Turkana Basic Model Posteriors

Turkana SLiM Model Building and Simulation Pipeline Review

- Build model with inferred parameters that incorporates selection
- Calculate observed summary stats on observed data (Amanda)
- Simulate forward in time simulations of model by sampling s and t
- Calculate summary stats on simulated data (s, pi, TajimaD, EHH)
    - Output of each run is appended to master summary stat csv
        - S value | T value | Simulation ID# | variants | pi | TD | EHH.up | EHH.down
- Run ABC algorithm to determine proportion of accepted trajectories given observed data and tolerance threshold
    - Accepted trajectories are stored to create posteriors of s and t

The Approximate Bayesian Computation (ABC) algorithm is detailed below
This overview details parameter selection of s and t using the rejection method

ABC OVERVIEW

ABC → a family of approximate inference methods - ABC is considered a post processing tool where simulations and the calculation of summary stats that summarize the simulations is performed elsewhere

ABC is appropriate to use when a likelihood function may be impossible to compute and the simulation of the model is straightforward

The rejection algorithm works as follows for parameter selection:
After parameter values are drawn from a prior distribution and corresponding datasets are simulated and each simulation is converted to a vector of summary statistics → a distance between the simulated and the summary statistics of the observed data is calculated.
Parameters producing [euclidean] distances below some threshold [tolerance] are accepted and form a sample from an approximation to the posterior distribution
The rejection method assumes that there is a sufficient number of simulations, a threshold that is insensitive to the summary statistics, summary statistics that are informative to the parameters of interest, and a weighted distance function (euclidean distance).

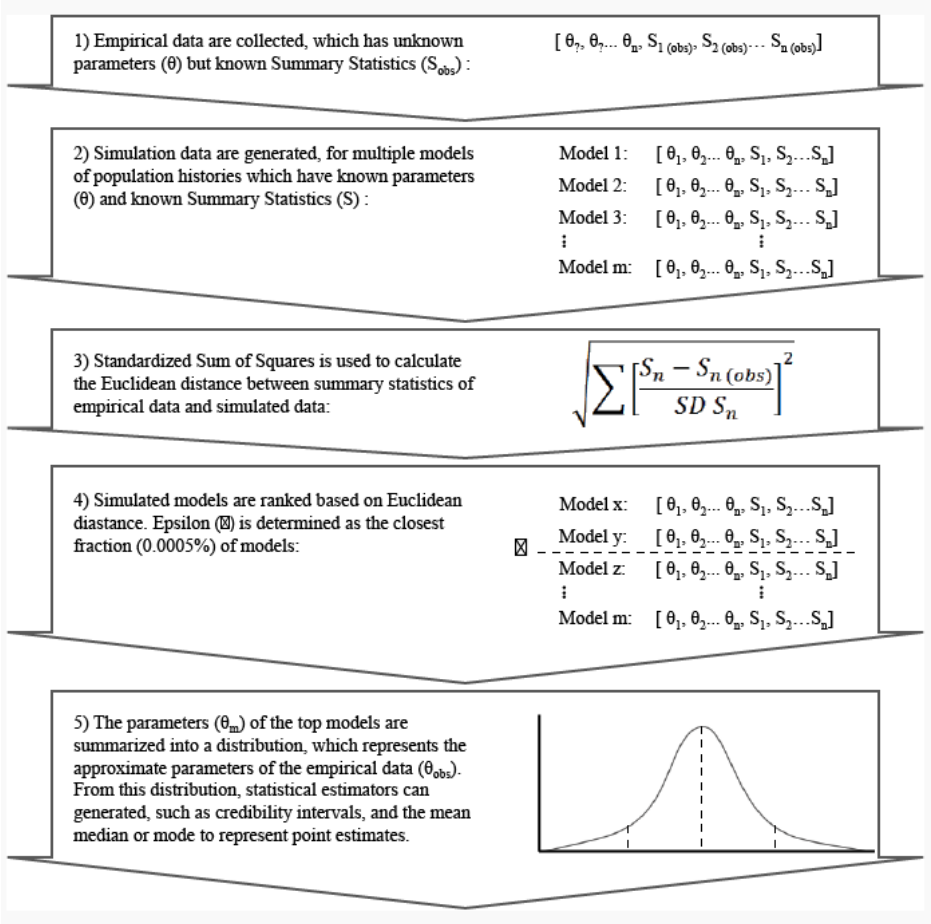
Euclidean Distance and Tolerance
Tolerance rate → the percentage of accepted simulations
Euclidean Distance → Each summary statistic is standardized by a robust estimate of the standard deviation (the median absolute deviation). If the distance between the set of observed and simulated statistics is less than the given threshold, the parameter value is accepted
This distance measure is essentially a ranking based on closeness.

