Supplemental Information for *Fast Estimation of Recombination Rates Using Topological Data Analysis*

D.P.H., M.R.M., M.M., and A.J.B.

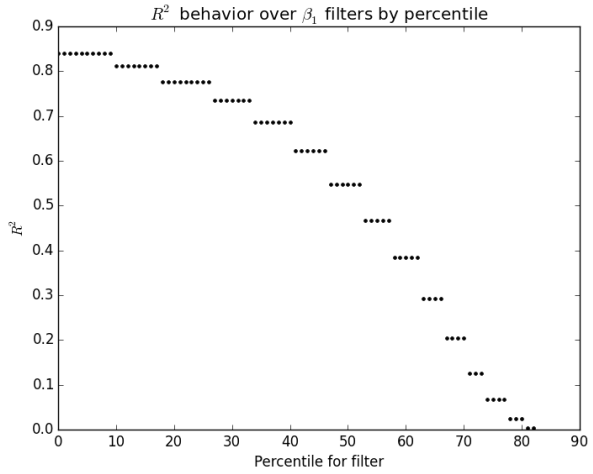

Fig. 1. R^2 values decrease as we filter out bars of increasing lengths in terms of percentiles of the overall distribution of bar lengths across all simulated datasets.

Filtering β_1. Some authors have suggested that short-lived bars in a persistence diagram may arise due to topological noise (1). It is unclear whether a universal threshold exists to filter out potentially noisy bars, however. In the case of our own study, we tested whether filtering out short-lived bars improves the relationship between existing topological models of recombination and β_1, specifically using Camara et al's ρ_{ph} for goodness of fit. We find that the filtration of short-lived bars only lowers the R^2 values for this model applied to our simulated datasets, thus we opt not to use any filtration of these bars in practical applications (Figure 1).

Missing Data. As large quantities of missing data are frequently encountered in empirical datasets, we investigated the performance of both ψ and β_1 under various scenarios involving different amounts of missing data using the same datasets we simulated for our main coalescent investigation. Specifically, we took these existing datasets, duplicated them, and then converted randomly chosen sites or blocks of sites to N, in order to simulate missing data. For each dataset, a total of 10% of each sequence (and therefore, 10% of the total alignment) was converted to Ns. Since in all cases, 10% missing data was introduced into the alignments, we note that our specific interest here is in how robust each topological feature is to either a) a random distribution of missing sites, as would be common

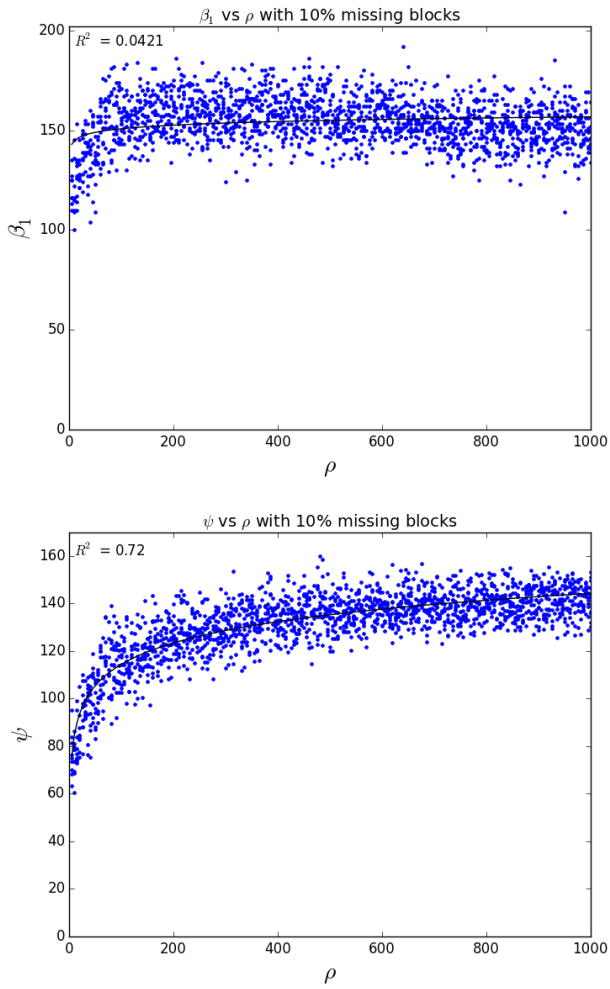

Fig. 2. Expectations of ψ and β_1 given ρ when 10% of data is missing in large tandem blocks. In the case of β_1, the R^2 value has dropped from 0.9 in the case of no missing data to 0.04, suggesting that this feature is highly sensitive to minimal missing data. In comparison, ψ maintains an R^2 of 0.76, suggesting greater robustness to loss of information in sequence data.

