the time of divergence. New polymorphisms can only show up in one species while legacy polymorphic sites can be polymorphic in one or both species. A natural extension of the DPRS table is a 2×3 contingency table, called the DOHRS table, that has columns for two different types of polymorphisms. Specifically, we define K_s as the number of silent sites that are fixed differences between a pair of species in a sample (that is, monomorphic within samples but polymorphic between samples), O_s as the number of silent sites that are polymorphic in only one sample, and H_s as the number of silent sites that are polymorphic in both samples [13]. We use K_r, O_r, and H_r as the corresponding counts for amino acid replacement sites. Let m and n denote the number of aligned DNA sequences from the two species (which we assume to have the same haploid effective population size N_e) and let t_{div} be the scaled divergence time since the time that the two populations diverged. For each locus, we use θ_s and θ_r to represent, respectively, the scaled synonymous and non-synonymous mutation rates and $\gamma_r=\gamma$ the scaled selection coefficient of a non-synonymous mutation. Synonymous mutations are assumed to be selectively neutral, i.e. $\gamma_s=0$, and unaffected by hitchhiking and other linkage-mediated effects. The parameters t_{div}, θ_{si}, θ_{ri}, and γ_i (where we now indicate the locus explicitly) are all scaled in terms of the haploid effective population size N_e [13]. Assuming independence among sites, constant and equal effective population sizes N_e for both species, and no migration between species, the counts K_s, O_s, H_s, K_r, O_r, and H_r are independent Poisson random variables with means given, in a united form, by

$$E(K)=\frac{\theta}{s(1)}\Big(\int_0^1 (I(x,m)K(x,n)+I(x,n)K(x,m))(s(1)-s(x))m(dx) + 2\Big(t-\int_0^t\int_0^1 \Big(\lim_{x\to 0}\frac{p(u,x,y)}{s(x)}\Big)s(y)m(dy)du\Big)+L(m)+L(n)\Big) \tag{1}$$

$$E(O)=\frac{\theta}{s(1)}\int_0^1 (2-x^m-(1-x)^m-x^n-(1-x)^n-2J(x,m)J(x,n))(s(1)-s(x))m(dx) \tag{2}$$

$$E(H)=\frac{\theta}{s(1)}\int_0^1 J(x,m)J(x,n)(s(1)-s(x))m(dx) \tag{3}$$

where

$$I(x,m)=\frac{s(1)-s(x)}{s(1)}-\int_0^1 p(t,x,y)\Big(1-(1-y)^m-\frac{s(y)}{s(1)}\Big)m(dy)$$

$$J(x,m)=\int_0^1 p(t,x,y)(1-y^m-(1-y)^m)m(dy),\quad J(x,1)=0$$

$$K(x,m)=\frac{s(x)}{s(1)}+\int_0^1 p(t,x,y)\Big(y^m-\frac{s(y)}{s(1)}\Big)m(dy)$$

$$L(m)=\int_0^1 x^m(s(1)-s(x))m(dx)-\int_0^1\int_0^1 p(t,x,y)y^m(s(1)-s(x))m(dy)m(dx)$$

with $s(x)=(1-e^{-\gamma x})/\gamma$, $m(dx)=e^{\gamma}dx/(x(1-x))$ for replacement sites and $s(x)=x$, $m(dx)=dx/(x(1-x))$ for silent sites.

The function $p(t,x,y)$ in these expressions is a smooth symmetric function of its arguments such that, for any continuous function $f(x)$ on $0\le x\le 1$, the integral $u(t,x)=\int_0^1 p(t,x,y)f(y)m(dy)$ is the solution of the following diffusion equation

$$\frac{\partial}{\partial t}u(t,x)=L_x u(t,x),\qquad t>0,\ 0<x<1 \tag{4}$$

$$u(t,0)=u(t,1)=0,\qquad u(0,x)=f(x)$$

where $L_x=x(1-x)d^2/dx^2+\gamma x(1-x)d/dx$ (see details in [13]).

