A simple procedure to detect, test for the presence of stuttering, and cure stuttered data with spreadsheet programs: application to parasites and vectors

Thierry de Meeûs, Intertryp, IRD, Cirad, Univ Montpellier, Montpellier, France, thierry.demeeus@ird.fr.
Camille Noûs, Cogitamus laboratory, France, camille.nous@cogitamus.fr.

Author for correspondence: thierry.demeeus@ird.fr

Running title: Stuttering detection and cure

Keywords: Microsatellite markers; Short tandem repeats (STRs); Simple sequence repeats (SSRs); Polymerase chain reaction (PCR); Genotyping errors.

Commenté [FH1]: useful to mention it ? I would delete it

Code de champ modifié

Code de champ modifié

Abstract

Microsatellite are powerful markers for empirical population genetics, but may be affected by amplification problems like stuttering that produces heterozygote deficits between alleles with one repeat difference. In this paper, we present a simple procedure that aims at detecting stuttering for each locus overall subsamples and only requires the use of a spreadsheet interactive application on any operating system. We compare the performances of this procedure with the one of MicroChecker on simulations of dioecious pangamic populations, monoecious selfing populations and clonal populations with or without stuttering, and on real data of vectors and parasites. We also propose a cure for loci affected and compare the results with those expected without stuttering. In sexual populations (dioecious or selfers), the new procedure appeared more than three times more efficient than MicroChecker. Cure was able to restore Wright's F_{IS} of stuttered data to the expected value, and particularly so in selfing simulations. In clones, lack of segregation artificially increased false stuttering detection, and only highly significant stuttering tests and loci strongly deviating from others, could be usefully cured, in which case F_{IS} estimate could be much improved. In doubt, and whenever possible, removal of affected and not curable loci may help to shift population genetics parameter estimates towards more reliable values.

Supprimé: still useful

Supprimé: This is particularly true for non-model and small organisms as parasites and vectors.

Supprimé: dioecious

Supprimé: detection works well and

Supprimé: cure

Supprimé: improves

Supprimé: parameter estimates but not perfectly so. In selfers, detection and cure work well, providing other confounding factors as null alleles do not interfere.

Supprimé: can

Supprimé: *parameter estimates*

Mis en forme : Police :Italique

Supprimé: may

Supprimé: Cures are kept only if parameter estimates are improved as well as the behavior of their variation in relation to other factors as number of missing data (null alleles) or specific criteria for clonal populations.

Introduction

Despite the recent democratization of NGS based techniques, microsatellite loci are still very useful markers, in particular for empirical population genetics of non-model and small organisms, as many parasites and/or their vectors, which are difficult (or impossible) to study with direct methods as direct observation (as for birds) or as mark-release-recapture approaches. Sequencing and single nucleotide polymorphism markers (SNPs) still represent expensive alternatives in time, money and expertise, which lies beyond the reach of many laboratories and most of the time at the expense of sample sizes. Three decades ago, microsatellite markers were presented as the most powerful genetic markers (Jarne & Lagoda, 1996). However, researchers began to detect the different problems that can arise and developed different kinds of cures. The last kind of detection tools and cures only arose very recently (Wang et al., 2012; De Meeûs, 2018; Manangwa et al., 2019; De Meeûs et al., 2021).

Stuttering is the result of the Taq polymerase slippage during the PCR amplification of the targeted DNA strand. This generates several PCR products that differ from each other by one repeat and can cause difficulties when discriminating between fake and true homozygotes, such as heterozygous individuals for dinucleotide microsatellite allele sequences with a single repeat difference (De Meeûs et al., 2021). Stuttering produces heterozygote deficits as compared to Castle-Weinberg (CW) expected genotypic proportions (Castle, 1903; Weinberg, 1908), also known as Hardy-Weinberg (HW) expectations (please have a glance at (De Meeûs et al., 2021) for an explanation why we prefer using CW instead of HW). This phenomenon is locus specific and the deviation produced, as measured by wright's F_{IS} (Wright, 1965), proportional to the intensity with which each locus is affected.

Today, and to our knowledge, the only procedure to detect stuttering is the one used in MicroChecker (Van Oosterhout et al., 2004). Though it works well enough, it only studies each locus one by one, which is fine because stuttering presence and intensity are expected to be locus specific. Nonetheless, MicroChecker tests for stuttering in each subsample separately, though a global test might be more powerful. Furthermore, MicroChecker was developed under Microsoft® Windows in 2003 (Windows XP), and it begins to display incompatibility issues with most current systems. A simple alternative, which can detect and test for the presence of stuttering at each locus overall subsamples on any platform kind would thus be welcome and timely.

In this paper, we present a very simple procedure that only requires the use of a spreadsheet interactive computer application such as Apache® OpenOffice Calc or

Supprimé: s

Supprimé: , especially so for non-model organisms

Commenté [FH2]: reference to Guichoux et al. 2011
Balloux, F., Lehmann, L., & Meeûs, T. de. (2003). The Population Genetics of Clonal and Partially Clonal Diploids. *Genetics*, *164*(4), 1635–1644.

Supprimé: It is probable that SNPs will experience the same fate. Null alleles are known to exist in SNPs (Vignal et al., 2002) and these markers may display frequent allelic dropouts (Bayerl et al., 2018). Ascertainment bias represent a very serious issue for non-model organisms (Garvin et al., 2010). The number of SNPs needed is at least 200 to compete with microsatellite information (Séré et al., 2017), which raises new problems regarding some testing procedures as neutrality tests or linkage disequilibrium (LD) tests. LD tests are classically undertaken between pairs of loci. Given the very low power of individual LD tests, which is aggravated by false discovery rate (FDR) procedures, such as Benjamini and Yekutieli's (Benjamini & Yekutieli, 2001), required on non-independent tests series, the number of genotyped individuals required to reach a reasonable power might often look prohibitive for most laboratories. This is particularly true for parasitic organisms and their vectors as *Ixodes scapularis* the vector of Lyme disease in North America (De Meeûs et al., 2021), and even more critical regarding the actors of neglected tropical diseases, among which fasciolasis and African trypanosomiasis cause much economic and health issues and represent significant constraint on development (Solano et al., 2010; Hunter, 2014; Correa et al., 2017).

Commenté [FH3]: It could have been interesting to review the different pitfalls, and highlight that stuttering is still an issue. Other biais can be corrected, e.g. null alleles (Chapuis et al. 2007)
Chapuis, M.-P., & Estoup, A. (2007). Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, *24*(3), 621–631. doi: 10.1093/molbev/msl191

Supprimé: the

Supprimé: and robust

Supprimé: .

Supprimé: In the most recent version (Windows 10) running the setup program can be tricky. On some computers, the setup program stops at some percent of the process and the user must end it and run again the setup program as many times as necessary (e.g. six times). We may fear that next versions of Windows will be completely incompatible with this program. For other operating systems, MicroChecker requires some specific emulators to work

Microsoft® Excel. We compare the performances of this new procedure with the one implemented in MicroCheker on simulated data without (null hypothesis) or with (alternative hypothesis) stuttering, in dioecious populations of various sizes with random mating, hermaphrodites with selfing or clonal populations. We also checked how the cure proposed in De Meeûs et al. (2021) of loci with stuttering signature restore the values expected for some parameters. We finally reanalyzed four real data sets on vectors and/or their parasite: the tick *Ixodes scapularis* in North America (De Meeûs et al., 2021); *Glossina palpalis palpalis*, vector of African trypnosomiasis in Côte d'Ivoire (Berté et al., 2019); the snail *Galba truncatula* and the fluke it transmits, *Fasciola hepatica* in France (Correa et al., 2017); and *Trypanosoma brucei gambiense*, the agent of sleeping sickness in Guinea and Côte d'Ivoire (Koffi et al., 2009). On these datasets, we checked if more loci with stuttering could be diagnosed, cured the loci with suspicion of stuttering, following the technique proposed recently (De Meeûs et al., 2021) to verify if some conclusions could be changed.

Supprimé: (

Supprimé: ,

Material and Methods

Simulations

Simulations were undertaken with EASYPOP (v. 2.0.1) (Balloux, 2001). We simulated random mating dioecious populations (pangamy), like what probably occurs in the wild for ticks (De Meeûs et al., 2021), Nematocera flies (Prudhomme et al., 2020), Hemipteran bugs (Gomez-Palacio et al., 2013), or tsetse flies (Berté et al., 2019). We also simulated selfing monoecious populations, as flukes and water snails (Correa et al., 2017). The total size of populations was N_T=10,000 individuals subdivided into either n=100 subpopulations of N=100 individuals, or n=500 and N=20, with an even sex ratio (dioecy) or with selfing rate s=0.3 (monoecy). The model of migration followed an Island model with migration rate m=0.01. We simulated 20 independent loci with a mutation rate of u=0.0001 that followed a mixed model with 70% of mutations following a stepwise mutation model (SMM) and 30% following a KAM model. The maximum possible number of alleles was K=20. Each simulation started with maximum variability and was run for 10,000 generations. At the end of each simulation, 20 individuals (10 males and 10 females in dioecious populations), when N≥20, were randomly sampled in 10 subpopulations. As can

Supprimé: s

be seen with the real datasets reanalyzed in the present work, such a sampling design approximately represents what is classically obtained for most parasites or vectors studies.

