

MINI REVIEW OPEN ACCESS

On the Use of the Jackknife to Compare R_{XY} Statistics in Conservation Genomics Research

Oliver Patrick Stuart

School of Biological, Earth, and Environmental Sciences, University College Cork, Cork, Ireland

Correspondence: Oliver Patrick Stuart (ostuart@ucc.ie)

Received: 30 October 2025 | **Revised:** 13 April 2026 | **Accepted:** 22 April 2026

Keywords: biological statistics | bootstrap | confidence intervals | conservation genetics | conservation genomics | jackknife

ABSTRACT

Many useful quantities in conservation genomics are point estimates that summarise genetic variation at the population or species level. This presents a challenge to the application of null hypothesis testing frameworks, which require some measure of uncertainty to test the authenticity of measured group differences. The jackknife is a convenient statistical resampling tool which sees regular use to this end: producing confidence intervals for point estimates. Unfortunately, the slightly obscure nature of the jackknife coupled with its superficial resemblance to the bootstrap has led to its misuse, specifically in the analysis of R_{XY} statistics, a recently formulated quantity designed to measure relative mutation load. In this review, I summarise the application of the jackknife as a method to produce confidence intervals and review its uses in the conservation/population genomics literature. I outline how it is similar to and different from the bootstrap, and how the treatment of jackknife pseudo-values as bootstrap pseudo-values produces inflated confidence in otherwise negligible group differences. I then provide suggestions for the future use of the jackknife in conservation genomics.

1 | Introduction

In conservation and population genomics we are usually interested in summarising some aspect of genetic variation at the population or species level. To this end, there are summary statistics at our disposal which can be used to measure population differentiation and inbreeding (Malécot 1948; Wright 1943, 1949), genetic load (Do et al. 2015; Grossen et al. 2020), introgression (Durand et al. 2011; Green et al. 2010; Patterson et al. 2012; Rosenzweig et al. 2016), and myriad other biological processes and quantities. Calculating these statistics requires summing or multiplying values across many individuals and/or loci, for example in statistics that take population allele frequencies as inputs. As a result, large datasets comprising complex genetic polymorphism data from multiple individuals and populations are often distilled into point estimates. Abstraction and simplification are obviously necessary for comparable inferences to be made across systems. However, such estimates present a challenge to working scientists who overwhelmingly favour null

hypothesis testing frameworks to evaluate the authenticity of group differences. These tests require some measure of uncertainty, for example a sample standard error which allows the use of Student's t distribution for a test of differences in means. To obtain uncertainty measures, statistical resampling of individuals or loci is often employed.

The jackknife, a resampling procedure with superficial similarities to but that predates the bootstrap, has risen in popularity in the last decade as a convenient method for calculating the variance in genome-wide summary statistics. Jackknife resampling is routinely used in the analysis of introgression (Patterson et al. 2012), effective population size (Do et al. 2014; Waples and Do 2008), and genetic load (Do et al. 2015). In this latter case, Do et al. (2015) used a block jackknife procedure to estimate confidence intervals for R_{XY} , a complex ratio statistic that measures relative genetic load among populations. They noted that they used the jackknife replicates to calculate standard errors “using jackknife theory”, although they did not describe exactly what

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

formulae were used. Since then, numerous studies have used jackknifing improperly, treating the distribution of jackknife pseudo-values as if it represents the empirical error or variance of R_{XY} (e.g., Chen, Feng, et al. 2025; Dussex, Kurland, et al. 2023; Furni et al. 2025; Mathur et al. 2022; von Seth et al. 2022). In one case, it appears as if jackknife pseudo-values were used themselves as samples for statistical testing (Grossen et al. 2020) resulting in inflated significance of group differences. It is high time that we have some updated reference which guides the correct use of this statistical procedure, in the context of R_{XY} and otherwise.

In this article, I briefly outline the use of the jackknife as a statistical procedure specifically in the context of conservation genetics/genomics. I use simulations to demonstrate its properties and to compare it to the bootstrap. I then describe how it has been misused in the literature, mostly in the analysis of R_{XY} statistics. I do not aim to overturn the results of any of the studies cited herein, but simply to draw attention to the improper interpretation of the jackknife and to strengthen the analytical rigour with which conservation genomics research is conducted. I then demonstrate the correct use of the jackknife to estimate standard errors using a published example of R_{XY} statistics (Stuart et al. 2025) and compare it to a bootstrap approach.

2 | The Jackknife in Practice

The jackknife is a non-random leave-one-out procedure originally proposed by Quenouille (1956) and further developed by Tukey (1958) which can be used to estimate the bias or variance of a statistical estimator (which I hereafter refer to simply as a statistic). Here, I focus primarily on its use in estimating the sample variance of a statistic i.e., the construction of confidence intervals. To perform the jackknife on a dataset of n observations, one simply removes each observation in turn and recalculates the statistic with the $n-1$ observations, resulting in n pseudo-values. The mean of these pseudo-values is an accurate estimator of the true sample mean (Miller 1974) but modified formulae are required for the variance and other higher order moments. Tukey proposed in a conference abstract (Tukey 1958) that these pseudo-values could be treated as independent and identically distributed random variables. The pseudo-values would therefore have an approximate t distribution with $n-1$ degrees of freedom (or an approximate Z distribution for large enough n). The jackknife could then be used for the estimation of confidence intervals for otherwise intractable statistics:

$$SE_{\text{jackknife}} = \sqrt{\frac{n-1}{n} \sum_{j=1}^n (x_{-j} - \bar{x})^2}$$

where n is the sample size, x_{-j} is the jackknife pseudo-value of the statistic omitting the j^{th} observation, and $\bar{x}$ is the average of estimators across the jackknife pseudo-values. This formula resembles the standard deviation (which is used to calculate the standard error of a statistic from by bootstrapping; see below) except for the sample size correction of $\frac{n-1}{n}$. The exact notation used here is based on an essay by McIntosh (2016). While Tukey (1958) is credited with inspiring the use of the jackknife

for interval estimation, he himself did not publish any specific equations to this purpose.

