Phylogenetic model of stabilizing selection is more informative about site specific selection than extrapolation from laboratory estimates.

CEDRIC LANDERER[1,2,*], BRIAN C. OMEARA[1,2], AND MICHAEL A. GILCHRIST[1,2]

[1]Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610

[2]National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996-3410

[*]Corresponding author. E-mail: cedric.landerer@gmail.com

Version dated: October 6, 2018

Introduction

Incorporation of selection into phylogenetic frameworks has already been a long lasting endeavor. Early models focused the influence of selection on the substitution rate between a resident and a mutant (**???**). These models however, lack site specific equilibrium frequencies. The importance of site specific equilibrium frequencies has long been noted (**??**). **?** first introduced a framework to incorporate the site specific equilibrium frequencies of amino acids. However, they had to concede that their model was to parameter rich and therefore intractable for biological data sets without simplifying assumptions.

- Incorporating selection into phylogenetic frameworks is already a long lasting endeavor.
  - Phylogenetic inference of sequence relationship was long focused on rates of substitutions.
  - No site specific equilibrium frequencies until (HB98, Bloom2014, ...).
  - Such models however, tend to be unfeasible as they are very parameter rich.
  - The type of selection on a protein is not always clear, or differs between proteins phylogenetic models also have to make generalizing assumptions.
  - Incorporating selection from experimental sources therefore seems like an attractive option.
  - Incorporating empirical fitness has some important features.
    * It allows for site specific amino acid preferences, acknowledging the heterogeneity of selection along the protein sequence.
    * It greatly reduces the number of parameters that have to be estimated from the data.
    * It allows for the fitting more complex models
  - However, the incorporation of empirical fitness also has some important short comings.

    * Loss of generality.
    * DMS experiments are limited to proteins and organisms that can be manipulated under laboratory conditions.
    * But even in the case of TEM, the applied selection pressure is limited to the defense against a specific antibiotic.
    * TEM, however, has evolved to compete against conspecifics and other microbes using secreted metabolites to gain an advantage.
    * Furthermore, DMS relies on a library of mutants and therefore on a heterogeneous population with competing genotypes.
    * Therefore, it is important to ask how adequate such experiments reflect natural evolution.

- In this study we will assess how adequate DMS inference of site specific selection on amino acids, using TEM and provide an alternative, more generally applicable solution.
    - Simulations using DMS inferred site specific selection on amino acids show that observed TEM variants are unexpected; revealing the inadequacy of DMS.
    - Models fits achieved by the incorporation of DMS experiments can be improved upon using a hierarchical phylogenetic framework of stabilizing selection: SelAC.
    - Extrapolating site specific selection on amino acids between sequences (TEM and SHV) with related function can be inadequate.

Results

Site Specific Selection on Amino Acids Improves Model Fit

We compared the models *phydms* (**?**) and *SelAC* (**?**), models of stabilizing site specific amino acid selection, to 281 other codon and nucleotide models by fitting them to 49 sequences of

Model	L	n	AIC	ΔAIC	AICc	ΔAICc
SelAC	-1498	374	3744	0	3766	6
SelAC +DMS	-1768	111	3758	14	3760	0
phydms	-2060	105	4331	586	4326	566
SYM+R2	-2230	102	4663	919	4694	934

Table 1: L, number of model parameters n, AIC, and ΔAIC., Full table has 231 models

the β-lactamase TEM. Models with site specific selection on amino acids improved model fits by 917 to 1483 AICc units over codon or nucleotide models without site specific selection (Table **??**). In addition, *SelAC* does outperform *phydms* by 560 to 566 AICc units.

SelAC utilizes a hierarchical model framework and estimates 263 site specific parameters, $\sim 5\%$ of the 4997 parameters necessary to fully describe the site specific selection on amino acids. In contrast, *phydms* does not infer any site specific parameters, but utilizes site specific selection on amino acids estimated from deep mutation scanning experiments. Incorporating site specific selection on amino acids estimated from deep mutation scanning experiments into *SelAC* (*SelAC* +DMS) yields similar a AICc value to *SelAC* without that information. This is solely due to a decrease in the number of parameters estimated, as the L decreases from -1498 to -1768 (Table **??**).

Laboratory inferences of selection are inconsistent with observed sequences.

