Sifan Tao and David Bass
Professor Xiwei Tang
STAT 6160: Experimental Design
9 May 2022

Examining coalescent times of evolutionary simulations in *msprime*

Problem description

We seek to study the relationship between various factors in population genetics and the coalescent time, the time that has passed since individuals in the current population evolved from their most recent common ancestor (MRCA). When estimated from data about a current population, coalescent time is a measure of the complexity of the evolution experienced by a population since its divergence from its MRCA, with greater coalescent times indicating greater complexity. We use the population genetics simulator *msprime* to perform ancestry simulations of standard populations under various levels for three quantitative factors: population size (`pop_size`), sequence length (`sequence_length`), and recombination rate (`recombination_rate`). We simulate each population under one of two different models, the discrete-time Wright–Fisher (DTWF) model and the Hudson model (`dtwf_vs_hudson`), and under one of two different selection scenarios, no selective sweep versus a selective sweep to fixation (`selective_sweep`). As an estimator of coalescent time, we use the maximum root time (`max_root_time`) of a population, which is the greatest length of time in any branch of the phylogenetic trees that *mpsrime* uses to describe an ancestry simulation.

We expect greater values of each of the quantitative factors to be correlated with greater maximum root time, as larger populations, larger genomes, and more recombination should add to the complexity of evolution. While the discrete-time Wright–Fisher and Hudson models are commonly applied in population genetics simulations, we cannot immediately formulate a hypothesis as to their effect on coalescent time. The presence of a selective sweep to fixation, in which a highly beneficial mutation occurs and proceeds to spread via natural selection to the entire population, is expected to introduce greater complexity of evolution, and thus result in a greater maximum root time, than the absence of a selective sweep, in which evolution is primarily driven by random drift.

We ran 216 simulations, two for each combination of the three levels of `pop_size` (10, 100, 1,000), three levels of `sequence_length` (1,000, 10,000, 100,000), three levels of `recombination_rate` (10^{-7}, 10^{-8}, 10^{-9}), two levels of `dtwf_vs_hudson` (0=DTWF, 1=Hudson), and two levels of `selective_sweep` (0=no selective sweep, 1=selective sweep to fixation). As our dependent variable, we calculated `max_root_time` for each simulation.

Designing and studying experiments for the study of this data will allow us to better understand the impacts of important factors in a population on the complexity of the evolution that it experiences.

To view the original data set and the code used to generate it, see our GitHub repository at https://github.com/DavidB256/STAT-6160-project. More information about *msprime* can be found in the online manual at https://tskit.dev/msprime/docs/stable/intro.html. Comprehensive information about the biology referenced in this report can be found in John H. Gillespie's *Population Genetics: A Concise Guide* (2nd ed.)

Experiment analysis

Randomization and blocking

Because we have generated our data set to our specifications with full knowledge of how *msprime* functions, there is no need to consider randomization or blocking of the data. The two trials performed for each of the 108 treatment combinations allow for us to capture the variation present, as *msprime* simulations are not deterministic.

2^2 factorial design

Keeping the values of the quantitative factors fixed at intermediate values (`pop_size`=100, `sequence_length`=10,000, `recombination_rate`=10^{-8}), we will examine the effects and interaction of the two binary categorical factors, `dtwf_vs_hudson` and `selective_sweep` with a 2^2 factorial design. We are especially interested in the interaction effect of these factors, as the DTWF and Hudson models may handle selective sweeps in distinct ways.

ANOVA

We will start by performing 3-factor ANOVA on the quantitative factors, `pop_size`, `sequence_length`, and `recombination_rate`. Because the levels of each of these factors differ by multiples of 10, we will perform a logarithmic transform on the data in order to make it conform to ANOVA's normality assumption. We will keep the categorical factors fixed at the baseline levels of `dtwf_vs_hudson`=1 (Hudson model, which is the default in *msprime*) and `selective_sweep`=0 (no selective sweep).

Using the same transformed data, we will also perform 5-factor ANOVA on all factors.