The use of the jackknife in population genetics dates originally to Mueller (1979) who compared it to the delta method (which involves taking a Taylor series expansion of an expectation) for estimating the variance of Nei's genetic distance (Nei 1972). He found that, in general, the jackknife performed favourably to the delta method in terms of reducing the bias associated with sampling small numbers of loci. The method was then shortly taken up to estimate the variance in co-ancestry coefficients (Reynolds et al. 1983) and F -statistics (Weir and Cockerham 1984). Crucial to the resampling process is the assumption of linkage equilibrium among loci (Weir and Cockerham 1984). In the context of genome-wide polymorphism data, the physical linkage of loci (under which condition total linkage equilibrium is unlikely) motivates block resampling. Loci in close physical proximity on a chromosome share similar evolutionary histories as they tend to be inherited together as long haplotypes. Resampling of large segments of chromosome (and the loci contained therein) aims to reduce some of the resultant correlation among loci by treating a block of linked loci as a single observation. In this case, the n observations are grouped into g blocks of contiguous loci. The jackknife proceeds similarly in this case, where the statistic is recalculated without the loci contained in block g , resulting in g pseudo-values. Each pseudo-value is then weighted by the proportion of n used to calculate it. This modified procedure, known as the weighted block jackknife, was originally proposed in a more general context by Busing et al. (1999). They provided modified equations for mean and sample standard error of an estimator using the weighted block jackknife:

$$\hat{\theta}_{\text{weighted block jackknife}} = g\hat{\theta} - \sum_{j=1}^g \frac{(n-m_j)\hat{\theta}_{-j}}{n}$$

$$SE_{\text{weighted block jackknife}} = \sqrt{\frac{1}{g} \sum_{j=1}^g \left(\frac{n-m_j}{m_j} \right) (\hat{\theta} - \hat{\theta}_{-j})^2}$$

where g is the number of blocks, n is the total number of observations (i.e., the total number of loci summed across all blocks), $\hat{\theta}$ is the value of the statistic calculated using the entire dataset, $\hat{\theta}_{-j}$ is the value of the statistic omitting the j^{th} block, and m_j is the number of observations (loci, in our case) in the j^{th} block.

The best documented examples of the weighted block jackknife in the population genetics literature come from the study of archaic introgression in human and Neandertal populations. Reich et al. (2009) used the jackknife standard error to calculate a Z -score for f -statistics (not to be confused with F -statistics) estimated from single nucleotide polymorphism data comparing contemporary Indian populations. The authors warned that the assumption of normality in Z -scores from jackknife standard errors is often violated when $|Z| > 2$, meaning that large Z -scores can be treated as statistically significant but that standard Z tables cannot be used to convert these into p values (Reich et al. 2009; Thorburn 1977). The authors did not show any results to this effect. The weighted block jackknife was then used in the analysis of the first draft Neandertal genome (Green et al. 2010) to obtain standard errors for D -statistics measuring

the extent of introgression between Neandertals and contemporary human populations. They varied the number of blocks and found that a minimum block size of 2Mb was large enough to obtain stable jackknife standard errors. Reich et al. (2009) instead used 5cM blocks rather than measuring by physical distance. Green et al. (2010) likely chose to use physical distance to define blocks because the extent of linkage disequilibrium (LD) in the Neandertal genome was not known at the time.

To my knowledge, there is no literature concerning how best to define blocks for resampling genomic datasets, and Green et al. (2010) are the only authors to present an investigation of the effect of varying genomic block sizes for jackknifing. The optimal size will probably vary depending on the physical extent of LD in the genomes of the populations being studied. Patterson et al. (2012) used the weighted block jackknife to compute standard errors for several quantities, D and f -statistics among them (now incorporated into the software ADMIXTOOLS), and used a variety of block sizes: 5cM, 10cM, even whole chromosomes. The use of whole chromosomes as blocks was taken up by later authors (e.g., Grossen et al. 2020) although it is unclear how this approach differs from using even block sizes. It is possible to mathematically correct for the unevenness of blocks (Busing et al. 1999) but extreme length variation and small chromosome numbers in some species may not yield robust standard errors using the weighted block jackknife. A physically explicit (in the chromosomal sense) simulation study would be illuminating to this point: how much does chromosome structure alter the robustness of a whole-chromosome jackknife procedure?

Another area where the jackknife sees regular use is the study of genetic load. There has been a boom in the study of deleterious mutations in the conservation genetics literature over the last decade, largely driven by the declining costs of sequencing technologies (Dussex, Morales, et al. 2023; Robinson et al. 2023; van Oosterhout 2020), and it is now *de rigueur* to calculate some measure of the genetic load for the populations being studied. A statistic which has seen rising use to this effect is R_{XY} (Do et al. 2015) which was formulated explicitly to compare deleterious mutation load among human populations. The calculation of R_{XY} follows from counts of alleles in two populations (Do et al. 2015). A reference population Y is chosen to which a focal population X is compared. The expected probability of sampling an allele in X but not Y given a random sample of one haploid genome from each population (averaging over all possible pairs of samples) is calculated and summed across i sites:

$$L_{X, \text{not } Y} = \sum_i \left(\frac{d_X^i}{n_X^i} \right) \left(1 - \frac{d_Y^i}{n_Y^i} \right)$$

where n is the number of sampled haploid genomes and d is the count of alleles from that sample. Alleles can be polarised to the ancestral state or unpolarised, but of course the interpretation will differ depending on whether d refers to a count of derived versus non-reference alleles. R_{XY} is then calculated as a ratio of the $L_{X, \text{not } Y}$ statistic to its inverse, where the focal and reference populations are flipped:

$$R_{XY} = L_{X, \text{not } Y} / L_{Y, \text{not } X}$$

If the probability of sampling an allele (derived or non-reference) at a given site is the same in both X and Y, then $R_{XY} = 1$. Values less than or greater than 1 are taken as evidence for a depletion or enrichment in population X compared to population Y, and therefore as evidence of increased or reduced efficacy of selection acting on the sites included. Meaningful comparisons are made by comparing values among populations or mutation classes.

Do et al. (2015) used a weighted block jackknife, dividing the human genome into 100 contiguous blocks and estimating standard errors for R_{XY} from the jackknife pseudo-values. In the conservation genetics literature, the use of R_{XY} also traces back to Xue et al. (2015) who referred to it as $R_{A/B}$ and also used a jackknife procedure to estimate the variance of the statistic. Crucially, neither of these papers contained equations for their standard error calculations, nor did they cite literature where such equations might be found. For the other cases discussed above, f -statistics and D -statistics, the jackknife procedure is built into existing software tools, sidestepping individual user error. This is not the case for R_{XY} , a statistic which remains relatively underexamined. As a result, numerous subsequent studies have used the jackknife to produce confidence intervals for R_{XY} which are implausibly narrow, likely because they use incorrect equations (in some cases, available computer code indicates that this is indeed the case). I describe the specifics of this error below in section “The jackknife is not the bootstrap”.