Improved model fits with phydms are deceiving. The site specific selection inferred by the deep mutation scanning experiment is inconsistent with the observed TEM sequences. We find that the sequence of selectively favored amino acids has only 49 % sequence similarity with the observed consensus sequence (Figure **??**). This is in contrast to the 99 % of sequence similarity with the sequence of selectively favored amino acids estimated by *SelAC*.

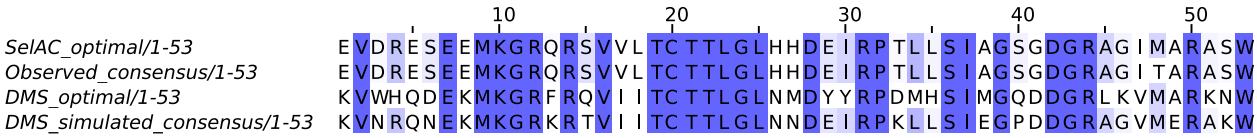

Figure 1: Every 5th residue. DMS and simulation based on DMS do not reflect natural sequences

Simulations of codon sequences under the experimentally inferred site specific selection for amino acids reveals that we would not expect to see the observed TEM sequences. We simulated under a wide range of effective population sizes N_e, and find that the experimentally inferred site specific selection is very strong. Only when N_e is on the order of 10^0 drift is overpowering the efficacy of selection. With realistic values for N_e, we expect that the observed sequences show sequence similarity of $\sim 70\%$ with the sequence of selectively favored amino acids inferred by the deep mutation scanning experiment (Figure **??**a). Similarly, we expect that the genetic load of the observed sequences should be half of the observed genetic load (Figure **??**b).

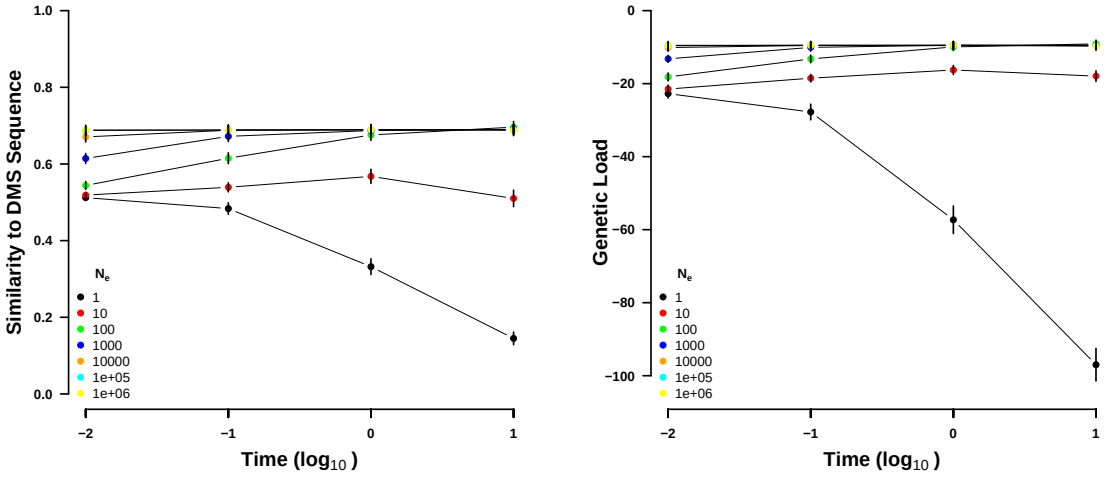

Figure 2: Sequences simulated under various values of N_e and for various times (expected substitutions per site). TODO: Add starting point and observed values.

Secondary Structure	Mean G	SE G	Mean Genetic Load	SE Genetic Load
Helix	206.1	12.4	0.18×10^{-7}	0.13×10^{-7}
Beta Sheet	238.6	15.8	6.8×10^{-8}	2.9×10^{-8}
Unstructured	224.8	11.4	0.19×10^{-7}	8.1×10^{-8}
Active Sites	300	0	0	0

Table 2: L, number of model parameters n, AIC, and ΔAIC., Full table has 231 models, TODO: Update numbers

SelAC Model Adequacy

When assessing model adequacy, we find that *SelAC* better explains the observed TEM sequences. The observed consensus sequence has a very high sequence similarity with the sequence of selectively favored amino acids estimated by *SelAC* (99 %). Furthermore, assuming the site specific selection estimated by *SelAC*, the observed sequences only show a minimal genetic load (Figure **??**).

