Revista Física no Campus, v. 3, n. 1, p. 14-22, Campina Grande - PB (2023)

Universidade Estadual da Paraíba

http://novo.revista.uepb.edu.br/fisicanocampus

Artigo Original

ISSN: 2764-5924

Relação entre tempo de fixação de um alelo e tamanhos populacionais em haploides assexuados por meio de um modelo de simulação no R

Relationship between allele fixation time and population sizes in asexual haploids using a simulation model in R

[1]Elias Dias Coelho Neto

[1]Universidade Estadual da Paraíba, Campina Grande, PB, Brasil.

Resumo: Por meio de um modelo de simulação implementado em linguagem R, investigamos os tempos de fixação de um alelo em populações de haploides assexuados. Consideramos diversos tamanhos populacionais, embora pequenos. As forças evolutivas empregadas foram mutação e deriva genética. Consideramos diferentes números de alelos iniciais. No início das simulações, as frequências dos alelos são uniformes. Utilizando box-plots, os resultados revelaram diferenças significativas entre os logaritmos dos tempos de fixação quando os tamanhos populacionais diferem, pelo menos, na razão de quatro para um. Para populações de tamanho fixo, comparamos taxas de mutação pequenas e grandes, o que levou a resultados equivalentes. Diferentes números de alelos iniciais não influenciaram os tempos de fixação.

Palavras chave: Simulações no R, populações de haploides assexuados, mutação, deriva genética.

Abstract: Using a simulation model implemented in R language, we investigated the fixation times of an allele in populations of asexual haploids. We consider diverse population sizes, albeit small ones. The evolutionary forces employed were mutation and genetic drift. We consider different numbers of initial alleles. At the beginning of the simulations, the allele frequencies are uniform. Using box-plots, the results revealed significant differences between logarithms of fixation times when population sizes differ by at least a four to one ratio. For fixed-size populations, we compared small and large mutation rates, which led to equivalent results. Different numbers of initial alleles did not influence fixation times.

eliasdcn@servidor.uepb.edu.br

Keywords: Simulations in R, asexual haploid populations, mutation, genetic drift.

eliasdcn@servidor.uepb.edu.br

Introdução

A física da matéria condensada é o campo da física que se ocupa com estudos das propriedades físicas macroscópicas e microscópicas da matéria. Na fase condensada normalmente a matéria se mostra em unidades organizadas em um sistema em que a interação é uma regra. Este é um campo muito vasto da física contemporânea. Dentro dele o subcampo de nosso interesse é a genética populacional, que se ocupa em estudar as diferenças genéticas dentro (ou entre) de populações. A genética populacional moderna engloba estudos teóricos, laboratoriais, de campo e, nas últimas décadas, computacionais. Uma das principais aplicações dos modelos genéticos populacionais são as previsões de sequências de DNA [1].

O diferencial da genética populacional de outros campos das ciências para modelar a evolução é a ênfase nos fenômenos aleatórios, como exemplo podemos citar mutação pontual e deriva genética. Isto a torna uma ciência apropriada para comparação de dados de genomas populacionais [2–5]. Nesse sentido, o computador tem desempenhado um papel indispensável na análise de dados do genoma e no desenvolvimento de modelos genéticos para visualização, simulação, análises numéricas e estatísticas de um grande leque de fenômenos evolutivos que, normalmente, são abordados em laboratórios para resolver problemas da genética de populações ou da evolução das espécies [6]. O benefício de simular cenários evolutivos, dos mais simples até os mais complexos, para geração de dados em uma diversidade de estudos tem impulsionado o desenvolvimento de dezenas de programas computacionais confiáveis, com bom desempenho e de ampla aplicação [7, 8]. Todavia, é notado grande dificuldade na compreensão, principalmente para o aluno de graduação, dos atuais códigos abertos como os implementados nas ferramentas de programação R e Python, isto se deve à grande complexidade dos fenômenos genéticos abordados.