Statistical analysis

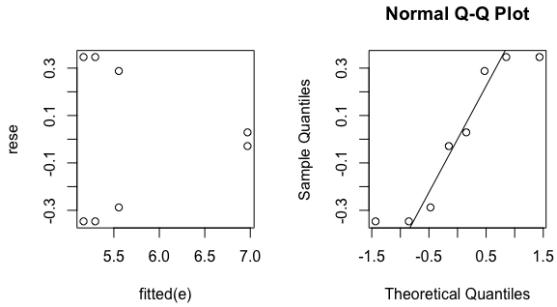
Unless stated otherwise, we will proceed with a significance level of $\alpha=0.05$ in this report. Because the levels of the quantitative factors differ by multiples of 10, we applied a logarithmic transform on the response variable, `max_root_time`.

2^2 factorial design for `dtwf_vs_hudson` and `selective_sweep`

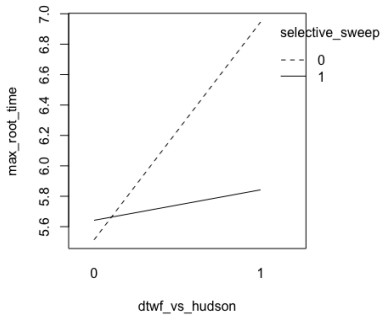
```
Analysis of Variance Table

Response: max_root_time
                            Df  Sum Sq Mean Sq F value  Pr(>F)
dtwf_vs_hudson               1 2.12612 2.12612 13.1298 0.02229 *
selective_sweep              1 0.82773 0.82773  5.1116 0.08660 .
dtwf_vs_hudson:selective_sweep  1 1.19136 1.19136  7.3572 0.05340 .
Residuals                    4 0.64772 0.16193
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA table R output shows that `dtwf_vs_hudson` is a significant predictor of `max_root_time`, while `selective_sweep` and their interaction term are not significant predictors. Note that, in order to isolate the effects of the categorical factors, we are only considering the eight observations at intermediate values of each of the quantitative factors. The low *p*-value of the interaction term indicates that the term could very possibly be deemed significant if more data were generated. The potential significance of the interaction term would be supported by Hamada and Wu's effect heredity principle, as one of the first-order terms is significant in our experiment.



The residual plot and Q–Q plot raise concerns of the model adequacy. The residuals exhibit "fanning-in," indicating heteroscedasticity. However, the balanced design of our experiment, with precisely two observations per treatment combination, makes the *F*-tests in ANOVA robust against nonconstant variance among groups. The Q–Q plot shows slight, but not problematic, deviations from normality, as the "S" shape of the points around the central line indicate a slightly leptokurtic distribution of `max_root_time`. These issues may be due to the small subset, only eight observations, of the data used in this experiment.



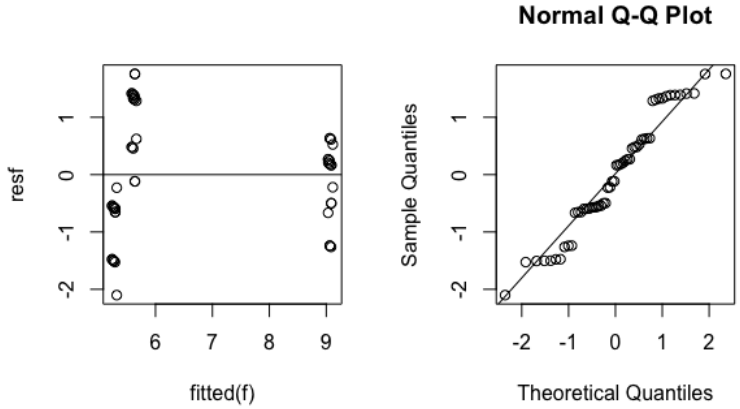
The interaction plot shows that the interaction between `dtwf_vs_hudson` and `selective_sweep` appears to be significant qualitatively, as the lines are not parallel, supporting our previous conjecture that the interaction of the categorical factors may be significant in a different context.

3-way ANOVA for quantitative factors

```
Analysis of Variance Table

Response: max_root_time
                  Df  Sum Sq Mean Sq  F value Pr(>F)
pop_size           1 158.600 158.600 145.0848 <2e-16 ***
sequence_length    1   0.010   0.010   0.0094 0.9231
recombination_rate 1   0.023   0.023   0.0213 0.8844
Residuals         50  54.658   1.093
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA table R output shows that `pop_size` is a highly significant predictor of `max_root_time` in the 3-way ANOVA model, with an *p*-value below R’s underflow threshold, while the `sequence_length` and `recombination_rate` are insignificant as predictors.



