

## REVIEW

# A brief history and popularity of methods and tools used to estimate micro-evolutionary forces

Jonathan Kidner<sup>1</sup>  | Panagiotis Theodorou<sup>1</sup> | Jan O. Engler<sup>2</sup> | Martin Taubert<sup>3</sup> | Martin Husemann<sup>1,4</sup>

<sup>1</sup>General Zoology, Institute for Biology, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

<sup>2</sup>Terrestrial Ecology Unit, Department of Biology, Ghent University, Ghent, Belgium

<sup>3</sup>Aquatic Geomicrobiology, Institute for Biodiversity, Friedrich Schiller University Jena, Jena, Germany

<sup>4</sup>Centrum für Naturkunde, University of Hamburg, Hamburg, Germany

## Correspondence

Jonathan Kidner, General Zoology, Institute for Biology, Martin Luther University Halle-Wittenberg, Hoher Weg 8, 06120 Halle (Saale), Germany.  
Email: jonathan.kidner@gmail.com

## Funding information

Funding for this has been from: the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig funded by the German Research Foundation (FZT 118), and JOE was funded by the FWO Flanders PostDoc Fellowship Program (grant number: 12G4317N).

## Abstract

Population genetics is a field of research that predates the current generations of sequencing technology. Those approaches, that were established before massively parallel sequencing methods, have been adapted to these new marker systems (in some cases involving the development of new methods) that allow genome-wide estimates of the four major micro-evolutionary forces—mutation, gene flow, genetic drift, and selection. Nevertheless, classic population genetic markers are still commonly used and a plethora of analysis methods and programs is available for these and high-throughput sequencing (HTS) data. These methods employ various and diverse theoretical and statistical frameworks, to varying degrees of success, to estimate similar evolutionary parameters making it difficult to get a concise overview across the available approaches. Presently, reviews on this topic generally focus on a particular class of methods to estimate one or two evolutionary parameters. Here, we provide a brief history of methods and a comprehensive list of available programs for estimating micro-evolutionary forces. We furthermore analyzed their usage within the research community based on popularity (citation bias) and discuss the implications of this bias for the software community. We found that a few programs received the majority of citations, with program success being independent of both the parameters estimated and the computing platform. The only deviation from a model of exponential growth in the number of citations was found for the presence of a graphical user interface (GUI). Interestingly, no relationship was found for the impact factor of the journals, when the tools were published, suggesting accessibility might be more important than visibility.

## KEYWORDS

drift, migration, mutation, population genetics, selection, software, user bias

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Ecology and Evolution* published by John Wiley & Sons Ltd.

## 1 | INTRODUCTION

The modern synthesis (Mayr & Provine, 1980) has revolutionized our perception of micro-evolution by providing a conceptual and theoretical framework to investigate its processes. In this synthesis, four main forces driving evolution were described as contributing to changes in allele frequencies: mutation ( $\mu$ ), gene flow ( $m$ ), drift (estimated as its inverse effective population size ( $N_E$ )), and selection ( $s$ ). The modern synthesis provides the theoretical framework to predict the effects of each of these forces in various settings. A chronology is apparent with many of the tools (particularly the older programs) still implementing algorithms developed earlier during the modern synthesis, alongside more recently developed ones adopted for different marker systems.

The types of population genetic markers have broadened substantially from its initial emergence; the earlier studies used allozymes to estimate population differentiation (for a review, see Allendorf, 2016). These were largely replaced by genetic markers, (Sunnucks, 2000), such as random primer binding (RAPD), restriction site polymorphisms (RFLP, AFLP), fragment length variation of satellite DNA (e.g., microsatellites), or sequence polymorphisms (single nucleotide polymorphisms, SNPs). Two main categories of molecular markers can be defined: dominant markers, absent of the ability to determine homo- or heterozygosity (AFLP, RFLP); and codominant, that can be used to determine homo- or heterozygosity (microsatellites, allozymes, nucleotide variation). Codominant markers have consequently received increasing attention among researchers due to their greater information content.

While analyzing a specific allozyme, or microsatellite would provide a single locus in a study, the development of mass sequencing and SNP-typing has enabled the creation of massive datasets involving thousands to millions of loci. However, the use of high-throughput sequencing (HTS) techniques introduced a whole new range of analytical problems for researchers studying genetic diversity, while increasing the scope for testing the predictions of the modern synthesis (Davey & Blaxter, 2010). This has required the development of new techniques to incorporate the issues introduced through mass sequencing, reducing false positives, while maintaining sensitivity (Cavalli-Sforza, 1998). These new methods now populate the software tools for estimating population genetic parameters alongside techniques that are developed for single loci. This requires the researchers using these software suites to be able to discern their needs, and the provisions of the individual programs/tools.

The first population genetic programs were based on goodness of fit tests (chi-square, or F-based tests: Allendorf, 2016; Excoffier & Heckel, 2006; Labate, 2000). However, these approaches became increasingly intractable as the null models (and the understanding of the complexity of the data) increased in complexity. Approaches such as maximum likelihood (ML) were developed as an alternative, while allowing for the efficient exploration of more complicated multi-dimensional parameter space. Later on, the "Bayesian Revolution" led to further developments in genetic data analyses that allowed for even more complex models in inferring the

distributions of estimates of micro-evolutionary forces (Beaumont & Rannala, 2004); as well as avoiding some critiques leveled at the ML approaches (Stigler, 2007). The subsequent use of computational techniques to simulate parameter distributions helped to reduce the computational demand involved in Bayesian testing, leading to the development of approximate Bayesian computation (ABC, Csillery et al., 2012) among others.

All of these means of inference are maintained within the current community of software and tools to analyze population genetics, and multiple comparisons of the tools and algorithms have been performed (Excoffier & Heckel, 2006; Labate, 2000; Putman & Carbone, 2014). However, a comprehensive overview of population genetic software (including more recent publications) and an analysis of their usage is currently missing.

Here, we provide a short overview of the methods, programs, and resources used to estimate mutation, gene flow, drift (in terms of  $N_E$ ), and selection. We furthermore provide an extensive list of approximately 100 programs for population genetic analyses, and briefly discuss their functionalities and differences. On these programs, we investigate popularity and usage patterns within the scientific community, our null model being that the probability of finding a new citation for a particular program is dependent on the previous total number of citations observed (of the program, Taylor, 1961). This leads to citation distributions described by log-normal, and series models, which can be more formally described through Taylor's power law mean-variance relationships. Subsequently, we also investigate the relationship between niche breadth and range size of a program. We use niche to refer to how research developments lead to new opportunities to expand research areas, for example, the development of HTS, or machine learning; range is used to describe the occupation of different niches by individual programs. These results are discussed in terms of program complexity, user friendliness, year of publication, impact factor (IF) of the journal at the time of publication, computing platform, and the type and number of population genetic parameters estimated.

## 2 | METHODS

We generated an extensive list of population genetic software by using a combination of Google searches, Google Scholar, Web of Knowledge, and the bioinformatic search engine omictools.com. For each program, we recorded the evolutionary force(s) or proxy of force(s) it is able to estimate, the operating system it runs on, the year of publication, presence-absence of a graphical user interface (GUI), and the journal's IF at the time of publication. Further, we downloaded citation records for all articles from ISI web of knowledge (25.02.2019), which were used as a proxy for citation bias. In total, we report 101 programs and scripts. For 96 of these, where papers were available, we downloaded citation records. From these 96, we only used those published before 2014 (to reduce the impact of recently emerged tools on our analysis), which resulted in 80 papers (software packages) included in our analyses. Analysis of citation

records was performed in R v. 3.6.3 (R Core Team, 2018) using customized versions of the scripts published by Keil et al. (2018). These are provided in the Appendix S1. The analyses included the use of the R packages: *plyr* (Wickham, 2011), *ggplot2* (Wickham, 2016), *reshape2* (Wickham, 2007), and *MASS* (Venables & Ripley, 2002).

In our analyses, we first investigated the programs using species abundance distributions (SADs); these were analyzed with log-normal and log-series models being fitted to citation data collected from ISI web of knowledge. Assuming that each use and citation of the software occurs independently, these models provide a good description of the abundance in the citation records (Baldrige et al., 2016). We used these two models as they are simple descriptors of SADs, the log-series fitting better when more singletons (software with single citations) are observed in the dataset, compared to the log-normal. The log-series model was implemented with the *sads* package in R, while the logged mean and variance were used for the log-normal model. An identical approach to sampling as in Keil et al. (2018) was done, with the average of 500 model runs taken for generating the expected values. For Taylor's power law, we adopted the number of times the software was cited as its citation bias, according to its publication record on ISI web of knowledge. The citation bias in our data was tested for deviations from a purely random model of citation growth with a linear model on log-transformed means and variances. This model describes how the process of accumulating new citations is based on the exponential growth of the number of citations (Keil et al., 2018; Taylor, 1961). A process that can be described by a chaotic, random walk that holds true when the mean–variance relationship is approximately equal to 2. When the mean–variance relationship is less than 2 additional parameters are required in the growth model. Further, we tested the effects of operating system (Microsoft Windows, Linux, macOS), parameter (mutation, drift, migration, selection), and GUI on citation bias using nonparametric Kruskal–Wallis tests implemented in R. For testing the effect of the IF of the journal, we used the robust regression model from the *MASS* package (Venables & Ripley, 2002). All programs included in the analyses are listed in Table 1 together with their references; download links and DOIs are provided in the Appendix S1, where available.

### 3 | RESULTS AND DISCUSSION

#### 3.1 | Estimating the rate and effect of mutation

Mutations (i.e., single nucleotide mutations; insertions/deletions and chromosome rearrangements) are the ultimate source of genetic variation, and the precise estimation of mutation rates ( $\mu$ ) is important to understand the mechanisms of evolution (Lynch, 2010). However, estimating mutation rates remains relatively difficult (Kondrashov & Kondrashov, 2010), mainly due to the randomness of the process and the generally low mutation rates observed in eukaryotes.

The most direct techniques to estimate mutation rates, applicable only to microorganisms or model species with short

generation times, are fluctuation tests in bacteria and yeast, and mutation accumulation experiments in nonmammalian model systems (Foster, 2006; Luria & Delbrück, 1943; Lynch et al., 2008; Rosche & Foster, 2000). In fluctuation experiments, a low number of wild-type cells are used to inoculate large numbers of parallel cultures under nonselective conditions; these are then moved to selective media to identify mutants. Even though the fluctuation experiment is conceptually simple, the mathematics to estimate the rate of mutation from the frequency of mutants remain challenging (Zheng, 2017). The most widely used statistical method to estimate  $\mu$  from fluctuation experiments is the Lea–Coulson method of the median (LC; Lea & Coulson, 1949). However, more sophisticated statistical analyses have emerged that can estimate mutation rates from fluctuation experiments more accurately, incorporating several statistical estimators. The MSS-MLE method implemented in the web tool FALCOR uses an initial estimate of  $\mu$  to generate the probability of observing  $n$  mutants on a selective medium and uses the complete dataset from a fluctuation experiment, rather than just summary statistics, providing more statistical power (Hall et al., 2009). In comparison, the web tool BZ-RATES employs a generating function estimator, allowing the calculation of  $\mu$  while controlling for differential growth rates (Gillet–Markowska et al., 2015). As an alternative, the R package RSALVADOR (Zheng, 2017) provides various methods for computing ML estimates of  $\mu$  with likelihood ratio-based confidence intervals.

Mutation accumulation experiments (MA) represent another method to directly estimate mutation rates (Luria & Delbrück, 1943; Lynch et al., 2008). In such experiments, isogenic lines of model organisms randomly accumulate mutations through several generations of inbreeding. The resulting loss of fitness (e.g., in terms of growth rate and reproductive success) compared to control lines ( $\Delta M$ ) and the fitness variance ( $\Delta V$ ) among the lines is then used to infer mutation rates, for example, using Bateman–Mukai (BM, Bateman, 1959; Mukai, 1964), ML (Wloch et al., 2001), or minimum distance methods (MD, Garcia-Dorado, 1997). The performance of BM, ML, and MD methods has been assessed using simulation datasets and concluded that MD methods produce estimates with the lowest bias and sampling variance (Garcia-Dorado & Gallego, 2003). A more thorough review on the performance of the above methods can be found in Garcia-Dorado and Gallego (2003).