Considerando o que foi descrito acima e as dificuldades enfrentadas na implementação de bibliotecas e de códigos para estudos nos campos das ciências, em especial o da genética populacional, e o engajamento entre ensino e pesquisa que deve existir no ambiente universitário, e a iniciativa de professores em propor estudos de temas extracurriculares para seus alunos (ajudando impulsionar suas carreiras), o nosso objetivo aqui é o de propor uma relação entre os tempos de fixação de um alelo que segrega em uma população de indivíduos haploides assexuados em resposta ao tamanho populacional, em que deriva genética e mutação pontual são as forças evolutivas envolvidas na dinâmica. Não há sobreposição de gerações. As populações consideradas são de tamanho pequeno e constantes no tempo, ou seja, é um jogo de soma zero, em que um novo indivíduo surge somente após um evento de morte. A relação que estamos propondo é obtida da análise estatística de dados de simulações computacionais que foram geradas em linguagem de programação R [9]. A dinâmica é iniciada com um número fixo de alelos com frequências uniformes.

Metodologia ou Procedimento Experimental/Prática

Elaboramos um programa em linguagem R para realizar simulações computacionais da dinâmica de segregação de alelos em um único locus do genoma de indivíduos de uma população de tamanho fixo e constante no tempo, em que não há sobreposição de gerações. As forças genética evolucionárias envolvidas são: deriva genética e mutação. Cada simulação é realizada sobre uma população de tamanho fixo e constante no tempo, denominada pela letra N, em que $N \in \{2^1, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8, 2^{10}\}$,

Para incluir mutações ao programa estipulamos uma taxa de mutação U.

eliasdcn@servidor.uepb.edu.br

Note que N é sempre de tamanho pequeno. A deriva genética é uma força evolucionária que faz as frequências dos alelos variarem probabilisticamente de uma geração para outra. O que causa a deriva genética são duas coisas, as variações aleatórias do número de proles de diferentes indivíduos e a natureza estocástica da produção dos gametas. O termo mutação é atribuído à mudança de um alelo para outro. Isto, de fato, altera diretamente as frequências dos alelos. Mutação é uma força evolutiva que sempre está presente em modelos evolucionários clássicos. A taxa de mudanças nas frequências dos alelos devido à mutação é da mesma ordem como a taxa de mutação por locus, algo em torno de 10^{-6} por geração [1]. Para as nossas simulações utilizamos cinco taxas de mutação diferentes, U, quais sejam: 10^{-4}, 5 x 10^{-4}, 10^{-5}, 5 x 10^{-5} e 10^{-6}.

Nosso programa funciona do seguinte modo: Fixemos a nossa atenção em um determinado locus, para uma população de tamanho constante no tempo, N, e para U e número de alelos iniciais fixos, definimos na geração zero e nas gerações seguintes a frequência de cada alelo dada por $F_t(i) = n_t(i)/N$, em que $i = 1, \ldots$ e $n_t(i)$ é o número de indivíduos com o alelo i na geração t; 1 representa o alelo A_1 e 2 o alelo A_2 e assim por diante. Na geração 0 o número de alelos iniciais são uniformemente distribuídos nos N indivíduos. Por exemplo, se iniciarmos com dois alelos, então $F_0(1) = 1/2$ e $F_0(2) = 1/2$. Em seguida, iniciamos um laço perpétuo, que se mantêm perpétuo enquanto $N_a > 1$, em que N_a é o número de alelos em uma geração qualquer. Nesse laço as novas gerações são obtidas. Uma nova geração surge após uma rodada de N reproduções dos progenitores, seguida da morte dos progenitores, em que a deriva genética está presente e, raramente, uma mutação pode ocorrer, como segue: 1) A geração 1 é obtida sorteando-se um número aleatório uniformemente distribuídos no intervalo $(0, 1)$, denominados com x_1. Definimos $F_0(0) = 0$, $S_0(0) = 0$e:

$$S_0(i) = \sum_{k=0}^{i} F_0(k).$$

Se $x_1 > U$ e $S_0(i-1) < x_1 < S_0(i)$, então o primeiro progenitor com alelo i teve sucesso em gerar uma prole também com alelo i; senão, se $x_1 < U$ então um novo alelo surge, uma mutação; 2) Repetimos 1 outras N-1 vezes até obtermos N proles, assim ao final contamos o número de alelos nessa geração, N_a e calculamos $F_1(i)$ para $i = 1, \ldots$ Se $N_a = 1$, o laço perpétuo é interrompido, possibilitando medir o tempo, em gerações, da geração zero até a homozigose populacional. As figuras 1 e 2 resumem o nosso programa.

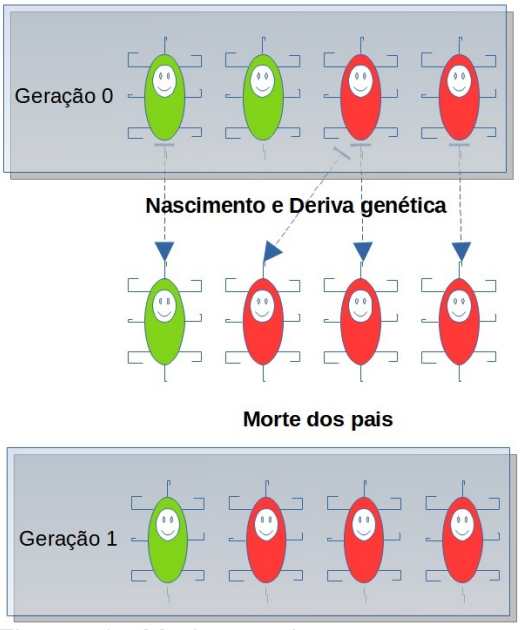

Figura 1. Mudança de geração para uma população de tamanho N = 4, com dois alelos iniciais (em verde o alelo A1 e em vermelho o A2). Neste caso, não ocorreu mutação.

eliasdcn@servidor.uepb.edu.br

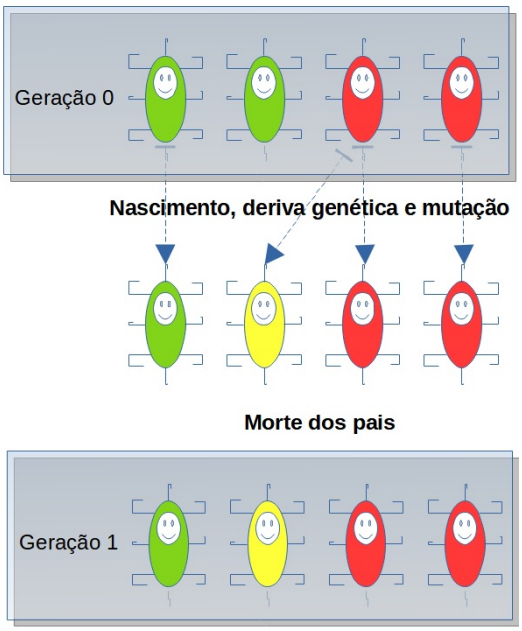

Figura 2. Mudança de geração para uma população de tamanho N = 4, com dois alelos iniciais (em verde o alelo A1 e em vermelho o A2). Neste caso ocorreu mutação no segundo nascimento (em amarelo) e temos um terceiro alelo, A3.

Para cada valor de N citado, fixamos uma taxa de mutação (valores supracitados) e fixado um número inicial de alelos, realizamos 1.000 simulações, exceto para o caso em que $N = 2^{10}$, que realizamos 500 simulações devido ao custo computacional elevado. Ao final de cada simulação, ou seja, quando $N_a = 1$, é anotado o número de gerações até este ponto, denominado por T, que é o nosso tempo de fixação. Foram registrados em um banco de dados o total de 36000 = 1000x1x2x1 + 1000x8x2x2 + 500x1x2x2 tempos de fixação, que são nosso objeto de análise.