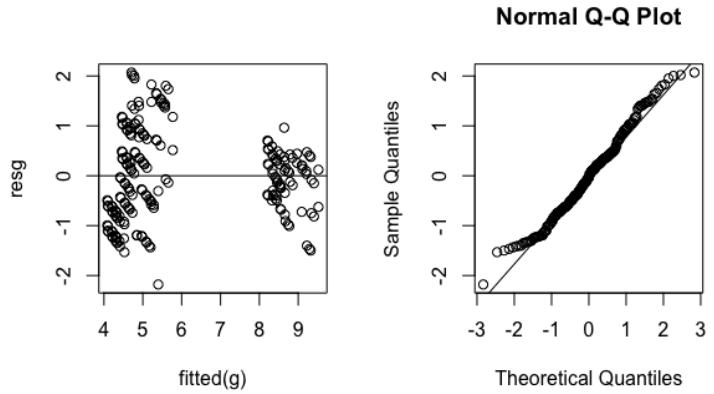
The residual plot posts some concerns of the model adequacy. The residual plot shows that the constant variance assumption of ANOVA may be violated in this model. However, we can repeat the same justifications for these issues mentioned in the discussion of the 2^2 factorial model in order to proceed with our ANOVA results.

5-way ANOVA for all factors and interaction between categorical factor

```
Analysis of Variance Table

Response: max_root_time
                                           Df Sum Sq Mean Sq  F value    Pr(>F)
pop_size                                    1 748.59  748.59 959.2879 < 2.2e-16 ***
sequence_length                             1   1.75    1.75   2.2424   0.13578
recombination_rate                          1   1.91    1.91   2.4505   0.11900
factor(dtwf_vs_hudson)                      1  19.51   19.51  25.0003 1.212e-06 ***
factor(selective_sweep)                     1   4.25    4.25   5.4513   0.02050 *
factor(dtwf_vs_hudson):factor(selective_sweep)  1   3.91    3.91   5.0098   0.02626 *
Residuals                                 209 163.09    0.78
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA table shows that `sequence_length` and `recombination_rate` are insignificant in this model, while all the other three first-order factors and the interaction term are significant. This is consistent with the findings from the previous 2^2 factorial design and 3-way ANOVA.



As in the previous two models, heteroscedasticity and a lack of normality continue to be potential issues for consideration.

Conclusion

After considering all three models and their diagnostics, we choose to fit a regression model with OLS, using `pop_size`, `dtwf_vs_hudson`, `selective_sweep`, and the interaction between `dtwf_vs_hudson` and `selective_sweep` as predictors.

```
Call:
lm(formula = max_root_time ~ pop_size + factor(dtwf_vs_hudson) +
    factor(selective_sweep) + factor(dtwf_vs_hudson) * factor(selective_sweep),
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.93755 -0.67868 -0.01482  0.52147  2.12956