We then simulated codon sequences forward in time for various length of time to assess the sequence similarity, assuming the *SelAC* inferred site specific selection for amino acids. Using the ancestral state of the 49 TEM variants as initial sequence we find that

To further demonstrate the consistency of *SelAC*, we utilized random codon sequences as starting points. We find that the sequence similarity increases with effective population size N_e. The random sequences start of with a similarity of $5\% - 8\%$ which increases with N_e to $26\% - 34\%$. The same initial sequences under the site specific selection inferred by the deep mutation scanning experiment increase only to $X\% - Y\%$ in sequence similarity.

Site specific estimates of Selection on Amino Acids

We find that the whole sequence is under strong selection and near the optimum with a small genetic load. We find stronger selection and a lower genetic load in beta sheets than in helices and unstructured regions (Table **??**).

The Active sites appear to be under the strongest selection, with no accumulated genetic load. This is in concordance with the experimental estimates.

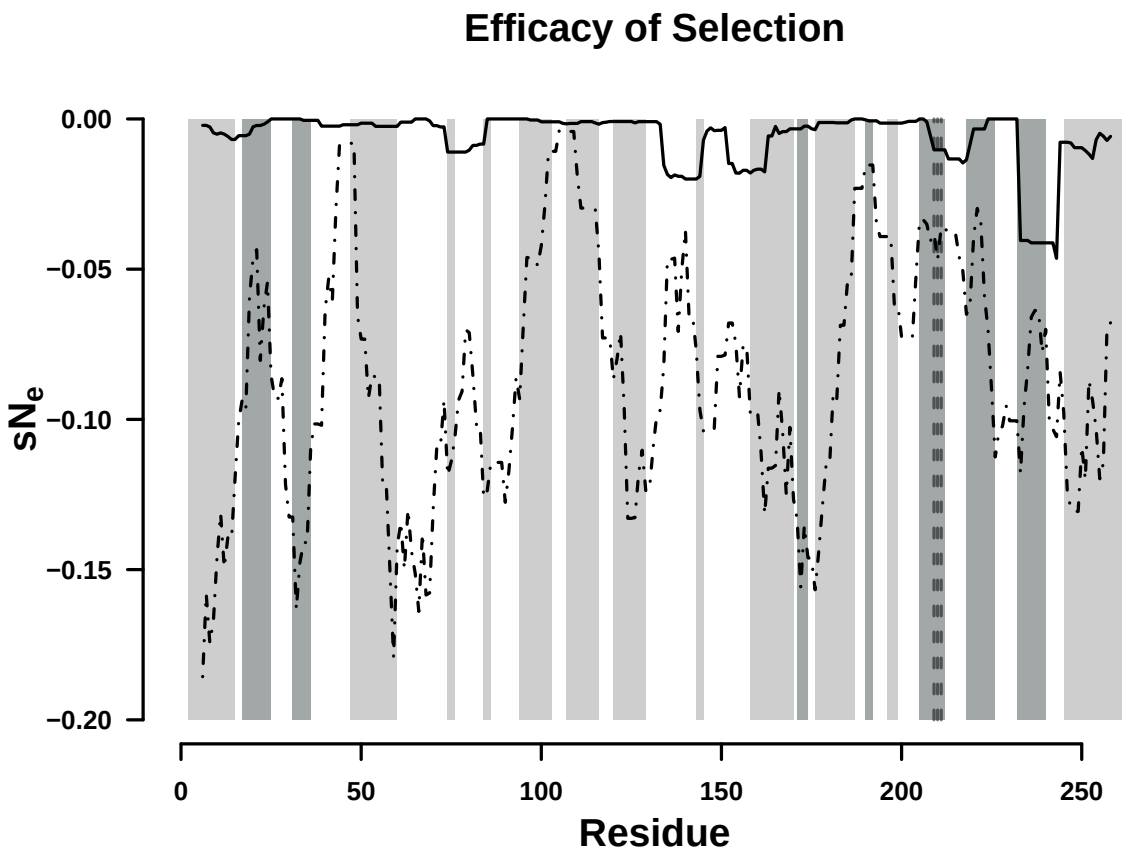

Figure 3: TEM, bars are different secondary structure elements. Dashed dotted line is DMS, solid is SelAC sNe, all lines are means of all sequences, sliding window of 10 sites. vertical lines are active/binding sites.,