Simulations of monoecious populations with 30% of selfing allowed checking the interaction of stuttering detection in inbred populations with a high expected heterozygote deficit (here 18%).

A subset of simulations with *n*=100, *N*=100 (same values for other parameters as above) but with 100% clonal propagation, was finally undertaken, to fit with diploid clonal pathogens as trypanosomes (Koffi et al., 2009) or yeasts as *Candida albicans* (Nébavi et al., 2006).

Each parameter set was replicated 10 times.

In the supplementary material, we provided an example with the results files of the first simulation, with the root name "TestStutterDioeciousNoStutter-n1000N100-1" and extensions "txt", "equ", "dat", and "gen", for the parameters used, the statistics along the simulation (all generations) and the resulting data files in Fstat and genepop formats, respectively.

Generating stuttering

Data were analyzed with Fstat 2.9.4 (Goudet, 2003) updated from (Goudet, 1995) to get information on the identity of alleles kept at the end of simulations. For each simulation, input files were imported into a spreadsheet keeping each allele of each locus separated in a single column. We arbitrarily considered that 10% of possible alleles affected by stuttering was enough. This means that two alleles (out of 20 possible ones) needed to be recoded for stuttering. Because of genetic drift, not all the 20 possible alleles were present at the end of each simulation. For each locus, among the allele still present, only the first two alleles separated by a single repeat were concerned. For each individual carrying one of these alleles as the second allele, if different by a single repeat from the first allele, the second allele was recoded as identical to the first one. Let us assume that the first two alleles separated by a single repeat were, for instance, allele 5 and 6 for the first locus (Locus1). If allele 5 was in cell B2 and allele 6 in cell C2 in the spreadsheet, the command for generating stuttering in a cell from a free zone (e.g. the first free column after the last column of the data) would be:

=B2, for the first allele (no change), and

=IF(ABS(B2-C2)=1,IF(OR(B2=5,B2=6),B2,C2),C2), for the second allele of that locus. This way, individuals 4/5, 5/6, or 6/7 are recoded as homozygotes for the first allele.

This command was then copied and pasted to transform the whole locus.

Supprimé: identitiy

Commenté [FH4]: output files from EasyPop , no ?

Supprimé: , which

Commenté [FH5]: add a sentence here : For mimicking stuttering we proceeded as follow :

Supprimé: in English or¶ =SI(ABS(B2-C2)=1;SI(OU(B2=5;B2=6);B2;C2);C2) in French

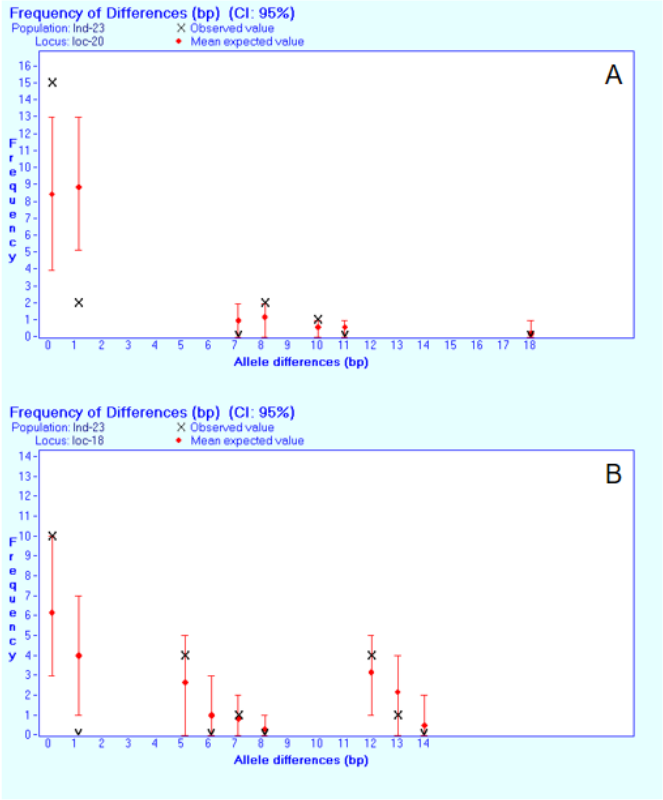
A template, using the first simulation, can be found in the supplementary material files as the spreadsheet file "TestStutterDioecious-n1000N100-1-10%Stuttering.xlsx".

Because of drift, some loci in some subsamples did not display allele with one repeat difference. Such manipulation thus generated data with 0% to 100% of alleles displaying stuttering for all of the 20 loci, but with various intensity from one locus to the other, and from one subsample to the other, as expected in real situations. This also allowed checking the kind of variance stuttering can generate on parameter estimates (see below).

Detection of stuttering and testing with MicroChecker

All datasets (raw and with stuttering) were analyzed with MicroChecker with 10,000 randomizations. All loci were considered as mononucleotidic, as simulated by Easypop. Stuttering was detected when the observed heterozygosity for alleles with one repeat difference was below the 95% confidence interval (95%CI) for random mating expectation. This was observed from the graphic outputs of MicroChecker (Figure 1).

Figure 1: Examples of significant stuttering tests using MicroChecker graphic outputs. Black crosses represent the number of observed genotype of a given class and red diamonds stand for the corresponding values expected under the null hypothesis (random mating). The abscissa are the different genotypic classes in terms of size differences between the two alleles within an individual (e.g. 0 stands for homozygous genotypes). 1-A: an example with a significant homozygous excess and a significant deficit of heterozygotes with one repeat size difference between the two alleles. 1-B: an example where only the deficit of heterozygotes with one repeat size difference between the two alleles was significant. In the present paper, both situations are considered significant while only the first one is for MicroChecker,



For each locus, we summed the number of times MicroChecker found a significant heterozygote deficit probably due to stuttering over the 10 subsamples. We compared this quantity with the expected 5% under the null hypothesis with a one sided exact binomial test with R (R-Core-Team, 2020) (command "binom.test"). The alternative hypothesis was that there are more than 5% significant tests. This test was repeated 20 times across the different loci. To take into account this repetition of independent tests, we used the Benjamini and Hochberg's (BH) false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995) with R (R-Core-Team, 2020) (command "p.adjust") to identify which tests are really significant (see (De Meeûs et al., 2009)).

Alternative method to detect and test for stuttering

We needed to compute the expected frequency of individuals heterozygous for two alleles with one repeat difference, for each locus over all subsamples. All allele frequencies outputted and sorted by Fstat were copied in a spreadsheet. Let us assume, for instance, that subsample size was in cell B3, that the size of the smallest allele of the first locus of the first subsample was in cell A4 and its frequency in cell B4, and allele size and allele frequency of the second allele was in cells A5 and B5 respectively. Then the expected frequency of individuals heterozygous for two alleles with one repeat difference was obtained by typing the following command in, for instance, cell C4:

=IF(ABS($A4-$A5)=1,2*B4*B5*B$3,0)

Supprimé: in English and¶ =SI(ABS($A4-$A5)=1;2*B4*B5*B$3;0) in French.

As can be seen, for two successive alleles with more than one repeat difference, this expected frequency was set to 0. Please, note that for a dinucleotidic locus the difference in size must be two (e.g. ABS($A4-$A5)=2). For imperfect dinucleotidic loci, the conditional command would be of the form

=IF(OR(ABS($A4-$A5)=1, ABS($A4-$A5)=2),2*B4*B5*B$3,0), to include both the cases of one base difference, which may also generate stuttering, and of two bases (one repeat) difference.

Supprimé: in English and¶ =SI(OU(ABS($A4-$A5)=1; ABS($A4-$A5)=2);2*B4*B5*B$3;0) in French

Now, if the penultimate allele is on line 10 of the spreadsheet, the sum of all expected heterozygotes with one repeat difference for the concerned locus and subsample was obtained by typing the following command in, for instance, cell C12:

=SUM(C4:C10)

Supprimé: in English and¶ =SOMME(C4:C10) in French

Finally, if this sum for the last subsample is in column U, then, the total number of expected heterozygotes with one repeat difference across all subsamples for that locus was obtained with:

=SUM(C12:U12)

Supprimé: in English and¶ =SOMME(C12:U12) in French

Then, we needed to compute the observed frequency of such heterozygotes. For this, we copied the raw data (one allele per column) in a spreadsheet. Let us assume that the first allele of the first locus of the first individual was in cell B2 and the last allele of the last locus of the first individual was in cell AO2. In cell AQ2 we typed:

=IF(ABS(B2-C2)=1,1,0)

Please, note again that for dinucleotidic loci, the difference in size would be two, and it should be one and two for imperfect dinucleotidic loci. We then copied this command in all remaining cells corresponding to the rest of the dataset. In the cell AQ202 (below the last line of the data), to compute the total of observed heterozygous individuals for alleles with one repeat difference, we typed:

=SUM(AQ2:AQ201)

A template, for the first simulation, is available in the spreadsheet file "TestStutterDioeciousNoStutter-n1000N100-1-FstatRes.xlsx".