The final major area where the jackknife is used for confidence interval estimation is in the analysis of effective population size (N_e). Here, the resampling approaches differ slightly to the previous examples. Single sample approaches to contemporary N_e estimation with molecular data typically make use of LD (reviewed by Waples 2024). Non-independence among measurements is inherent to the calculation of LD using more than two loci: among L loci there are $\frac{L(L-1)}{2}$ unique pairs but only $\frac{L}{2}$ non-overlapping pairs. In the LDNE software, Waples and Do (2008) implemented a jackknife which operates over pairs of loci rather than individual loci, and used this to estimate the effective sample size of pairs based on equations from Hill (1981). This effective sample size is the degrees of freedom required to calculate confidence intervals for the average LD among locus pairs.

The estimation of confidence intervals for N_e is one of the few areas of population/conservation genomics where the use of the jackknife as a statistical procedure has been recently evaluated. Jones et al. (2016) suggest that jackknifing over locus pairs is not jackknifing at all as no recalculation is performed, rather the existing LD calculations are re-averaged. They used simulations to show that jackknifing over individuals produces confidence intervals that more effectively capture true N_e under realistic data conditions such as when working with massive single-nucleotide polymorphism datasets. This discrepancy, between the jackknife over individuals and what Jones et al. (2016) refer to as the “pseudo-jackknife” over loci, may also apply to other published examples of the jackknife, including the weighted block jackknife. This label, the “pseudo-jackknife” (although I will continue to refer to resampling of loci/blocks as a jackknife), does not necessarily imply that the method cannot be used to produce viable confidence intervals. The conclusions of Reynolds et al. (1983), that the jackknife does indeed capture

the true variance in co-ancestry coefficients, were based on a comparison to the true variances obtained from simulated data, although they did not opt to resample individuals as a comparison to resampling of loci. To my knowledge, the only examples aside from Jones et al. (2016) that resample individuals rather than loci do so in the context of single locus analyses, e.g., van Dongen and Backeljau's (1995) bootstrap tests of inbreeding coefficients. In this case, resampling of individuals is the only option.

This problem is not new and was discussed (although in slightly different terms) by Weir (1996) and briefly by Weir and Cockerham (1984). The question at stake here is: exactly which variance is being explored when resampling is performed? Weir (1996) referred to two sources of variance, genetic variance and statistical variance. Genetic variance refers to variance across loci attributable to stochasticity in the evolutionary process (Weir and Cockerham (1984) focused on randomness in gamete formation). Statistical variance refers to the sampling variance that affects our ability to adequately characterise a given realisation of the evolutionary process. It would be informative to compare the effects of jackknifing over individuals rather than loci or blocks of loci in the estimation of quantities like R_{XY} or D -statistics, indeed any statistic that is calculated using many loci simultaneously. This would be intractable for some workflows, for example when using *angsd* (Korneliussen et al. 2014) which would require the generation of large intermediate analysis files for each resample. However, genomic analyses that operate on hard-called genotypes may be amenable to this approach.

3 | Wherefore the Jackknife?

The primary alternative to the jackknife is the bootstrap, another resampling procedure which was originally proposed by Efron (1979). In the original proposition, Efron (1979) showed that the jackknife is in fact a linear approximation of the bootstrap, and so the two are closely related. Bootstrapping proceeds by randomly resampling observations from a dataset with replacement to produce a new dataset of equivalent size to the original, recalculating the statistic of interest with each resampled dataset to produce a distribution of pseudo-values. In this way it superficially resembles the jackknife, although the properties of the two differ.

Efron (1979) suggested that the bootstrap is more generalisable and reliable than the jackknife. Indeed, other authors have cautioned that although the jackknife is convenient, its performance as an estimator for confidence intervals is not assured (Miller 1974). A convenient property of the bootstrap is that the standard deviation of the distribution of pseudo-values is an estimator of the sample standard error (Efron 1979), allowing the easy calculation of confidence intervals for summary statistics. The standard deviation is a familiar formula for anyone who has taken even an elementary statistics course. Why, then, should anyone prefer the more mathematically unwieldy jackknife?

The simple answer is computational tractability. In the case of block resampling of genomic datasets, each block may consist of thousands to tens of thousands of loci and the generation of

each bootstrap resample may be time consuming. Compared to the jackknife, where each resampled dataset is produced by simply removing one of the observations, the bootstrap may take some time. For effective population size estimates from LD statistics, this may be even more intractable given that the number of unique locus pairs grows much faster than the number of loci.

Setting aside the size of the resampled datasets, the total number of bootstraps to perform need not be that large. If the criterion is a stable estimate of the standard error of a statistic, then this could be easily verified on an *ad hoc* basis by tracking the coefficient of variation as the number of bootstraps increases. Indeed, the introduction of such stopping criteria has been proposed in the analysis of phylogenetic trees (Pattengale et al. 2010), wherein 100–500 bootstraps were deemed adequate for stable node support estimates in maximum likelihood phylogenetic inference. The number of bootstraps required to produce stable results (regardless of the statistic being calculated) is rarely studied in population/conservation genetic research.

Finally, the issue of “pseudo-jackknifing” over loci and not samples will also apply to bootstrap resampling. However, if a bootstrap of genomic windows or loci is considered more costly than a jackknife, then a bootstrap of individuals may exceed even that in computation time.

4 | The Jackknife is Not the Bootstrap

In my opinion, the superficial similarity of the bootstrap to the jackknife, in that they both involve recalculation of a statistic on subsets of data to produce a distribution of pseudo-values, has led to the misuse of the jackknife in the conservation genomics literature. Specifically, this has led to false confidence in the precision of R_{XY} estimates. It is my concern that this false confidence may lead to unfounded conclusions regarding the mutation load in threatened and vulnerable populations, and that this practice may spread to the use of other statistics.

From a verbal description alone, we can tell that the distribution of pseudo-values obtained using the jackknife will be much tighter than for the bootstrap. If we were to take a sample of 100 measurements from a population and perform a jackknife to estimate our confidence in the mean, then each jackknife resample will contain 99% of the original data and produce a pseudo-value closely resembling the true sample mean. Simple simulations can be used to demonstrate that the standard deviation of jackknife pseudo-values will systematically underestimate the sample standard error (of the mean, in this example). This bias increases with the number of observations following a simple power law relationship of $y = x^{-0.5}$, where y is the relative underestimation and x is the number of samples used for the jackknife (Figure 1, top). With 100 observations, the standard deviation of jackknife pseudo-values will be one tenth of the true standard error (Figure 1, top). The bootstrap is not biased in this way. The standard deviation of bootstrap pseudo-values is an accurate estimator of the sample standard error with precision commensurate to the number of resamples taken (Figure 1, middle). While the standard deviation of jackknife pseudo-values cannot be directly substituted for the sample standard error in the same way as for the bootstrap, the formula above