The majority of tools, for estimating mutation rate, applicable to a taxonomically wider range of organisms are based on genetic data. These methods can compare a variety of neutral homologous sequences from related species with calibrated divergence times to infer mutation rates, based on the simple assumption that the rate of neutral sequence divergence is equal to the mutation rate (Kimura, 1968). For the user, the choice of appropriate neutral markers and availability of appropriate sister species for comparison are among the biggest challenges involved in using these methods (Kondrashov & Kondrashov, 2010). Many software suites can be used to estimate divergence rates and hence to infer mutation rates indirectly. One of such programs is BEAST (Bouckaert et al., 2014), a software that uses sampling across phylogenetic tree space to infer

TABLE 1 List of population genetics software

Program name	Year	$\mu$	m	$N_E$	s	Windows	Mac	Linux	IF	GUI	Reference
2mod	1999			Yes		Yes			NA	NA	Ciofi et al. (1999)
abc	2012		Yes	Yes		Yes	Yes	Yes	5.924	No	Csillery et al. (2012)
ABCtoolbox	2010		Yes	Yes		Yes	Yes	Yes	3.029	No	Wegmann et al. (2010)
adegenet	2008		Yes			Yes	Yes	Yes	4.328	No	Jombart (2008)
AgeStructure	2010			Yes		Yes			5.659	NA	Wang et al. (2010)
ANGSD	2013				Yes			Yes	2.672	No	Korneliusen et al. (2014)
Arlequin3	2005		Yes		Yes	Yes			NA	Yes	Excoffier and Schneider (2005)
Arlequin35	2010		Yes	Yes	Yes	Yes			1.631	Yes	Excoffier and Lischer (2010)
Ballet	2014			Yes	Yes	Yes		Yes	7.528	No	DeGiorgio et al. (2014)
Bayenv	2010			Yes	Yes		Yes	Yes	4.866	No	Gunther and Coop (2013)
BayesAss	2003		Yes		Yes	Yes	Yes	Yes	4.276	No	Wilson and Rannala (2003)
Bayescan	2012			Yes	Yes	Yes	Yes	Yes	4.002	Yes	Foll and Gaggiotti (2008)
BayesFST	2004			Yes	Yes	Yes	Yes	Yes	2.261	No	Balding (2003)
BayPass	2015			Yes	Yes	Yes	Yes	Yes	4.644	No	Gautier (2015)
BEAST	2007	Yes		Yes		Yes	Yes	Yes	4.091	Yes	Drummond and Rambaut (2007)
BEAST17	2012	Yes		Yes	Yes	Yes	Yes	Yes	10.353	Yes	Drummond et al. (2012)
BEAST2	2014	Yes		Yes	Yes	Yes	Yes	Yes	NA	Yes	Bouckaert et al. (2014)
BIMr	2008		Yes		Yes	Yes	Yes	Yes	4.002	Yes	Faubet and Gaggiotti (2008)
bz_rates	2015	Yes			Yes	Yes	Yes	Yes	2.91	Yes	Gillet-Markowska et al. (2015)
Colony	2009			Yes	Yes	Yes	Yes	Yes	1.259	NA	Jones and Wang (2009)
CoNe	2004			Yes	Yes	Yes			4.289	NA	Anderson (2005)
DetSel	2013			Yes	Yes	Yes	Yes	Yes	1.707	No	Vitalis et al. (2003)
diCal	2013			Yes	Yes	Yes	Yes	Yes	4.866	NA	Sheehan et al. (2013)
DIYABC	2008		Yes	Yes	Yes	Yes	Yes	Yes	4.328	NA	Cornuet et al. (2008)
DIYABC2	2014		Yes	Yes	Yes	Yes	Yes	Yes	4.981	Yes	Cornuet et al. (2014)

(Continues)

TABLE 1 (Continued)

Program name	Year	$\mu$	m	$N_E$	s	Windows	Mac	Linux	IF	GUI	Reference
dlik	1998			Yes		Yes			NA	NA	O'Ryan et al. (1998)
DnaSP	2009				Yes	Yes			4.926	Yes	Librado and Rozas (2009)
EggLib	2014			Yes	Yes	Yes	Yes	Yes	2.808	No	De Mita and Sjol (2012)
EigenGWAS	2015			Yes	Yes	Yes	Yes	Yes	3.801	No	Chen et al. (2015)
Estim	2001		Yes	Yes		Yes			NA	NA	Vitalis and Couvet (2001)
FALCOR	2009	Yes			Yes	Yes	Yes	Yes	4.926	Yes	Hall et al. (2009)
FLK	2010			Yes	Yes	Yes	Yes	Yes	4.087	No	Bonhomme et al. (2010)
Garcia_script	1998	Yes			Yes				0.863	No	Garcia-Dorado and Marin (1998)
GenAIEX	2006		Yes		Yes	Yes	Yes	Yes	4.894	Yes	Peakall and Smouse (2006, 2012)
Geneland	2005		Yes		Yes	Yes	Yes	Yes	1.219	No	Guillot et al. (2005)
Genetree	2000	Yes	Yes		Yes	Yes	Yes	Yes	1.833	NA	Bahlo and Griffiths (2000)
Gimlet	2002			Yes		Yes			NA	Yes	Valière (2002)
GONE	2012			Yes		Yes			7.432	Yes	Coombs et al. (2012)
HacDivSel	2017				Yes		Yes	Yes	2.766	No	Carvajal-Rodriguez (2017)
hapbin	2015			Yes		Yes		Yes	13.649	No	Maclean et al. (2015)
hapflk	2013			Yes			Yes	Yes	4.866	No	Fariello et al. (2013)
IM	2001	Yes	Yes	Yes		Yes	Yes	Yes	4.803	NA	Nielsen and Wakeley (2001)
Ima	2007	Yes	Yes	Yes		Yes	Yes	Yes	9.598	No	Hey and Nielsen (2007)
IMa2	2010	Yes	Yes	Yes		Yes	Yes	Yes	5.51	No	Hey (2010)
Ima2p	2015	Yes	Yes	Yes		Yes	Yes	Yes	5.298	No	Sethuraman and Hey (2016)
Keightly_script	1998	Yes			Yes	Yes	Yes	Yes	4.45	No	Keightley (1998)
Lamarc	1999		Yes	Yes		Yes	Yes		4.221	No	Beerli and Felsenstein (1999)
Lamarc2	2006	Yes	Yes	Yes		Yes	Yes	Yes	4.894	NA	Kuhner (2006)
LDNe	2008			Yes		Yes	Yes	Yes	NA	Yes	Waples and Do (2008)
Lositan	2008				Yes	Yes	Yes	Yes	3.781	NA	Antao et al. (2008)

(Continues)

TABLE 1 (Continued)

Program name	Year	$\mu$	m	$N_E$	s	Windows	Mac	Linux	IF	GUI	Reference
McDonald-Kreitman_online	2008				Yes	Yes	Yes	Yes	6.878	Yes	Egea et al. (2008)
MCLEEPS	2000			Yes			Yes		4.687	No	Anderson et al. (2000)
mdiv	2001		Yes			Yes			4.803	No	Nielsen and Wakeley (2001)
MEAdmix	2006			Yes		Yes			4.242	No	Wang (2006)
MEGA6	2013				Yes	Yes	Yes	Yes	14.308	Yes	Tamura et al. (2013)
MigEst	2014		Yes			Yes			6.494	NA	Wang (2014)
MIGRATE_N	2013	Yes	Yes			Yes	Yes	Yes	4.087	No	Beerli and Palczewski (2010)
MIMAR	2007		Yes	Yes					5.169	No	Becquet and Przeworski (2007)
MISAT	1997			Yes		Yes			4.275	No	Nielsen (1997)
MLNe	2001		Yes	Yes		Yes			2.317	NA	Wang (2001)
mlRho	2010	Yes					Yes	Yes	6.457	No	Haubold et al. (2010)
MSMC	2014			Yes			Yes	Yes	29.352	No	Schiffels and Durbin (2014)
Nb_HetEx	2008			Yes		Yes			1.775	NA	Zhdanova and Pudovkin (2008)
NeEstimator	2004			Yes		Yes	Yes	Yes	NA	Yes	Peel et al. (2004)
NeEstimator2	2014			Yes		Yes	Yes	Yes	3.712	Yes	Do et al. (2014)
neighbor	2002		Yes			Yes			3.014	NA	Burczyk et al. (2002)
NM+	2010		Yes			Yes			1.631	NA	Chybicki and Burczyk (2010)
nSI	2014				Yes		Yes		9.105	No	Ferrer-Admetlla et al. (2014)
Omegamap	2006				Yes	Yes	Yes		4.242	No	Wilson and McVean (2006)
OmegaPlus	2010				Yes	Yes		Yes	5.323	No	Alachiotis et al. (2012)
ONeSamp	2008			Yes		Yes	Yes	Yes	NA	No	Tallmon et al. (2008)
OutFLANK	2015				Yes	Yes	Yes	Yes	3.148	No	Whitlock and Lotterhos (2015)
pcadapt	2017				Yes	Yes	Yes	Yes	7.059	No	Luu et al. (2017)
pegas	2010		Yes			Yes	Yes	Yes	4.877	No	Paradis (2010)
Pool_hmm	2012				Yes	Yes	Yes	Yes	10.353	No	Boitard et al. (2013)

(Continues)

TABLE 1 (Continued)

Program name	Year	$\mu$	m	$N_E$	s	Windows	Mac	Linux	IF	GUI	Reference
PopABC	2009	Yes	Yes	Yes		Yes	Yes	Yes	4.926	No	Lopes et al. (2009)
poggene	1997				Yes	Yes			0.231	Yes	Yeh and Boyle (1997)
PSMC	2011			Yes				Yes	36.28	No	Li and Durbin (2011)
rehh	2012				Yes	Yes	Yes	Yes	5.468	No	Gautier and Vitalis (2012)
rehh2	2017			Yes		Yes	Yes	Yes	7.059	No	Gautier et al. (2017)
RpackageNB	2015			Yes		Yes	Yes	Yes	4.644	No	Hui and Burt (2015)
SAMOVA	2002		Yes			Yes		Yes	3.014	NA	Dupanloup et al. (2002)
SelectionHapStats	2018			Yes			Yes	Yes	3.564	No	Harris et al. (2018)
Selecton	2007			Yes		Yes	Yes	Yes	6.954	No	Stern et al. (2007)
SelEstim	2014			Yes		Yes	Yes	Yes	5.963	No	Vitalis et al. (2014)
selscan	2014			Yes		Yes	Yes	Yes	9.105	No	Szpiech and Hernandez (2014)
SNeP	2015			Yes		Yes	Yes	Yes	NA	No	Barbato et al. (2015)
SPAM	2004		Yes			Yes			1.511	Yes	Debevec et al. (2000)
spatbg	2016			Yes			Yes	Yes	6.086	No	Gompert (2016)
Structure	2000		Yes			Yes	Yes	Yes	4.687	Yes	Pritchard et al. (2000)
SweeD	2013			Yes			Yes	Yes	14.308	No	Pavlidis et al. (2013)
Sweep	2005			Yes		Yes	Yes	Yes	30.432	Yes	Sabeti et al. (2002)
sweepfinder	2006			Yes			Yes	Yes	10.139	No	Nielsen et al. (2005)
SweepFinder2	2016			Yes		Yes	Yes	Yes	7.307	Yes	DeGiorgio et al. (2016)
TempoFs	2007			Yes		Yes		Yes	4.001	No	Jorde and Ryman (2007)
Tess	2007		Yes			Yes	Yes	Yes	1.257	No	Chen et al. (2007)
ThetaCurve	2012	Yes				Yes	Yes	Yes	NA	No	n.a.
tm3	2002			Yes		Yes			4.483	NA	Berthier et al. (2002)
tmvp	2003			Yes		Yes			4.276	NA	Beaumont (2003)
TreeSelect	2011			Yes	Yes	Yes	Yes	Yes	10.603	No	Bhatia et al. (2011)
WFABC	2015		Yes	Yes	Yes	Yes	Yes	Yes	5.298	No	Foll et al. (2015)

Note: 101 Programs for predicting population genetic parameters. Each software/Program is described by the types of parameters it estimates (mutation— $\mu$ , migration— $m$ , effective population size (drift)— $N_E$ , or selection— $s$ ), as well as the platform it is available on (Windows, Mac, or Linux) and the reference and date of publication. Links to the programs are provided in the Appendix S1.

a variety of population parameters; the locus specific mutation rate being one of which.

The development of HTS technologies helped to minimize the limitations of both the direct (e.g., fitness-based assays) and indirect methods (e.g., small number of loci) and provided novel means to estimate mutation rates. The application of HTS technologies in fluctuation and mutation accumulation experiments has allowed a direct estimate of mutation rates irrespective of phenotypic or fitness effects (Katju & Bergthorsson, 2019; Nishant et al., 2009). In addition, long-term mutation accumulation experiments are no longer required; data on few generations suffices, expanding the possibility to obtain accurate measures of mutations in nonmodel species. Furthermore, the development of HTS technologies allowed the detection of germline mutation rates by comparing genomic sequences between subsequent generations, as well as estimating mutation rates by sequencing individuals of pedigreed families (e.g., parent–offspring). Pedigree sequencing is a very promising method applicable to any organism; however, such method requires a large number of individuals (Keightley et al., 2014) and special care should be taken during the analysis to avoid false variant calls and the risk for increased false-positive rates.

### 3.2 | Estimation of gene flow

Migration, also a creative force introducing new variants into a population, is synonymous with gene flow in population genetics and refers to the spatial movement of alleles (Broquet & Petit, 2009). The parameter estimated can either be the effective number of migrants ( $N_E M_E$ ), or the effective migration rate ( $M_E$ ).  $N_E M_E$  estimates the migration rate in terms of the effective population size of migrants, while the  $M_E$  estimates the net migration rate per generation (Waples & Gaggiotti, 2006). Both parameters can be estimated based on  $F_{ST}$  (Wright, 1931, 1940), yet doing so generally entails issues due to the presence of confounding factors (Cayuela et al., 2018; Whitlock & McCauley, 1999) of  $F_{ST}$  estimation. A brief flavor of which can be described through  $F_{ST}$  also being the product of drift ( $1/N_E$ ) and hence consequently being affected by all factors contributing to  $N_E$ , including demographic and adaptive processes. With many alternative tools existing, the determination of the most appropriate method for analyzing gene flow will generally depend on the type and amount of available data to inform the models.