$$d(\boldsymbol{s}, \boldsymbol{s}_{\text{obs}}) = \left[\sum_{i=1}^{m} \left(\frac{s_i - s_{\text{obs},i}}{\sigma_i}\right)^2\right]$$

d = distance between set of simulated stats to observed
S = summary statistic
$\sigma i \ = standard\ deviation\ of\ the\ summary\ stats\ from\ a\ simulation$

Cross validation → To evaluate the effect of the choice of the tolerance rate on the quality of the estimation with ABC
ABC can implement a leave-one-out cross-validation to evaluate the accuracy of parameter estimates and the robustness of the estimates to the tolerance rate.
To perform cross-validation, the ith simulation is randomly selected as a validation simulation, its summary statistic(s) are used as pseudo-observed summary statistics, and its parameters are estimated with the function abc using all simulations except the ith simulation. The process is repeated n times, where n is the number of simulations (so- called n-fold cross-validation). The prediction error is calculated to determine sensitivity to the tolerance rate.

ABC Visual: Basic steps



Considerations and Results

Summary of number of simulations

Recall a priori ranges of s and t

s ~ 0 - 0.4 (by 0.01)

t ~ 50 - 2500 generations (by 5 generations)

= 20459 possible pairs of s and t X 4 replicates of each pair

Pairs are replicated to emulate stochasticity

Total forward in time simulations under basic model	81835
Accepted simulations based on 0.01 tolerance	819

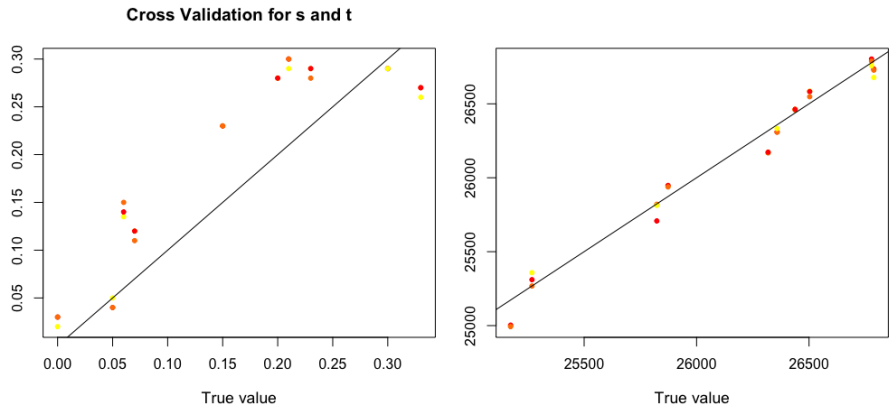
Cross validation and tolerance

Before moving to the inference step, I assessed if ABC was able to estimate the parameters (s and t) at all. Here I can determine the accuracy of ABC and the sensitivity of estimates to the tolerance rate. The following code evaluates the accuracy of estimates of s and t under three tolerance rates using the rejection method (I tested 0.005, .01, 0.05).

Prediction error based on a cross-validation sample of 10

	s	t
0.005	0.1949608	0.1470498
0.01	0.2001260	0.1496692
0.05	0.2082283	0.1660849

The summary above shows the prediction error under the three tolerance rates.
Points of the cross-validation plot are scattered around the identity line indicating that S and T can only be well estimated using the three summary statistics ~80% of the time. Estimates were slightly inaccurate for s and t, but also insensitive to the tolerance rate. The prediction error is not as low as I would like it, and the error is not totally independent of the tolerance rate. This tells me that either I do not have enough simulations or my summary stats are not a good measure for predicting s and t.