We then compared the observed and expected frequencies with a one sided exact binomial test with R, the alternative hypothesis being "there are less heterozygote observed with 1 repeat difference than expected". This provided 20 independent *p*-values that we corrected for False Discovery Rate with the Benjamini and Hochberg's procedure.

With selfing, natural homozygosity increase may artificially enhance stuttering detection. Nevertheless, since there is no easy way to estimate precisely the selfing rate in case of stuttering, we did not adapt stuttering detection.

For clonal propagation, full clonality exhibits specific signature regarding genetic diversity, F_{IS} and linkage disequilibrium (De Meeûs & Balloux, 2004, 2005; De Meeûs et al., 2006; Séré et al., 2014). Nevertheless, as shown by the simulations we undertook, the absence of segregation makes it impossible to predict the expected frequency of specific heterozygous classes. This led us to entirely modify stuttering detection in that case. Knowing that the expected proportion of homozygous sites is $Q_I=1/K$ in clonal populations (De Meeûs, 2015), the expected total heterozygosity should be $H_I=(K-1)/K$ (K is the total number of possible alleles). The quantity K is never known, so we considered the total number of alleles observed K_o, as an underestimate, with $H_I'=(K_o-1)/K_o$. Some heterozygote kinds are expected to be more frequent than others. We thus considered frequencies p_i and p_j of alleles i and j, to weight expected values by $2p_ip_j$ and built an "expected" heterozygote frequency between alleles i and j as:

$$H_{expij} = \frac{2p_ip_j\left(\frac{K_o-1}{K_o}\right)}{\left(\sum_{i,j\neq i}^{K_o} 2p_ip_j\right)} = \frac{2p_ip_j\left(\frac{K_o-1}{K_o}\right)}{\left(1-\sum_{i}^{K_o} {p_i}^2\right)}$$

Supprimé: in English or¶ =SI(ABS(B2-C2)=1;1;0) in French

Supprimé: or

Supprimé: in English and¶ =SOMME(AQ2:AQ201) in French

Commenté [FH6]: add reference to Stoeckel & Masson 2014
Stoeckel, S., & Masson, J.-P. (2014). The Exact Distributions of FIS under Partial Asexuality in Small Finite Populations with Mutation. *PLoS ONE*, *9*(1), e85228. doi: 10.1371/journal.pone.0085228

We can see that this way the sum of all H_{expij} indeed reaches H_i'. For each locus, H_{expij} was computed for each heterozygous class with one repeat difference within each subpopulation. We then summed all these expected frequencies over all heterozygote classes with one repeat difference and multiplied it by 20 (subsample size) to obtain the expected number of heterozygotes with one repeat difference for a given subsample. We then summed the results obtained across all the 10 subsamples to obtain the total number of expected heterozygotes with one repeat difference and compared it, for each locus, to the one observed in the simulation. Since $K_o \leq K$, these expectations may be under-estimates of the real expected frequencies. We might thus expect a deficiency in stuttering detection. Alternatively, since drift should favor particular heterozygous classes by chance, we also expect a total lack of other heterozygous classes, which may lead to strongly significant spurious stuttering signatures.

Commenté [FH7]: This looks like a stuttering wording pattern (SSSS) ;-)

Estimation of Fixation indices and linkage disequilibrium

Supprimé: estimation

Wright's *F*-statistics (Wright, 1965) estimated with Weir and Cockerham's estimators (Weir & Cockerham, 1984) were computed. F_{IS} measures inbreeding of individuals relative to inbreeding of their subpopulation and F_{ST} measures inbreeding of subpopulations relative to the total inbreeding. We also computed the 95% confidence intervals of bootstrap over loci of these statistics. These were estimated and computed with Fstat 2.9.4 (Goudet, 2003) updated from Fstat 1.2 (Goudet, 1995).

Amplification problems can increase the variance of *F*-statistic estimation across loci, and this affects more the F_{IS} than the F_{ST} (De Meeûs, 2018). We used the jackknife over loci estimate of the standard error of F_{IS} and F_{ST} (StdrdErrFIS and StdrdErrFST) of Fstat to measure the effect of stuttering on parameter variation across loci. In particular, in case of null alleles, StdrdErrFIS is at least twice StdrdErrFST (De Meeûs, 2018). We thus measured the ratio R_{SE}=StdrdErrFIS/StdrdErrFST.

Supprimé:

Linkage disequilibrium can be favored by allele miscoring (De Meeûs et al., 2021). We thus tested linkage disequilibrium between all pairs of loci with the *G*-based randomization test of Fstat over all subsamples because it is the most powerful for combining tests across subsamples (De Meeûs et al., 2009). The False Discovery Rate for dependent tests series was computed following Benjamini and Yekutieli procedure (Benjamini & Yekutieli, 2001) with R (command p.adjust).

Supprimé: (BY)

Statistical comparisons of method performances

Performance of tests were compared with the Fisher exact test with R-commander package (Fox, 2005; Fox, 2007) for R.

We also undertook generalized linear mixed models with the package lme4 (Bates et al., 2015) of R to explain the number of times a test appeared significant. We used a Poisson distribution with a log link. The models were of the form

NSig~*n*+*N*+Mating+Stuttering+Mating:Stuttering+(1|Rep)

where NSig was the number of loci that outputted a significant stuttering test or the number of locus pairs that appeared in significant LD, *n* is the total number of subpopulations, *N* was the size of subpopulations, Mating was the mating system (either pangamic dioecy or Monoecy with 30% selfing), Stuttering was either 0 (no stuttering) or 10 (10% stuttering), ":" stood for interaction between two variables, and (1|Rep) was the random effect of replicates.

Cured data sets

Stuttering correction was made for loci that appeared with a significant stuttering at the BH level with the new method described in the present paper. We used the rules described in (De Meeûs et al., 2021): for each incriminated locus, all alleles with one repeat difference were pooled together. Each group of pooled alleles contained an allele the frequency of which was at least 0.05. The main principle behind this rule is that rare alleles should keep small weights in the data. Pooling rare alleles together may artificially create a fairly frequent artificial allele, with a strong though artificial weight. Pooling rare alleles with a reasonably frequent one is supposed to attenuate this problem. If no frequent allele was available, then two solutions were chosen. If the sum of the frequencies of these alleles remained below 0.05, these were not pooled. Otherwise, to minimize the impact that these successive alleles may jointly have on the heterozygote deficit, and to avoid pooling rare alleles together, they were pooled with the closest allele with frequency above or equal to 0.05, even if more than one repeats distant from the closest one.

In rare cases, all alleles were one repeat different. To prevent obtaining a monomorphic locus in that case, we pooled alleles two by two, taking care of one of the two alleles pooled displayed a frequency of at least 0.05. In case of uneven number of alleles, the last allele was not left alone and pooled with the previous pair in allele size.

Cured data were reanalyzed and statistics compared with the results expected under the null hypothesis (without stuttering). The efficiency of the correction was checked for each locus in each replicate of each simulation. We retained only the corrected loci for

Commenté [FH8]: awkward sentence. Change to : Each group of pooled alleles contained one allele with a frequency of at least 0.05%.

Supprimé: ,

Supprimé: i

Supprimé: ;

Supprimé: o

which the correction produced a lower F_{IS} as compared to the value obtained with the uncured locus. The average F_{IS} and 95% bootstraps confidence intervals were thus computed on data sets with efficiently cured loci and uncured remaining loci. These values were compared to those obtained under the null hypothesis (no stuttering), and to expected value for Wright's F_{IS}, F_{IS_exp}. For pangamic dioecious populations we used equation 8 from Balloux (2004) (Balloux, 2004): $F_{IS\_exp}=-1/(2N+1)$. For selfing populations, we used the classic $F_{IS\_exp}=s/(2-s)$ (e.g. (De Meeûs et al., 2007) page 213).

Real data sets

Five data sets were reanalyzed: two regarding dioecious species, two regarding monoecious species and the last regarding a clonal species.

The first real data set reanalyzed was on the tick *Ixodes scapularis* the vector of Lyme disease in western USA (De Meeûs et al., 2021). We used the data cured for short allele dominance (SAD) as explained in the originator paper (De Meeûs et al., 2021) but uncured for stuttering.

The second data set concerned the tsetse fly *Glossina palpalis palpalis*, an important vector of sleeping sickness in Côte d'Ivoire (Berté et al., 2019), for which loci X55-3 and pGp23 displaying uncured SAD and locus GPCAG, obviously under selection, were removed. Because some loci were X-linked, only females were kept. Three dinucleotidic loci (pGp20, pGp24, B3) displayed some discrepancies of allele sizes and were marked as mononucleotidic for MicroChecker. For these loci, heterozygotes with single and double nucleotide differences in size were checked for heterozygote deficit due to stuttering.