- Site Specific Selection on Amino Acids Improves Model Fit
  - phyDMS improved model fit to 49 TEM sequences by 917 AICc units
  - Number of parameters estimated from data comparable to GY94 and others despite complex description of fitness landscape because of experimental estimates.
  - Model selection shows that SelAC outperforms phydms (Table **??**).
- Lab inferences of selection (DMS) are inconsistent with observed sequences.
  - The inferred fitness landscape is inconsistent with observed sequences.

    - The optimal amino acid sequence inferred by DMS only shows 49% sequence similarity with the observed sequences (Figure **??**).
  - Observed sequences unlikely under the lab inferred fitness landscape (Figure **??**a,b).
    - We would expect about half of the observed fitness burden.
    - Sequence similarity is expected to be about $\sim 70\%$.
- *SelAC* Model Adequacy.
  - Model adequacy assessed by sequence similarity shows that SelAC better represents the observed sequences.
- Application of SelAC to TEM.
  - Site specific estimates of aa fitness (Figure **??**).
    - Fitness burden based on SSE.
    - Fitness burden at binding sites
    - Most sites show the estimated optimal amino acid.
    - We find that selection against used amino acids is clustered and locally confined.
    - Role of secondary structures?
  - Application of SelAC to TEM and comparison to TEM
    - Site specific G terms for TEM and SHV are only weakly correlated ($\rho = 0.17$), despite similar α_G (Figure **??**a).
    - Greatest difference is observed in the physicochemical properties, specifically α (which PC is that?) (Figure **??**b).

Discussion

- We evaluated how well experimental selection estimates from DMS experiments explain natural sequence evolution and compared it to a novel phylogenetic framework, SelAC.
  - Previous work has shown that DMS selection estimates can improve model fit over classical approaches like GY94 and our work confirms this.
  - Model selection favored the SelAC model fit and the corresponding fitness estimates over the DMS estimates using both, SelAC and phyDMS (Table **??**).
- Adequacy of the DMS selection has previously not been assessed.
  - The amino acid with the cumulative highest fitness experimentally estimated with DMS only has 49% concordance with the observed alignment.
  - In contrast, the SelAC estimate has 99% concordance (Figure **??**).
  - Estimates of selection coefficients do not represent evolution.
    * Due to artificial selection environment; Heterogeneous population, very large s.
    * Only one antibiotic used, maybe a mixture of antibiotics would better reflect natural evolution.
    * Lack of repeatability between labs introduces further problems (Firnberg et al 2014 vs. Stifler et al. 2016).
- Assuming that the DMS selection inference adequately reflects natural evolution, the observed TEM sequences are either mal-adapted or where unable to reach a fitness peak.
  - *E. coli* has a large effective population size, estimates are on the order of 10^8 to 10^9 (Ochman and Wilson 1987, Hartl et al 1994).

- The large N_e would allow *E. coli* to effectively "explore" the sequence space, thus suggesting that the TEM sequences are mal-adapted according to the DMS estimates.
- Our simulations of sequence evolution with various N_e values and the DMS fitness values in contrast show that we would expect higher adaptation even with much smaller N_e (Figure **??**).

- Estimates of selection coefficients do not represent evolution.
  - Due to artificial selection environment; Heterogeneous population, very large s.
  - Only one antibiotic used, maybe a mixture of antibiotics would better reflect natural evolution.
  - Lack of repeatability between labs introduces further problems (Firnberg et al 2014 vs. Stifler et al. 2016).

- DMS estimates of the observed TEM variants predict them to be mal-adapted while SelAC predicts most TEM variants to be well adapted.
  - Given *E. coli*'s large effective population size, the efficacy of selection should be very large.
  - We therefore expect the observed sequence variants to be at the selection-mutation-drift barrier, which in turn can expected to be near the optimum.
  - We find the majority of sequences near the optimum, therefore the SelAC estimates are consistent with theoretical population genetics results.
  - In contrast, finding strong selection against the observed TEM variants indicates that DMS is not consistent with theoretical population genetics expectations.
  - This is consistent when thinking about that DMS only reflects the selection on the TEM sequence with regards to one antibiotic, which seems appropriate to

model selection in modern hospital environments but not when the interest lies in the natural evolution of TEM.