At each locus, the theoretical expectations Eqs.(1)–(3) of the six Poisson counts (K_s, O_s, H_s, K_r, O_r, and H_r) for a single DOHRS table depend on four parameters (t_{div}, θ_s, θ_r, and γ), where t_{div} is a global parameter shared by all loci and the rest are locus specific parameters. The goal of this study is to use Bayesian methods to estimate genetic parameters based on a set of DOHRS tables of aligned gene sequences from a pair of closely related species. We assume that all non-synonymous mutant nucleotides at the ith locus have the same selection coefficient γ_i. Across loci, the γ_i are normally distributed with mean μ_γ and variance σ^2. In our Bayesian framework (as in [4]), we assign an inverse-gamma-normal distribution as a joint prior distribution of the mean μ_γ and variance σ^2, gamma distributions with given parameters as prior distributions of the two types of mutation rates θ_{si} and θ_{ri}, and a uniform distribution for the divergence time t_{div}. In standard Bayesian notation,

$$\frac{1}{\sigma^2} \sim \Gamma(\alpha_0,\beta_0) \tag{5}$$

$$\mu_\gamma \sim N(\mu_0,\sigma^2/n_0) \tag{6}$$

$$\theta_{s,i} \sim \Gamma(\alpha_s,\beta_s) \tag{7}$$

$$\theta_{r,i} \sim \Gamma(\alpha_r,\beta_r) \tag{8}$$

$$t_{\rm div} \sim U(0,T) \tag{9}$$

All hyperparameters α_0, β_0, α_s, β_s, α_r, β_r, μ_0, and n_0 are chosen to be small (~ 0.001) so as to be "uninformative" and T is a fixed large value. The full likelihood, based on the sampling formulas in Eqs.(1)–(3) and the prior distributions in Eqs.(5)–(9), is given explicitly by

$$\begin{aligned} & L(\mu_\gamma,\sigma,\gamma_i,\theta_{si},\theta_{ri},K_{si},O_{si},H_{si},K_{ri},O_{ri},H_{ri}) \\ = \;& \Pi_{i=1}^{N}\{\phi(\gamma_i,\mu_\gamma,\sigma)\Gamma(\theta_{si},\alpha_s,\beta_s)\Gamma(\theta_{ri},\alpha_r,\beta_r) \\ & \times Poi_1(\theta_{si},0,t_{\rm div},K_{si},m_i,n_i)Poi_2(\theta_{si},0,t_{\rm div},O_{si},m_i,n_i) \\ & Poi_3(\theta_{si},0,t_{\rm div},H_{si},m_i,n_i)\times Poi_1(\theta_{ri},\gamma_i,t_{\rm div},K_{ri},m_i,n_i) \\ & Poi_2(\theta_{ri},\gamma_i,t_{\rm div},O_{ri},m_i,n_i)Poi_3(\theta_{ri},\gamma_i,t_{\rm div},H_{ri},m_i,n_i)\} \\ & \times \Gamma(\frac{1}{\sigma^2},\alpha_0,\beta_0)\phi(\mu_\gamma,\mu_0,\frac{\sigma}{\sqrt{n_0}})u(t_{\rm div},0,T) \end{aligned} \tag{10}$$

where N is the number of loci, $\phi(y,\mu,\sigma)$, $\Gamma(y,\alpha,\beta)$, and $u(y,0,T)$ are respectively normal, gamma, and uniform densities, and

$$Poi_j(\theta,\gamma,t,c_j,m,n)=\frac{e^{-\lambda_j}(\lambda_j)^{c_j}}{c_j!} \quad j=1,2,3; \quad \begin{cases} c_1=K,\lambda_1=E(K) \\ c_2=O,\lambda_2=E(O) \\ c_3=H,\lambda_3=E(H) \end{cases}$$

Integrals involving the transition density $p(t,x,y)$ are estimated by Crank-Nicolson method ([12,15]). Gauss-Legendre quadrature [15] is used for all other integrals. Finally, the posterior distributions of the genetic parameters given the Poisson counts in the DOHRS tables are obtained by means of Markov chain Monte Carlo simulations. In the implementation of MCMC simulations, convergence is assessed using traceplots as well as Gelman-Rubin statistics <1.01 [16].

Results

Behavior on simulated data

We simulated 23 data sets each containing 30 genes as follows. For each data set, fixed values were assigned to the global parameters μ_γ, σ^2, and $t_{\rm div}$. The locus specific parameters γ, θ_s, θ_r, m, and n were generated from probability distributions. Specifically, at the ith locus, the selection coefficients γ_i was sampled from the normal distribution with mean μ_γ and variance σ^2, the two types of mutation rates θ_{si} and θ_{ri} were drawn from two continuous uniform distributions with given ranges, and the number of DNA alignments m_i and n_i were taken from two discrete uniform distributions with specified ranges. For each locus, expected values for the Poisson counts K_s, O_s, H_s, K_r, O_r, and H_r were calculated using Eqs.(1)–(3) with the given parameters. We then sampled six numbers from the Poisson distributions with calculated means to make up entries of each DOHRS table. Each simulated data set has 30 DOHRS tables.