The third and the fourth data sets concerned the highly selfing snail *Galba truncatula* and its parasite *Fasciola hepatica*, also monoecious but almost panmictic, in France (Correa et al., 2017). For both species, microsatellite profiles did not fit with the expected pure dinucleotidic motives and, again, stuttering was considered between alleles of 1 base and two bases differences in size.

Finally, the clonal species studied was *Trypanosoma brucei gambiense* 1 in Western African foci of sleeping sickness (Koffi et al., 2009), for which one locus (Trbpa1/2), suspected of being under selection, was removed. For this data, we used a derived version of Séré et al superimposition criterion (Séré et al., 2014). In pure clonal populations, the expected value for F_{IS} is $F_{IS\_exp}=-(1-H_S)/H_S$, where H_S is Nei's estimator of local genetic diversity (measured within subsamples) (Nei & Chesser, 1983). This criterion can only be used for sufficiently polymorphic loci with $H_S \geq 0.5$. To express the goodness of

fit of observed F_{IS} towards this value, we designed a superimposition index S_C=|F_{IS}-F_{IS_exp}|/max(|F_{IS}|,| F_{IS_exp}|), where "max" means the maximum value of the two F_{IS}'s absolute values.

Results and discussion

A synthetic view of simulation results, averages detail computations and test tables are available in the supplementary file S1 for sexual simulations and in the supplementary file S2 for clonal simulations.

Detection of stuttering in sexual populations

The results of these analyses are summarized in Table 1. Stuttering detection per locus was weak in general, with 0‰ and 5‰ significant tests for MicroChecker and the alternative methods, respectively, under the null hypothesis (H0: there is no stuttering) in monoecious populations. For Microchecker, the total proportion of significant tests over all loci and subsamples was 2‰ in that case. No test stayed significant after Benjamini and Hochber correction. In populations with 30% selfing, but still under H0 (i.e. no stuttering), these proportions increased to 10% for the alternative method only. It dropped to 2.5‰ after Benjamini and Hochberg correction. For MicroChecker over all loci and subsamples, 2% only appeared significant under H0 with selfing. With stuttering (H1), the number of significant tests reached 7% and 24% in dioecious populations for MicroChecker and alternative methods respectively (5% and 17% respectively with Benjamini and Hochberg). This reached 14% and 47% respectively with 30% selfing (10% and 38% respectively with Benjamini and Hochberg).

Supprimé: The performance of s

Supprimé: BH

Supprimé: BH

Supprimé: BH

Supprimé: BH

Supprimé: In Table 1, we can see that, except under H0 in pangamic dioecious populations, the alternative method appeared much more performant than MicroChecker for the detection of stuttering.¶

Table 1: Statistical comparisons between MicroChecker and the alternative method to detect stuttering under the null hypothesis with no stuttering or with 10% of alleles affected, for different system of mating (dioecious pangamy or 30% of selfing). The *p*-values obtained correspond to a Fisher exact test comparing number of significant (S) and not significant (NS) tests (before Benjamini and Hochberg correction) between the two methods.

Stuttering	Mating	Method	S	NS	*p*-value
0	Pangamy	MicroChecker	0	600	0.2494
		Alternative method	3	597	
	30% Selfing	MicroChecker	0	400	<0.0001
		Alternative method	41	359	
10%	Pangamy	MicroChecker	43	557	<0.0001
		Alternative method	146	454	
	30% Selfing	MicroChecker	56	344	<0.0001
		Alternative method	187	213	

As can be seen from Table 2, selfing significantly increased stuttering detection, even under H0, where it significantly appeared above the 5% threshold (*p*-value<0.0001).

Table 2: Statistical comparisons between dioecious pangamic populations with monoecious populations with 30% selfing to detect stuttering with the alternative method under the null hypothesis with no stuttering or with 10% of alleles affected. The *p*-values obtained correspond to a Fisher exact test comparing number of significant (S) and not significant (NS) tests.

Stuttering	Mating	S	NS	*p*-value
0	Pangamy	3	597	<0.0001
	30% Selfing	41	359	
10%	Pangamy	146	454	<0.0001
	30% Selfing	187	213	

Expectedly, stuttering was much more easily detected in populations with stuttering than under H0 (Table 3).

Commenté [FH9]: Coud you find a way to combine the three tables. It is the same test applied to all contingency tables.

Supprimé: ¶

Supprimé: ,

Supprimé: the power to detect

Table 3: Statistical comparisons of stuttering detection, with the alternative method, between cases with no stuttering or with 10% of alleles affected, in dioecious pangamic populations or in monoecious populations with 30% selfing. The *p*-values obtained correspond to a Fisher exact test comparing number of significant (S) and not significant (NS) tests.

Mating	Stuttering	S	NS	*p*-value
Pangamy	0	3	597	<0.0001
	10%	146	454	
30% Selfing	0	41	359	<0.0001
	10%	187	213	

The stuttering proportion used here (10%) was relatively small, since the realized actual proportion of alleles affected at the end of simulations was in general much lower on average. This also explains why the power of stuttering detection appeared quite small. With higher values, we may expect that the method proposed here will be very accurate, especially in inbred populations (selfers).

Fixation indices and linkage disequilibrium

The results for *F*-statistics are presented in Table 4. With 10% stuttering, we observed a significant heterozygote deficit of 4% in pangamic dioecious populations. With 30% selfing, F_{IS} expectedly grew much more and reached 20%. Here, the difference between 0 and 10% of allele submitted to stuttering was not significant (95% CI overlap).

Table 4: Results obtained for F_{IS}, its 95% confidence interval (95%CI, 5000 bootstraps over loci) and for the ratio of jackknife over loci standard error between F_{IS} and F_{ST} (R_{SE}), for 0 or 10% of stuttering and for pangamic dioecious populations or monoecious populations with 30% selfing.

Stuttering	Mating	F_{IS}	95%CI F_{IS}	R_{SE}
0	Pangamy	0.0002	[-0.0211, 0.0214]	1.0756
	30% selfing	0.1761	[0.1573, 0.1947]	0.9912
10%	Pangamy	0.0408	[0.0086, 0.0768]]	1.8172
	30% selfing	0.2202	[0.1892, 0.2542]	1.6735

Supprimé: performances

Commenté [FH10]: yes but compared to the inital value, the increase is only 4.4 %. Not very different from the 4% in pangamic population.

Supprimé: seemed

The ratio between standard errors of F_{IS} and F_{ST} was 1.04 on average under H0 in 95% CI=[0.95, 1.14] and reached 1.77 in 95%CI=[1.52, 2] with stuttering. It thus increased slightly with stuttering but rarely reached R_{SE}=2 on average, as was observed in case of null alleles (De Meeûs, 2018).

The proportion of locus pairs in significant LD varied between 8% and 53% depending on the population structure and mating system (average 18%). With the Benjamini and Yekutieli False Discovery Rate correction, this varied between 0% (majority of cases) and 15% (average 3%). The effect of stuttering on LD was never significant, whatever the mating system (all *p*-values>0.388). This could be expected since in our simulations, stuttering was not correlated between loci. In real datasets, however, it may occur that stuttering happen in samples with issues (poor preservation, mutations affecting the zone of primers' anchorage, low DNA concentration). In that case, several loci of the same individuals will be affected together, then producing fake significant LDs, as was observed for the tick *I. scapularis* (De Meeûs et al., 2021).

Generalized linear mixed models

The generalized linear mixed models confirmed the results seen above with more accuracy.

For the number of significant tests, the results figure in Table 5. All parameters appeared to display a significant effect that stayed so after BH correction. The most important parameters were stuttering (positive effect), mating system (selfing increases the effect) and their interaction (more effect of stuttering in random mating dioecious populations). Number of subpopulations and subpopulation sizes displayed a rather weak (though significant) negative effect, but this is probably an artefact due to inconsistencies of results as a function of *n* or *N* (see supplementary File S1). For instance, *N*=50, under H0 with random mating, provided the smallest numbers of significant stuttering while *N*=100 provided more significant results than *N*=200. In the same framework, for subpopulation numbers, it was *n*=500 that provided the smallest number of significant tests, followed by 100 and 50. Similar observations can be done for 10% of alleles affected by stuttering and/or in monecious populations with 30% selfing (see supplementary File S1).

Table 5: Summary of the generalized linear mixed model for the number of loci found with a significant stuttering (response variable) with the new alternative multi-subsamples method. Explanatory variables were: *n* (number of subpopulations), *N* (subpopulation size), mating system (dioecious pangamy or monoecy with 30% selfing) and stuttering intensity (0 or 10%). In case of qualitative variables (mating system), the modalities with least positive effects are compared to the one with the most positive effect (not shown in the output of the analysis). ":" stands for the interaction between two variables. Coefficient estimates (Estimate), standard error (SE), the *Z* statistic and its *p*-value are given.