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     4.2432110  0.1309293  32.408  < 2e-16 ***
pop_size                                        0.0041648  0.0001353  30.777  < 2e-16 ***
factor(dtwf_vs_hudson)1                         0.8701314  0.1710876   5.086 8.06e-07 ***
factor(selective_sweep)1                       -0.0116072  0.1710876  -0.068   0.9460
factor(dtwf_vs_hudson)1:factor(selective_sweep)1 -0.5381309  0.2419544  -2.224   0.0272 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.889 on 211 degrees of freedom
Multiple R-squared:  0.8232,    Adjusted R-squared:  0.8198
F-statistic: 245.6 on 4 and 211 DF,  p-value: < 2.2e-16
```

The R output yields the following linear regression model:

$$\texttt{max\_root\_time} = 4.2432 + 0.00416 \times \texttt{pop\_size} + 0.8701 \times \texttt{dtwf\_vs\_hudson} - 0.0116 \times \texttt{selective\_sweep} - 0.5381 \times \texttt{dtwf\_vs\_hudson} \times \texttt{selective\_sweep}$$

As `selective_sweep` was deemed a significant predictor in previous testing, we neglected to perform the Wald test that would suggest removing it from this linear regression model with a *p*-value of 0.9460. This model features a positive correlation between `pop_size` and `max_root_time`. It gives that the Hudson model corresponds to greater maximum root times than the DTWF model and that the presence of a selective sweep corresponds to lesser maximum root times. The interaction term gives that the combination of the DTWF model and the presence of a selective sweep also corresponds to a lesser maximum root time. Since we used maximum root as an estimator of coalescent time, we ultimately see that

These conclusions from the model support our initial hypothesis that greater population size contributes to evolutionary complexity, increasing coalescent time and thus maximum root time. The insignificance of recombination rate as a predictor of maximum root time is a surprise, as chromosomal recombination in a population is often a dominant contributor to evolutionary phenomena like genetic hitchhiking. It is also surprising to see such a significant difference in maximum root time between the DTWF and Hudson models, as they are often used interchangeably in the study of population genetics. The negative correlation between `selective_sweep` and `max_root_time` also challenges our initial hypothesis that the presence of a selective sweep should complicate evolution, thus increasing coalescent time.

In considering several single-factor *F*-tests consecutively, we have incurred a family-wise error rate that is worth considering. While we did not perform enough hypothesis tests to warrant a strategy as conservative as the Bonferroni correction, it would be worthwhile to reconsider our conclusions with Tukey’s method or the Benjamini–Hochberg procedure.

A further model that we considered analyzing is a mixed-effects model containing all five factors, but with levels of the quantitative factors selected randomly. This would allow for us to better study the variance of coalescent times in populations. However, it would be a contrived model, as we are already able to study more powerful fixed-effects models with the current data, which was generated to our specifications.

From a biological perspective, a potential further step would be to examine the differences between the DTWF and Hudson models in greater detail. Is our finding that `dtwf_vs_hudson` is a significant predictor of `max_root_time` in every model that we tested supported by the theoretical differences between the models? Additionally, it would be interesting to examine a different estimator of coalescent time than maximum root time; like how maximum likelihood estimation only considers the mode of a likelihood function, the maximum root time only considers the longest branch of a phylogenetic tree, but conveys no information about the distribution of shorter trees.

While our sample size was severely restricted by the computational expense of performing simulations in *msprime*, more observations per treatment group could be easily obtained by utilizing a supercomputer like UVA's Rivanna cluster. To add more levels of the factors in our experiments without significantly increasing computational needs, we could have constructed incomplete designs like a BIBD or Latin hypercube design.

Works cited

- Dominic Nelson, Jerome Kelleher, Aaron P. Ragsdale, Claudia Moreau, Gil McVean and Simon Gravel (2020), Accounting for long-range correlations in genome-wide simulations of large cohorts, PLOS Genetics 16(5): e1008619. https://doi.org/10.1371/journal.pgen.1008619
- Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E Castedo Ellerman, Jared G Galloway, Ariella L Gladstein, Gregor Gorjanc, Bing Guo, Ben Jeffery, Warren W Kretzschumar, Konrad Lohse, Michael Matschiner, Dominic Nelson, Nathaniel S Pope, Consuelo D Quinto-Cortés, Murillo F Rodrigues, Kumar Saunack, Thibaut Sellinger, Kevin Thornton, Hugo van Kemenade, Anthony W Wohns, Yan Wong, Simon Gravel, Andrew D Kern, Jere Koskela, Peter L Ralph and Jerome Kelleher (2022), Efficient ancestry and mutation simulation with msprime 1.0, Genetics, Volume 220, Issue 3. http://doi.org/10.1371/journal.pcbi.1004842
- Gillespie, John H. Population Genetics: A Concise Guide. 2nd ed. Johns Hopkins University Press, 2004.
- Jerome Kelleher, Alison M Etheridge and Gilean McVean (2016), Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes, PLOS Comput Biol 12(5): e1004842. http://doi:10.1371/journal.pcbi.1004842
- Kelleher J, Etheridge AM, McVean G (2016) Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Computational Biology 12(5): e1004842. https://doi.org/10.1371/journal.pcbi.1004842
- Montgomery, Douglas C., and Academic, General Engineering & Project Administration Knovel. Design and Analysis of Experiments. Eighth ed. John Wiley & Sons, Inc, 2013.
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/