In a general approach to investigating hypotheses of migration, an initial step is to assign samples to a population. Thereby generating a population structure within which the migration rates can be estimated. Multiple approaches exist to do so, one of which is the simulation style approach of Markov Chain Monte Carlo as used in BAYESASS (Wilson & Rannala, 2003), a second is the likelihood approach as in MIGEST/POPCLUSTER (Wang, 2014). While these tools produce estimates of population structure to estimate gene flow, the tools STRUCTURE and SPAM provide more detailed estimates and measures of population structure in samples (Debevec et al., 2000; Pritchard et al., 2000). SPAM provides the ability to identify the

contribution of multiple populations to a single sample (it was originally developed for use with fisheries data; Debevec et al., 2000). In contrast, STRUCTURE is most frequently used to identify stratification and (sub-) divisions within populations, providing sample assignment probabilities (Pritchard et al., 2000). Unfortunately, STRUCTURE performs less well with more continuously defined populations; in these scenarios, SAMOVA can provide an alternative approach, apportioning genetic variance maximally between groups of populations (Dupanloup et al., 2002).

Many software suites offer opportunities to investigate population structure further, including the provision of more detailed statistics and many more graphical representations surrounding population assignment. Similarly to SPAM and STRUCTURE—GENALEX, TESS3, and GENELAND (Chen et al., 2007; Guillot et al., 2005; Peakall & Smouse, 2006)—provide statistics on individual assignment probabilities and distributions. GENALEX offers many different types of analysis and graphical representations, providing a high level of detail on the population assignment statistics, as well as the allelic breakdown over multiple spatial dimensions. This is similar to TESS3 and GENELAND. However, these programs use different techniques for inferring the individual assignment probabilities and provide different summary statistics: the ancestry coefficients produced by TESS3 additionally allow for investigations of neutrality among the available loci, when population substructuring is strong. Both GENELAND and TESS3 offer plotting features, that integrate maps and assignment probabilities to provide a fully geographical analysis of populations.

The assignment approaches mentioned in the previous paragraphs are also used in BAYESASS, MIGEST/POPCLUSTER, and ARLEQUIN (Excoffier & Schneider, 2005; Excoffier & Lischer, 2010; Wang, 2014; Wilson & Rannala, 2003, respectively). However, these assignment approaches are used to produce estimates for rates of recent migration, providing a valuable tool for ecology and conservation research, but may not reflect the wider demographics of a species/population. This leads to a fundamental division between the remaining tools for investigating migration rate: those for investigating the degree of population admixture (typically for evolutionary, or population genetic related questions) and those used for investigating temporal, or recent migration (of more relevance to conservation/ecological questions), although, this is not a firm division.

One of the more widespread and evolutionary styled approaches for investigating population admixture is the coalescent. This is employed in: MIGRATE/LAMARC (Beerli & Felsenstein, 1999; 2001; Kuhner, 2006; Beerli & Palczewski, 2010), GENETREE (Bahlo & Griffiths, 2000), MDIV (Hey & Nielsen, 2004; Nielsen & Wakeley, 2001), IMA/IMA2 (Hey, 2010; Hey & Nielsen, 2007), DIYABC2 (Cornuet et al., 2014), ABCTOOLBOX (Wegmann et al., 2010), and BEAST /BEAST2: (Bouckaert et al., 2014; Ewing et al., 2004; Ewing & Rodrigo, 2006). The coalescent can be used to simulate genealogies to create the probability distributions of the parameters of interest (Kuhner & Smith, 2006), or as a model description of the sampled populations (providing the population parameters for estimation; Nielsen & Wakeley, 2001). The coalescent is especially helpful for describing complex population structures, for example,



as described in human populations (Bhatia et al., 2011). Humans have a hierarchical population structure that leads to increased allelic correlations among closely related subpopulations. This introduces biases into the migration estimates. Hence, incorporating hierarchical population structure is necessary to provide accurate estimates, as is available in IMA (Hey, 2010; Hey & Nielsen, 2007; Nielsen & Wakeley, 2001; Sethuraman & Hey, 2016) and MIMAR (Becquet & Przeworski, 2007). These tools use phylogenetic structures to inform on population structure; IMA estimates the genealogical structure of the population from sequence data from multiple loci (Sethuraman & Hey, 2016), while MIMAR uses a phylogenetic structure to identify the ancestral loci states (it required the presence of an outgroup—unfortunately MIMAR is no longer maintained; Becquet & Przeworski, 2007).

Alternatives to the above involve methods using allelic, or genotype frequencies with alternative inferences: for example, the use of method of moments estimators with genotype frequencies in ESTIM (Vitalis & Couvet, 2001) for calculating likely migration rates, the use of Mantel tests in ADGENET to test for isolation between subpopulations (Jombart, 2008), or the traditional calculation of  $F_{ST}$  statistics available in a wide range of tools described here (i.e., ARLEQUIN, PEGAS, POPGENE—EXCOFFIER & LISCHER, 2010; Paradis, 2010; Yeh & Boyle, 1997). A final consideration for users is the ease with which various tools can be employed to analyze HTS data, of those programs previously mentioned DIYABC and MIGEST/POPCLUSTER seem best suited for use with computing clusters (Cornuet et al., 2014; Wang, 2014). Alternatives such as PEGAS and ADEGENET may also read in HTS data, but are limited in terms of the number of loci to an order of around  $10^5$ , which may not be suitable for all HTS studies.

### 3.3 | Effective population size and drift

Drift describes fluctuations in allelic frequencies caused by random processes in populations with finite size and it is rather difficult to estimate directly. Instead, the magnitude of drift is described by the effective population size ( $N_E$ ), both being negatively correlated with each other (Wang et al., 2016).  $N_E$  represents a key parameter in conservation and evolutionary biology (Charlesworth, 2009; Habel et al., 2014; Husemann et al., 2016; Lanfear et al., 2014; Luikart et al., 2010) and a variety of methods have been developed to estimate different types of  $N_E$  in natural populations (Barker, 2011; Luikart et al., 2010; Palstra & Ruzzante, 2008; Wang et al., 2016). In conservation biology,  $N_E$  is used as a measure of the susceptibility of a population to stochastic processes and inbreeding and hence can estimate the viability of a population (Hare et al., 2011). Besides the different types of  $N_E$ , one can also distinguish between methods estimating the contemporary population size and those estimating demographic changes over time.

2MOD was the only software found to directly estimate the presence of drift, which, however, is no longer supported and the interpretation of results was not intuitive (Ciofi et al., 1999). However, for many conservation applications and management decisions the

contemporary  $N_E$  is a more important parameter. Its estimation usually involves a single panmictic population; deviations from the “optimal” population, for example, resulting from migration (e.g., Beerli & Palczewski, 2010), overlapping generations (Coombs et al., 2012; Waples et al., 2014), population subdivision (Ryman et al., 2014) and lack of clearly defined units in continuously distributed populations (Neel et al., 2013) lead to problems for estimating  $N_E$ . Some of these limitations have been addressed conceptually or practically (Nunney, 2016; Wang et al., 2016; Waples et al., 2014), and some solutions have been implemented in software packages (Wang & Whitlock, 2003).

Estimates for the effective population size ( $N_E$ ) are broadly grouped into three categories (Wang et al., 2016): the variance  $N_E$ , which is estimated from fluctuations in allele frequencies in genetic times series; the inbreeding  $N_E$ , representing an estimate of  $N_E$  derived from linkage between sampled loci; and the coalescent  $N_E$ , that either uses simulations, or estimates for branch lengths to infer  $N_E$ . Variance  $N_E$ , and some of the coalescent  $N_E$  methods, require sampling of the same population at least twice at different points in time, preferably several generations apart. This represents a main limitation, as such samples are rarely available in nonmodel taxa and organisms with long generation times (Habel et al., 2014).

Estimates of variance  $N_E$  can, for example, be obtained using TEMPOFS (Jorde & Ryman, 2007), a moment-based estimator with relatively low accuracy (Wang et al., 2016). Similarly, MLNE and GONE represent moment-based inference methods for variance  $N_E$  (Wang, 2001; Wang & Whitlock, 2003), the latter providing a modification to adjust for age structure in populations (Coombs et al., 2012). However, as GONE requires estimates of age-specific survival and birth rates, it is difficult to apply to many datasets (Coombs et al., 2012). MLNE in addition has the capability to estimate migration rates. Several other tools, such as TM3—(Berthier et al., 2002), TMVP—(Beaumont, 2003), and CoNE—(Anderson, 2005), also use temporal sampling, but estimate coalescent  $N_E$ . Overall, for most programs estimating variance  $N_E$ , the main limitation is the restriction in estimation accuracy to populations with small effective sizes (and the computationally intensive nature of the approaches). The RPACKAGE ND (Hui & Burt, 2015) overcomes some of these limitations (to the applicable population sizes) by using a hidden Markov model to reduce computational load and raise the upper bounds (of  $N_E$ ) to several million individuals. MCLEEPS (Anderson et al., 2000) implicitly involves a Markov Chain and uses a Monte Carlo algorithm to overcome the computationally intensive nature of generating probability distributions for the effective population size. Despite the complexities involved, the temporal method makes fewer assumptions and is considered more robust for real populations (Wang et al., 2016).

Inbreeding  $N_E$ , based on the linkage disequilibrium (LD) between sampled loci, is most commonly estimated using LDNE (Waples, 2006; Waples & Do, 2008) and only requires a single temporal sample. Recently, extensions of the LD method have been developed to allow the estimation from many loci across the genome (Waples et al., 2014) and to incorporate the effects of linkage for genomic data, which is implemented in LINKNE (Hollenbeck et al., 2016).

Further, the LD method has been implemented in SNEP (Barbato et al., 2015) to estimate recent  $N_E$  trajectories from genome-wide SNP data.

Alternative single sample estimators are the heterozygosity excess method, which employs the relationship between the  $N_E$  of a parental population and the amount of heterozygosity excess in the offspring population (Luikart & Cornuet, 1999; Pudovkin et al., 1996; Robertson, 1975; Wang, 1996; Wang et al., 2016) as implemented in  $N_E$ ESTIMATOR (Do et al., 2014), or the sibship method, which uses inferred sibship/parentage frequencies as implemented in COLONY (Jones & Wang, 2009). The heterozygosity excess method, however, does not seem to perform well on empirical data (Wang et al., 2016).

The coalescence  $N_E$  is based on the assumption that two random gene copies will coalesce with a chance  $1/2N$  within one generation (Wakely & Sargsyan, 2009; see Wang et al., 2016 for a summary). The theory has been implemented in several software packages, such as TM3—(Berthier et al., 2002), TMVP—(Beaumont, 2003), and CoNE—(Anderson, 2005) which all use temporal sampling. These programs use maximum likelihood or pseudo-likelihood to estimate coalescent  $N_E$  (for a more in depth review, see Wang et al., 2016). As many of the other methods, these tools are performing better with small  $N_E$  when drift is strong and have problems estimating  $N_E$  in large populations. Further, they are computationally relatively demanding and are limited to smaller datasets.

As with the other parameters, Bayesian approaches have brought more flexibility and allow more complexity for parameter estimation. While it is often not clear which type of  $N_E$  (inbreeding, variance, or coalescent) is estimated, the Bayesian approaches often not only estimate  $N_E$ , but often trace changes in  $N_E$  over time. ABC methods (Csilléry et al., 2010) allow the implementation of demographic models and are now implemented in a variety of programs (WFABC—Foll et al., 2015; PopABC—Lopes et al., 2009; DIYABC—Cornuet et al., 2008; ABCtoolbox—Wegmann et al., 2010; and the abc R-Package—Csilléry et al., 2012), several of which employ the coalescent (Cornuet et al., 2008; Lopes et al., 2009). Yet, so far, ABC models have not been evaluated in comparison to more classic  $N_E$  estimation methods (Wang et al., 2016), although comparisons among various ABC implementations have been made. The issue of limited numbers of implementable summary statistics in traditional ABC has been solved using a kernel-based approach (Nakagome et al., 2013). Kernelization allows for an increased number of summary statistics, while maintaining high performance. Other alternatives include implementing machine learning methods to avoid problems with the high number of dimensions in posterior estimation (Cornuet et al., 2014). Additionally, Markovian models are gaining attention in inferring historic population sizes from whole genome sequences. The initial pairwise sequential Markov chain model (PSMC, Li & Durbin, 2011) has been improved by the multiple sequential Markov chain model (MSMC, Schiffels & Durbin, 2014); a further alternative is provided by the DiCAL model, which represents a generalization of the PSMC model applicable to multiple sequences (Sheehan et al., 2013). A more empirical distinction can be made with WFABC, as it is the only program that does not assume neutrality and enables

the estimation of both  $N_E$  and  $S$ , potentially even allowing estimation of  $N_E$  when  $S = 1$  (Bollback et al., 2008; Foll et al., 2015; Malaspinas et al., 2012; Mathieson & McVean, 2013).