As a check of accuracy of the time-dependent PRF model, estimated values of the parameters for each data set were compared with the given values. As shown in Figure 1, estimates of μ_γ and $t_{\rm div}$ lie closely to their given values. The differences between estimates and true values for σ are small for small values of σ and increase as σ goes large. The σ estimates may get improved by increasing the number of loci contained in each data set. Estimation errors of locus specific parameters are presented in Figure 2. These are histograms of $\gamma_{ij}-\hat{\gamma}_{ij}$, $\theta_{s,ij}-\hat{\theta}_{s,ij}$, and $\theta_{r,ij}-\hat{\theta}_{r,ij}$ for $1\le i\le 23$ and $1\le j\le 30$ respectively. The results show that the point estimates for the two types of mutation rates are quite accurate. For the selection coefficients, the 95% posterior credible intervals obtained via MCMC algorithm cover the true parameters most of the time though the point estimates look less precise. Note that each γ_{ij} estimate is based on a single DOHRS table.

Results on real data

We next applied our method to data of Pröschel et al. [14]. This consists of the coding sequences of $n=7$ to $n=12$ alleles of each of 91 autosomal genes in *Drosophila melanogaster* collected from a population near Lake Kariba, Zimbabwe. A single highly-inbred line of *D. simulans* ($m=1$) from Chapel Hill, North Carolina was used as a comparison of divergence [17]. After disregarding the first 20,000 burn-in iterations of MCMC simulations, estimates of parameters are obtained from 10,000 samples taken every 10 iterations. Scaled to the haploid population size, point estimates (median) and 95% credible intervals for the global parameters are $\mu_\gamma=1.98$ (0.89,3.37), $\sigma=3.44$ (2.39,4.77), and $t_{\rm div}=2.61$ (2.41,2.87). Selection coefficients for the 91 genes are estimated by medians of their posterior distributions. The medians and corresponding 95% credible intervals appear in Figure 3, with the loci sorted by the medians.

Among the 91 *Drosophila* genes, 73 have their median $\gamma>0$ and 13 credible intervals are entirely positive (do not overlap 0). Although the mean amino acid replacement mutation that could contribute to polymorphism or divergence in *Drosophila* is beneficial, the magnitude of the selective intensity is small. Based on our estimates, 48% of the non-synonymous mutations have $\gamma<1$, 84% have $\gamma<3$, and 99% have $\gamma<5$. Assuming a haploid effective population size of $N_e=0.645\times 10^6$ years [3], our estimate of $t_{\rm div}=2.61$ implies a species divergence time of 1.68 Myr (million years) between *D. melanogaster* and *D. simulans*. This value is almost in the middle of a range 0.8–3Myr [18,19]. In contrast, the time-independent fixed effects model of [4] estimates 4.46 (median) with 95% credible interval (4.06, 5.00) for this data set and the time-independent random effects model of [5,6] yields 4.47 and (4.06, 4.93).

Based on the difference of gene expression level between males and females (or between testes and ovaries), Pröschel et al. divided the data set into 33 male-biased, 28 female-biased, and 30 sex-unbiased genes [14]. We applied the time-dependent model to the three types of genes and means and standard deviations of the posterior distributions of the scaled selection coefficients are presented using their medians and 95% credible intervals. They are $\mu_M=2.98$ (2.02,4.26), $\sigma_M=0.17$ (0.05,1.94), $\mu_F=1.70$ (-0.26,4.50), $\sigma_F=3.98$ (2.12,7.99), and $\mu_{Un}=0.37$