	Estimate	SE	*Z*	*p*-value
(Intercept)	1.2942	0.2787	4.644	<0.0001
n	-0.0024	0.0007	-3.541	0.0004
N	-0.0027	0.0013	-2.043	0.0411
Mating[T.Dioecious]	-2.8398	0.5995	-4.737	<0.0001
Stuttering	0.1518	0.0172	8.8	<0.0001
Mating[T.Dioecious]:Stuttering	0.2367	0.0608	3.892	<0.0001

For LD, results are presented in Table 6. The main effects were hold by *n*, with a positive impact, and mating system with a strong negative impact of random mating as compared to selfing. Subpopulation sizes seemed to play a weaker though significant role. Nevertheless, the pattern of simulations explored introduced a strong collinearity between *n* and *N*. If *n* is removed from the model, then *N* become highly significant with a much stronger (as expected) negative impact (Coefficient of estimate=-0.006). Stuttering did not influence at all the occurrence of significant linkage disequilibrium between pairs of loci. These conclusions did not change when the Benjamini and Hochberg procedure was applied to these series of *p*-values.

Supprimé: BH

Table 6: Summary of the generalized linear mixed model for the number of locus pairs found with a significant linkage disequilibrium. Explanatory variables were: *n* (number of subpopulations), *N* (subpopulation size), mating system (dioecious pangamy or monoecy with 30% selfing) and stuttering intensity (0 or 10%). In case of qualitative variables (mating system), the modalities with least positive effects are compared to the one with the most positive effect (not shown in the output of the analysis). ":" stands for the interaction between two variables. Coefficient estimates (Estimate), standard error (SE), the *Z* statistic and its *p*-value are given.

	Estimate	SE	*Z*	*p*-value
(Intercept)	3.1325	0.1754	17.858	<0.0001
n	0.0040	0.0004	11.164	<0.0001
N	-0.0020	0.0009	-2.15	0.0316
Mating[T.Dioecious]	-0.5683	0.1029	-5.522	<0.0001
Stuttering	-0.0056	0.0104	-0.543	0.5872
Mating[T.Dioecious]:Stuttering	0.0067	0.0134	0.501	0.6162

Clonal populations

The proportion of significant stuttering detection was very large, even under H0: 53% of significant tests, 50% with Benjamini and Hochberg. Under H1, these proportions increased slightly (61 and 57.5 % respectively), but not significantly so (for the uncorrected tests, the Fisher exact test outputted a *p*-value=0.157). Wilcoxon rank sum tests found no significant difference between data without stuttering and data with 10% stuttering for F_{IS} or R_{SE} (*p*-value=0.2481 and *p*-value=0.4698, respectively). The expectation for F_{IS} following Balloux et al (2003)'s Equation 10 (Balloux et al., 2003), for an infinite allele model, and set for full clonality, gives:

$$F_{IS} = \frac{\gamma[q_s - \gamma(q_s - q_d)]}{2N(1-\gamma)[\gamma(q_s - q_d) - 1] - \gamma[q_s - \gamma(q_s - q_d)]}$$

with $\gamma=(1-u)^2$; $q_s=(1-m)^2+m^2/(n-1)$, and $q_d=(1-q_d)/(n-1)$.

With the actual parameters, this gave F_{IS_exp}=-0.3284. The averages of F_{IS} and 95% bootstrap confidence intervals (95%CI) were F_{IS}=-0.3649 in 95%CI=[-0.409, -0.315] without stuttering; and F_{IS}=-0.3443 in 95%CI=[-0.3962, -0.2845], with 10% stuttering. Both confidence intervals largely overlapped and contained F_{IS_exp}.

Clonal populations generated very high proportions of false stuttering detections. This was because without segregation of alleles, only some classes of heterozygotes

Supprimé: signatures
Supprimé: big
Supprimé: BH
Supprimé: These
Supprimé: outputed
Supprimé: were consistent with this expectation:
Déplacé (insertion) [2]

propagate by chance. Heterozygote classes with one repeat difference are quite rare. Indeed, if all 20 alleles were present (which never happened), there would be 19 heterozygote classes with one repeat difference amongst $\binom{20}{2}$=190 possible heterozygous states, hence a proportion of 0.1. Nevertheless, not all allele combinations were kept by drift, which means that in many instances, even if some alleles were one repeat different at the end of simulations, no such heterozygotes were kept by drift. This produced many tests with very small *p*-values<0.0001, but also many with very high *p*-values>0.5: 43 % and 53 % respectively. Put it another way, many simulations ended with no individuals heterozygous for two alleles with one repeat difference, and many others with too many of those (more than 20%): 40 % and 20 % respectively. This also explains why stuttering did not have much impact on the global results, such as F_{IS} estimates. It means that much more than 10% stuttering will be needed to significantly affect parameter estimates in clonal organisms. In such situations, stuttering detection will need being considered with the necessary prudence (see the conclusion) in fully clonal populations. In partial clones, and given the lack of accuracy of the expected number of heterozygotes with one repeat difference, using the panmictic expectations will probably display better performances.

Supprimé: tried out

Commenté [FH11]: I am not convinced, especially because partial but high clonal rate (above 0.9) would lead to a very twisted distribution of FIS values across loci (see for ex fig 3 in Stoeckel & Masson 2014).

Cured data

The alleles that were indeed pooled are highlighted with the same color and can be found in excel files that are contained in the supplementary file S3 (zipped file) "PoolingProtocolCureSupFileS3.zip ".

Supprimé: The strategy used to pool close in size alleles followed what was written in the Material and Methods section.

In dioecious pangamic simulations, cured data did not entirely fix the stuttering problem. Indeed, using the new method for stuttering detection, eight tests (1.3 %) remained significant after BH correction, which is significantly more than the initial absence of significant test (Fisher exact test, *p*-value=0.0076), before stuttering was introduced (H0). In monoecious populations with 30% selfing, two tests remained significant (0.5 %), which is not significantly more than the initial result (one significant test) under H0 (*p*-value=1).

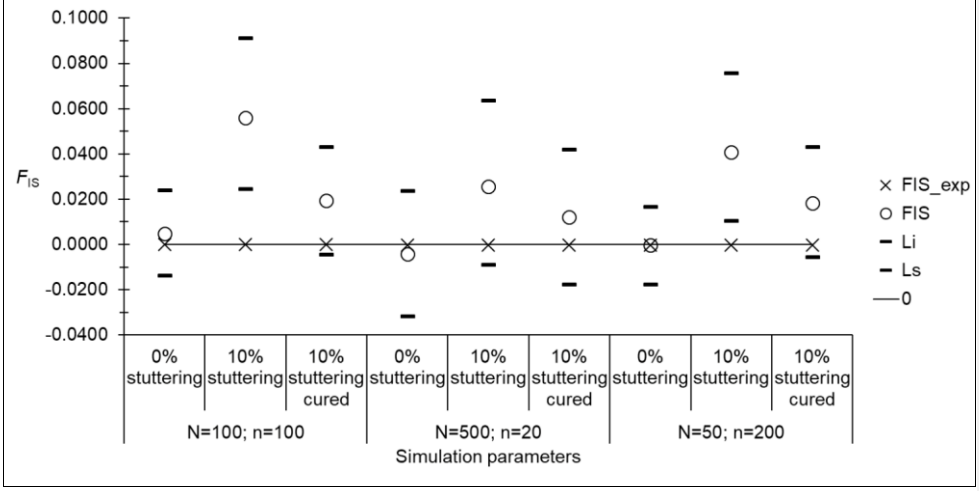
Regarding F_{IS} estimates, in pangamic simulations, the fit between the expected values and the one observed in simulations under the null hypothesis, was not very good and better for larger *n*'s (not shown). We thus preferred the more complex but more accurate equation 11 in (Vitalis, 2002):

Supprimé: bigger

$$F_{IS\_exp} = -\frac{\gamma[q_s - \gamma(q_s - q_d)]}{2N[1 - \gamma(q_s - q_d)] + \gamma[q_s - \gamma(q_s - q_d)]}$$

With this equation, the fit was very good, as can be seen in Figure 2.

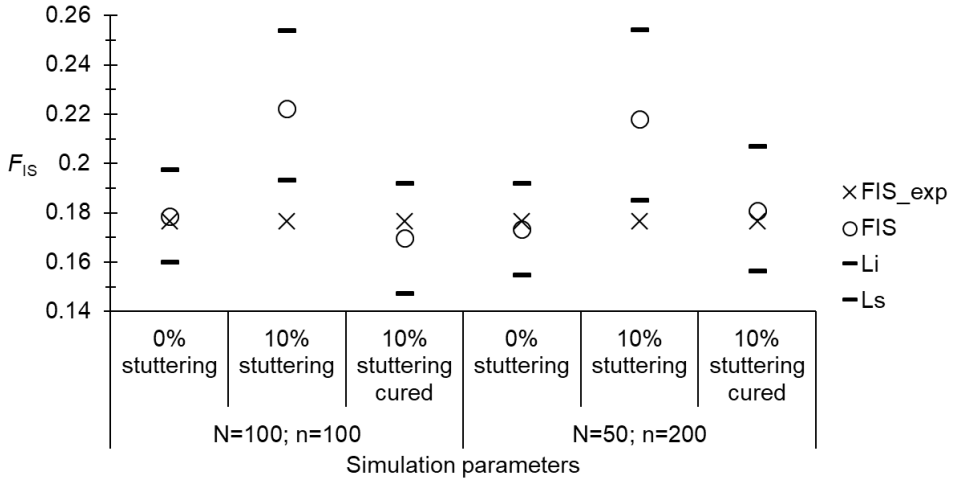
Figure 2: Comparison of Wright's F_{IS} estimates obtained in different simulations in dioecious pangamic populations, without stuttering, with 10% stuttering and with 10% stuttering cured data. The F_{IS_exp}, according to Vitalis (2002), is represented, as are the line of F_{IS}=0, and 95% bootstraps confidence intervals (Li and Ls).