Para as análises, optamos por aplicar inicialmente técnicas descritivas, no nosso caso utilizamos diagrama de dispersão e diagramas de caixa (ou *box-plots*) para representar os dados de tempos de fixação em cada valor de N, U e número de alelos iniciais [12]. Buscamos aplicar um modelo de regressão linear simples na modelagem da relação tempo de fixação versus N [13].

Desenvolvimento e Resultados

Relação entre tempo de fixação e tamanhos populacionais

Apresentamos na figura 3 o diagrama de dispersão de um caso particular dos dados dos tempos de fixação versus os tamanhos populacionais transformados pelo log_2, uma vez que aumentamos os valores de N exponencialmente na base 2, facilitando assim a visualização. A primeira característica que notamos aqui é heterocedasticidade, ou seja, o aumento da variabilidade de T quando aumentamos $log_2(N)$. Se focarmos a nossa atenção nos valores máximos de T, notamos um comportamento que lembra um crescimento exponencial.

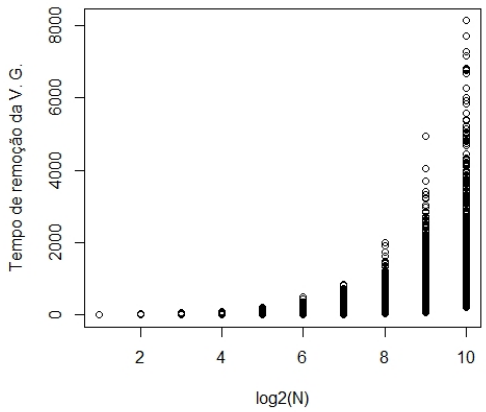

Figura 1: Tempos de fixação versus $log_2(N)$ dos tamanhos populacionais para U = 0,0001 e dois alelos iniciais.

Este comportamento dos dados também é revelado para o outro valor de U, e para quatro alelos iniciais nas duas taxas de mutação, cujas figuras omitiremos aqui para não alongar.

A literatura nos sugere a aplicação de uma transformação logarítmica na variável resposta quando heterocedasticidade está presente, ou quando há esse indicativo. Após a transformação de T para $log(T)$, em

eliasdcn@servidor.uepb.edu.br

que log é o logaritmo natural, obtemos as figuras 3, 4 e 5, que são os *box-plots* dos logaritmos do tempo de fixação para diferentes valores de $log_2(N)$. Como era esperado, a transformação logarítmica de T corrige a heterocedasticidade. Notamos também a existência de outliers, que são os pequenos círculos abaixo e acima dos "bigodes" das caixas. Sabemos que os outliers influenciam a média aritmética e a variância, assim optamos pela mediana como a medida que melhor representa o valor central dos dados. A grande presença de outliers foi o que nos motivou a apresentar os resultados em forma de box-plots, isto porque a mediana não é afetada por valores extremos, ao contrário da média aritmética.

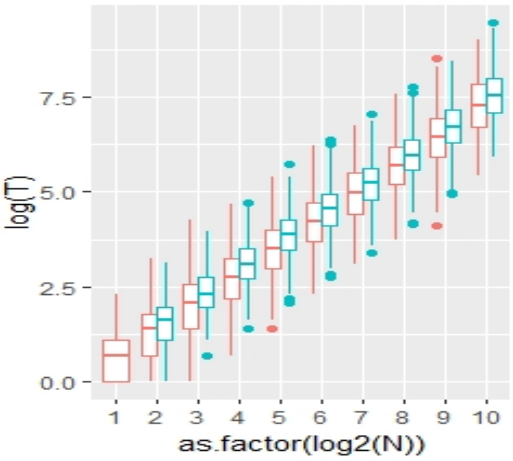

Figura 3. Relação logaritmo do tempo de fixação versus log2 dos tamanhos populacionais para U = 10^{-4}. Em vermelho dois alelos iniciais, em verde quatro.