with sequencing error, or b) tandemly linked missing sites, as common in indels. We find that, overall, ψ is much more robust to missing data than is β_1 when predicting ρ. Specifically, we looked at the expectation of β_1 and ψ given ρ as in Camara et. al's model, to see how well the model fits each topological

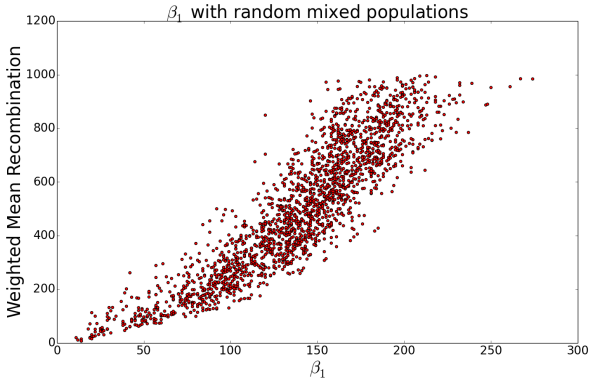

Fig. 3. β_1 versus $\bar{\rho}$ rate from randomly mixed populations of different recombination rates.

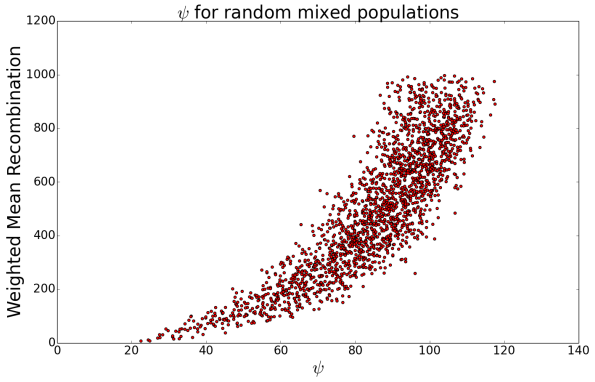

Fig. 4. ψ versus $\bar{\rho}$ from randomly mixed populations of different recombination rates.

feature under each missing data scenario 2. We did this in order to evaluate the strength of the relationship without a prior expectation for the expectation of ψ, and found a greater fit to the Camara model describing the expectation of ψ in place of β_1 given ρ in these circumstances than their original expectation of β_1 given ρ.

When we invert the graph to see how ψ or β_1 behave as predictors of ρ, we lose the fit to the Camara model. However, we observe similar differences in the variances of the two predictors, specifically noting that the variance in β_1 remains higher and so poses greater uncertainty than does ψ.

Mixing Populations. We explored the robustness of ψ and β_1 under mixed populations of varying recombination rate. In particular, we randomly sampled N individuals from a population of known recombination rate ρ_1, along with M individuals from a population of known recombination rate ρ_2. We kept $N + M = 160$ constant while varying N, M, ρ_1, and ρ_2. These experiments were all done on the simulated data, and we fixed the population mutation rate to $\theta = 25$.

For each randomly concatenated population we computed the weighted mean recombination rate $\bar{\rho} = \frac{N}{160} * \rho_1 + \frac{M}{160} * \rho_2$. The results of comparing our main topological summary statistics ψ and β_1 to the weighted mean recombination rate are presented in figures 3-4. We see that under randomly mixed populations ψ maintains a tight exponential relationship with $\bar{\rho}$. In comparison, in this setting the relationship between β_1 and $\bar{\rho}$ becomes noisier. The nice behavior of ψ is expected as mixing two distant populations only adds one non-informative coalescence event between the populations, and as a result has little effect on ψ other than averaging the recombination parameters of the samples.

LASSO Weights. In order to gain intuition for which barcode statistics would be the best predictors for recombination we first ran a LASSO regression analysis using 15 barcode statistics as input and analyzed the resulting weight vector. We used the LASSO weight vectors as a proxy for the predictive power of each barcode statistic.

The barcode statistics we studied were the Betti number (β_i), mean barcode length (ψ_i), medium barcode length (m_i), maximum barcode length (M_i), and the thresholded Betti number (β_i^T) in dimensions 0, 1, and 2. Here, the thresholded Betti number refers to the number of bars whose bar length is greater than a specified cutoff, where the cutoff is set as a percentage of maximum bar length M_i. In these experiments we tested thresholds corresponding to $10 - 60\%$ of the maximum bar length.

We ran LASSO on the simulated data with fixed sample size using Scikit-learn (2) in Python 2.7. For each threshold, we ran LASSO 20 times on randomly selected training data. Initially we used all 15 barcode summary statistics as input and then analyzed the LASSO weight vectors for all 20 runs across the varying thresholds. These weight vectors are visualized in a heat map in Figure 5.