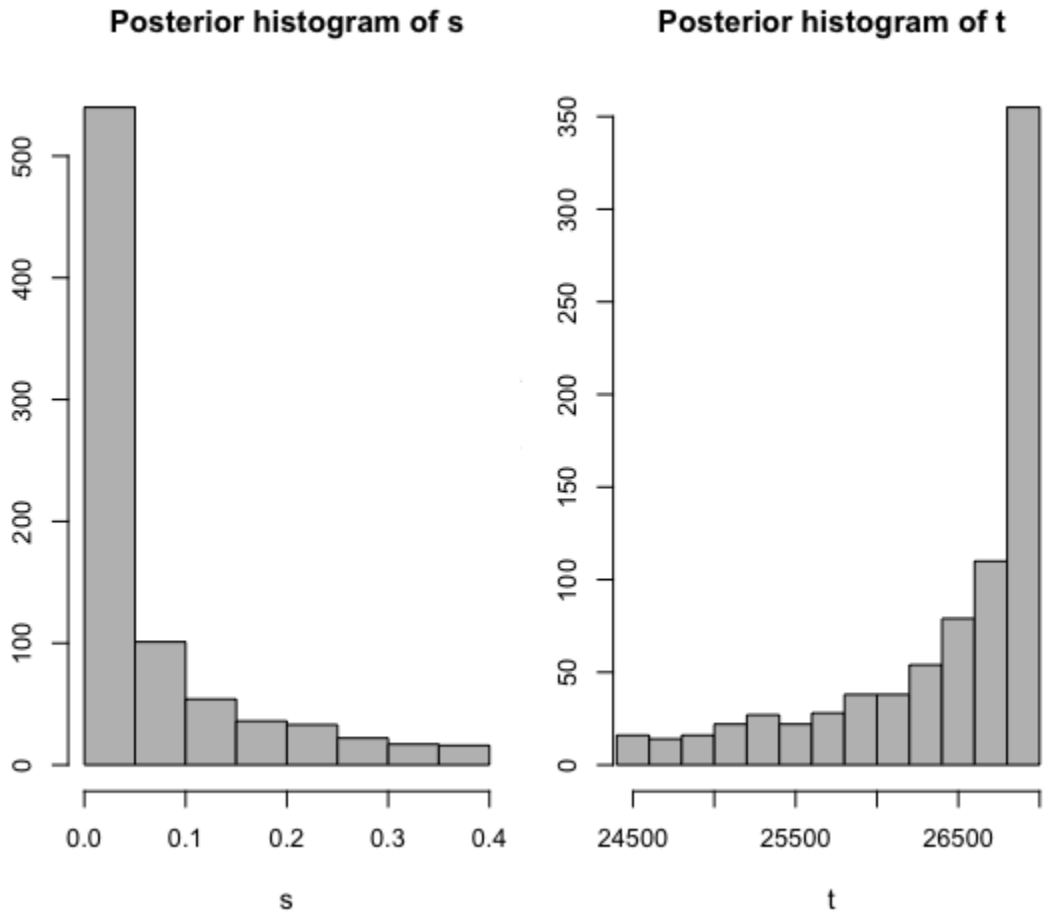


Parameter inference estimation summary

The posterior density is representative of our uncertainty of s and t. The estimation of s and t using ABC in indicative of our acceptance that we cannot know the exact values of s and t that models our observed data. Instead, by generating a distribution of parameters derived from the selected simulations by the rejection algorithm, it is assumed that the proportion in which a parameter value occurs in the distribution reflects the likelihood of it occurring in the natural population, and thus its posterior probability. Below, s and t are estimated from the distribution of the simulated values as an the 95% confidence interval, the median (with 50% probability, the value is below this), the mean (the expected value) and the mode (the most probable value).

	s	t
Min.:	0.0000	24478
2.5% Perc.:	0.0000	24710
Median:	0.0200	26733
Mean:	0.0654	26415
Mode:	**0.0068**	**26875**
97.5% Perc.:	0.3400	26963
Max.:	0.4000	26968

Posteriors distributions



Discussion and suggested changes

The posteriors of s and t are skewed to give [modal] estimates of s {~<0.006} and t {~2000 years) - meaning this focal allele would have been under weak selection in very recent human history. Though, the confidence intervals for both s and t encapsulate most of the prior ranges - so I would not consider this model to be estimating these parameters accurately. I would not rush to say that the summary stats are not providing a good summarization of the data yet. I believe that the main issue is the frequency conditional and the number of simulations. Even though the model is simple and there may be more to add, I think it would be best to revisit this and to rerun it with some minor changes before we add in major elements.

In the model currently - I have it such that the minimum frequency our focal allele has to reach is 0.15 - though I do not program that the final output of that focal allele frequency has to be 15%. Therefore, after the focal allele reaches 15% (which may take many iterations), the allele frequency is free to deviate - as well as be lost or fix. In my attempts to calculate the EHH summary stat I realized this is a huge issue, because at least half of my simulations do not contain the focal allele due to loss or fixation. This is the main reason I did not include EHH as a summary stat for this ABC analysis. This discrepancy is also the reason why I believe this basic model is struggling to estimate s and t.

It is clear that there are not enough simulations to accurately predict s and t given the cross validation measure. In my readings, I think having a posterior distribution of 10000 simulations would be the most beneficial. To do this, I will need to run ~1,000,000 forward in time simulations at a 0.01 tolerance threshold. I will also change the step size and replicate count for the s and t ranges to get more of a sampling of the parameter values.