From this figure, we can see that stuttering had a significant impact on F_{IS} estimates, driving it sometimes quite far above the expected value. Cured data, though always providing values higher as compared to H0, always included the expected value in their 95% confidence intervals.

As can be seen from Figure 3, in monoecious simulations with 30% of selfing, the fit is almost perfect between F_{IS_exp} and the value obtained under H0 (no stuttering). Stuttering expectedly significantly increased observed F_{IS}, while cured data presented values that were almost superimposed with expected ones.

Figure 3: Comparison of Wright's F_{IS} estimates obtained in different simulations in monoecious populations with 30% selfing, without stuttering, with 10% stuttering and with 10% stuttering cured data. The $F_{IS\_exp}=s/(2-s)$, is represented, as are the 95% bootstraps confidence intervals (Li and Ls) around observed values.



In conclusion, 10% stuttering significantly increased F_{IS} and the cure used reasonably restored the F_{IS} expected under the null hypothesis, and particularly so in inbred populations (with 30% of selfing).

Real datasets

For *I. scapularis* from Western USA, MicroChecker test only found one locus with significant stuttering out of nine (i.e. 11%), after exact binomial tests and Benjamini and Hochberg correction. Alternatively, the new method developed here found three loci out of nine loci with a significant stuttering (33%). Cured data provided results in agreement with a pangamic reproductive strategy, as was already observed by the authors (De Meeûs et al., 2021).

For *Glossina palpalis palpalis* in Côte d'Ivoire no binomial tests provided a significant result with MicroChecker statistics. For the new method presented here, two loci (B3 and XB110) out of seven (22.22%) displayed a significant stuttering. Cured data set was obtained by pooling alleles following De Meeûs et al (2019)'s rules (De Meeûs et al., 2021) (see Appendix 2, A2.1). This cure provided a slightly lower F_{IS}=0.221 (instead of 0.231) for locus B3, but a higher one for locus XB110 (F_{IS}=0.277 instead of 0.252),

Mis en forme : Police :Non Italique

Mis en forme : Police :Non Italique

Supprimé: , the results were similar between those obtained with our method and the published results with MicroChecker because these were submitted neither to exact binomial tests nor to subsequent BH correction. Nevertheless

Supprimé: 7

Supprimé: tests

Supprimé: the 171 that could be made

Supprimé: 4.09

Supprimé: After exact binomial tests, only a single locus gave a significant stuttering test with MicroChecker (11%). All these tests stayed significant after BH correction

Supprimé: , no stuttering was initially found because author used a different criterion (a significant global heterozygote deficit was necessary to validate a one-repeat heterozygote deficit observed in the graphics of MicroChecker) as compared to what we used here. Nevertheless, only two tests out of 98 (2.04%) gave a significant stuttering with MicroChecker, and

Supprimé: , even after BH correction

Déplacé vers le bas [1]: For Locus B3 allele 203 was recoded 201; allele 207 was recoded 205; allele 211 was recoded 209; and alleles 215 was recoded 213. For Locus XB110, alleles 179 was recoded 177; allele 183, was recoded 181; alleles 187 to 199 were recoded 185, allele 203 was recoded 201 and all rare alleles from 205 to 277 were also recoded 203.

703 meaning that the heterozygote deficit at this locus was better explained by another
704 phenomenon (i.e. null alleles).
705 In *G. truncatula*, after Benjamini and Hochberg correction, 50% of loci displayed a
706 significant stuttering, while all six loci were significant with the new method (all *p*-
707 BH<0.0104). Allele pooling for curing the data was as described in Appendix 2 (A2.2). This
708 cure lowered the F_{IS} for three loci, Lt9 (0.778 to 0.776), Lt16 (0.958 to 0.006), and Lt24
709 (0.966 to 0.947) and increased it for the others. However, missing data (assumed null
710 homozygotes) explained almost 50% of F_{IS} variation, with a highly significant Spearman's
711 rank correlation between F_{IS} and number of blank genotypes (ρ=0.9411, *p*-value=0.0025),
712 while with the data set cured for loci Lt9, Lt16 and Lt24, the correlation dropped to a non-
713 significant value (*p*-value=0.2589). When removing the correction for locus Lt16, which
714 provided an outlier as compared to other loci, the correlation became significant again (*p*-
715 value=0.0103), but with a smaller correlation (ρ=0.8804) and only 41% of F_{IS} variation
716 explained by missing genotypes.
717 For *F. hepatica*, no locus was significant with MicroChecker, while only one locus
718 (Fh28) presented a significant stuttering signature after Benjamini and Hochberg
719 correction (*p*-BH=0.0001). Pooling of alleles at this locus is described in Appendix 2
720 (A2.3). The F_{IS} of cured data dropped from 0.644 to 0.536. However, we knew that null
721 alleles explained well most of the observed heterozygote deficit (Correa et al., 2017), and
722 the correlation between missing data and F_{IS}, when excluding one locus (Fh25) that
723 displayed too many missing genotypes, was initially significant (ρ=1, *p*-value=0.0417), with
724 83 % of the F_{IS} variation explained by blank genotypes. In the cured data, it dropped to a
725 not significant relationship (ρ=0.8, *p*-value=0.1667), with 72 % of F_{IS} variation explained by
726 missing genotypes.
727 These discrepancies strongly suggested that stuttering detection in *G. truncatula*
728 and *F. hepatica* corresponded to type error I, due to the fact that null alleles better explain
729 the data and probably interacted with our stuttering detection test. It shows that several
730 checks need being undertaken before deciding that a locus is significantly affected by
731 stuttering and requires being cured, especially in selfing species.
732 For the clonal *T. brucei gambiense* 1, we avoided using MicoChecker (for obvious
733 reasons). With the method expounded in the present paper, three loci displayed a
734 significant stuttering signature: micbg1, misatg4 and misatg9. These loci were cured as in
735 Appendix 2 (A2.4). These loci presented a lower F_{IS} when cured: from -0.647 to -818, from
736 -0.579 to -0.72, and from -0.471 to -0.496 for micbg1, misatg4 and misatg9, respectively.
737 Moreover, several observations suggested an improvement of the quality of the data after

Supprimé: displayed a

Supprimé: stuttering signature after BH correction

Supprimé: followed

Supprimé: : for locus Lt9, 203 with 202, 210 and 212 with 208; for Lt16, 231 to 233 with 230; for Lt21, 107 with 105, 112 with 111, 115 and 116 with 114; for Lt24, 208 and 210 with 207, 215 to 217 with 214, 220 and 221 with 219; for Lt36, 187 to 190 and 192 with 185; and for Lt37, 115 with 113, and 123 with 212

Supprimé: was the following:

Supprimé: Alleles 182 to 183 with 180, allele 188 with 186, and alleles192 to 194 with 190

Supprimé: indeed

Supprimé: t

Supprimé: , according to the criterion defined in the Material and Methods section

Supprimé: Curing the data implied the

Supprimé: following: for micbg1, allele 164 was pooled with 162, and allele 194 with 192; for misatg4, allele 117 with 115, and 145 with 143; and for misatg9, alleles 130 and 128 with 126, allele 186 with 184, and alleles 194 and 192 with 190

Supprimé: indeed

the cure for stuttering. The proportion of significant linkage disequilibrium between locus pairs increased from 53% (uncured data) to 80% (data cured for the three loci), and the superimposition S_C increased from 0.9554 to 0.9782, from 0.9800 to 0.9975, and from 0.9539 to 0.9663 for micbg1, misatg4 and lisatg9 respectively. Three loci with missing genotypes could be suspected to display null alleles, i.e. misatg4, misatg9 and m6c8. When these loci were removed, averaged superimposition increased to S_C=0.9908. Here again, we can see that some checks allowed deciding that stuttering corrections were valid and significantly improved the quality of the data. After stuttering cure, removing other loci with suspected null alleles (i.e. with missing genotypes) drove the superimposition index defined in the Material and Methods section to almost unity, i.e. a perfect fit with the expected value under full clonality known to occut in that species (Weir et al., 2016).

Conclusions

The new method developed here appeared at least three times more efficient (and often much more) than MicroChecker. Moreover, the use of spreadsheet programs makes its portability universal for any microcomputer.