Observe, consistently ψ_0 (or ψ), m_0, and β_0^T are among most heavily weighted inputs, with ψ_0 having the most influence overall. This motivated our rigorous exploration of ψ as an estimator for recombination rate. Also note that β_0 has zero weight since β_0 is precisely the sample size, which is constant in this experiment. Consequently, β_0^T only varies as M_0 varies.

In contrast to the dimension-0 statistics, the dimension-1 barcode statistics have negligible weights across all model runs and all thresholds. Moreover, almost all of the dimension-2 barcode statistics have low weights across all model runs, with the exception of M_2 whose weight increases as the threshold for β_i^T increases.

In an attempt to extract the best dimension-1 topological predictor for recombination, we re-ran the 20 runs of LASSO for varying thresholds using only the dimension-1 topological summary statistics as input. The results of this refined analysis are presented in Figure 6. Here we see that out of all the dimen-

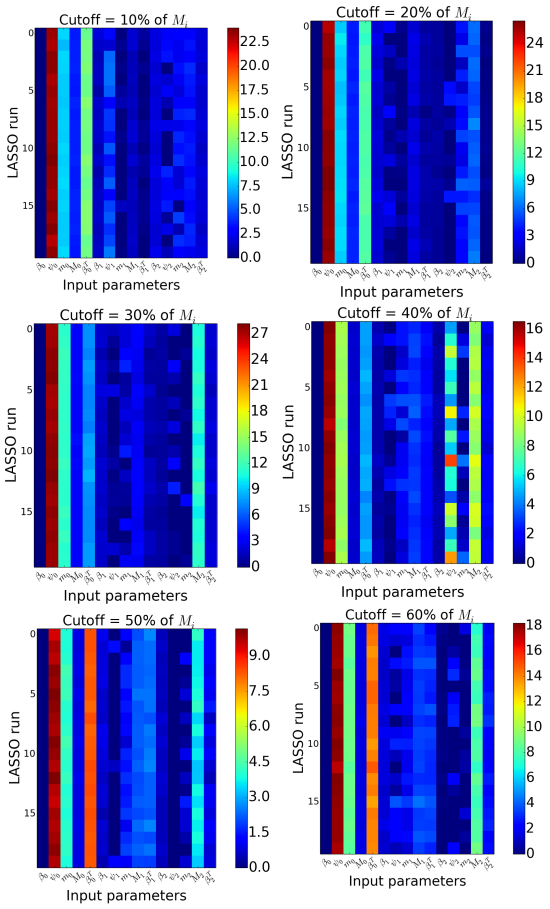

Fig. 5. Absolute value of LASSO weight vectors across 20 model runs on randomly selected 15-dimensional training data for varying thresholds for β_i^T.

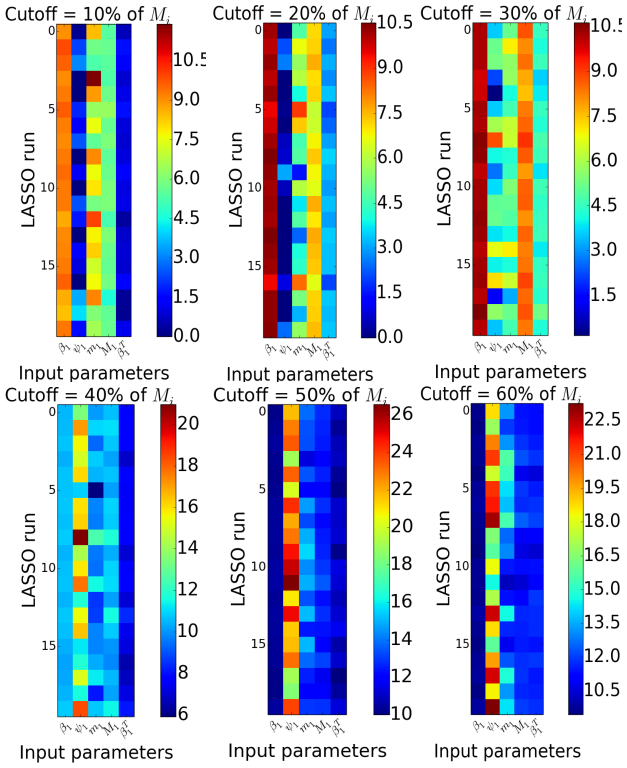

Fig. 6. Absolute value of LASSO weight vectors across 20 model runs on randomly selected dimension 1 training data for varying thresholds for β_i^T.

