

Paper reference

Stephen F. Schaffner, Catherine Foo, Stacey Gabriel, David Reich, Mark J. Daly and David Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* 15:576-1583, 2005

Abstract

Population genetics models are useful for obtaining background expectations about genetic variation. In this work, the authors calibrate such a model using empirical data sets, specifically *Single Nucleotide Polymorphism* (SNP) data from three populations (West Africa, East Asia and Europe). The authors use a total of 15 measures to compare the simulated data to the empirical observations. For each measure they compute the root-mean-square error (RMSE) of the simulated value with respect to the mean empirical value. The overall goodness of fit of a particular model is obtained by calculating the total RMS discrepancy of all measures.

The calibration of the model starts from a standard neutral model that includes the separation between the African and the non-African population, and the subdivision of the later into the European and East Asia populations. This model has a root-mean-square-error (RMSE) of 4.7. The model is refined using a stepwise approach: a set of parameters is added to the model and those parameters are optimized in order to minimize the error. This step is then repeated with additional parameters. In the first of these steps, the RMSE for the single-locus measures is reduced to 1.15 by increasing the fraction of low-frequency alleles and adding population bottlenecks and migration populations to the model. The recombination rate is then optimized to match the observed heterozygosity. The recombination model is improved in order to obtain a larger, and thus more realistic, level of linkage disequilibrium (LD) than with the neutral model. This is done by adding large-scale variations as well as fine-scale variations such as localized hotspots to the model.

The best-fitting model has an overall RMSE of 1.35 with respect to the mean empirical values. The model is evaluated by generating predictions for the X chromosome of the same population, which was not used during the calibration. The calibrated model performs significantly better than the neutral model (RMSE of 0.97 instead of 1.51). The calibrated model is also shown to be able to simulate haplotype blocks significantly better than the neutral model. Finally, the authors show that the variations found in a set of 100 genes are reproduced by the calibrated model and can thus be explained without having to hypothesize positive selections.

Discussion

The calibration process described in this article allows to have, for the first time, a simulation of population genetics that faithfully reproduces a wide range of properties observed in data from multiple populations. In particular, this model is able to reproduce the LD patterns accurately, which isn't the case with the standard neutral model, thus providing a baseline for studies where LD is important. The authors clearly stress out the benefits of this work, as well as possible limitations of their approach. In particular, they note that their method was not meant to be exhaustive, and that better parameter choices and better models are likely to exist. Furthermore, the emphasis of the work was on producing more realistic simulations, not on getting information about population evolution from the parameters: as the authors say, they are values that happen to generate useful simulations. However, as shown in the last example of the paper, having a calibrated model is helpful for obtaining random instances of population evolution that are consistent with our current empirical knowledge. Empirically observed variations can be compared to these instances, and if there is no significant difference, then the model captures these differences and there is no need to formulate additional hypotheses. Having a more precise model leads to a more precise null hypothesis and therefore increases the discriminative power of such comparisons.

The main issue with any calibration approach is validation. In this work, the available data is split into a training set (all autosomes) and a validation set (the X chromosome). These sets are, however, significantly different, for example in their size. As indicated in the Methods section, scaling between the learned parameters and the parameters used in the validation was necessary. The exact scaling and correction, and its influence on the results, are not explained. It is difficult to find better means of validation. While the differences between the X chromosome and the autosome require parameter tuning, using a cross-validation on the autosomes is likely to result in test sets that are not independent enough from the respective training set. It is however important to note that, in this context, being able to obtain good results on the training data is already significant. The structure of the model and the tuning of the parameters are clearly defined and plausibly justified, and the risk of overfitting is thus limited compared to a method that would learn a complete model.

Having a calibrated model is useful for a wide variety of studies in population genetics. In particular, it provides a way of detecting variations in the experimental observations that are already explained by the assumptions made in this model. Population genetics model allow for a wide range of plausible parameters. It is therefore important for any work based on this calibrated model to clearly understand the assumptions that were made and which data was used for the calibration.