In dioecious pangamic populations, like ticks and tsetse flies, detection works well and cure improves population genetics parameter estimates but not perfectly so, which means that, for instance, F_{IS} and F_{ST} will still be slightly overestimated in datasets cured for stuttering. So, whenever possible, removal of affected loci may help to shift such estimates towards (slightly) more accurate values.

In monoecious selfers, detection works well and cure works very well, providing other confounding factors as null alleles do not interfere, in which case avoiding stuttering cure and correct for null alleles appear more appropriate. In doubt, and for subdivision measures, curing for null alleles may be achieved by the elimination of involved loci for strong selfers, or applying the INA correction (Chapuis & Estoup, 2007), for reasonably panmictic populations.

In clones, only highly significant stuttering tests with a significant higher F_{IS} as compared to other loci will need being cured, in which case parameter estimates may be much improved.

In any way, cure must only be kept if F_{IS} values are improved (lower than it was initially) and special care must be devoted to the behavior of their variation in relation to

Supprimé: . The average for cured data was then S_C=0.9718, which is very important

Déplacé (insertion) [3]

Supprimé: this

Supprimé: Discussion

Supprimé: The power of stuttering detection was weak, except in populations with an increased inbreeding coming from a closed system of mating. In this case, BH correction allowed reducing the false discovery rate to a reasonable level.¶

Supprimé: Selfing significantly increased power detection as discussed above.

Supprimé: . The stuttering proportion used here was relatively small, since the realized actual proportion of alleles affected at the end of simulations was in general much lower on average. Indeed, in many subsamples, the two first consecutive allele chosen for stuttering must be present and preceded or followed by an allele with one repeat difference. Otherwise, no stuttering occurs. Nevertheless, this was enough to generate significant heterozygote deficits under the pangamic dioecious mating system. Alternatively, this may also explain why stuttering had no significant impact on the F_{IS} of monoecious selfers. Nevertheless, stuttering detection significantly increased with local inbreeding.¶
Stuttering had a quite small impact on the ratio between the standard error of F_{IS} and F_{ST}, which did not reach the minimum two-fold threshold met with null alleles (De Meeûs, 2018)

Supprimé: . Stuttering did not affect linkage disequilibrium. This might have been expected here since how stuttering was generated did not affect more any individual than another. In real life, we may expect amplification problems to affect preferentially some individuals, at several loci, due to several possible and nonexclusive reasons: bad preservation of several DNA extracts, genetically distant individuals (genetically distant populations, subspecies, or cryptic species), and bad amplification conditions. In that case, we may expect some correlation between the affected loci thus generating an increase in linkage disequilibrium between those, as this appeared to be the case for the tick *I. scapularis* in USA (De Meeûs et al., 2021).¶
Subpopulation sizes and number of subpopulations did not affect much stuttering discovery. There is indeed no reason why the effect of stuttering would be increased or decreased by those parameters. The lim...

Déplacé vers le haut [2]: . Clonal populations generated very high proportions of false stuttering detections. This was because without segregation of alleles, only some classes of heterozygotes propagate

Supprimé: The strategy to cure the data suffering from stuttering did not allow to entirely restore a completely immune set of population genetics results in pangamic dioecious populations, though the 95% ...

Déplacé vers le haut [3]: After this cure, removing other loci with suspected null alleles (i.e. with missing genotypes) drove the superimposition index defined in

Supprimé: ¶
¶
Conclusion¶

Supprimé: ;

other factors as number of missing data (null alleles) or superimposition index (clonal populations).

The stuttering detection and cure strategies proposed in the present paper will help interpreting microsatellite data with more accuracy and at the lowest cost. This will be particularly helpful in non-model organisms, as parasite-vector systems, for which microsatellite markers still represent the best cost benefit ratio.

Supprimé: Such measures

Supprimé: in

Supprimé: at the lowest cost, which will be of particular interest in projects involving non-model, neglected, and small organisms that often happened to be economically and medically highly relevant

Author's contributions

All authors read, amended and/or approved the final manuscript.

Supprimé: , except JBR who could not check the last versions

Conceptualization: Thierry de Meeûs, Camille Noûs.
Data analyses: Thierry de Meeûs.
Design of figures: Thierry de Meeûs.
Writing of the original draft: Thierry de Meeûs.
Supervision: Thierry de Meeûs.

Data availability

Scripts used are available in the Appendix 1.

Parameter for the first simulation with Easypop are in the file "TestStutterDioeciousNoStutter-n1000N100-1.txt"; The file "TestStutterDioeciousNoStutter-n1000N100-1.equ gives population genetics parameters measured at each generation of the first replicate; "TestStutterDioeciousNoStutter-n1000N100-1.gen" and "TestStutterDioeciousNoStutter-n1000N100-1.dat" are the genepop and Fstat files generated by this simulation, respectively. The file "TestStutterDioecious-n1000N100-1-10%Stuttering.xlsx", is a template for generating 10% stuttering in the first replicate of the first Easypop simulation; and the file "TestStutterDioeciousNoStutter-n1000N100-1-FstatRes.xlsx" is a template of stuttering detection (together with LD tests and other measured done by Fstat), for this first replicate of the fists simulation with 10 % stuttering.

Synthetic results for dioecious and monoecious simulations are in Excel format in the supplementary file S1 "SynthesisStutteringTestSexualsSupFileS1.xlsx".

Synthetic results for clonal simulations are in Excel format in the supplementary file S2 "StutteringClonesSynthesisSupFileS2.xlsx".

Pooling protocols for curing simulated data are in Excel files that were compressed in a zipped file, supplementary file S3 "PoolingProtocolCureSupFileS3.zip".

Conflict of interest disclosure

The authors declare that they have no financial conflict of interest with the content of this article. Thierry de Meeûs is one of the PCI Infections administrators.

Acknowledgements

This work was made possible by the support of IRD (French National Institute for Sustainable Development).

References

Balloux, F. (2001) EASYPOP (version 1.7): A computer program for population genetics simulations. *Journal of Heredity*, **92**, 301-302.

Balloux, F. (2004) Heterozygote excess in small populations and the heterozygote-excess effective population size. *Evolution*, **58**, 1891-1900.

Balloux, F., Lehmann, L., De Meeûs, T. (2003) The population genetics of clonal and partially clonal diploids. *Genetics*, **164**, 1635-1644.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1-48.

Bayerl, H., Kraus, R.H.S., Nowak, C., Foerster, D.W., Fickel, J., Kuehn, R. (2018) Fast and cost-effective single nucleotide polymorphism (SNP) detection in the absence of a reference genome using semideep next-generation Random Amplicon Sequencing (RAMseq). *Molecular Ecology Resources*, **18**, 107-117. 10.1111/1755-0998.12717

Benjamini, Y., Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, **57**, 289–300.

Benjamini, Y., Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165–1188.

Berté, D., De Meeus, T., Kaba, D., Séré, M., Djohan, V., Courtin, F., N'Djetchi, K.M., Koffi, M., Jamonneau, V., Ta, B.T.D., Solano, P., N'Goran, E.K., Ravel, S. (2019) Population genetics of *Glossina palpalis palpalis* in sleeping sickness foci of Côte d'Ivoire before and after vector control. *Infection Genetics and Evolution*, **75**, 103963.

Castle, W.E. (1903) The laws of heredity of Galton and Mendel, and some laws governing race improvement by selection. *Proceedings of the American Academy of Arts and Sciences*, **39**, 223-242.

Chapuis, M.P., Estoup, A. (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, **24**, 621-631. 10.1093/molbev/msl191