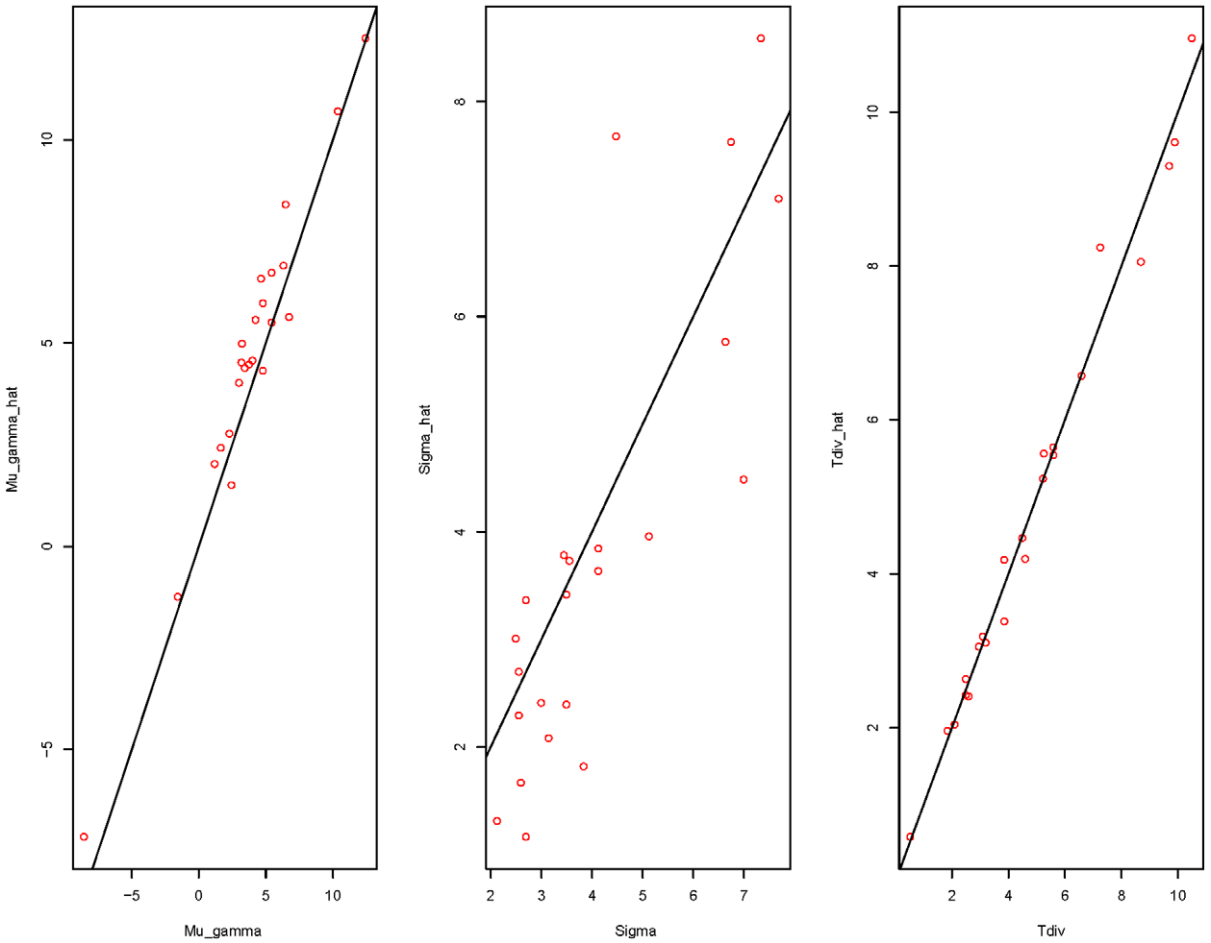

Figure 1. Comparisons of the estimated parameters with their corresponding true values based on 23 simulated data sets. Three plots are $\hat{\mu}_\gamma$ vs μ_γ, $\hat{\sigma}$ vs σ, and $\hat{t}_{\rm div}$ vs $t_{\rm div}$ respectively. The selection coefficient γ is assumed to be normally distributed with mean μ_γ and variance σ^2 and $t_{\rm div}$ is the species divergence time.
doi:10.1371/journal.pone.0034413.g001

$(-1.41,3.19)$, $\sigma_{Un}=3.24$ $(0.86,6.43)$ for male-biased, female-biased, and sex-unbiased genes respectively. Selection coefficients for individual genes of the three types are presented side by side, in Figure 4, using their median estimates and 95% credible intervals. According to our estimates, there is strong evidence that positive selection occur more often among sex-biased genes (both male and female) than among sex-unbiased genes. Specifically, the selection coefficients γ_i for male-biased genes, with an estimated normal distribution of mean 2.98 and standard deviation 0.17, show an almost uniform signal of adaptive selection. However, since we do not have information about linkage disequilibrium between these genes, we cannot exclude that this is simply a consequence of linkage between genes. In contrast, female-biased genes experience more variance in the direction of selection based on their estimated selection coefficients which vary from -4.52 to 6.70. On average, the selective effect for sex-unbiased genes are nearly neutral with a moderate size of variation $(-4.68\sim3.18)$. In their original paper, Pröschel et al. estimated average strength of selection for non-synonymous mutations within each group of genes using a time-independent fixed effects PRF model [14]. After excluding all low-frequency (singleton) polymorphisms, the mean selection parameter γ were estimated to be 2.0 and 1.8 for male- and female- biased genes respectively, while the mean γ for sex-unbiased genes was -0.1. These results are quite consistent with our estimates. Later, Baines et al. studied effects of X-linkage on sex-biased gene evolution using a time-independent random effects PRF model [20]. They analyzed DNA sequence polymorphism and divergence in 45 X-linked genes for which 17 are male-biased, 13 are female-biased, and 15 are sex-unbiased genes and found evidence for adaptive evolution in both group of sex-biased genes. The estimated mean selection coefficients for male-biased, female-biased, and sex-unbiased genes are respectively 4.7, 2.5, and -0.8 using all polymorphic sites and 7.4, 2.1, and 0.5 after removal of singleton polymorphisms.