The new number of simulations will be

s ~ 0 - 0.4 (by 0.001)
t ~ 50 - 2500 generations (by 5 generations)
= ~200000 possible pairs of s and t X 5 replicates of each pair

Total forward in time simulations under basic model	1,000,000
Accepted simulations based on 0.01 tolerance	10000

With a rerun of this model under these new additions and variables, the CI may be smaller to run posterior predictive checks.

Next steps

1. Confirm ancestral and derived states / chose new SNP / recalc ss(obs)
2. Change frequency conditional in SLiM script
3. Edit pipeline to reflect new number of simulations, ranges, and replicates
4. Rerun SLiM
5. Calculate summary stats on new data including EHH
6. Run cross validation and ABC rejection algorithm
7. Perform posterior predictive checks

Sources

Aeschbacher, Simon, Mark A. Beaumont, and Andreas Futschik. “A Novel Approach for Choosing Summary Statistics in Approximate Bayesian Computation.” *Genetics* 192, no. 3 (November 2012): 1027–47. https://doi.org/10.1534/genetics.112.143164.

Beaumont, Mark A., Wenyang Zhang, and David J. Balding. “Approximate Bayesian Computation in Population Genetics.” *Genetics* 162, no. 4 (December 1, 2002): 2025–35. https://www.genetics.org/content/162/4/2025.

Burr, Tom, and Alexei Skurikhin. “Selecting Summary Statistics in Approximate Bayesian Computation for Calibrating Stochastic Models.” Research Article. BioMed Research International. Hindawi, September 1, 2013. https://doi.org/10.1155/2013/210646.

Csilléry, Katalin, Michael G. B. Blum, Oscar E. Gaggiotti, and Olivier François. “Approximate Bayesian Computation (ABC) in Practice.” *Trends in Ecology & Evolution* 25, no. 7 (July 1, 2010): 410–18. https://doi.org/10.1016/j.tree.2010.04.001.

Csilléry, Katalin, Olivier François, and Michael G. B. Blum. “Abc: An R Package for Approximate Bayesian Computation (ABC).” *Methods in Ecology and Evolution* 3, no. 3 (2012): 475–79. https://doi.org/10.1111/j.2041-210X.2011.00179.x.

darrenjw, Author. “Introduction to Approximate Bayesian Computation (ABC).” *Darren Wilkinson’s Blog* (blog), March 31, 2013. https://darrenjw.wordpress.com/2013/03/31/introduction-to-approximate-bayesian-computation-abc/.

Fearnhead, Paul, and Dennis Prangle. “Constructing Summary Statistics for Approximate Bayesian Computation: Semi-Automatic Approximate Bayesian Computation [with Discussion].” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 74, no. 3 (2012): 419–74. https://www.jstor.org/stable/41674639.

Pacchiardi, Lorenzo, Pierre Künzli, Marcel Schöngens, Bastien Chopard, and Ritabrata Dutta. “Distance-Learning For Approximate Bayesian Computation To Model a Volcanic Eruption.” *Sankhya B*, January 24, 2020. https://doi.org/10.1007/s13571-019-00208-8.

Prangle, Dennis. “Adapting the ABC Distance Function.” *ArXiv:1507.00874 [Stat]*, December 15, 2015. http://arxiv.org/abs/1507.00874.

Sunnåker, Mikael, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. “Approximate Bayesian Computation.” *PLOS Computational Biology* 9, no. 1 (January 10, 2013): e1002803. https://doi.org/10.1371/journal.pcbi.1002803.

Turner, Brandon M., and Trisha Van Zandt. “A Tutorial on Approximate Bayesian Computation.” *Journal of Mathematical Psychology* 56, no. 2 (April 1, 2012): 69–85. https://doi.org/10.1016/j.jmp.2012.02.005.

Veeramah, Krishna R., Daniel Wegmann, August Woerner, Fernando L. Mendez, Joseph C. Watkins, Giovanni Destro-Bisol, Himla Soodyall, Leslie Louie, and Michael F. Hammer. “An Early Divergence of KhoeSan Ancestors from Those of Other Modern Humans Is Supported by an ABC-Based Analysis of Autosomal Resequencing Data.” *Molecular Biology and Evolution* 29, no. 2 (February 2012): 617–30. https://doi.org/10.1093/molbev/msr212.

Villanea, Fernando. “Approximate Bayesian Computation.” Accessed January 12, 2021. https://hub.wsu.edu/fernandovillanea/approximate-bayesian-computation/.