Several studies have aimed to evaluate the performance of different methods to estimate  $N_E$  from genetic data. However, no gold standard has been reached, as performance strongly depends on the demographics of the studied populations. Many studies find that temporal estimates (variance  $N_E$ ) tend to be larger than single sample estimates (inbreeding  $N_E$ ) (Barker, 2011; Husemann et al., 2015; b). Different single sample estimators, in turn, frequently yield relatively congruent results (Álvarez et al., 2015), but can prove sensitive to the study system (Holleley et al., 2014; Gilbert, Whitlock & Lotterhos, 2015). Hence, it remains important to use different estimators within the same study to gain confidence in the estimates. This is facilitated by the implementation of multiple estimators in single commonly used programs such as  $N_E$ ESTIMATOR, which can accommodate more than 45,000 di-allelic SNPs in the current version and hence is applicable to a wide variety of datasets.

### 3.4 | Selection

In tests for selection, two different distinctions can be made in the applicability and approaches used. The first distinction that we make here is in the applicable use of the methods: those traditional statistics that are appropriate for single loci, and those for HTS datasets. The second distinction we make is within the analyses for HTS datasets: We distinguish between those that involve generating estimates of  $F_{ST}$  across populations and the genome, compared to those analyzing sequences for selective sweeps (looking for extended regions of homozygosity in the genome sequences).

In traditional population genetic analyses, Watterson's theta provides a measure of genetic diversity (Watterson, 1975), that can allow for a limited approach to testing selection through comparisons of genetic diversity with the pairwise estimate ( $\pi$ ). These two measures are used within Tajima's D for estimating selection; this test essentially compares the distributions of variants quantifying either an excess, or dearth of rare variants (Tajima, 1989). Hence, deviations in Tajima's D can be used to infer different types of selection. However, this statistic is strongly influenced by demographic factors, such as population expansions, bottlenecks, and contractions. Two similarly traditional tests offer solutions to these problems, both dependent on the availability of additional information. The HKA test (Hudson, Kreitman & Aguade, 1987) uses information from multiple loci and species (a minimum of two), assuming that under neutral theory divergence between the loci should be equivalent to the difference in polymorphism. The McDonald-Kreitman (MK) test predominantly does not use multiple loci (although variations exist), but does require the use of an outgroup as it uses comparisons of sequence divergence and diversity to estimate deviations from neutrality and directionality (Eyre-Walker, 2006). These tests are widely available within several packages (MEGA6—Tamura et al., 2013; POPGENE—Yeh & Boyle, 1997; Tajima's D in ARLEQUIN—Excoffier & Schneider, 2005;

and its namesake in MK—Egea et al., 2008). Differentiation between these implementations are minor, the larger differences lie within the data formats. For MEGA, file conversions are necessary and are available within sequence viewing/analysis tools (as well as within MEGA for a large variety of formats). ARLEQUIN and MK can ultimately read in sequence formats (MK from multiple loci). In contrast POPGENE primarily reads in a variety of different allele/marker encodings (as does ARLEQUIN), providing for analysis of dominant and codominant markers.

Differences between these traditional methods of estimating selection involve the range of genetic sequence information available. Typically, they require the use of sequence data and are not appropriate for use with dominant markers. Tajima's D requires data from a single population, while MK requires the presence of a single suitable outgroup (preferably a sister species). The HKA test is arguably more demanding in the requirement of needing sequences from at least two loci. All of these tests run into problems with either longer sequences or HTS datasets, in that they are susceptible to producing elevated levels of error (Andolfatto, 2008). To reduce the rates of error from a lack of control of the influence of recombination on the test statistics, sliding window approaches can be employed (Nielsen et al., 2005); although difficulties in interpreting significance may remain.

To deal with either longer sequences or HTS datasets, several approaches exist. One of the popular “families” of approaches is derived from the Lewontin–Krakauer (LK) test (Lewontin & Krakauer, 1973). This uses the  $F_{ST}$  statistic calculated across multiple loci and populations to test for deviations in the estimates. These deviations are then used to infer selection through outlier analyses (Bonhomme et al., 2010). While these methods are appropriate for testing recent and ongoing selection (as well as variation in selection between populations), they are also generally beset by problems arising from correlations in  $F_{ST}$  due to hierarchical population structure (typical for human populations) and linkage between loci. The tools currently available employ a variety of different techniques to circumvent or avoid these issues. FLK (Bonhomme et al., 2010) uses a hierarchical estimate of the population structure through a kinship matrix to avoid interference from the population structure. To estimate this, FLK requires the presence of a population that can act as an outgroup (it can also incorporate previous estimates of genetic distance). HAPFLK also avoids interference from population structure and is more robust against interference on the estimates from linkage, yet requires a known pedigree for assessing population structure (Fariello et al., 2013). In comparison, TREESELECT has no requirement on data informing on the population structure; unfortunately, the marker sets for analysis must be shared across all populations (Bhatia et al., 2011). A solution that might be limiting for species, where strong population differentiation might exist. The other methods in this LK family mainly differentiate through the techniques of inference. DETSEL (Vitalis et al., 2003) uses simulated distributions of differentiation using allele counts across multiple populations to produce a null distribution (and avoid issues with independence between samples), while HACDIVSEL (Carvajal-Rodríguez, 2017) uses

two complimentary approaches for inferring selection (haplotype and outlier based). HACDIVSEL is also one of the few tools that accepts sequence data for analysis (most of the alternative approaches require some processing of HTS data before analyses can be run). In comparison, the LK-based OUTFLANK (Whitlock & Lotterhos, 2015) uses a matrix of  $F_{ST}$  values as its input data and analyses of selection are run on these values.

$F_{ST}$  approaches are also used by other programs, of those previously mentioned ARLEQUIN is one of the more commonly cited tools where an  $F_{ST}$  test is implemented (Excoffier & Lischer, 2010). ARLEQUIN, along with BAYESCAN, BAYESFST, BAYENV, and SELESTIM all use outlier based tests for inferring the presence of selection with  $F_{ST}$  (respectively, Balding, 2003; Foll & Gaggioti, 2008; Gunther & Coop, 2013; Vitalis et al., 2014). Alternatively, eigenvector based regression models can be used in the tools EIGENGWAS or PCADAPT (Chen et al., 2015; Luu et al., 2017). For the outlier tests, all the tools use allelic count data from multiple populations for generating  $F_{ST}$  estimates. Differentiating itself from the other tools BAYENV can optionally include the degree of covariance between the environments of the different populations to inform predicted correlations in allelic frequencies (Gunther & Coop, 2013). Otherwise, differentiation between these tools is largely limited to the possible input data: The majority can use SNP counts (biallelic in BAYENV, ARLEQUIN, BAYESFST, SELESTIM; and multiallelic with BAYESCAN), BAYESCAN can use dominant markers, while SELESTIM can use read counts in place of allele counts for input (Foll & Gaggioti, 2008; Vitalis et al., 2014).

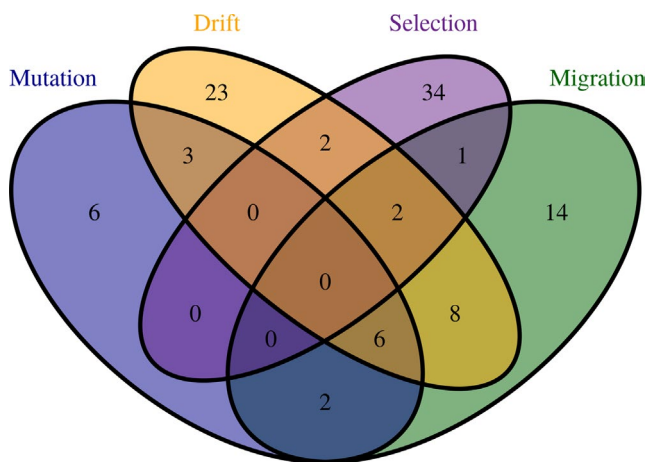
For more information regarding differences between these available tools, a variety of reviews are available. Most of these comparisons have been performed upon the publication of a new method: Vitalis et al. (2014), for example, demonstrated the advantages provided by SELESTIM against BAYESCAN. DeGiorgio et al. (2014) in a more stark contrast investigated the ability of tools to estimate balancing selection, demonstrating that BALLET (designed purely for testing balancing selection) provided improvements beyond those present in HAPFLK compared to SELESTIM (Fariello et al., 2013).

The second “family” of approaches we consider are the selective sweep-based approaches. These can be further subdivided into the methods that can detect hard and soft sweeps. Hard sweeps define those situations where a new allele has reached fixation in a population, leading to extensive regions of homozygosity. The common framework for this test is haplotype homozygosity statistics (HH). This statistic identifies the probability that a core haplotype is identical by descent between two randomly chosen chromosomes (Sabeti et al., 2002). The second approach is suitable for detecting selection on preexisting mutations or incomplete selective sweeps (soft sweeps), through analyzing distortions in the allele frequencies, away from those expected under neutrality. These analyses (particularly the hard sweep tests) are generally predicated on population-wide selection. Historical and recent selection events can be tested by either of these approaches, whereas ongoing processes are ascertainable only through tools developed for detecting soft selective sweeps.

Among those programs analyzing HTS data for hard sweeps, SWEEP was one of the earlier ones employing the extended haplotype

homozygosity (EHH) statistic and using a coalescent model to generate parameter probabilities (Sabeti et al., 2002). All of the later models that use the HH statistics (SELSCAN, REHH, iHS, XPEHH, and HAPBIN) incorporate the long range haplotype (LRH), integrated HH (iHS), and cross-population (XP-EHH) statistics (respectively, Szpiech & Hernandez, 2014; Gautier & Vitalis, 2012; Voight et al., 2006; Pickrell et al., 2009; and Maclean et al., 2015). Differences between these programs involve changes to the algorithms in the utilization of modern technologies to improve performance, reducing run times (Szpiech & Hernandez, 2014). SELSCAN, for example, offers significant speed improvements over REHH, iHS, and XPEHH through multi-threading (Szpiech & Hernandez, 2014), while HAPBIN reportedly demonstrates even greater improvements over SELSCAN (~3,000 fold speed increase, Maclean et al., 2015). It should be noted that both HAPBIN and SELSCAN require additional mapping information and processing for the loci, when compared to the R-package REHH.

From the software detecting hard sweeps, only SELSCAN also offers methods to analyze soft sweeps, using another recently developed statistic ( $nS_L$ , Szpiech & Hernandez, 2014). This statistic being originally developed for the program  $nS_L$  by Ferrer-Admetlla et al. (2014), provides more reliable estimates when variation in recombination rates is present in the data and fixation of alleles in haplotypes is not expected ( $nS_L$  itself is no longer supported). Other techniques for analyzing soft sweeps include investigations of the distortion of linkage disequilibrium (developed from the statistics of Kim & Nielsen, 2004), as provided in OMEGAPLUS (Alachiotis et al., 2012), or using site frequency spectra (the distribution in allele frequencies at a set of loci), as available in the SWEEPfinder programs (including SWEEP: Nielsen et al., 2005; DeGiorgio et al., 2016; and Pavlidis et al., 2013). While the files describing the input data should

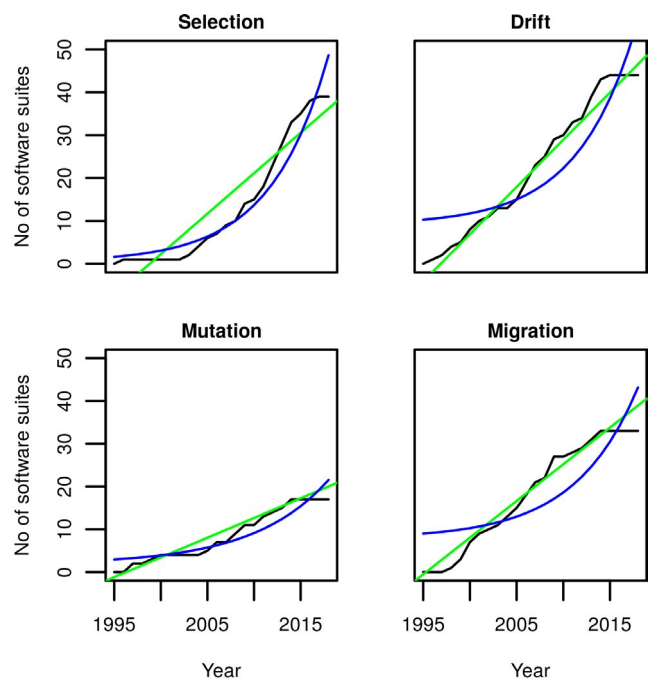


**FIGURE 1** Distribution of the programs estimating the various population genetic parameters. While the software suites for estimating mutation, gene flow (migration) and drift can be observed to follow the same distributions (hypergeometric tests, probability of coming from the same distribution  $p > .5$ ), the same cannot be said of selection (hypergeometric test,  $p < 10^{-08}$ ). This may suggest that the development of software for detecting selection comes from a different research community compared to the other parameters

all follow the same format (frequencies of a binary description of alleles/marker data), SWEEPfinder offers distinct speed advantages in regards to HTS data (Pavlidis et al., 2013). This approach carries the cost of requiring increased preprocessing of data from the raw reads. However, it can offer advantages compared to earlier approaches for detecting hard sweeps. Excluding SELSCAN, data input is similar for all of these programs: they require allele frequency data. The SWEEPfinder programs can also utilize data on background selection to improve estimates.