Discussion

The classical Bayesian model [4–6], fixed effect or random effects, assumes that the two daughter populations are immediately at mutation-selection-drift equilibrium after species divergence. Knowing that this assumption may be biologically unrealistic, we apply a previously developed time-dependent Poisson random field model to DNA sequences data to make inferences about selection, mutation, and species divergence. The results of this study suggest that a majority of newly-arisen non-synonymous mutations observed as polymorphisms is beneficial, although the magnitude of selection is very small, which is consistent with the conclusion drawn by Bustamante et al. [4] where a time-independent fixed effects PRF model was applied to 34 genes

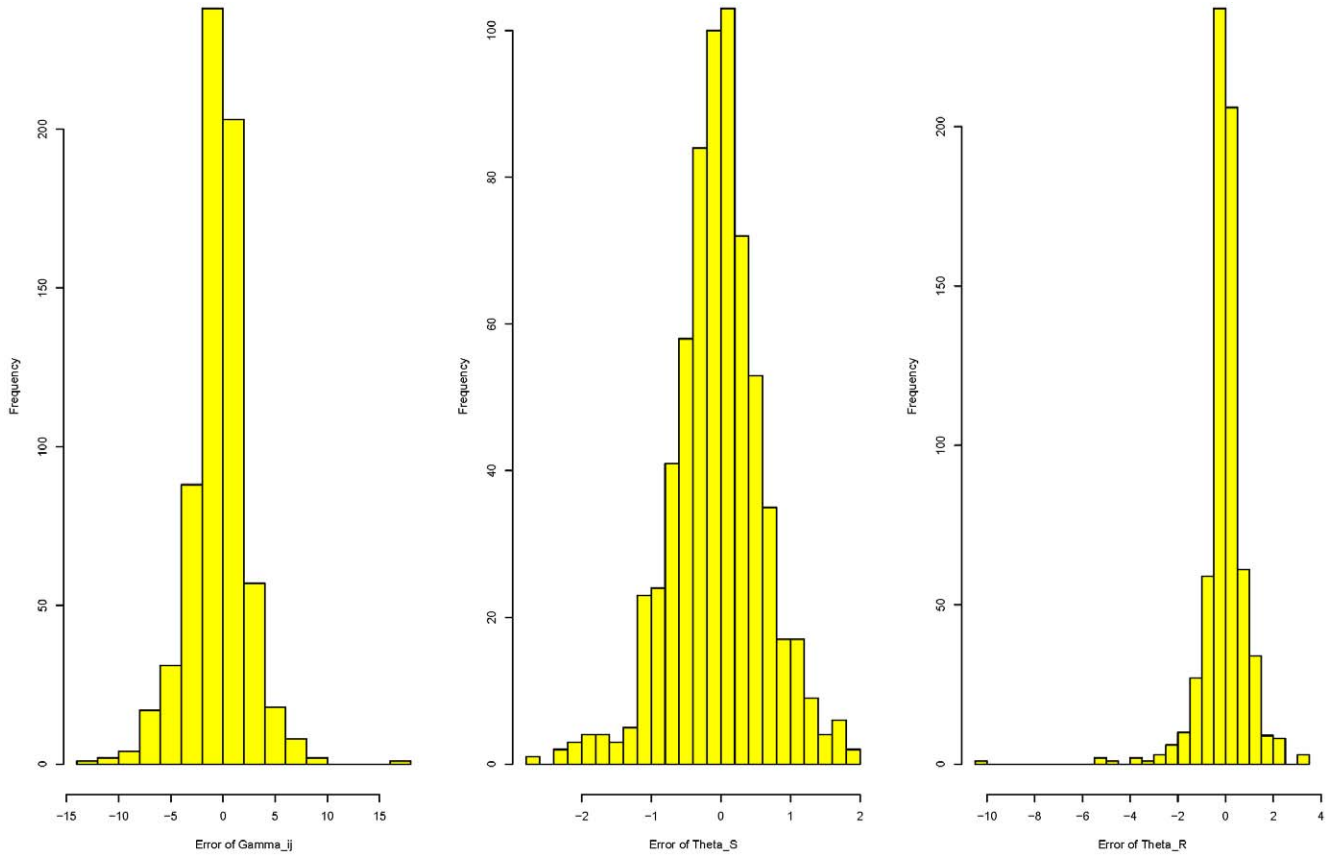

Figure 2. Histograms of estimation errors for selection coefficient (γ), silent mutation rate (θ_s), and replacement mutation rate (θ_r) using 23 simulated data sets each containing 30 genes.
doi:10.1371/journal.pone.0034413.g002

from *D. melanogaster* and *D. simulans*. Based on the results of Markov chain Monte Carlo simulations, they estimated the selection coefficient γ for each individual gene and concluded that "the average amino-acid replacement that is polymorphic or fixed in *Drosophila* is beneficial". The set of 91 *D. melanogaster* genes studied here has previously been analyzed in a time equilibrium random