Focando a atenção nas medianas nas figuras 3 e 4, que são os traços no interiores das "caixas", por exemplo nas caixas vermelhas, notamos que um modelo de regressão linear simples parece se ajustar muito bem aos dados. Isto sugere que E[logT] cresce linearmente a medida que aumentamos os valores de log2N. Na próxima subseção apresentamos uma análise de regressão linear simples.

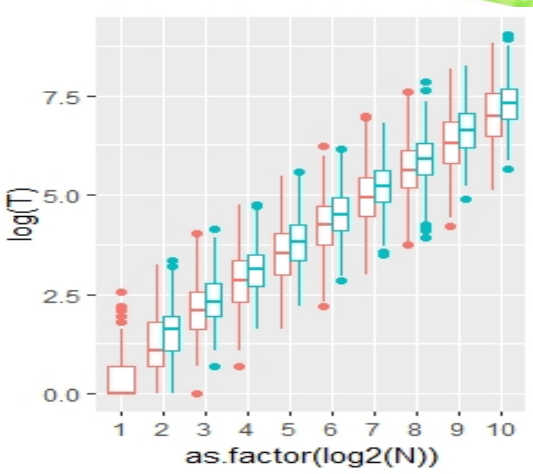

Figura 4. Relação logaritmo do tempo de fixação versus log2 dos tamanhos populacionais para U igual a 10^{-6}. Em vermelho dois alelos iniciais, em verde quatro.

Por vez, fixando nossos olhos em um único tamanho populacional, notamos que o número de alelos iniciais (dois e quatro alelos) não tem influência significativa nos tempos de fixação, porque as projeções das duas "caixas" (vermelha e verde) no eixo das ordenadas, $proj_y\left(T\left(log_2(N)\right)\right)$, resulta na interseção dos intervalos constituídos pelas projeções,

$$\left[proj_y(Q_1), proj_y(Q_3)\right],$$

em que Q_1 e Q_3 são, respectivamente, o primeiro e o terceiro quartil de $log\left(T\left(log_2(N)\right)\right)$. Esta interseção nos mostra que não há diferença significativa entre os tempos de fixação ao iniciar a dinâmica com dois alelos, ou com quatro alelos, para um valor de N fixado. O que explica esse fenômeno é a deriva genética sobre pequenas populações. Em pequenas populações a deriva genética tem grande efeito na variabilidade genética no sentido de remover a diversidade de alelos. Nesse nível de tamanho populacional é impossível haver equilíbrio entre geração e remoção de alelos, a remoção irá prevalecer. Esse fenômeno se repete para todos os valores de N.

Na figura 5 notamos que o mesmo fenômeno ocorre quando comparamos

uma taxa de mutação grande, 10^{-4} com uma taxa de mutação da literatura, 10^{-6}. A justificativa disso é a mesma do parágrafo anterior. Mais uma vez, um modelo de regressão linear simples parece se ajustar bem aos dados.

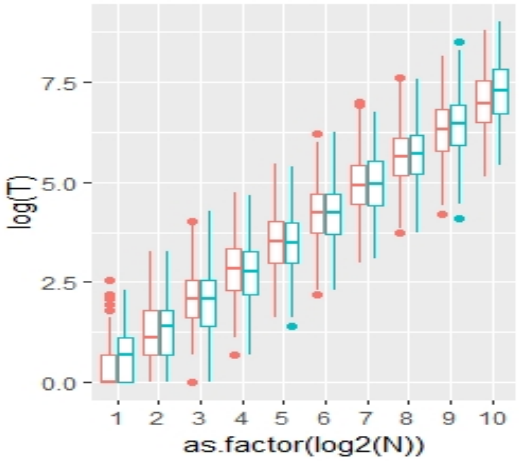

Figura 5. Relação logaritmo do tempo de fixação versus log2 dos tamanhos populacionais para dois alelos iniciais. Em vermelho U = 10^{-4} e em verde U = 10^{-6}.