### 3.5 | Citation bias

In addition to the short reflection and documentation of software developments to estimate population genetic (micro-evolutionary) parameters, we were interested in the usage bias of such programs. As no direct estimate of user numbers is available for many of the software packages, we used the frequency with which the programs are cited as a proxy of user frequency. With the citation records available on ISI databases, we could gather a fairly standardized



**FIGURE 2** Comparison of linear (green) to exponential (blue) growth of the number of software suites for estimating the various parameters, plotted against the data (black). The data for selection, fit to the nonlinear model better than to the linear (linear-MSS = 14.42, nonlinear-MSS = 7.88;  $n = 38$ ), although this difference is considerably weaker than for the other parameters. For mutation (linear-MSS = 2.39, nonlinear-MSS = 6.74;  $n = 17$ ), migration (linear-MSS = 3.37, nonlinear-MSS = 31.43;  $n = 33$ ) and drift (linear-MSS = 2.48, nonlinear-MSS = 26.14;  $n = 44$ ), the linear models all fit considerably closer to the observed data. These differences are taken to illustrate a more rapid rate of growth in the development of software analyzing/estimating selection, over drift, mutation, and migration

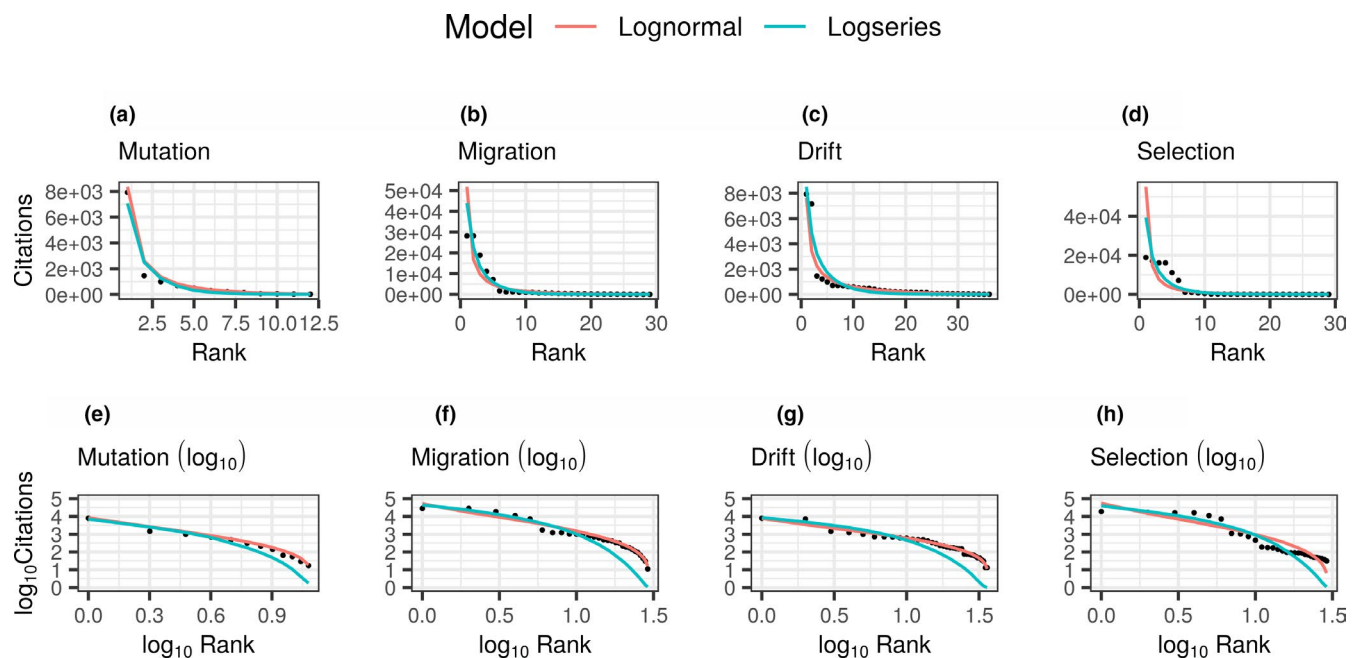
measure for the frequency of program/software usage. Of the 101 programs and scripts investigated (Table 1), some programs were not published in peer-reviewed journals. Furthermore, of those published, only those before 2014 were investigated. The pattern of growth in the citation record was visually inspected for this subset (plus the total number of citations for the four most cited programs: MEGA, ARLEQUIN, DNASP, and STRUCTURE, Figure S1). As a consequence the geometric mean was used for measuring the average number of citations per year (for all relevant tests).

We found clear differences between the programs developed to analyze the different parameters: Generally, more programs can be found to estimate selection and migration. Furthermore, a more distinct research community appears to exist for investigating selection, than for the other parameters (there is an absence of overlap of programs estimating selection with the other micro-evolutionary parameters (hyper-geometric tests,  $p < 10^{-08}$  for selection, compared to  $p > .5$  for migration, mutation, drift; Figure 1)). This may also translate through to the observed different ages/growth in the number of programs for testing the different parameters (Figure 2), with programs measuring selection undergoing exponential growth (the numbers for the other parameters showing a linear increase over time). This points toward a growing interest in estimating selection in the last years.

In an attempt to disambiguate the potential role of measured factors on the citation bias, Kruskal–Wallis rank sum tests were performed over the platforms and parameters estimated. While no significant effects were observed for the computing platforms

( $\chi^2 = 8.86$ ,  $df = 7$ ,  $p > .2$ ,  $N = 86$ ), differences were observed according to the number of estimated parameters ( $\chi^2 = 23.96$ ,  $df = 10$ ,  $p < .01$ ,  $N = 86$ ). These differences, however, were due to either single programs, or suites from a single group/development (respectively, ARLEQUIN v.3 and v.3.5, and BEAST v.1, v.1.7 and v.2). Dropping these records as outliers, no significant differences were observed ( $\chi^2 = 15.19$ ,  $df = 8$ ,  $p > .05$ ,  $N = 81$ ). For all parameters, though, there was a high disparity in the distribution of citations: over all programs more than half the citations (>62%) belonged to the four most cited (top five percent of the most cited) programs; respectively, these are STRUCTURE, MEGA6, ARLEQUIN v.3, and DNASP (all of which have been in development beyond the average development time observed here, 2008).

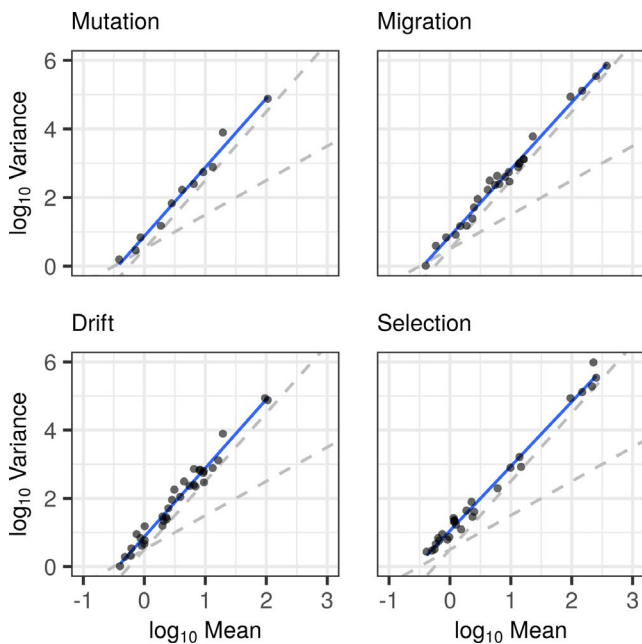
With such a highly skewed distribution, we fitted log-normal and log-series models to our citation dataset (Figure 3). Irrespective of the parameter that the software or program is designed to estimate, the citation records strongly fitted these log-models. These distributions appear to fit the ecological law described by Taylor (Taylor, 1961) in Figure 4, whereby the frequency with which a program is observed is based on random processes, namely, its previous observed frequency. We tested this using a log-regression model ( $b = 1.93 \pm 0.03$  (SE),  $df = 84$ ,  $N = 86$ ,  $p < 2 \times 10^{-16}$ ;  $\log(\text{var}) \sim \log(\text{mean})$ ), which shows a value deviating slightly from 2.0. This indicates that while there may be some processes other than Taylor's power law influencing the usage of the software tools, a simple exponential model of growth provides a powerful explanation of our observed results. Where the number of citations is the best predictor for future citations. A fact



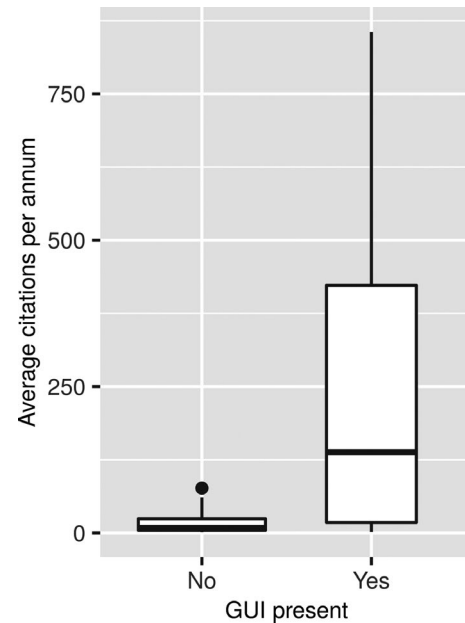
**FIGURE 3** Log-normal and log-series plots of the citation records for the different micro-evolutionary parameters. In these series, the number of citations reported on ISI web of knowledge site are displayed as dependent on the citation rank. These results are reported as raw counts (a–d), or as plots of the log (to base 10, e–h) for both the dependent and independent variables. From all of these plots, the log-normal model fits better to the observed data, independent of the parameter the programs estimate. Generally, the log-series fits data with inflated single observations, here single citations, a situation that is unlikely to be common here. A situation that does not reflect the data where the greatest deviation from the log-normal was observed (d, h)

that is not due to an over abundance of very infrequently cited tools as evidenced in Figure 3 (where better fits to log-normal models are observed).

To investigate whether this deviance from Taylor's power law is being influenced by other parameters, we analyzed the journal IF (where the program was published in) and presence/absence of a GUI. There was no effect based on the regression of the geometric mean of citations (to remove the impact of the life span of the program) on IF (Figure S2), using a robust linear model (R-Cran MASS package: value = 0.74, SE = 0.42,  $t$ -value = 1.76). Although it should be noted that this result is difficult to interpret, correlations between manuscript/software quality and journal IF can be expected. Whether the lack of observed significance for the IF indicates the lack of an effect for either role remains unclear (Figure S2). In contrast, a strong effect from the presence/absence of a GUI (Figure 5, Wilcoxon rank sum test,  $W = 186$ ,  $p$ -value  $< 4.8 \times 10^{-04}$ ,  $N = 65$ ) suggests an important role for the "ease of use" for the adoption of programs within the biological community. Complications to this conclusion arise from the fact that some of the older suites have incorporated GUIs since their initial publication, as well as the development of some tools within previously existing frameworks (R-Cran; R Core Team, 2018).



**FIGURE 4** Taylors power law graphs illustrating the relationship between the variance and mean in distribution of citations. Here, the pattern of the citation bias is plotted using the log of the means (to base 10) as a predictor of the variance. With the slope of the temporal variance and mean equal to 2 (the steeper dashed line, in all four graphs), the process of bias in citations follows a simple power law where the variation in citation follows previous frequencies of citation. Hence, alternative factors affecting the distribution of citation bias can be discarded. This relationship is observed for all four population genetic parameters (Mutation, upper-left; Migration, upper-right; Drift lower-left; and Selection, lower-right)



**FIGURE 5** Effect of existence of a GUI on the citation rate. The differences in the geometric mean in citation rate according to the presence or absence of a GUI for software for estimating any of the four parameters considered here (mutation, drift, migration, or selection). Much greater range sizes are observed in the citation records for those software suites where a GUI has been developed, with this difference being highly significant (Wilcoxon rank sum test,  $W = 186$ ,  $p$ -value  $4.8 \times 10^{-04}$ ,  $N = 65$ )

Investigating further factors that could underlie the distribution of software usage contains additional difficulties, including confounding factors for many hypotheses. This can be illustrated through the concept of "service provision" in the software "marketplace". The "marketplace" represents the diversity of software and algorithms that are available to researchers, each individual program offering a service within the "marketplace". With the advent of new technologies (as occurred with HTS), a range of new niches arose allowing for new/more "service provision" from new and existing software. Any early developments in such an area gain an advantage through a higher degree of "service provision". This encourages further developments and software within the field, each subsequent development (excluding the expansion of niches), in turn, offering fewer novel services. Hence, the "service provision" of software depends on the population of pre-existing software and cannot be considered temporally independent. Another factor difficult to measure is the degree to which researchers continue to utilize tools due to familiarity, as this repetition of use could also reflect the quality of the software. Similarly, the ability of software developers to gain attention from the research community may be related to the required investment from the users to develop the skills and data to employ the new software. This might also depend on the degree of "overlap" between the diversity of software tools that are available, for example, the wide range of programs providing LK tests for detecting selection, or  $F_{ST}$  variance tests for population size/drift. In cases of high overlap, the usability will play a major role compared to niches, which are only occupied by single, or few programs.