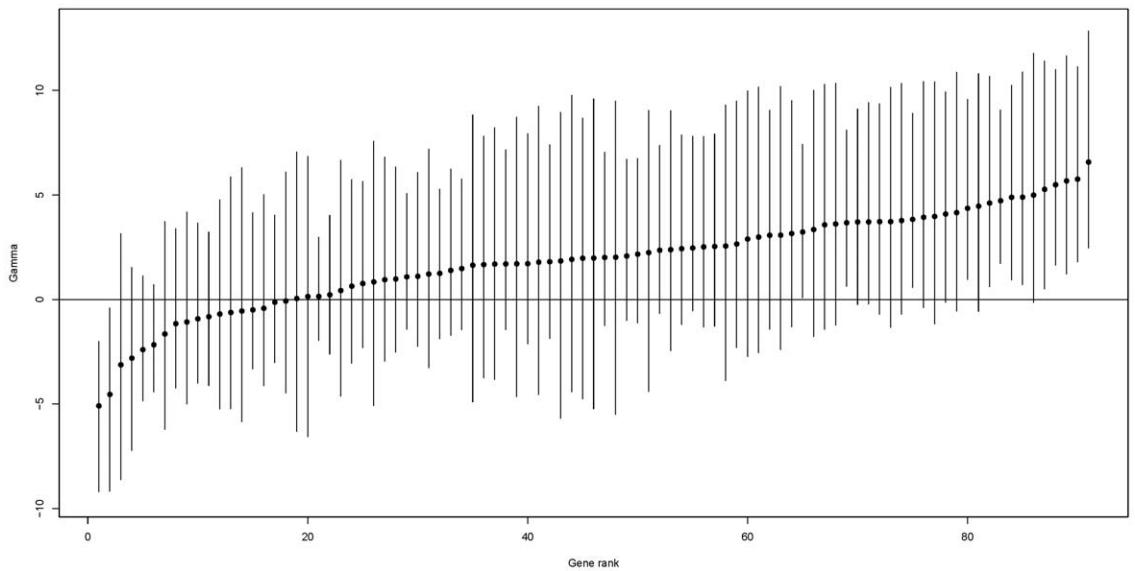

Figure 3. Estimated selection parameter (γ) for each gene with the loci sorted by the values of the estimates (medians). Error bars represent 95% credible intervals.
doi:10.1371/journal.pone.0034413.g003