1014 Correa, A.C., De Meeûs, T., Dreyfuss, G., Rondelaud, D., Hurtrez-Boussès, S. (2017)
1015 *Galba truncatula* and *Fasciola hepatica*: Genetic costructures and interactions with
1016 intermediate host dispersal. *Infection Genetics and Evolution*, **55**, 186-194.
1017 10.1016/j.meegid.2017.09.012
1018 De Meeûs, T. (2015) Genetic identities and local inbreeding in pure diploid clones with
1019 homoplasic markers: SNPs may be misleading. *Infection Genetics and Evolution*, **33**, 227–
1020 232.
1021 De Meeûs, T. (2018) Revisiting F_{IS}, F_{ST}, Wahlund effects, and Null alleles. *Journal of*
1022 *Heredity*, **109**, 446-456. 10.1093/jhered/esx106
1023 De Meeûs, T., Balloux, F. (2004) Clonal reproduction and linkage disequilibrium in
1024 diploids: a simulation study. *Infection Genetics and Evolution*, **4**, 345-351.
1025 De Meeûs, T., Balloux, F. (2005) F-statistics of clonal diploids structured in numerous
1026 demes. *Molecular Ecology*, **14**, 2695-2702.
1027 De Meeûs, T., Chan, C.T., Ludwig, J.M., Tsao, J.I., Patel, J., Bhagatwala, J., Beati, L.
1028 (2021) Deceptive combined effects of short allele dominance and stuttering: an example
1029 with *Ixodes scapularis*, the main vector of Lyme disease in the U.S.A. *Peer Community*
1030 *Journal*, **1**, e40. https://doi.org/10.24072/pcjournal.34
1031 De Meeûs, T., Guégan, J.F., Teriokhin, A.T. (2009) MultiTest V.1.2, a program to
1032 binomially combine independent tests and performance comparison with other related
1033 methods on proportional data. *BMC Bioinformatics*, **10**, 443.
1034 De Meeûs, T., Lehmann, L., Balloux, F. (2006) Molecular epidemiology of clonal diploids:
1035 A quick overview and a short DIY (do it yourself) notice. *Infection Genetics and Evolution*,
1036 **6**, 163-170.
1037 De Meeûs, T., McCoy, K.D., Prugnolle, F., Chevillon, C., Durand, P., Hurtrez-Boussès, S.,
1038 Renaud, F. (2007) Population genetics and molecular epidemiology or how to "débusquer
1039 la bête". *Infection Genetics and Evolution*, **7**, 308-332.
1040 Fox, J. (2005) The R commander: a basic statistics graphical user interface to R. *Journal*
1041 *of Statistical Software*, **14**, 1–42.
1042 Fox, J. (2007) Extending the R commander by "plug in" packages. *R News*, **7**, 46–52.
1043 Garvin, M.R., Saitoh, K., Gharrett, A.J. (2010) Application of single nucleotide
1044 polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*,
1045 **10**, 915-934. 10.1111/j.1755-0998.2010.02891.x
1046 Gomez-Palacio, A., Triana, O., Jaramillo, N., Dotson, E.M., Marcet, P.L. (2013) Eco-
1047 geographical differentiation among Colombian populations of the Chagas disease vector

1048 *Triatoma dimidiata* (Hemiptera: Reduviidae). *Infection Genetics and Evolution*, **20**, 352-
1049 361. 10.1016/j.meegid.2013.09.003
1050 Goudet, J. (1995) FSTAT (Version 1.2): A computer program to calculate F-statistics.
1051 *Journal of Heredity*, **86**, 485-486.
1052 Goudet, J. (2003) Fstat (ver. 2.9.4), a program to estimate and test population genetics
1053 parameters. Available at http://www.t-de-meeus.fr/Programs/Fstat294.zip, Updated from
1054 Goudet (1995).
1055 Hunter, P. (2014) Tropical diseases and the poor: Neglected tropical diseases are a public
1056 health problem for developing and developed countries alike. *EMBO Reports*, **15**, 347-350.
1057 10.1002/embr.201438652
1058 Jarne, P., Lagoda, P.J.L. (1996) Microsatellites, from molecules to populations and back.
1059 *Trends in Ecology & Evolution*, **11**, 424-429.
1060 Koffi, M., De Meeûs, T., Bucheton, B., Solano, P., Camara, M., Kaba, D., Cuny, G., Ayala,
1061 F.J., Jamonneau, V. (2009) Population genetics of *Trypanosoma brucei gambiense*, the
1062 agent of sleeping sickness in Western Africa. *Proceedings of the National Academy of*
1063 *Sciences of the United States of America*, **106**, 209-214.
1064 Manangwa, O., De Meeûs, T., Grébaut, P., Segard, A., Byamungu, M., Ravel, S. (2019)
1065 Detecting Wahlund effects together with amplification problems : cryptic species, null
1066 alleles and short allele dominance in Glossina pallidipes populations from Tanzania.
1067 *Molecular Ecology Resources*, **19**, 757-772. 10.1111/1755-0998.12989
1068 Nébavi, F., Ayala, F.J., Renaud, F., Bertout, S., Eholié, S., Moussa, K., Mallié, M., De
1069 Meeûs, T. (2006) Clonal population structure and genetic diversity of *Candida albicans* in
1070 AIDS patients from Abidjan (Côte d'Ivoire). *Proceedings of the National Academy of*
1071 *Sciences of the United States of America*, **103**, 3663-3668.
1072 Nei, M., Chesser, R.K. (1983) Estimation of fixation indices and gene diversities. *Annals of*
1073 *Human Genetics*, **47**, 253-259.
1074 Prudhomme, J., De Meeûs, T., Toty, C., Cassan, C., Rahola, N., Vergnes, B., Charrel, R.,
1075 Alten, B., Sereno, D., Bañuls, A.L. (2020) Altitude and hillside orientation shapes the
1076 population structure of the *Leishmania infantum* vector *Phlebotomus ariasi*. *Scientific*
1077 *Reports - Nature*, **10**, 14443. 10.1038/s41598-020-71319-w
1078 R-Core-Team (2020) R: A Language and Environment for Statistical Computing, Version
1079 3.6.3 (2020-02-29) ed. R Foundation for Statistical Computing, Vienna, Austria,
1080 http://www.R-project.org.
1081 Séré, M., Kabore, J., Jamonneau, V., Belem, A.M.G., Ayala, F.J., De Meeûs, T. (2014)
1082 Null allele, allelic dropouts or rare sex detection in clonal organisms : simulations and

application to real data sets of pathogenic microbes. *Parasites and Vectors*, **7**, 331. 10.1186/1756-3305-7-331

Séré, M., Thévenon, S., Belem, A.M.G., De Meeûs, T. (2017) Comparison of different genetic distances to test isolation by distance between populations. *Heredity*, **119**, 55-63. 10.1038/hdy.2017.26

Solano, P., Ravel, S., De Meeûs, T. (2010) How can tsetse population genetics contribute to African trypanosomiasis control? *Trends in Parasitology*, **26**, 255-263.

Van Oosterhout, C., Hutchinson, W.F., Wills, D.P.M., Shipley, P. (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, **4**, 535-538.

Vignal, A., Milan, D., SanCristobal, M., Eggen, A. (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, **34**, 275-305.

Vitalis, R. (2002) Sex-specific genetic differentiation and coalescence times: estimating sex-biased dispersal rates. *Molecular Ecology*, **11**, 125-138.

Wang, C., Schroeder, K.B., RosenbergO, N.A. (2012) A maximum-likelihood method to correct for allelic dropout in microsatellite data with no replicate genotypes. *Genetics*, **192**, 651-669. 10.1534/genetics.112.139519

Weinberg, W. (1908) Über den Nachweis der Verebung beim Menschen. *Jahresheft des Vereins fur Vaterlundische Naturkunde in Wurttemberg*, **64**, 368-382.

Weir, B.S., Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358-1370.

Weir, W., Capewell, P., Foth, B., Clucas, C., Pountain, A., Steketee, P., Veitch, N., Koffi, M., De Meeûs, T., Kaboré, J., Camara, M., Cooper, A., Tait, A., Jamonneau, V., Bucheton, B., Berriman, M., MacLeod, A. (2016) Population genomics reveals the origin and asexual evolution of human infective trypanosomes. *eLife*, **5**, e11473.

Wright, S. (1965) The interpretation of population structure by F-statistics with special regard to system of mating. *Evolution*, **19**, 395-420.

Appendices

Appendix 1: R-scripts and R-commander menu used to analyse the different datasets

Regarding R-scripts, most are very small and basic: `binom.test`, for the exact binomial test; `p.adjust`, for adjusting the *p*-values in a series of p-values.

For generalized linear mixed models with the package lme4, the commands were, for the number of significant stuttering tests (for instance):

```
Dataset <-
read.table("C:/DeMeeus/thierry/StutteringTest/DataMixedModelStutterSimul.txt",he
ader=TRUE)
SigStut1<-glmer(StutterSigSperLocus~n+N+Mating+Stuttering+(1|Rep),
family=poisson(link = "log"), data=Dataset)
summary(SigStut1)
```

The Fisher exact tests were undertaken with R-commander, in the menu Statistics-Contingency table-Enter and analyze two-way table

Appendix 2: Pooling of alleles used to cure real datasets

A2.1. For *Ixodes scapularis*

For Locus B3 allele 203 was recoded 201; allele 207 was recoded 205; allele 211 was recoded 209; and alleles 215 was recoded 213. For Locus XB110, alleles 179 was recoded 177; allele 183, was recoded 181; alleles 187 to 199 were recoded 185, allele 203 was recoded 201 and all rare alleles from 205 to 277 were also recoded 203.

Déplacé (insertion) [1]

A2.2. For *Galba truncatula*

For locus Lt9, 203 with 202, 210 and 212 with 208; for Lt16, 231 to 233 with 230; for Lt21, 107 with 105, 112 with 111, 115 and 116 with 114; for Lt24, 208 and 210 with 207, 215 to 217 with 214, 220 and 221 with 219; for Lt36, 187 to 190 and 192 with 185; and for Lt37, 115 with 113, and 123 with 212.

A2.3. For Fasciola hepatica

Alleles 182 to 183 with 180, allele 188 with 186, and alleles192 to 194 with 190.

A2.4. For *Trypanosoma brucei gambiense* 1
for micbg1, allele 164 was pooled with 162, and allele 194 with 192; for misatg4, allele 117 with 115, and 145 with 143; and for misatg9, alleles 130 and 128 with 126, allele 186 with 184, and alleles 194 and 192 with 190.