Stochastic processes have been shown to be present in other human systems, such as the use/success of Linux distributions (Keil et al., 2018) and law and order (Hanley, Lewis & Ribeiro 2016). While the frequency that a software tool is observed might still be related to the age/development of the tool, such data were not available in the current study. The presence of multiple articles and iterations of the different software tools additionally makes such analyses considerably more difficult. Despite this, one can ascertain distinctions between the different communities involved in software development (Figure 1) and factors leading to differences in citation frequency (Figure 5). Further investigations within the use of population genetics software within the scientific community may best be advanced through the collection of more data on the tools, or more thorough investigations of some of the specific communities. For instance, more recent studies on selection have employed multiple different statistics to identify candidate sites using Tajima's D, LK tests (HAPFLK) and the iHS (IHS) (Harpur et al., 2019). Investigation of the potential use of complementary approaches in population genetic analyses might be of much wider interest to the research community. This could be investigated through the robustness of tools, such as HADIVSEL (Carvajal-Rodríguez, 2017), that incorporate multiple statistics compared to tools that employ a single family of techniques.

## 4 | CONCLUSIONS

A large assortment of software exists for the estimation of population genetic parameters; between these parameters the development environments appear to be distinct (Figures 1 and 2). (Although this should be taken with some caution as this may be the result of low usage frequencies observed before 2002/03.) While not every software suite offers unique functionality, the majority offer unique capabilities. A clear distinction for this can be found in the discussed software: differences in capability can be seen in GENALEX that estimates migration rates, and STRUCTURE that analyses admixture and population stratification, assigning samples to populations. In contrast, ARLEQUIN offers the functionality of offering multiple different capabilities, that, however, are not unique and do not offer the user different capabilities. Despite the large number of available solutions, it appears, that in the majority of cases a much more limited range of software suites are actually utilized, potentially suggesting an under-exploitation of a large variety of bioinformatic resources within molecular population genetics. Key to determining the degree of exploitation of the bioinformatic resources is the presence of a GUI, the only factor identified to predict high citation frequencies. The range of under-exploitation is, however, observed for all the population genetic parameters investigated here (Figure 3). The effect potentially being more prevalent for those software suites estimating selection and migration, as greater ranges in citation frequencies were observed for these parameters.

Future studies should also implement the analyses of custom scripts, which are more frequently used in the advent of genomic studies. Further, hybridization and recombination have been

established as important additional evolutionary forces and need to be included in future analyses of population genetic software.

## ACKNOWLEDGMENT

This work was supported by EvoHack, the inaugural science hackathon of the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig funded by the German Research Foundation (FZT 118). JOE was funded by the FWO Flanders PostDoc Fellowship Program (grant number: 12G4317N). Open Access funding enabled and organized by Projekt DEAL.

## CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## AUTHOR CONTRIBUTIONS

**Jonathan Kidner:** Data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); software (equal); visualization (lead); writing—original draft (equal); writing—review and editing (equal). **Panagiotis Theodorou:** Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); software (equal); supervision (equal); visualization (supporting); writing—original draft (equal); writing—review and editing (equal). **Jan O. Engler:** Data curation (equal); formal analysis (supporting); investigation (equal); methodology (equal); software (equal); writing—original draft (equal); writing—review and editing (equal). **Martin Taubert:** Data curation (equal); methodology (equal); supervision (equal); writing—original draft (supporting); writing—review and editing (equal). **Martin Husemann:** Conceptualization (lead); data curation (equal); formal analysis (supporting); funding acquisition (lead); investigation (equal); methodology (equal); project administration (lead); software (equal); supervision (equal); writing—original draft (equal); writing—review and editing (equal).

## DATA AVAILABILITY STATEMENT

All datasets generated and analyzed for this study are provided within the manuscript and Appendix S1. All the data used are additionally available on Dryad with the accession number/address <https://doi.org/10.5061/dryad.0p2ngf220>

## ORCID

Jonathan Kidner  <https://orcid.org/0000-0002-9849-9342>

## REFERENCES

- Alachiotis, N., Stamatakis, A., & Pavlidis, P. (2012). OmegaPlus: A scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics*, 28, 2274–2275. <https://doi.org/10.1093/bioinformatics/bts419>
- Allendorf, F. W. (2016). Genetics and the conservation of natural populations: Allozymes to genomes. *Molecular Ecology*, 26, 420–430. <https://doi.org/10.1111/mec.13948>
- Álvarez, D., Lourenço, A., Oro, D., & Veló-Anton, G. (2015). Assessment of census (N) and effective population size (N<sub>e</sub>) reveals consistency

- of  $N_e$  single-sample estimators and a high  $N_e/N$  ratio in an urban and isolated population of fire salamander. *Conservation Genetic Resources*, 7, 705–712.
- Anderson, E. C. (2005). An efficient Monte Carlo method for estimating  $N_e$  from temporally spaced samples using a coalescent-based likelihood. *Genetics*, 170, 955–967. <https://doi.org/10.1534/genetics.104.038349>
- Anderson, E. C., Williamson, E. G., & Thompson, E. A. (2000). Monte Carlo evaluation of the likelihood for  $N_e$  from temporally spaced samples. *Genetics*, 156, 2109–2118.
- Andolfatto, P. (2008). Controlling type-I error of the McDonald-Kreitman test in Genomewide scans for selection on noncoding DNA. *Genetics*, 180, 1767–1771. <https://doi.org/10.1534/genetics.108.091850>
- Antao, T., Lopes, A., Lopes, R., Beja-Pereira, A., & Luikart, G. (2008). LOSITAN: A workbench to detect molecular adaptation based on a  $F_{st}$ -outlier method. *BMC Bioinformatics*, 9, 323.
- Bahlo, M., & Griffiths, R. C. (2000). Inference from gene trees in a subdivided population. *Theoretical Population Biology*, 57, 79–95.
- Balding, D. J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology*, 63, 221–230.
- Baldrige, E., Harris, D. J., Xiao, X., & White, E. P. (2016). An extensive comparison of species-abundance distribution models. *PeerJ*, 4, e2823. <https://doi.org/10.7717/peerj.2823>
- Barbato, M., Orozco-terWengel, P., Tapio, M., & Bruford, M. W. (2015). SNeP: A tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in Genetics*, 6, 1–6. <https://doi.org/10.3389/fgene.2015.00109>
- Barker, J. S. F. (2011). Effective population size of natural populations of *Drosophila buzzatii*, with a comparative evaluation of nine methods of estimation. *Molecular Ecology*, 20, 4452–4471. <https://doi.org/10.1111/j.1365-294X.2011.05324.x>
- Bateman, A. J. (1959). The viability of near-normal irradiated chromosomes. *International Journal of Radiation Biology and Related Studies in Physics, Chemistry and Medicine*, 1, 170–180. <https://doi.org/10.1080/09553005914550241>
- Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164, 1139–1160. <https://doi.org/10.1093/genetics/164.3.1139>
- Beaumont, M. A., & Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5, 251–261. <https://doi.org/10.1038/nrg1318>
- Becquent, C., & Przeworski, M. (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, 17, 1505–1519. <https://doi.org/10.1101/gr.6409707>
- Becquet, C., & Przeworski, M. (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, 17, 1505–1519. <https://doi.org/10.1101/gr.6409707>
- Beerli, P., & Felsenstein, J. (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, 152, 763–773. <https://doi.org/10.1093/genetics/152.2.763>
- Beerli, P., & Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in  $n$  subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8), 4563–4568. <https://doi.org/10.1073/pnas.081068098>
- Beerli, P., & Palczewski, M. (2010). Unified framework to evaluate Panmixia and migration direction among multiple sampling locations. *Genetics*, 185, 313–326. <https://doi.org/10.1534/genetics.109.112532>
- Berthier, P., Beaumont, M. A., Cornuet, J.-M., & Luikart, G. (2002). Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: A genealogical approach. *Genetics*, 160, 741–751. <https://doi.org/10.1093/genetics/160.2.741>
- Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., Palmer, C. D., Adeyemo, A. A., Akyllbekova, E. L., Cupples, L. A., Divers, J., Fornage, M., Kao, W. H. L., Lange, L., Li, M., ... Price, A. L. (2011). Genome-wide comparison of African-ancestry population from CARE and other cohorts reveals signals of natural selection. *The American Journal of Human Genetics*, 89, 368–381.
- Boitard, S., Kofler, R., Francoise, P., Robelin, D., Schloetterer, C., & Futschik, A. (2013). Pool-hmm: A Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. *Molecular Ecology Resources*, 13, 337–340. <https://doi.org/10.1111/1755-0998.12063>
- Bollback, J. P., York, T. L., & Nielsen, R. (2008). Estimation of  $2N_e$ s from temporal allele frequency data. *Genetics*, 179, 497–502.
- Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., & SanCristobal, M. (2010). Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics*, 186, 241–262. <https://doi.org/10.1534/genetics.110.117275>
- Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4), e1003537. <https://doi.org/10.1371/journal.pcbi.1003537>
- Broquet, T., & Petit, E. J. (2009). Molecular estimation of dispersal for ecology and population genetics. *Annual Review of Ecology Evolution and Systematics*, 40, 193–216. <https://doi.org/10.1146/annurev.ecolsys.110308.120324>
- Burczyk, J., Adams, W. T., Moran, G. F., & Griffin, A. R. (2008). Complex patterns of mating revealed in a *Eucalyptus regnans* seed orchard using allozyme markers and the neighbourhood model. *Molecular Ecology*, 11(11), 2379–2391. <https://doi.org/10.1046/j.1365-294x.2002.01603.x>
- Carvajal-Rodriguez, A. (2017). HacDivSel: Two new methods (haplotype-based and outlier-based) for the detection of divergent selection in pairs of populations. *PLoS One*, 12(4), e0175944. <https://doi.org/10.1371/journal.pone.0175944>
- Cavalli-Sforza, L. L. (1998). The DNA revolution in population genetics. *Trends in Genetics*, 14, 60–65. [https://doi.org/10.1016/S0168-9525\(97\)01327-9](https://doi.org/10.1016/S0168-9525(97)01327-9)
- Cayuela, H., Rougemont, Q., Prunier, J. G., Moore, J. S., Clobert, J., Besnard, A., & Bernatchez, L. (2018). Demographic and genetic approaches to study dispersal in wild animal populations: A methodological review. *Molecular Ecology*, 27, 3976–4010. <https://doi.org/10.1111/mec.14848>
- Charlesworth, B. (2009). Effective population size and patterns of molecular evolution. *Nature Reviews Genetics*, 10, 195–205.
- Chen, C., Durand, E., Forbes, F., & Francois, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: A new computer program and a comparison study. *Molecular Ecology Notes*, 7, 747–756. <https://doi.org/10.1111/j.1471-8286.2007.01769.x>
- Chen, G. B., Lee, S. H., Zhu, Z. X., Benyamin, B., & Robinson, M. R. (2015). EigenGWAS: Finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity*, 117(1), 51–61.
- Chybicki, I. J., & Burczyk, J. (2010). NM+: Software implementing parentage-based models for estimating gene dispersal and mating patterns in plants. *Molecular Ecology Resources*, 10(6), 1071–1075. <https://doi.org/10.1111/j.1755-0998.2010.02849.x>
- Ciofi, C., Beaumont, M. A., Swingland, I. R., & Bruford, M. W. (1999). Genetic divergence and units for conservation in the Komodo dragon *Varanus komodoensis*. *Proceedings of the Royal Society London B*, 266, 2269–2274. <https://doi.org/10.1098/rspb.1999.0918>