- We find that SelAC produces similar selection against the observed TEM variants if we assume the fitness peaks (optimal AA) that are estimated by DMS.
  - This shows that DMS and SelAC can provide consistent estimates of selection against amino acids.
  - SelAC has the advantage that it can be applied to any protein coding sequence alignment.
  - This removes the need for extrapolation e.g. from TEM to SHV.
- SelAC has the advantage that it can be applied to any protein coding sequence alignment.
  - This removes the need for extrapolation e.g. from TEM to SHV.
- Difference in selection parameters between TEM and SHV indicate that extrapolation is not a good idea.
  - The difference in the site specific strength of selection shows that TEM and SHV are facing different selection pressures.
  - this is also highlighted by the differences in physicochemical weightings between the two proteins.
- SelAC outperforms DMS, but is not without flaws itself
  - Like DMS and most phylogenetic models, SelAC assumes site independence.
  - SelAC is a model of stabilizing selection, in contrast to e.g. GY94 which is a model of frequency dependent selection.
    - Since TEM plays a role in the chemical warfare with conspecifics and other microbes, some sites may be under negative frequency dependent selection.

- SelAC assumes the same G distribution across all sites.
    - Different G distribution for each type of secondary structure
    - active sites may not follow distribution.
- SelAC assumes that selection is proportional to distance in physicochemical space.
    - We used Grantham (1974) properties, however many other distances are available which may an even better model fit.

- Low sequence variation in the TEM may be cause for concern as it could be misinterpreted by the model as stabilizing selection because of the short branches.
    - However, provided our simulations support that TEM is actually under stabilizing selection

- In conclusion, DMS experiments have been proposed to supplement information on selection on amino acids in phylogenetic studies.
    - This study shows that information on selection can be extracted from alignments of protein coding sequences.
    - This highlights the limitations of DMS to explain natural evolution.

Materials and Methods

Phylogenetic Inference and Model selection

TEM sequences were obtained from **?** and already aligned. Experimentally fitness values for TEM were taken from **?**. We chose the highest concentration (2500 $\mu g/mL$) treatment of ampicillin for our comparison. We modified the experimental fitness such that the amino acid with the highest fitness at each site has a value of one.

SelAC (version 1.6.1) was fitted to the TEM alignment using R (version 3.4.1) (**?**) with and without site specific selection on amino acids estimated from deep mutation scanning

experiments. *phydms* (version 2.0.5) was fitted using site specific selection on amino acids estimated from deep mutation scanning experiments from **?**. All other models were fitted using IQTree (**?**).

We report each model's log(*likelihood*), AIC, and AICc. Models were selected based on the AICc values.

Sequence Simulation

Sequences were simulated by stochastic simulations using a Gillespie algorithm (**?**) that was model independent. The simulation followed **?** to calculate fixation probabilities. Mutation rates were taken from the *SelAC* or *SelAC* +DMS fit. The initial sequences were either a random sample of codons or the ancestral sequence reconstructed using FastML (**?**) (last accessed: 30.09.2018). Each sequence was simulated 10 times and we report average genetic load and sequence similarity as well as the standard error.

Model Adequacy

Model adequacy was assessed by comparing the observed sequences and simulations under the site specific selection inferred by the deep mutation scanning experiment or *SelAC*. First, similarity between the sequence of selectively favored amino acids and the observed TEM sequences was assessed. Sequence similarity was measured as the number of differences in the amino acid sequence. Second, the genetic load of the observed and the simulated sequences was calculated using either the site specific selection inferred by the deep mutation scanning experiment or *SelAC*.

Genetic load was calculated as

$$L = \frac{w_{max} - w_i}{w_{max}} \tag{1}$$

were w_{max} is the fitness of the sequence of selectively favored amino acids estimated using

the site specific selection inferred by the deep mutation scanning experiment or $SelAC$. w_i represents the fitness of the ith observed or simulated sequence.