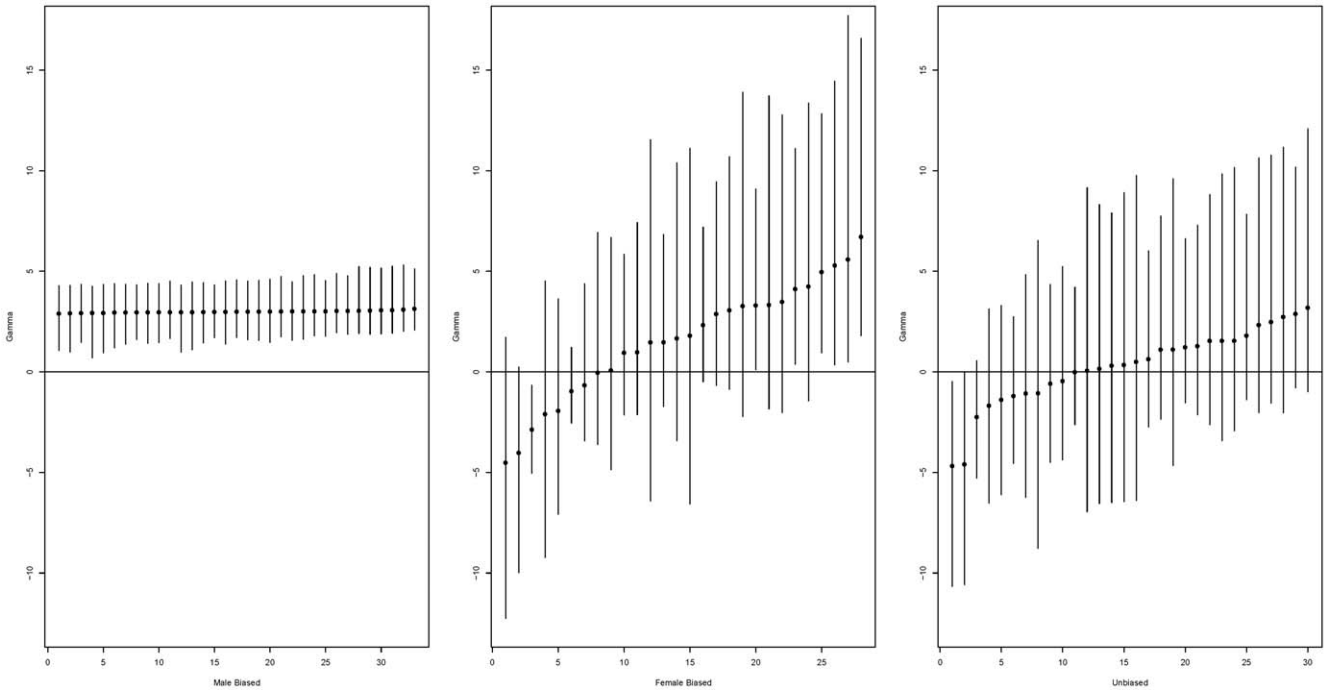

Figure 4. Estimated selection parameter (γ) for male-biased, female-biased, and sex-unbiased genes with the loci sorted by the values of the estimates (medians). Error bars represent 95% credible intervals. Because selection coefficients are fitted jointly to a Gaussian distribution, uncertainties can be highly correlated. This is particularly visible for male-biased genes, where the uncertainty on the mean selection coefficient is larger than the estimated standard deviation.
doi:10.1371/journal.pone.0034413.g004

effects PRF model [6]. Specifically, they assumed that the selective effect of each non-synonymous mutation, y_i, is normally distributed with mean γ_i and variance σ_w^2 but the mean selection coefficient γ_i within a gene varies from one gene to the next, according to a normal distribution with mean μ_γ and variance σ_b^2. Scaled to the diploid population size, they estimated the mean selection coefficient $\mu_\gamma = -5.7 \pm 15.5$, within- and between-locus standard deviations $\sigma_w = 3.5 \pm 5.7$ and $\sigma_b = 2.1 \pm 2.2$ respectively. Their analysis suggested that 95% of all replacement mutations that could contribute to polymorphism or divergence are deleterious. On the other hand, majority of fixed differences between species are positively selected. The difference between our estimate of the mean selective effect of newly arisen non-synonymous mutations and that of Sawyer et al [6] is due to the assumption imposed on the distribution of the selective effects within a locus. It is biologically more realistic to model selective effect within a gene as random variable, as in [6], instead of constant. However assuming mutation-selection-drift equilibrium is artificial and it may bias the estimates of selective effects. In contrast, we put the species divergence time explicitly into the model to also make it biologically more reasonable. The Bayesian framework that we have applied in this study assumes that the selection intensity γ is same for each coding sequence but distributed normally with a fixed mean and variance across loci. This assumption is somewhat artificial but it is still meaningful for the original purpose of inferring polymorphism and divergence based on the newly proposed time-dependent PRF model. To conquer the disadvantages of the two models as well as to be able to estimate the fraction of amino acid fixations that are driven by positive selection, we are developing a more sophisticated time-dependent random effects model and its application to simulated data as well as to real data will appear in a future publication.

Because the time of divergence is explicitly built into the model, we can estimate the value of the divergence time precisely and hence, it will help us to distinguish between fixations of beneficial mutations in a short period of time and fixations of deleterious mutations over a long period of time. The PRF model was derived under the assumption of independence among nucleotide sites. Due to the fact that high levels of recombination between nucleotides results in nearly independent assortment, whereas tight linkage is caused by low rates of recombination, it is equivalent to assume that nucleotide sites are at linkage equilibrium. For estimates of the mean selection coefficients, simulations have shown that methods based on Poisson random field for multi-locus data are relatively robust to the violation of this assumption([7,9,21,22]). The effect of linkage on the overall shape, in particular, the variance, of the distribution of the selective coefficients needs to be analyzed as part of the model validation. The model also assumes that individual species have constant and equal population sizes. However changes in varying recombination rates and demographic history of the population such as population expansion and bottlenecks may result in changes of population size that could affect the parameter estimates and hence confound the interpretation of polymorphism and divergence ([2,23–26]). The use of African *Drosophila* sample can avoid some of the demographic complexity ([27,28]). As we mentioned earlier, one highly-inbreed line from *D. simulans* was used as a comparison of divergence. Although the high inbreeding ratio contradicts with the model assumption of equal population sizes, the use of a single line from second species minimizes the effect caused by this contradiction. Further study need to be conducted to check the robustness of the model to deviations from the assumptions.