O resultado mais importante que nós notamos, em todos os casos, é a existência de diferenças significativas entre tempos medianos de fixação somente quando comparamos tamanhos populacionais cujo valor da fração N_i/N_j, com i > j, é igual ou superior a 4. Ou seja, quando comparamos os tempos de fixação de duas populações, sendo uma o dobro do tamanho da outra, notamos que não há diferença significativa, exceto quando N = 2 e N = 4, que mostra a existência de diferença significativas em casos particulares. Assim, só será garantido haver diferenças significativas entre dois valores de $log(T)$ quando, pelo menos, quadruplicarmos os tamanhos populacionais.

Análise de regressão linear simples

Procuramos ajustar um modelo de regressão linear simples (MRLS) aos dados transformados pelos logaritmos. A sugestão que nos foi dada na seção anterior é a de que um MRLS pode ser ajustado aos dados transformados pelos logaritmos.

Primeiramente realizamos a análise de resíduos e o que nos foi revelado é interessante. Mostramos a presença de heterocedasticidade, como foi revelado pelo teste de Breusch-Pagan [9], ao nível de confiança α = 0,05, que resultou em um p-valor igual a 0,0001587. Assim, com 95% de probabilidade, o pressuposto de homogeneidade das variâncias dos $\log(T_k)$, k = 1, …, 10, é violado para o ajuste de um modelo de regressão linear simples. A heterocedasticidade pode ser melhor visualizada na figura 6, que é o gráfico dos resíduos versus os $log_2(N)$.

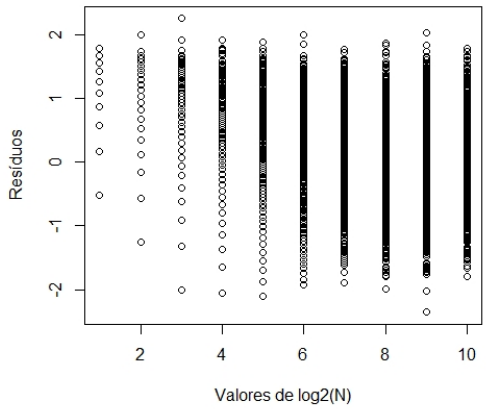

Figura 7. Gráfico dos resíduos de versus $log_2(N)$ para U = 0,0001 e dois alelos iniciais.

Mostramos o gráficos Quantil-quantil (Qqplot) para os resíduos na figura 7. No QQplot, notamos que somente os quantis dos resíduos com valores entre -1,2 e 1,2 que são valores muitos próximos aos quantis da normal padrão. Muitos quantis amostrais fora desse intervalo são distantes dos quantis normais. Isto mostra que é violado o pressuposto de normalidade dos resíduos para o ajuste de um modelo de regressão linear simples.

eliasdcn@servidor.uepb.edu.br

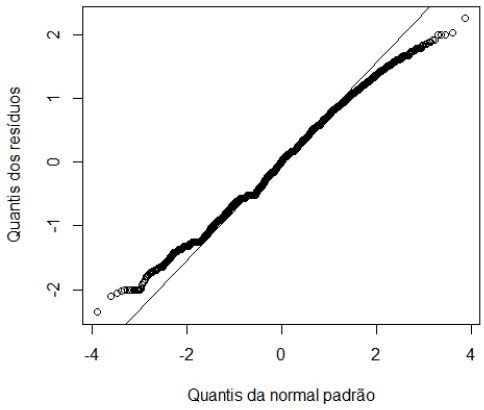

Figura 3. Qqplot dos resíduos de $log(T)$ versus $log_2(N)$ para U = 0,0001 e dois alelos iniciais.