Figures

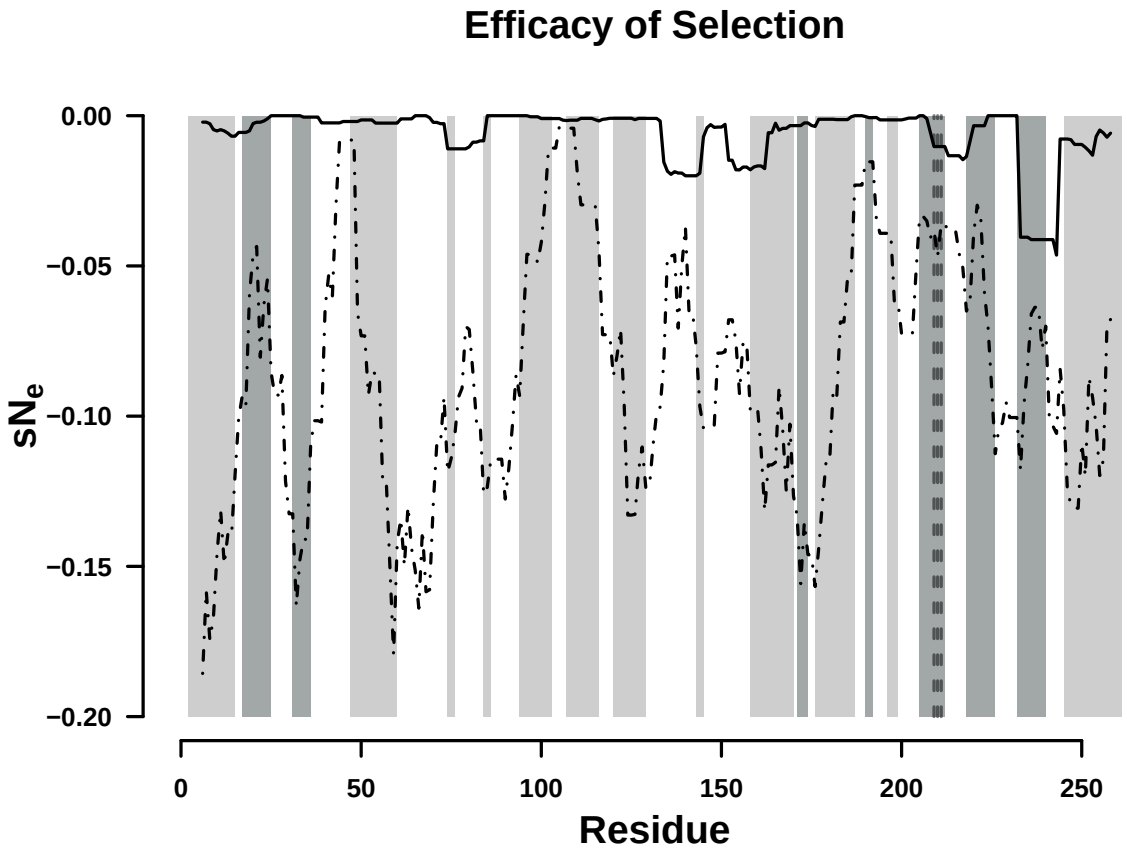

Figure 4: TEM, bars are different seconday structure elements. Dashed dotted line is DMS, solid is SelAC sNe, all lines are means of all sequences, sliding window of 10 sites. vertical lines are active/binding sites.