- Coombs, J. A., Letcher, B. H., & Nislow, K. H. (2012). GON<sub>c</sub>: Software for estimating effective population size in species with generational overlap. *Molecular Ecology Resources*, 12, 160–163.
- Cornuet, J. M., Pudlo, P., Veysier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., Marin, J.-M., & Estoup, A. (2014). DIYABC v2.0: A software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, 30(8), 1187–1189.
- Cornuet, J. M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J. M., Balding, D. J., Guillemaud, T., & Estoup, A. (2008). Inferring population history with DIY ABC: A user-friendly approach to approximate Bayesian computation. *Bioinformatics*, 24, 2713–2719.
- Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, 25, 410–418. <https://doi.org/10.1016/j.tree.2010.04.001>
- Csilléry, K., François, O., & Blum, M. G. B. (2012). abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3, 475–479. <https://doi.org/10.1111/j.2041-210X.2011.00179.x>
- Davey, J. W., & Blaxter, M. L. (2010). RADSeq: Next-generation population genetics. *Briefings in Functional Genomics*, 9, 416–423. <https://doi.org/10.1093/bfpg/elq031>
- De Mita, S., & Siol, M. (2012). EggLib: Processing, analysis and simulation tools for population genetics and genomics. *BMC Genetics*, 13, 27. <https://doi.org/10.1186/1471-2156-13-27>
- Debevec, E. M., Gates, R. B., Masuda, M., Pella, J., Reynolds, J., & Seeb, L. W. (2000). SPAM (Version 3.2): Statistics program for analyzing mixtures. *Journal of Heredity*, 91(6), 509–510.
- DeGiorgio, M., Huber, C. D., Hubisz, M. H., Hellmann, I., & Nielsen, R. (2016). SweepFinder2: Increased sensitivity, robustness, and flexibility. *Bioinformatics*, 32, 1895–1897. <https://doi.org/10.1093/bioinformatics/btw051>
- DeGiorgio, M., Lohmueller, K. E., & Nielsen, R. (2014). A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genetics*, 10, e1004561. <https://doi.org/10.1371/journal.pgen.1004561>
- Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillett, B. J., & Ovenden, J. R. (2014). N<sub>e</sub>Estimator v2: Re-implementation of software for the estimation of contemporary effective population size (N<sub>e</sub>) from genetic data. *Molecular Ecology Resources*, 14, 209–214.
- Drummond, A. J., & Rambaut, A. (2007). "BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7, 214.
- Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8), 1969–1973. <https://doi.org/10.1093/molbev/mss075>
- Dupanloup, I., Schneider, S., & Excoffier, L. (2002). A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology*, 11(12), 2571–2581. <https://doi.org/10.1046/j.1365-294X.2002.01650.x>
- Egea, R., Casillas, S., & Barbadilla, A. (2008). Standard and Generalized McDonald-Kreitman Test: A website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research*, 36, w157–162. <https://doi.org/10.1093/nar/gkn337>
- Ewing, G., Nicholls, G., & Rodrigo, A. (2004). Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics*, 168, 2407–2420.
- Ewing, G., & Rodrigo, A. (2006). Estimating population parameters using the structured serial coalescent with Bayesian MCMC inference when some demes are hidden. *Evolutionary Bioinformatics*, 2, 227–235.
- Excoffier, G. L., & Schneider, S. (2005). Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, 1, 47–50.
- Excoffier, L., & Heckel, G. (2006). Computer programs for population genetics data analysis: A survival guide. *Nature Reviews Genetics*, 7, 745–758. <https://doi.org/10.1038/nrg1904>
- Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10, 564–567. <https://doi.org/10.1111/j.1755-0998.2010.02847.x>
- Eyre-Walker, A. (2006). The genomic rate of adaptive evolution. *Trends in Ecology & Evolution*, 21, 569–575. <https://doi.org/10.1016/j.tree.2006.06.015>
- Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., & Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, 193, 929–941. <https://doi.org/10.1534/genetics.112.147231>
- Faubet, P., & Gaggiotti, O. E. (2008). A new Bayesian method to identify the environmental factors that influence recent migration. *Genetics*, 178, 1491–1504. <https://doi.org/10.1534/genetics.107.082560>
- Ferrer-Admetlla, A., Liang, M., Korneliusen, T., & Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, 31(5), 1275–1291. <https://doi.org/10.1093/molbev/msu077>
- Foll, M., & Gaggiotti, O. E. (2008). A genome scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*, 180, 977–993. <https://doi.org/10.1534/genetics.108.092221>
- Foll, M., Shim, H., & Jensen, J. D. (2015). WFABC: A Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources*, 15, 87–98. <https://doi.org/10.1111/1755-0998.12280>
- Foster, P. L. (2006). Methods for determining spontaneous mutation rates. *Methods in Enzymology*, 409, 195–213.
- García-Dorado, A. (1997). The rate and effects distribution of viability mutation in drosophila: Minimum distance estimation. *Evolution*, 51, 1130–1139. <https://doi.org/10.2307/2411042>
- García-Dorado, A., & Gallego, A. (2003). Comparing analysis methods for mutation-accumulation data: A simulation study. *Genetics*, 164, 807–819. <https://doi.org/10.1093/genetics/164.2.807>
- García-Dorado, A., & Marin, J. M. (1998). Minimum distance estimation of mutational parameters for quantitative traits. *Biometrics*, 54(3), 1097–1114. <https://doi.org/10.2307/2533860>
- Gautier, M. (2015). Genome-wide scan for adaptive differentiation and association analysis with population-specific covariables. *Genetics*, 201(4), 1555–1579.
- Gautier, M., Klassmann, A., & Vitalis, R. (2017). Genome-wide scan for adaptive differentiation and association analysis with population-specific covariables. *Genetics*, 201, 1555–1579.
- Gautier, M., & Vitalis, R. (2012). rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, 28(8), 1176–1177. <https://doi.org/10.1093/bioinformatics/bts115>
- Gillet-Markowska, A., Louvel, G., & Fischer, G. (2015). bz-rates: A web tool to estimate mutation rates from fluctuation analysis. *G3: Genes, Genomes, Genetics*, 5, 2323–2327.
- Gompert, Z. (2016). Bayesian inference of selection in a heterogeneous environment from genetic time-series data. *Molecular Ecology*, 25, 121–134.
- Guillot, G., Mortier, F., & Estoup, A. (2005). Geneland: A computer package for landscape genetics. *Molecular Ecology Resources*, 5(3), 712–715. <https://doi.org/10.1111/j.1471-8286.2005.01031.x>
- Gunther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, 196, 205–220. <https://doi.org/10.1534/genetics.113.152462>
- Habel, J. C., Husemann, M., Finger, A., Danley, P. D., & Zachos, F. E. (2014). The relevance of time series in molecular ecology and conservation biology. *Biological Reviews*, 89, 484–492.

- Hall, B. M., Ma, C., Liang, P., & Singh, K. K. (2009). Fluctuation Analysis CalculatOR: A web tool for the determination of mutation rate using Luria-Delbrück fluctuation analysis. *Bioinformatics*, *25*, 1564–1565.
- Hanley, Q. S., Dan, L., & Ribeiro, H. V. (2016). Rural to urban population density scaling of crime and property transactions in English and Welsh Parliamentary Constituencies. *PLoS One*, *11*(2), e0149546. <https://doi.org/10.1371/journal.pone.0149546>
- Hare, M. P., Nunney, L., Schwartz, M. K., Ruzzante, D. E., Burford, M., Waples, R. S., Ruegg, K., & Palstra, F. (2011). Understanding and estimating effective population size for practical application in marine species management. *Conservation Biology*, *25*(3), 438–449. <https://doi.org/10.1111/j.1523-1739.2010.01637.x>
- Harpur, B. A., Guarna, M. M., Huxter, E., Higo, H., Moon, K. M., Hoover, S. E., Ibrahim, A., Melathopoulos, A. P., Desai, S., Currie, R. W., Pernal, S. F., Foster, L. J., & Zayed, A. (2019). Integrative genomics reveals the genetics and evolution of the honey Bee's social immune system. *Genome Biology and Evolution*, *11*, 937–948. <https://doi.org/10.1093/gbe/evz018>
- Harris, A. M., Garud Nandita, R., & Michael, D. G. (2018). Detection and classification of hard and soft sweeps from unphased genotypes by multilocus genotype identity. *Genetics*, *210*(4), 1429–1452. <https://doi.org/10.1534/genetics.118.301502>
- Haubold, B., Pfaffelhuber, P., & Lynch, M. (2010). mlRho: A program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Molecular Ecology*, *19*, 277–284. <https://doi.org/10.1111/j.1365-294X.2009.04482.x>
- Hey, J. (2010). Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, *27*, 905–920. <https://doi.org/10.1093/molbev/msp296>
- Hey, J., & Nielsen, R. (2004). Multilocus methods for estimating population sized, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, *167*, 747–760.
- Hey, J., & Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(8), 2785–2790. <https://doi.org/10.1073/pnas.0611164104>
- Holleley, C. E., Nichols, R. A., Whitehead, M. R., Adamack, A. T., Gunn, M. R., & Sherwin, W. B. (2014). Testing single-sample estimators of effective population size in genetically structured populations. *Conservation Genetics*, *15*, 23–35. <https://doi.org/10.1007/s10592-013-0518-3>
- Hollenbeck, C. M., Portnoy, D. S., & Gold, J. R. (2016). A method for detecting recent changes in contemporary effective population size from linkage disequilibrium at linked and unlinked loci. *Heredity*, *117*(4), 207–216. <https://doi.org/10.1038/hdy.2016.30>
- Hudson, R. R., Kreitman, M., & Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, *116*(1), 153–159.
- Hui, T.-Y.-J., & Burt, A. (2015). Estimating effective population size from temporally spaced samples with a novel, efficient maximum-likelihood algorithm. *Genetics*, *200*, 285–293. <https://doi.org/10.1534/genetics.115.174904>
- Husemann, M., Nguyen, R., Ding, B., & Danley, P. D. (2015). A genetic demographic analysis of Lake Malawi rock-dwelling cichlids using spatio-temporal sampling. *Molecular Ecology*, *24*, 2686–2701. <https://doi.org/10.1111/mec.13205>
- Husemann, M., Zachos, F. E., Paxton, R. J., & Habel, J. C. (2016). Effective population size in ecology and evolution. *Heredity*. <https://doi.org/10.1038/hdy.2016.75>
- Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*, 1403–1405. <https://doi.org/10.1093/bioinformatics/btn129>
- Jones, O. R., & Wang, J. (2009). COLONY: A program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, *10*, 551–555. <https://doi.org/10.1111/j.1755-0998.2009.02787.x>
- Jorde, P. E., & Ryman, N. (2007). Unbiased estimator for genetic drift and effective population size. *Genetics*, *177*, 927–935. <https://doi.org/10.1534/genetics.107.075481>
- Katju, V., & Bergthorsson, U. (2019). Old trade, new tricks: Insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biology and Evolution*, *11*(1), 136–145. <https://doi.org/10.1093/gbe/evy252>
- Keightley, P. D. (1998). Inference of genome-wide mutation rates and distributions of mutation effects for fitness traits: A simulation study. *Genetics*, *150*(3), 1283–1293. <https://doi.org/10.1093/genetics/150.3.1283>
- Keightley, P. D., Ness, R. W., Halligan, D. L., & Haddrill, P. R. (2014). Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics*, *196*, 313–320.
- Keil, P., MacDonald, A. A., Ramirez, K. S., Bennett, J. M., García-Peña, G. E., Yguel, B., Bourgeois, B., & Meyer, C. (2018). Macroecological and macroevolutionary patterns emerge in the universe of GNU/Linux operating systems. *Ecography*, *41*, 1788–1800.
- Kim, Y., & Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics*, *167*, 1513–1524. <https://doi.org/10.1534/genetics.103.025387>
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, *217*, 624–626. <https://doi.org/10.1038/217624a0>
- Kondrashov, F. A., & Kondrashov, A. S. (2010). Measurements of spontaneous rates of mutations in the recent past and the near future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*, 1169–1176. <https://doi.org/10.1098/rstb.2009.0286>
- Korneliusson, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 356. <https://doi.org/10.1186/s12859-014-0356-4>
- Kuhner, M. K. (2006). LAMARC 2.0: Maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, *22*, 768–770. <https://doi.org/10.1093/bioinformatics/btk051>
- Kuhner, M. K., & Smith, L. P. (2006). Comparing likelihood and Bayesian coalescent estimation of population parameters. *Genetics*, *175*, 155–165. <https://doi.org/10.1534/genetics.106.056457>
- Labate, J. A. (2000). Software for population genetic analyses of molecular data. *Crop Science*, *40*, 1521–1528.
- Lanfear, R., Kokko, H., & Eyre-Walker, A. (2014). Population size and the rate of evolution. *Trends in Ecology & Evolution*, *111*, 189–199. <https://doi.org/10.1016/j.tree.2013.09.009>
- Lea, D. E., & Coulson, C. A. (1949). The distribution of the numbers of mutants in bacterial populations. *Journal of Genetics*, *49*, 264–285. <https://doi.org/10.1007/BF02986080>
- Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*(1), 175–195.
- Li, H., & Durbin, R. (2011). Inference of human population history from whole genome sequence of a single individual. *Nature*, *475*, 493–496.
- Librado, P., & Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, *25*, 1451–1452. <https://doi.org/10.1093/bioinformatics/btp187>
- Lopes, J. S., Balding, D., & Beaumont, M. A. (2009). PopABC: A program to infer historical demographic parameters. *Bioinformatics*, *25*, 2747–2749. <https://doi.org/10.1093/bioinformatics/btp487>
- Luikart, G., & Cornuet, J. M. (1999). Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics*, *151*, 1211–1216. <https://doi.org/10.1093/genetics/151.3.1211>