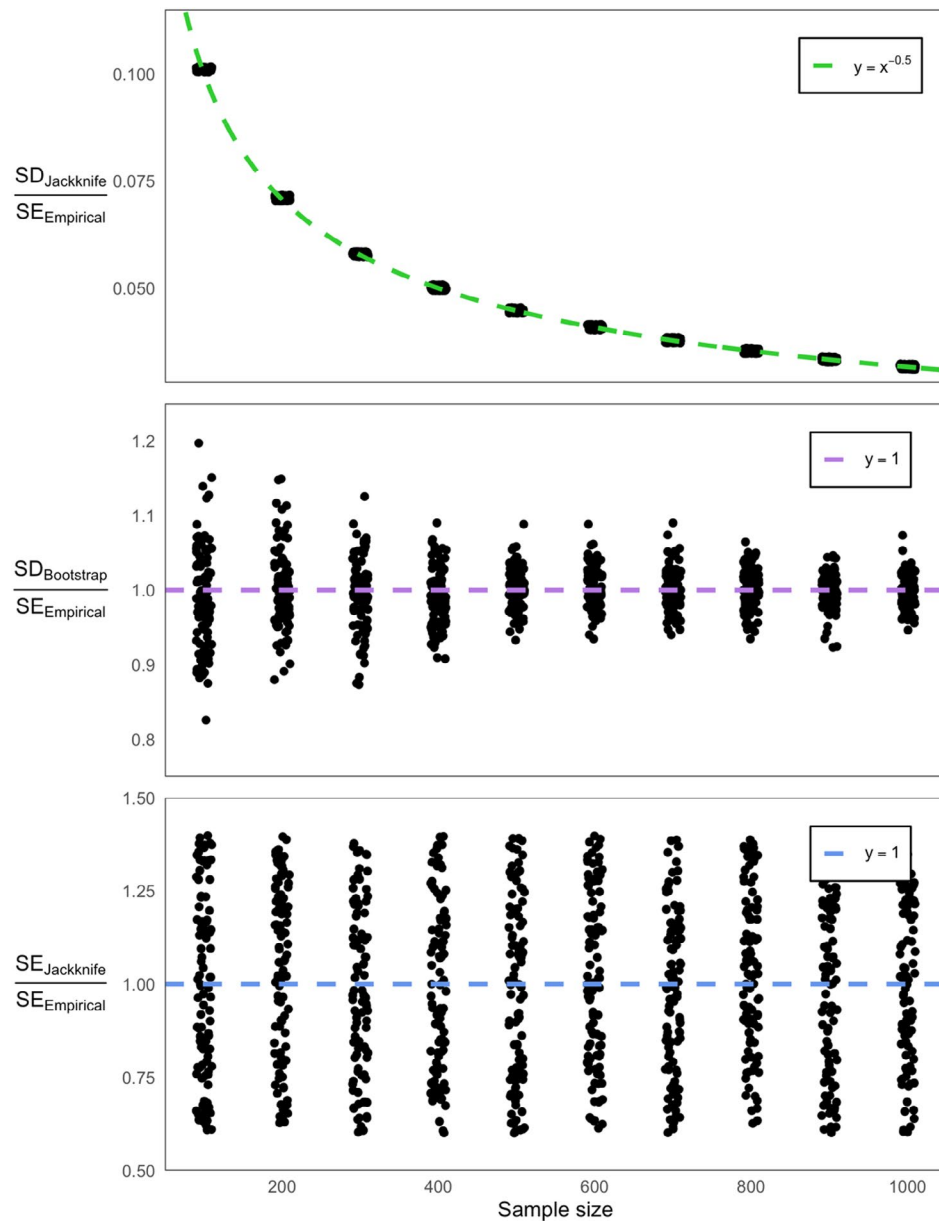


FIGURE 1 | Comparison of the jackknife and the bootstrap for estimating the precision of a summary statistic, in this case the mean. For all panels, samples were simulated from a normal distribution with mean and standard deviation equal to 10 and 1. For each random sample of size n (x-axis), I calculated estimators of the standard error based on the standard deviation of the jackknife (top), the standard deviation of the bootstrap (middle), and the standard error of the jackknife (bottom) following the calculation in McIntosh (2016) and then divided each by the empirical standard error to estimate bias.

produces estimates of the standard error that tend towards the true value with a precision that is insensitive to the sample size (Figure 1, bottom).

Many of the papers citing Do et al. (2015) for the use of the R_{XY} statistic also use the jackknife to estimate the statistical significance of such differences. Notably, Grossen et al. (2020) used this procedure to analyse R_{XY} values for mutations of different predicted impacts on protein function in bottlenecked populations of domestic alpine ibex (*Capra ibex*). To produce pseudo-values, they used whole chromosomes as blocks ($n = 29$ for the domestic goat assembly used; Dong et al. 2013) and performed tests of Tukey's Honest Significant Difference (HSD) comparing the jackknife pseudo-values of R_{XY} among different mutation

classes. They concluded that bottlenecked populations, relative to stable ones, purged highly deleterious mutations but accumulated weakly deleterious ones given that R_{XY} values calculated from pseudo-values were significantly lower or higher respectively than a non-coding mutation class. The use of jackknife pseudo-values as if they represented independent samples from the population violates the Tukey's HSD test assumptions and fails to recognise the distinct properties of the jackknife as a statistical procedure. This is also, as far as I am aware, the first instance of jackknife pseudo-values themselves being presented graphically (fig. 2e in Grossen et al. 2020). Since this study was published, numerous other studies have performed jackknife procedures for R_{XY} and presented boxplots of the pseudo-values or their mean $\pm 1-2$ standard deviations as if they represented the

sample variance of R_{XY} (e.g., Chen, Zhang, et al. 2025; Dussex, Kurland, et al. 2023; Furni et al. 2025; Lan et al. 2025). The result is a set of exceedingly narrow distributions/confidence intervals which are then taken as evidence for the enrichment/depletion of deleterious mutations in threatened populations. It is not my aim here to dispute the findings of the above cited studies which all presented multiple lines of evidence in support of their conclusions. However, as I have shown above with reference to the literature and my own simulations, the jackknife requires particular treatment to produce accurate estimates of uncertainty.