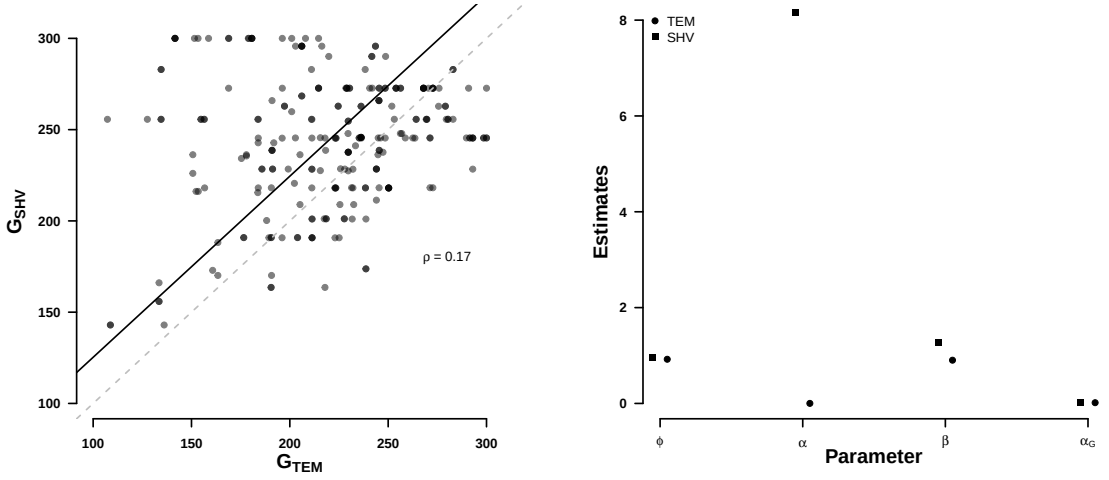

Figure 5: Comparisson of selection related parameters between TEM and SHV.

Supplementary Figures

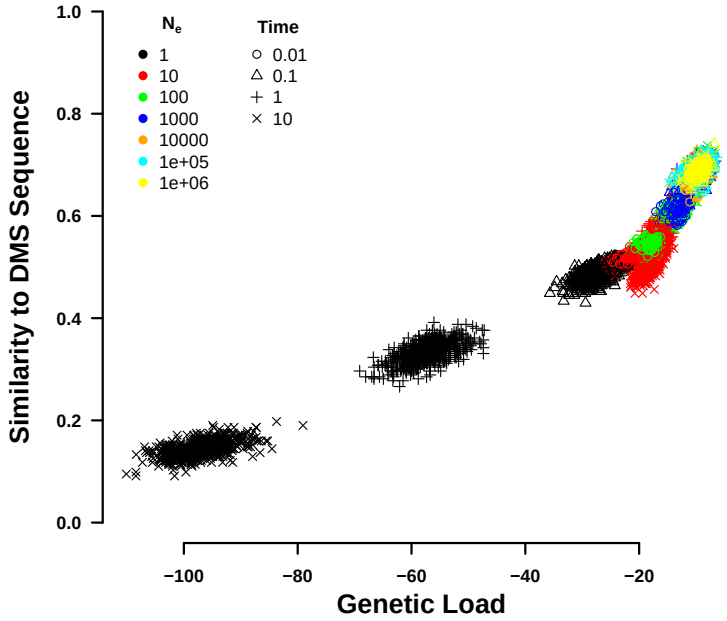

Figure S1: Suppl: Sequences simulated under various values of N_e and for various times. TODO: replace clouds by mean+sd bars

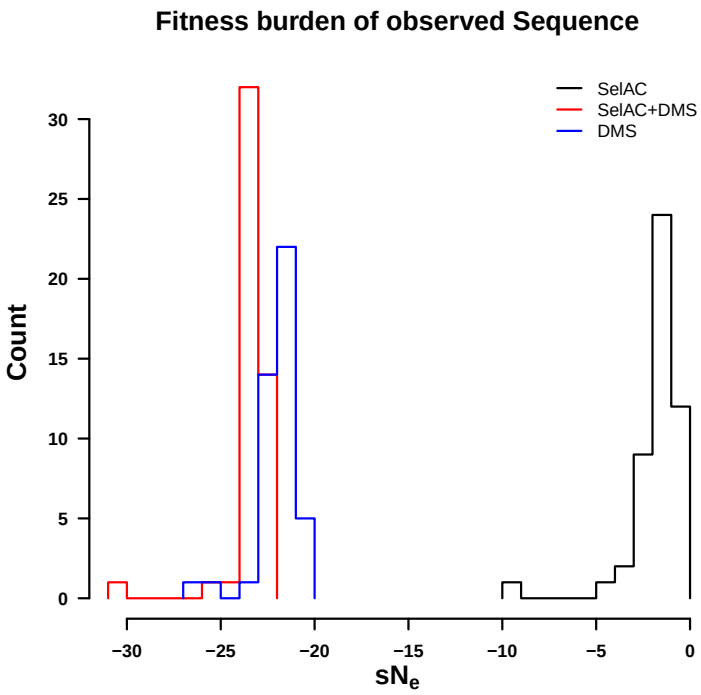

Figure S2: Suppl: sN_e of whole sequence, variation across tips. TEM