- Luikart, G., Ryman, N., Talmon, D. A., Schwartz, M. K., & Allendorf, F. W. (2010). Estimation of census and effective population sizes: The increasing usefulness of DNA-based approaches. *Conservation Genetics*, 11, 355–373. <https://doi.org/10.1007/s10592-010-0050-7>
- Luria, S. E., & Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28, 491–511. <https://doi.org/10.1093/genetics/28.6.491>
- Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: An R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, 17(1), 67–77.
- Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics: TIG*, 26, 345–352. <https://doi.org/10.1016/j.tig.2010.05.003>
- Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., Dickinson, W. J., Okamoto, K., Kulkarni, S., Hartl, D. L., & Thomas, W. K. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 9272–9277. <https://doi.org/10.1073/pnas.0803466105>
- Maclean, C. A., Hong, N. P. C., & Prendergast, J. G. D. (2015). hapbin: An efficient program for performing haplotype-based scans for positive selection in large genomic datasets. *Molecular Biology and Evolution*, 32(11), 3027–3029.
- Malaspina, A.-S., Malaspina, O., Evans, S. N., & Slatkin, M. (2012). Estimating allele age and selection coefficient from time-serial data. *Genetics*, 192, 599–607. <https://doi.org/10.1534/genetics.112.140939>
- Mathieson, I., & McVean, G. (2013). Estimating selection coefficients in spatially structured populations from series data of allele frequencies. *Genetics*, 193, 973–984.
- Mayr, E., & Provine, W. B. (1980). *The evolutionary synthesis: Perspectives on the unification of biology*. Harvard University Press.
- Mukai, T. (1964). The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. *Genetics*, 50, 1–19. <https://doi.org/10.1093/genetics/50.1.1>
- Nakagome, S., Fukumizu, K., & Mano, S. (2013). Kernel approximate Bayesian computation in population genetic inferences. *Statistical Applications in Genetics and Molecular Biology*, 12, 667–678. <https://doi.org/10.1515/sagmb-2012-0050>
- Neel, M. C., McKelvey, K., Ryman, N., Lloyd, M. W., Short Bull, R., Allendorf, F. W., Schwartz, M. K., & Waples, R. S. (2013). Estimation of effective population size in continuously distributed populations: There goes the neighborhood. *Heredity*, 111, 189–199. <https://doi.org/10.1038/hdy.2013.37>
- Nielsen, R. (1997). A maximum likelihood approach to population samples of microsatellite alleles. *Genetics*, 146, 711–716.
- Nielsen, R., & Wakeley, J. (2001). Distinguishing migration from isolation: A markov chain Monte Carlo approach. *Genetics*, 158(2), 885–896.
- Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research*, 15, 1566–1575. <https://doi.org/10.1101/gr.4252305>
- Nishant, K. T., Singh, N. D., & Alani, E. (2009). Genomic mutation rates: What high-throughput methods can tell us. *Bioassays*, 31(9), 912–920. <https://doi.org/10.1002/bies.200900017>
- Nunney, L. (2016). The effect of neighborhood size on effective population size in theory and in practice. *Heredity*, 117, 224–232. <https://doi.org/10.1038/hdy.2016.76>
- O’Ryan, C., Harley, E. H., Bruford, M. W., Mark, B., Wayne, R. K., & Cherry, M. I. (1998). Microsatellite analysis of genetic diversity in fragmented South African buffalo populations. *Animal Conservation*, 1(2), 85–94. <https://doi.org/10.1111/j.1469-1795.1998.tb00015.x>
- Palstra, F. P., & Ruzzante, D. E. (2008). Genetic estimates of contemporary effective population size: What can they tell us about the importance of genetic stochasticity for wild population persistence? *Molecular Ecology*, 17, 3428–3447. <https://doi.org/10.1111/j.1365-294X.2008.03842.x>
- Paradis, E. (2010). pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics*, 26, 419–420. <https://doi.org/10.1093/bioinformatics/btp696>
- Pavlidis, P., Zivkovic, D., Stamatakis, A., & Alachiotis, N. (2013). SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, 30(9), 2224–2234. <https://doi.org/10.1093/bioinformatics/btt499>
- Peakall, R., & Smouse, P. E. (2006). GENALEX 6: Genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, 6, 288–295. <https://doi.org/10.1111/j.1471-8286.2005.01155.x>
- Peakall, R., & Smouse, P. E. (2012). GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*, 28, 2537–2539. <https://doi.org/10.1093/bioinformatics/bts460>
- Peel, D., Ovenden, J. R., & Peel, S. L. (2004). *NeEstimator: Software for estimating effective population size*. Queensland Government, Department of Primary Industries and Fisheries.
- Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., & Pritchard, J. K. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, 19(5), 826–837. <https://doi.org/10.1101/gr.087577.108>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959. <https://doi.org/10.1093/genetics/155.2.945>
- Pudovkin, A. I., Zaykin, D., & Hedgecock, D. (1996). On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics*, 144, 383–387. <https://doi.org/10.1093/genetics/144.1.383>
- Putman, A. I., & Carbone, I. (2014). Challenges in analysis and interpretation of microsatellite data for population genetics. *Ecology and Evolution*, 4, 4399–4428.
- R Core Team (2018). *R: A language and environment for statistical computing*. Retrieved from <https://www.R-project.org>
- Robertson, A. (1975). Gene frequency distributions as a test of selective neutrality. *Genetics*, 81, 775–785. <https://doi.org/10.1093/genetics/81.4.775>
- Rosche, W. A., & Foster, P. L. (2000). Determining mutation rates in bacterial populations. *Methods (San Diego, Calif.)*, 20, 4–17.
- Ryman, N., Allendorf, F. W., Jorde, P. E., Laikre, L., & Høssjer, O. (2014). Samples from subdivided populations yield biased estimates of effective size that overestimate the rate of loss of genetic variation. *Molecular Ecology Resources*, 14, 87–99. <https://doi.org/10.1111/1755-0998.12154>
- Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419, 832–837. <https://doi.org/10.1038/nature01140>
- Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46, 919–925. <https://doi.org/10.1038/ng.3015>
- Sethuraman, A., & Hey, J. (2016). IM a2p-parallel MCMC and inference of ancient demography under the Isolation with migration (IM) model. *Molecular Ecology Resources*, 16, 206–215. <https://doi.org/10.1111/1755-0998.12437>
- Sheehan, S., Harris, K., & Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics*, 194, 647–662. <https://doi.org/10.1534/genetics.112.149096>

- Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E., & Pupko, T. (2007). Selecton 2007: Advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Research*, 35, W506–511. <https://doi.org/10.1093/nar/gkm382>
- Stigler, S. M. (2007). The epic story of maximum likelihood. *Statistical Science*, 22(4), 598–620. <https://doi.org/10.1214/07-STS249>
- Sunnucks, P. (2000). Efficient genetic markers for population biology. *Trends in Ecology and Evolution*, 15, 376–377. [https://doi.org/10.1016/S0169-5347\(00\)01825-5](https://doi.org/10.1016/S0169-5347(00)01825-5)
- Szpiech, Z. A., & Hernandez, R. D. (2014). selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, 31(10), 2824–2827. <https://doi.org/10.1093/molbev/msu211>
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585–595. <https://doi.org/10.1093/genetics/123.3.585>
- Tallmon, D. A., Koyuk, A., Luikart, G., & Beaumot, M. A. (2008). OneSamp: A program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources*, 8, 299–301.
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Ecology and Evolution*, 30, 2725–2729.
- Taylor, L. R. (1961). Aggregation, Variance and the mean. *Nature*, 189, 732–735. <https://doi.org/10.1038/189732a0>
- Valière, N. (2002). gimlet: A computer program for analysing genetic individual identification data. *Molecular Ecology Notes*, 2(3), 377–379. <https://doi.org/10.1046/j.1471-8286.2002.00228.x-i2>
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. Retrieved from <http://www.stats.ox.ac.uk/pub/MASS4>. ISBN 0-387-95457-0.
- Vitalis, R., & Couvet, D. (2001). Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics*, 157(2), 911–925. <https://doi.org/10.1093/genetics/157.2.911>
- Vitalis, R., Dawson, K., Boursot, P., & Belkhir, K. (2003). DetSel 1.0: A computer program to detect markers responding to selection. *Journal of Heredity*, 94, 429–431. <https://doi.org/10.1093/jhered/esg083>
- Vitalis, R., Gautier, M., Dawson, K. J., & Beaumont, M. A. (2014). Detecting and measuring selection from gene frequency data. *Genetics*, 196, 799–817. <https://doi.org/10.1534/genetics.113.152991>
- Voight, B. F., Kudravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, 4(3), e72.
- Wakeley, J., & Sargsyan, O. (2009). Extensions of the coalescent effective population size. *Genetics*, 181(1), 341–345. <https://doi.org/10.1534/genetics.108.092460>
- Wang, J. (1996). Deviation from Hardy-Weinberg proportions in finite populations. *Genetics Resources*, 68, 249–257. <https://doi.org/10.1017/S0016672300034224>
- Wang, J. (2006). A coalescent-based estimator of admixture from DNA sequences. *Genetics*, 173(3), 1679–1692. <https://doi.org/10.1534/genetics.105.054130>
- Wang, J. (2001). A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical Research*, 78, 243–257. <https://doi.org/10.1017/S0016672301005286>
- Wang, J. (2014). Estimation of migration rates from marker based parentage analysis. *Molecular Ecology*, 23(13). <https://doi.org/10.1111/mec.12806>
- Wang, J., Brekke, P., Huchard, E., Knapp, L. A., & Cowlshaw, G. (2010). Estimation of parameters of inbreeding and genetic drift in populations with overlapping generations. *Evolution*, 64, 1704–1718.
- Wang, J., Santiago, E., & Caballero, A. (2016). Prediction and estimation of effective population size. *Heredity*, 117, 193–206. <https://doi.org/10.1038/hdy.2016.43>
- Wang, J., & Whitlock, M. C. (2003). Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*, 163, 429–446. <https://doi.org/10.1093/genetics/163.1.429>
- Waples, R. S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked loci. *Conservation Genetics*, 7, 167–184.
- Waples, R. S., Antao, T., & Luikart, G. (2014). Effects of overlapping generations on linkage disequilibrium estimates of effective population size. *Genetics*, 197, 769–780. <https://doi.org/10.1534/genetics.114.164822>
- Waples, R. S., & Do, C. (2008). LDNe: A program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, 8, 753–756.
- Waples, R. S., & Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, 15, 1419–1439. <https://doi.org/10.1111/j.1365-294X.2006.02890.x>
- Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7, 256–276. [https://doi.org/10.1016/0040-5809\(75\)90020-9](https://doi.org/10.1016/0040-5809(75)90020-9)
- Wegmann, D., Leuenberger, C., Neuenschwander, S., & Excoffier, L. (2010). ABCtoolbox: A versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, 11, 116. <https://doi.org/10.1186/1471-2105-11-116>
- Whitlock, M. C., & Lotterhos, K. J. (2015). Reliable detection of loci responsible for local adaptation: Inference of a neutral model through trimming the distribution of  $F_{ST}$ . *The American Naturalist*, 186, s24–s36.
- Whitlock, M. C., & McCauley, D. E. (1999). Indirect measures of gene flow and migration:  $F_{ST} \neq 1/(4Nm + 1)$ . *Heredity*, 82, 117–125. <https://doi.org/10.1038/sj.hdy.6884960>
- Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, 21(12), 1–20. Retrieved from <http://www.jstatsoft.org/v21/i12/>
- Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1), 1–29. Retrieved from <http://www.jstatsoft.org/v40/i01/>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.
- Wilson, D. J., & McVean, G. (2006). Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics*, 172(3), 1411–1425. <https://doi.org/10.1534/genetics.105.044917>
- Wilson, G. A., & Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, 163(3), 1177–1191. <https://doi.org/10.1093/genetics/163.3.1177>
- Wloch, D. M., Szafraniec, K., Borts, R. H., & Korona, R. (2001). Direct Estimate of the mutation rate and the distribution of fitness effects in the Yeast *Saccharomyces cerevisiae*. *Genetics*, 159, 441–452.
- Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16, 97–159. <https://doi.org/10.1093/genetics/16.2.97>
- Wright, S. (1940). Breeding structure of populations in relation to speciation. *American Naturalist*, 74, 232–248. <https://doi.org/10.1086/280891>
- Yeh, F. C., & Boyle, T. B. J. (1997). *POPGENE, the user-friendly shareware for population genetic analysis*. Molecular Biology and Biotechnology Centre. University of Alberta, Canada.
- Zhdanova, O. L., & Pudovkin, A. I. (2008). Nb\_HetEx: A program to estimate the effective number of breeders. *Journal of Heredity*, 99(6), 694–695. <https://doi.org/10.1093/jhered/esn061>

Zheng, Q. (2017). rSalvador: An R package for the fluctuation experiment. *G3 Genes|Genomes|Genetics*, 7(12), 3849–3856. Genetics Society of America. <https://doi.org/10.1534/g3.117.300120>

#### SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Kidner, J., Theodorou, P., Engler, J. O., Taubert, M., & Husemann, M. (2021). A brief history and popularity of methods and tools used to estimate micro-evolutionary forces. *Ecology and Evolution*, 11, 13723–13743. <https://doi.org/10.1002/ece3.8076>