5 | A Worked Example

Here, I reanalyse an empirical dataset (Stuart et al. 2025) to demonstrate how (1) the plotting of jackknife pseudo-values and (2) the incorrect use of jackknife pseudo-values in a null hypothesis test can both lead to false confidence in group differences. The data are drawn from captive and wild populations of *Dryococcus australis*, the Lord Howe Island stick insect. Allele frequencies for mutations in both populations were estimated with low-coverage whole genome sequence data, and fitness effects of mutations were approximated by estimating their predicted impact on protein function with gene models from Stuart et al. (2023) using snpEff (Cingolani et al. 2012); a similar procedure to Grossen et al. (2020). Mutations are grouped into high, moderate, and low impact categories, with an additional category for non-coding mutations in intergenic regions. Impacts on protein function are assumed to have unconditionally negative fitness consequences. The question I am interested in is: “has the strong bottleneck at the onset of the captive breeding program resulted in the purging or enrichment of deleterious mutations?” I answered this by calculating R_{XY} for the four different mutation categories in the captive individuals, using the wild individuals as the reference population. I standardised these values using the intergenic mutations, such that if a mutation class is experiencing the same average effect of selection as intergenic mutations, then $R_{XY} = 1$. The whole-genome values strongly suggest that highly impactful mutations were purged ($R_{XY} = 0.73$), likely because of increased homozygosity in captivity exposing them to selection, and there may also be some accumulation of mutations in the low ($R_{XY} = 1.01$) and moderate ($R_{XY} = 1.01$) categories. This is almost identical to the patterns seen by Grossen et al. (2020).

To estimate intervals around these values, I divided the genome into 100 contiguous blocks of roughly equal numbers of loci (uneven chromosome/scaffold length means that exactly equal blocks cannot be defined). I then resampled these blocks using a jackknife procedure and calculated R_{XY} pseudo-values. I did the same for the 16 autosomes of the *D. australis* genome, using each chromosomal scaffold as a block. I used the Busing et al. (1999) equations to calculate means and standard errors. Finally, I used the 100 contiguous blocks to calculate a bootstrap estimate of the standard error, using standard equations for the weighted mean and standard deviation. I generated resampled datasets using 100, 500, and 1000 bootstraps to assess the stability of standard errors.

The distributions of R_{XY} from the equal size block jackknife are exceedingly narrow (Figure 2, top-left), but slightly less narrow

for whole chromosomes (Figure 2, bottom-left). For equal length blocks, the distributions of R_{XY} for both low and moderate impact mutations only slightly overlap 0. For the whole chromosome jackknife, this is also true for the moderate impact mutations but not the low impact mutations. Mean R_{XY} for low and moderate impact mutations was clearly less than one standard error away from 1. This was true for both the equal length block and chromosome jackknife (Figure 2, top-right, bottom-right). In all cases, neither the distributions nor the mean ± 1 standard error overlap 1 for the high impact mutations. This is strong evidence that the captive *D. australis* population has purged highly impactful mutations. The bootstrap estimates of the standard error reflect those of the block jackknife with the Busing et al. (1999) equations (Figure 3). Both mean R_{XY} and the size of the confidence intervals varied slightly with the number of bootstrap replicates, although the magnitude of these fluctuations was minor (Figure 3).

The calculation of standard errors (for the bootstrap and the jackknife) suggests that only high impact mutations have been affected by the captivity bottleneck, but based on the raw distributions of R_{XY} from jackknife replicates (Figure 2, left), it seems plausible that the captive population has simultaneously accumulated moderate and/or low impact mutations. How do these apparent visual differences stand up to a hypothesis test?

I performed two variations on a t -test to test the alternative hypothesis that R_{XY} for each mutation class was different from the neutral expectation. In some datasets, the chromosome number may be small, and so I deemed a t -test a more realistic test than a Z -test for these cases where the degrees of freedom would be low. In all cases, I used a two-sided test. First, I used the jackknife pseudo-values of R_{XY} as samples in a paired t -test. I compared each protein-coding mutation class to intergenic mutations. I can use a paired t -test because each respective pseudo-value for the two classes being compared was calculated using the same jackknife replicate. Next, I used the means and standard errors for both the jackknife and the bootstrap to test whether the standardised R_{XY} for each category was different to 1 in a single sample t -test.

The t -tests of equal sized block jackknife pseudo-values were uniformly significant, suggesting a depletion of highly impactful mutations and an enrichment of low and moderate impact mutations in captivity relative to intergenic mutations (Table 1). High and moderate impact mutations were significantly different from intergenic mutations when using chromosome jackknife pseudo-values (Table 1). In all cases, t values were much greater than when using the means and standard errors, regardless of whether the test was significant. This underscores the fact that using jackknife pseudo-values as samples will overestimate the significance of group differences, as suggested by the visual comparison (Figure 2). Only the high impact mutation class was significantly different from 1 when using the means and standard errors to perform the t -test (Table 2). Block and chromosome jackknife and bootstrap test statistics were more similar when using means and standard errors (Table 2) than when using the pseudo-values (Table 1), suggesting some of the bias produced by uneven block size and different block numbers can be corrected for. The number of bootstrap replicates used did not impact the test results, although there was a slight trend