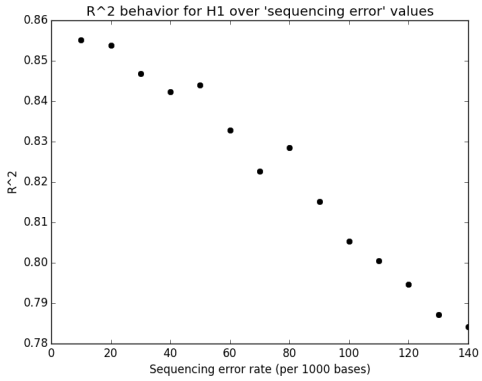

Fig. 7. R^2 values decrease as we include increasing sequence errors, but remains greater than 0.75 over all realistic and slightly more extreme possible error rates.

sion 1 topological statistics, β_1 is the most heavily weighted input feature when the threshold is set to $\leq 30\%$ of M_i. For increased threshold values the weight of β_1 decreases significantly and ψ_1 is the most heavily weighted input features for the dimension 1 barcode statistics, although its weights vary greatly across different model runs. This is consistent with the results presented in *Filtering* β_1, which suggest that filtering out short-lived bars hinders the predictive power of β_1 for recombination estimates.

The results of these LASSO analyses provided us with the motivation to focus on understanding the significance of lower dimensional topological statistics as predictors for recombination. Moreover, we used the results of the dimension-1 LASSO analysis to decide which higher dimensional statistics to use in tandem with ψ. We chose to exclude dimension-2 statistics as predictors due to their overall low weight vectors and the lack of biological significance.

Noise Experiments. We further tested whether topological noise may contribute to relationships we observed between β_1, ψ, and ρ by adding noise or randomizing sequences within datasets to obtain topological structures unrelated to recombination. In one experiment, we simulated realistic cases of sequencing error in each dataset by adding a random base over a range of error rates from 0 to 140 / 1000 bases (see Figure 7). As this error rate increases, we find a reduction in the fit of topological models in β_1 to the data, but the decrease in fit is slow over all realistically expected error rates (around 10% sequencing error, we see no reduction in R^2 for ρ_{ph}).

Other Relationships. We tested for relationships between our topological features ψ and β_1 and Watterson's θ, as well as for possible correlations between the topological features them-

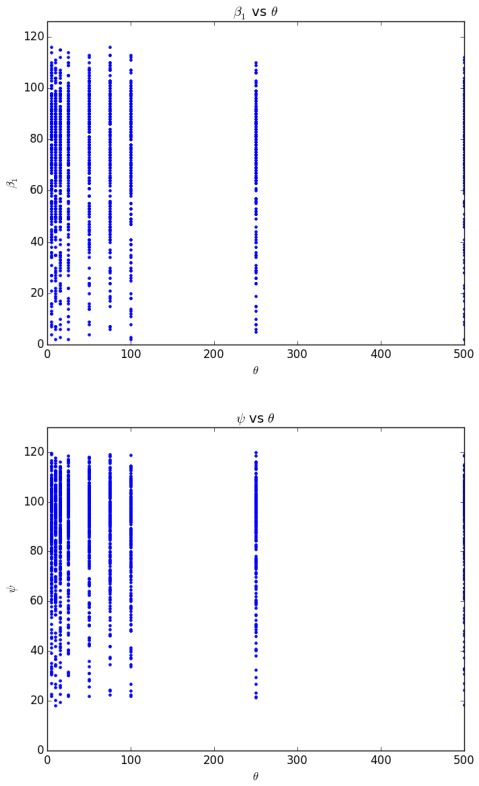

Fig. 8. ψ and β_1 are uncorrelated with variance in θ.

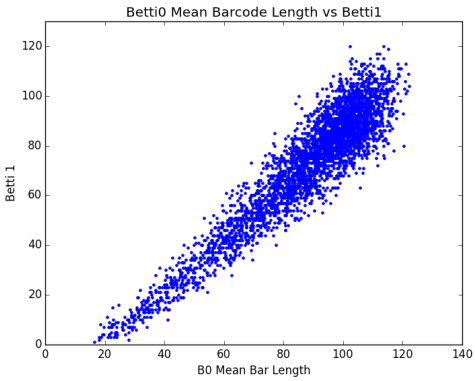

Fig. 9. ψ and β_1 are tightly correlated over variable values of ρ.

selves.

We computed each feature from the set of 3600 simulated sequence alignments with varying values of ρ and θ, and produced similar correlation plots showing relationships between each pair of variables. We find that neither β_1 nor ψ is correlated with Watterson's θ, thus we can confidently assert that this is not a confounding factor in our analyses (Figure 8).

In contrast to this, we do find that β_1 and ψ are correlated quite tightly in our work (Figure 9). While this correlation is expected since both statistics should be elevated in the presence of recombination events, we have noted in other experiments that ψ nevertheless adds unique information.

1. Carlsson G (2009) Topology and data. *Bulletin of the American Mathemaical Society* 46(2):255–308.
2. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12:2825–2830.