Não notamos valores discrepantes para os resíduos, isto é, valores maiores do que 2,5 ou menores do que -2,5. A maioria dos resíduos para $\log_2(N) = 1, 2, 3$ e 4 se concentram acima da linha y = 0; para $\log_2(N) > 4$, os resíduos se espalham aleatoriamente em torno do zero. Não procedemos o teste de normalidade de Shapiro-Wilk porque as amostras são maiores do que 5000 observações [10]. Uma alternativa é o teste de Kolmogorov-Smirnov [11], que apresentou resultado para o p-valor inferior a $2,2x10^{-16}$. Ou seja, significativamente falando, ao nível de 95% de probabilidade, rejeita-se que os resíduos sejam normalmente distribuídos com média zero e variância σ^2.

Diante do que foi notado acima, podemos afirmar que são violados os pressupostos de normalidade dos resíduos e homogeneidade das variâncias para um ajuste de um modelo de regressão linear simples aos dados de $\log(T_k)$ versus $\log2(N_k)$, k = 1, …, 10.

Considerações Finais

Neste trabalho nós elaboramos um programa em linguagem R para simular a dinâmica de segregação de um alelo em um único locus do genoma de indivíduos haploides assexuados em populações em que deriva genética e mutação são as forças evolucionárias envolvidas. Nós observamos uma relação crescente entre os $log(T)$ versus $log_2(N)$ que é aparentemente linear. Há outiliers e heterocedasticidade nessa relação, o que leva a impossibilidade de ajuste de um modelo de regressão linear simples. O principal resultado revelado aqui é a existência de diferença significativa entre os $log(T)$ quando há diferença entre os tamanhos populacionais de no mínimo quatro vezes. Os próximos passos dessa pesquisa será incorporar outros tamanhos populacionais para além de 2^{-10}, acasalamento e seleção natural.

Referências

[1] D. L. Hartl, A. G. Clark, ***"Principles of Population Genetics,"*** 4° Ed. Sinauer Associates, Inc. (2007).

[2] M. Kimura, ***"Rare Variant Alleles in the Light of the Neutral Theory,"*** Mol. Biol. Evol. 1(1), 84–93 (1983).

[3] W. C. FUNK et al., ***"Adaptive divergence despite strong genetic drift: genomic analysis of the evolutionary mechanisms causing genetic differentiation in the island fox (Urocyon littoralis),"*** Molecular Ecology, 25, 2176–2194, (2016).

[4] P. Cossu et al., ***"Influence of genetic drift on patterns of genetic variation: The footprint of aquaculture practices in Sparus aurata (Teleostei: Sparidae),"*** Molecular Ecology, 28, 3012–3024 (2019).

[5] Z. Gompert, A. Springer, M. Brady, S. Chaturvedi, L. K. Lucas, ***"Genomic time-series data show that gene flow***

maintains high genetic diversity despite substantial genetic drift in a butterfly species," Molecular Ecology, 30, 4991–5008, (2021).

[6] L. J. Revell, ***"learnPopGen: An R package for population genetic simulation and numerical analysis,"*** Ecology and Evolution, 9, 7896–7902, (2019).

[7] B. Peng et al., ***"Genetic Data Simulators and their Applications: An Overview,"*** Genetic Epidemiology, 39, No. 1, 2–10, (2015).

[8] A. J. Aberer, A. Stamatakis, ***"Rapid forward-in-time simulation at the chromosome and genome level,"*** BMC Bioinformatics, 14, 216, (2013).

[9] 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/.

[9] T. S. Breusch, A. R. Pagan, ***"A Simple Test for Heteroscedasticity and Random Coefficient Variation"***. Econometrica. 47 (5): 1287–1294 (1979).

[10] P. Royston. ***"An extension of Shapiro and Wilk's W test for normality to large samples,"*** Applied Statistics, 31, 115–124 (1982).

[11] W. J. Conover. ***"Practical Nonparametric Statistics,"*** New York: John Wiley & Sons. Pages 295–301, 309–314 (1971).

[12] Bussab, W. de O., Morettin, P. A. Estatística Básica. 9° ed. São Paulo: Saraiva (2017).

[13] Montgomery, D. C., Peck, E. A. Vining, G. G. Introduction to linear regression analysis. 5th ed. Wiley (2012).

eliasdcn@servidor.uepb.edu.br