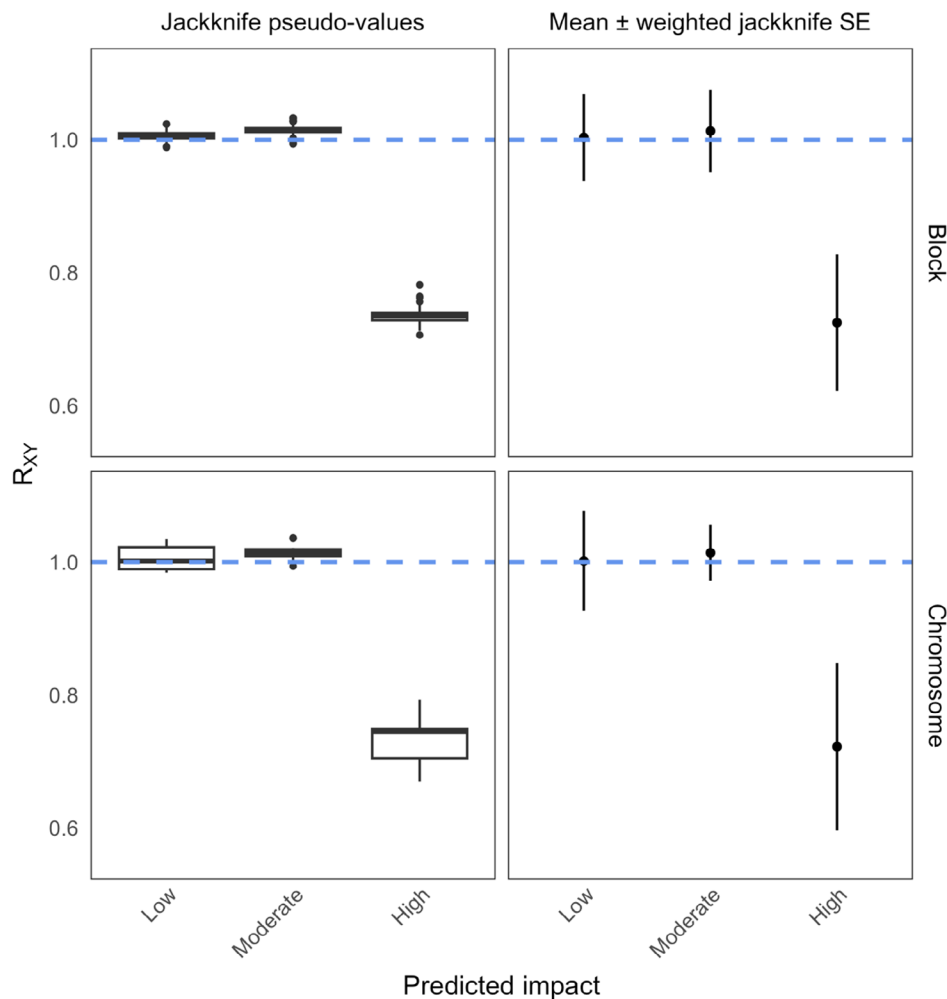


FIGURE 2 | Comparison of methods for estimating and visualising variance in R_{XY} . Left column: Boxplots showing the distribution of R_{XY} pseudo-values from each jackknife replicate. Right column: Mean $R_{XY} \pm 1$ standard error using the equations formulated for a weighted block jackknife. Top row: Jackknifing performed over 100 roughly equally sized contiguous blocks across the *D. australis* autosomal genome. Bottom row: Jackknifing performed using the 16 *D. australis* autosomes as blocks. The hashed blue line indicates $R_{XY} = 1$, the neutral expectation.

towards 0 in the t values for all impact classes with more bootstraps (Table 2).

It is worth mentioning that the small sample size of chromosomes in the *D. australis* genome (16 autosomes) makes for both an underpowered t -test and a less egregious underestimation of the sample standard error: a greater number of observations produces downward bias in the variance of jackknife pseudo-values (Figure 1). Therefore, the more chromosomes an organism has, the more downwardly biased a chromosome block jackknife procedure will be unless the correct equations are used. However, it is likely that this bias also occurs in the other direction: the fewer chromosomes an organism has, the more likely there is to be upward bias in the size of confidence intervals.

6 | Conclusions and Recommendations

I have demonstrated (rather, redemonstrated), using simple simulations and empirical data, that the jackknife can be used to estimate the sampling variance of genomic statistics and therefore

can be used to test for group differences. Specific equations (Busing et al. 1999) for the standard error and the mean are crucial for this. My hope is that readers will be more attuned to the subtleties of the jackknife as a variance estimation procedure, one which is convenient and robust (if sometimes conservative) when the pseudo-values are treated correctly. Alternatively, the bootstrap, an equally convenient and more robust procedure, stands ready for use without the need for specialised equations (except for the weighting for uneven block size). Historical issues regarding the computational tractability of the bootstrap become less concerning as the speed of computation grows. Researchers should not shy away from the bootstrap just because it may take more time than expected.

I put forward the following recommendations regarding the use of resampling procedures to test for group differences in genomic statistics:

1. Divide the genome into contiguous blocks with roughly equal numbers of loci, as in Do et al. (2015) when resampling loci in genomic datasets. Using chromosomes as blocks is convenient but inter/intraspecific variation in

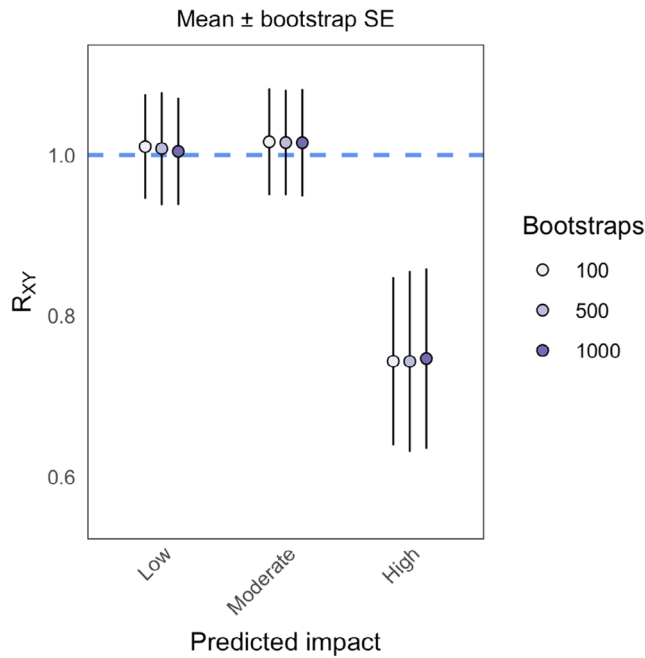


FIGURE 3 | Results of block bootstrapping for estimating variance in R_{XY} . Points represent weighted mean R_{XY} for each mutation impact category, error bars represent one standard error (calculated as the weighted standard deviation of R_{XY} across bootstrap pseudo-values), and colours denote the number of bootstrap resamples taken. The hashed blue line indicates $R_{XY}=1$, the neutral expectation.

TABLE 1 | Results of paired two-sided t -tests testing the alternative hypothesis that R_{XY} (captive individuals relative to wild individuals) for three protein-coding mutation types is different to intergenic mutations, indicating a relative enrichment ($t>0$) or depletion ($t<0$). Samples of pseudo-values were generated by a block jackknife, where the blocks were either 100 contiguous segments of chromosome containing roughly equal numbers of loci throughout the *D. australis* autosomes or the 16 autosomal scaffolds themselves. Bolded values indicate significant tests at $\alpha=0.05$.

Variant effect	Mean difference	SE difference	t	p
Block ($n=100$)				
Low	0.005	0.001	8.760	0.000
Moderate	0.012	0.001	21.411	0.000
High	-0.209	0.001	-235.422	0.000
Chromosome ($n=16$)				
Low	0.005	0.004	1.355	0.196
Moderate	0.012	0.002	5.007	0.000
High	-0.209	0.007	-29.12	0.000

chromosome number and size may affect the relative performance of the block jackknife.

2. Perform a bootstrap over the blocks and use the weighted mean and weighted variance to estimate standard errors for statistical testing.

TABLE 2 | Results of two-sided single-sample t -tests testing the alternative hypothesis that standardised R_{XY} (captive individuals relative to wild individuals, standardised using intergenic mutations) for a given protein-coding mutation class is significantly different from 1, indicating a relative enrichment ($t>0$) or depletion ($t<0$). Means and standard errors were calculated using a block jackknife/ bootstrap procedure. The blocks were either 100 contiguous segments of chromosome ("Block") containing roughly equal numbers of loci throughout the *D. australis* autosomes or the 16 autosomal scaffolds themselves ("Chromosome"). For the jackknife, equations from Busing et al. (1999) were used to calculate means and standard errors. For the bootstrap, the weighted mean and weighted variance were used.