Acknowledgments

We would like to thank the editor and two anonymous reviewers for their careful reading and detailed comments which improved the manuscript a lot.

Author Contributions

Conceived and designed the experiments: AA SS. Performed the experiments: AA SS. Analyzed the data: AA SS. Contributed reagents/materials/analysis tools: AA SS. Wrote the paper: AA.

References

1. Lewontin RC (1974) The Genetic Basis of Evolutionary Change. New York: Columbia University Press.
2. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351: 652–654.
3. Sawyer S, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.
4. Bustamante CD, Nielsen R, Sawyer SA, Purugganan MD, Olsen KM, et al. (2002) The cost of inbreeding: fixation of deleterious genes in Arabidopsis. Nature 416: 531–534.
5. Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL (2003) Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. Journal of Molecular Evolution 57: S154–S164.
6. Sawyer SA, Parsch J, Zhang Z, Hartl DL (2007) Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila. Proceedings of National Academy of Sciences of the United States of America 104: 6504–6510.
7. Abel HJ (2009) The role of positive selection in molecular evolution: Alternative models for withinlocus selective effects. Ph D thesis, Washington University in St.Louis. University Microfilms.
8. Bustamante CD, Nielsen R, Hartl DL (2003) Maximum likelihood and bayesian methods for estimating the distribution of selective effects among classes of mutations using dna polymorphism data. Theoretical Polulation Biology 63: 91–103.
9. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genetics.
10. Huerta-Sanchez E, Durrett R, Bustamante CD (2008) Population genetics of polymorphism and divergence under uctuating selection. Genetics 178: 325–337.
11. Wakeley J (2008) Polymorphism and divergence for island-model species. Genetics 163: 411–420.
12. Williamson S, Hernandez R, Alon AF, Zhu L, Nielsen R, et al. (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proceedings of National Academy of Sciences of the United States of America 102: 7882–7887.
13. Amei A, Sawyer S (2010) A time-dependent poisson random field model for polymorphism within and between two related biological species. Annals of Applied Probability 20: 1663–1696.
14. Proschel M, Zhang Z, Hartl DL (2006) Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. Genetics 174: 893–900.
15. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) Numerical recipes: the art of scientific computing. England: Cambridge University Press, 3rd edition. MR2371990 p.
16. Gelman A (1996) Inference and monitoring convergence. In: Gilks W, Richardson S, Spiegelhalter D, eds. Markov Chain Monte Carlo in Practice. London: Chapman and Hall. pp 131–143.
17. Meiklejohn CM, Kim Y, Hartl DL, Parsch J (2004) Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-lilklihood estimation. Genetics 168: 265–279.
18. Lemeunier F, David JR, Tsacas L, Ashburner M (1986) The genetics and biology of *Drosophila*. In: Ashburner M, Carson HL, eds. The *melanogaster* species group. New York: Academic Press. pp 147–256.
19. Caccone A, Amato GD, Powell JR (1988) Rates and patterns of scnDNA and mtDNA divergence within the *Drosophila melanogaster* subgroup. Genetics 118: 671–683.
20. Baines JF, Sawyer SA, Hartl DL, Parsch J (2008) Effects of x-linkage and sex-biased gene expression on the rate of adaptive protein evolution in *Drosophila*. Molecular Biology and Evolution 25: 1639–1650.
21. Bustamante CD, Wakeley J, Sawyer SA, Hartl DL (2001) Directional selection and the sitefrequency spectrum. Genetics 159: 1779–1788.
22. Bustamante CD, Wakeley J, Sawyer SA, Hartl DL (2005) A composite-likelihood approach for detecting directional selection from dna sequence data. Genetics 170: 1411–1421.
23. Fay J, Wyckoff GJ, Wu CI (2002) Testing the netrual theory of molecular evolution with genomic data from *Drosophila*. Nature 415: 1024–1026.
24. Eyre-Walker A (2002) Changing effective population size and the mcdonald-kreitman test. Genetics 162: 2017–2024.
25. Begunjh DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, et al. (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biology.
26. Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177: 2251–2261.
27. Glinka S, Ometto L, Mousset S, Stephan L, De Lorenzo D (2003) Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics 165: 1269C1278.
28. Ometto L, Glinka S, De Lorenzo D, Stephan W (2005) Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of dna variation. Molecular Biology and Evolution 22: 2119C2130.