Variant effect	Mean	SE	t	p
Block Jackknife ($n=100$)				
Low	1.003	0.065	0.056	0.956
Moderate	1.013	0.062	0.215	0.831
High	0.725	0.103	-2.679	0.009
Chromosome Jackknife ($n=16$)				
Low	1.002	0.075	0.026	0.980
Moderate	1.014	0.042	0.336	0.742
High	0.722	0.126	-2.200	0.044
Block Bootstrap ($n=100$)				
Low	1.011	0.065	0.167	0.868
Moderate	1.017	0.066	0.249	0.804
High	0.744	0.105	-2.443	0.016
Block Bootstrap ($n=500$)				
Low	1.008	0.070	0.116	0.908
Moderate	1.016	0.065	0.239	0.811
High	0.743	0.112	-2.284	0.023
Block Bootstrap ($n=1000$)				
Low	1.005	0.067	0.069	0.945
Moderate	1.016	0.067	0.232	0.816
High	0.747	0.112	-2.250	0.025

3. If using the block jackknife instead, use the equations presented above (Busing et al. 1999) to calculate means and standard errors.
4. Proceed normally through data visualisation and statistical testing using the means and standard errors. Do not use the jackknife pseudo-values; these will almost always artificially inflate precision and overestimate group differences.

Acknowledgements

I would like to thank Drs Sam Cox and Tom Reed who provided comments on an initial draft of the manuscript. I also extend my thanks to two anonymous reviewers, who provided comments which improved the quality of the manuscript, and the handling editor, Pr Joanna

Freeland, who provided helpful advice on the structure of the manuscript. Generative AI (ChatGPT-4) was used to assist debugging the R code used to define the boundaries of genomic blocks.

Funding

This work was supported by Grant-Aid Agreement no. PDOC/23/04/01, administered by the Marine Institute *Foras na mara* and funded under the Marine Research Programme by the Government of Ireland.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

No new data were generated in the production of this manuscript. All code and allele frequency data required to conduct the analyses and reproduce the figures herein are available at https://github.com/OliverPStuart/2025_Jackknife_Review/. The full code required to process the original raw sequence data, calculate allele frequencies, and estimate the effects of mutations on protein function is available at https://github.com/OliverPStuart/Daustalis_Purging_Manuscript/.

References

- Busing, F. M. T. A., E. Meijer, and R. V. D. Leeden. 1999. "Delete-m Jackknife for Unequal m." *Statistics and Computing* 9, no. 1: 3–8. <https://doi.org/10.1023/A:1008800423698>.
- Chen, Y., X. Feng, K. Reid, C. Zhang, A. Löytynoja, and J. Merilä. 2025. "Dynamics of Deleterious Mutations and Purifying Selection in Small Population Isolates." *Molecular Biology and Evolution* 42, no. 7: msaf110. <https://doi.org/10.1093/molbev/msaf110>.
- Chen, Y., S. Zhang, W. Zang, et al. 2025. "Balancing Selection of Climate Adaptive Loci Underlies the Success of Introduction of Eurasian Tree Sparrows." *Science Advances* 11, no. 47: eadw3797. <https://doi.org/10.1126/sciadv.adw3797>.
- Cingolani, P., A. Platts, L. L. Wang, et al. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff." *Fly* 6, no. 2: 80–92. <https://doi.org/10.4161/fly.19695>.
- Do, C., R. S. Waples, D. Peel, G. M. Macbeth, B. J. Tillett, and J. R. Ovenden. 2014. "NeEstimator v2: Re-Implementation of Software for the Estimation of Contemporary Effective Population Size (N_e) From Genetic Data." *Molecular Ecology Resources* 14, no. 1: 209–214. <https://doi.org/10.1111/1755-0998.12157>.
- Do, R., D. Balick, H. Li, I. Adzhubei, S. Sunyaev, and D. Reich. 2015. "No Evidence That Selection Has Been Less Effective at Removing Deleterious Mutations in Europeans Than in Africans." *Nature Genetics* 47, no. 2: 126–131. <https://doi.org/10.1038/ng.3186>.
- Dong, Y., M. Xie, Y. Jiang, et al. 2013. "Sequencing and Automated Whole-Genome Optical Mapping of the Genome of a Domestic Goat (*Capra hircus*)." *Nature Biotechnology* 31, no. 2: 135–141. <https://doi.org/10.1038/nbt.2478>.
- Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin. 2011. "Testing for Ancient Admixture Between Closely Related Populations." *Molecular Biology and Evolution* 28, no. 8: 2239–2252. <https://doi.org/10.1093/molbev/msr048>.
- Dussex, N., S. Kurland, R.-A. Olsen, et al. 2023. "Range-Wide and Temporal Genomic Analyses Reveal the Consequences of Near-Extinction in Swedish Moose." *Communications Biology* 6, no. 1: 1035. <https://doi.org/10.1038/s42003-023-05385-x>.
- Dussex, N., H. E. Morales, C. Grossen, L. Dalén, and C. van Oosterhout. 2023. "Purging and Accumulation of Genetic Load in Conservation." *Trends in Ecology & Evolution* 38, no. 10: 961–969. <https://doi.org/10.1016/j.tree.2023.05.008>.

- Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7, no. 1: 1–26. <https://doi.org/10.1214/aos/1176344552>.
- Furni, F., M. Bérubé, M. Suárez-Menéndez, et al. 2025. "Humpback Whale Genomes Reflect the Increased Efficiency of Commercial Whaling." *Science Advances* 11, no. 51: eady7091. <https://doi.org/10.1126/sciadv.ady7091>.
- Green, R. E., J. Krause, A. W. Briggs, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science* 328, no. 5979: 710–722. <https://doi.org/10.1126/science.1188021>.
- Grossen, C., F. Guillaume, L. F. Keller, and D. Croll. 2020. "Purging of Highly Deleterious Mutations Through Severe Bottlenecks in Alpine Ibex." *Nature Communications* 11, no. 1: 1–12. <https://doi.org/10.1038/s41467-020-14803-1>.
- Hill, W. G. 1981. "Estimation of Effective Population Size From Data on Linkage Disequilibrium." *Genetics Research* 38, no. 3: 209–216. <https://doi.org/10.1017/S0016672300020553>.
- Jones, A. T., J. R. Ovenden, and Y.-G. Wang. 2016. "Improved Confidence Intervals for the Linkage Disequilibrium Method for Estimating Effective Population Size." *Heredity* 117, no. 4: 217–223. <https://doi.org/10.1038/hdy.2016.19>.
- Korneliussen, T. S., A. Albrechtsen, and R. Nielsen. 2014. "ANGSD: Analysis of Next Generation Sequencing Data." *BMC Bioinformatics* 15, no. 1: 356. <https://doi.org/10.1186/s12859-014-0356-4>.
- Lan, T., H. Li, B. Liu, et al. 2025. "Revealing extensive inbreeding and less efficient purging of deleterious mutations in wild Amur tigers in China." *Journal of Genetics and Genomics* 52, no. 5: 641–649. <https://doi.org/10.1016/j.jgg.2024.12.004>.
- Malécot, G. 1948. *Les Mathématiques de l'Hérédité*. Masson.
- Mathur, S., J. M. Tomeček, L. A. Tarango-Arámbula, R. M. Perez, and J. A. DeWoody. 2022. "An Evolutionary Perspective on Genetic Load in Small, Isolated Populations as Informed by Whole Genome Resequencing and Forward-Time Simulations." *Evolution* 77: qpac061. <https://doi.org/10.1093/evolut/qpac061>.
- McIntosh, A. 2016. "The Jackknife Estimation Method." arXiv: 1606.00497. <https://doi.org/10.48550/arXiv.1606.00497>.
- Miller, R. G. 1974. "The Jackknife—A Review." *Biometrika* 61, no. 1: 1–15. <https://doi.org/10.2307/2334280>.
- Mueller, L. D. 1979. "A Comparison of Two Methods for Making Statistical Inferences on Nei's Measure of Genetic Distance." *Biometrics* 35, no. 4: 757–763. <https://doi.org/10.2307/2530107>.
- Nei, M. 1972. "Genetic Distance Between Populations." *American Naturalist* 106, no. 949: 283–292.
- Pattengale, N. D., M. Alipour, O. R. P. Bininda-Emonds, B. M. E. Moret, and A. Stamatakis. 2010. "How Many Bootstrap Replicates Are Necessary?" *Journal of Computational Biology* 17, no. 3: 337–354. <https://doi.org/10.1089/cmb.2009.0179>.
- Patterson, N., P. Moorjani, Y. Luo, et al. 2012. "Ancient Admixture in Human History." *Genetics* 192, no. 3: 1065–1093. <https://doi.org/10.1534/genetics.112.145037>.
- Quenouille, M. H. 1956. "Notes on Bias in Estimation." *Biometrika* 43, no. 3/4: 353–360. <https://doi.org/10.2307/2332914>.
- Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. 2009. "Reconstructing Indian Population History." *Nature* 461, no. 7263: 489–494. <https://doi.org/10.1038/nature08365>.
- Reynolds, J., B. S. Weir, and C. C. Cockerham. 1983. "Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance." *Genetics* 105, no. 3: 767–779. <https://doi.org/10.1093/genetics/105.3.767>.
- Robinson, J., C. C. Kyriazis, S. C. Yuan, and K. E. Lohmueller. 2023. "Deleterious Variation in Natural Populations and Implications for

Conservation Genetics.” *Annual Review of Animal Biosciences* 11: 93–114. <https://doi.org/10.1146/annurev-animal-080522-093311>.

Rosenzweig, B. K., J. B. Pease, N. J. Besansky, and M. W. Hahn. 2016. “Powerful Methods for Detecting Introgressed Regions From Population Genomic Data.” *Molecular Ecology* 25, no. 11: 2387–2397. <https://doi.org/10.1111/mec.13610>.

Stuart, O. P., R. Cleave, M. J. L. Magrath, and A. S. Mikheyev. 2023. “Genome of the Lord Howe Island Stick Insect Reveals a Highly Conserved Phasmid X Chromosome.” *Genome Biology and Evolution* 15, no. 6: evad104. <https://doi.org/10.1093/gbe/evad104>.

Stuart, O. P., R. Cleave, K. Pearce, M. J. L. Magrath, and A. S. Mikheyev. 2025. “Purging of Highly Deleterious Mutations Through an Extreme Bottleneck.” *Molecular Biology and Evolution* 42, no. 4: msaf079. <https://doi.org/10.1093/molbev/msaf079>.

Thorburn, D. 1977. “On the Asymptotic Normality of the Jackknife.” *Scandinavian Journal of Statistics* 4, no. 3: 113–118.

Tukey, J. 1958. “Bias and Confidence in Not-Quite Large Samples.” *Annals of Mathematical Statistics* 29, no. 2: 614.

van Dongen, S., and T. Backeljau. 1995. “One- and Two-Sample Tests for Single-Locus Inbreeding Coefficients Using the Bootstrap.” *Heredity* 74, no. 2: 129–135. <https://doi.org/10.1038/hdy.1995.19>.

van Oosterhout, C. 2020. “Mutation Load Is the Spectre of Species Conservation.” *Nature Ecology & Evolution* 4, no. 8: 1004–1006. <https://doi.org/10.1038/s41559-020-1204-8>.

von Seth, J., T. van der Valk, E. Lord, et al. 2022. “Genomic Trajectories of a Near-Extinction Event in the Chatham Island Black Robin.” *BMC Genomics* 23, no. 1: 747. <https://doi.org/10.1186/s12864-022-08963-1>.

Waples, R. S. 2024. “Practical Application of the Linkage Disequilibrium Method for Estimating Contemporary Effective Population Size: A Review.” *Molecular Ecology Resources* 24, no. 1: e13879. <https://doi.org/10.1111/1755-0998.13879>.

Waples, R. S., and C. Do. 2008. “Ldne: A Program for Estimating Effective Population Size From Data on Linkage Disequilibrium.” *Molecular Ecology Resources* 8, no. 4: 753–756. <https://doi.org/10.1111/j.1755-0998.2007.02061.x>.

Weir, B. S. 1996. *Genetic Data Analysis II*. Sinauer Associates Inc.

Weir, B. S., and C. C. Cockerham. 1984. “Estimating F-Statistics for the Analysis of Population Structure.” *Evolution* 38, no. 6: 1358–1370. <https://doi.org/10.2307/2408641>.

Wright, S. 1943. “Isolation by Distance.” *Genetics* 28, no. 2: 114–138. <https://doi.org/10.1093/genetics/28.2.114>.

Wright, S. 1949. “The Genetical Structure of Populations.” *Annals of Eugenics* 15, no. 1: 323–354. <https://doi.org/10.1111/j.1469-1809.1949.tb02451.x>.

Xue, Y., J. Prado-Martinez, P. H. Sudmant, et al. 2015. “Mountain Gorilla Genomes Reveal the Impact of Long-Term Population Decline and Inbreeding.” *Science* 348, no. 6231: 242–245. <https://doi.org/10.1126/science.aaa3952>.