Final-Year Practical Project Cover Sheet

Name of Student: Isabelle Divyabharathy Rajendiran

CID: 01852916

Degree Stream: Biological Sciences

Date: 07/06/2023

Word Count: 5990

Project Title: Comparing estimates of historical and contemporary effective population size (N_e)

I hereby declare that this Project has not been submitted, either in the same or different forms, to this or any other university for a degree. I also declare that this Project does not draw from any other work prepared under consultancy or other professional undertaking, by myself or jointly with other authors in any way other

Name of Supervisor: Dr Vassiliki Koufopanou

Name of Examiner 1: Dr Bhavin Khatri

Name of Examiner 2: Professor Vincent Savolainen

Final-Year Practical Project reports: assessment criteria

These criteria are to be used for all types of final-year practical project report: *Lab project, Field project and Computational project*. Outside reading is fundamental when writing up a research project, so is not mentioned explicitly in the criteria that follow. Most outside reading should be from the peer-reviewed scientific literature, including primary research papers. The expected format of the report is described in detail in the guidelines for **Practical project reports.**

Class	%	Criteria
1st	100	Report is of **sufficient quality to submit for publication to an international peer-reviewed journal** (assuming, ideally, that positive and negative results have equal merit).
	95	Report is close to a publishable standard, containing a **succinct survey of the most important primary literature** and an **accurate and logical account and justification of the methods used**. It presents the **results in a publishable format**, and knowledgeably **applies any necessary mathematical and/or statistical techniques**. Discussion of results demonstrates **high levels of rigour and critical ability** in the context of the relevant literature. Report **demonstrates an appreciation of the limitations** of the experimental or other procedures, shows **attention to detail** (in references, figures, *etc.*), and **shows clear and possibly novel insight** into the subject.
	90	
	85	Excellent report, meeting all of the criteria for a mark of 68 and **most but not all of the criteria for a mark of 90+**.
	80	
	76	Excellent report, meeting all the criteria for a mark of 68 and **one or a few of the criteria for a mark of 90+.**
	72	
2A	68	Very good, **well-structured** report **written in good scientific style** (*i.e.* in clear, concise, direct, precise, dispassionate scientific English, including all important details). It shows the following features: (i) an ability to **carry out experimental procedures successfully** to generate original results (which may be negative and need not be novel); (ii) a **very good understanding of the study design and the methods used** to generate and analyse the data; (iii) **appropriate – if not high-level – analyses**; (iv) **clear presentation of results**; (v) **sound knowledge of how the study fits in to the relevant literature**; (vi) **some critical interpretation** of the results and the study overall.
	65	
	62	
2B	58	Good report showing the following features: (i) an **ability to follow experimental procedures**; (ii) **basic understanding of the relevant concepts and methods**; (iii) **mostly logical structure and scientific style**; (iv) **reasonable interpretation** of the data or information collected; and (iv) a **reasonable attempt to relate the results to the latest literature**. Reports that are **too long, poorly written**, and/or that show **poor use of references** are unlikely to be marked above a 2B.
	55	
	52	
3rd	48	Acceptable report showing the following features: (i) **an ability to follow some experimental procedures**; (ii) a **weak grasp of most of the relevant concepts and methods**; (iii) **need for close guidance in design and interpretation**; and (iv) at best **limited relation of the results to the relevant literature**. Research projects in this bracket are likely to be marred by significant errors, important omissions, brevity and/or a failure to interpret the data critically.
	45	
	42	
Fail	38	Poor report showing the following features: (i) **understanding of less than half of the theoretical basis of the project**; (ii) **evidence of widespread difficulty following procedures** to generate and analyse data; (iii) need for **complete instruction in design and interpretation**; (iv) **does not relate the outcome of the experimental work to the literature**.
	35	
	30	
	25	Report **contains more than a few relevant sentences** but shows very little understanding of the background to the project, the project design, or the methods used to generate or analyse the data. Students in this bracket are unlikely to have been able to carry out even basic procedures, despite proper instruction.
	20	
	15	Report contains **only a few sentences relevant to the subject**, and does not contain any interpretable results.
	10	
	5	
	0	Report contains **nothing** relevant or was not submitted.

Practical Project research performance: assessment criteria

For students undertaking lab, field or computational final-year practical projects.

Class	%	Criteria
1st	100 95 90	Student **worked safely, confidently, diligently, and designed appropriate investigations**. Student developed a **high level of technical expertise**. Student **kept supervisor informed of progress, but consistently showed initiative** and did not require micromanagement. Student **contributed very positively to the research group.**
	85 80	Student met all of the criteria for a mark of 68 as well as **most of the criteria for a mark of 90+**
	76 72	Student met all of the criteria for a mark of 68 as well as **one or a few of the criteria for a mark of 90+.**
2A	68 65 62	Student's **research work was performed competently**. The student **contributed meaningfully to the experimental design**, worked reasonably hard, **picked up procedures well**, and was **able to work largely independently**.
2B	58 55 52	Student's research was **performed safely throughout**. The student had **some input into experimental design** and **worked reasonably hard**. The student was **able to work usefully with day-to-day supervision**.
3rd	48 45 42	Student showed **some ability to follow experimental procedures without close supervision** and **appreciated safety aspects**, but the **work was small in quantity and poorly executed**. Student's **input into experimental design was minimal**.
Fail	38 35 30	Student worked for **up to a half of the expected time** and **worked safely/adequately only when very closely supervised**. Student showed **very little or no initiative or independence**.
	25 20	Student **attended the laboratory or field site for up to a third of the expected time** and performed **some work safely/adequately but only when micromanaged**. Very **little useful work completed**.
	15 10 5	Student **attended the laboratory or field site** but either attended for less than a quarter of the expected time or worked in an unsafe or otherwise wholly unsatisfactory fashion despite proper instruction. **Negligible amount of work completed**.
	0	Student **did not attend** the laboratory or field site, was barred for preventable reasons (e.g., an unacceptable attitude to safety), or was found to have fabricated results.

Comparing estimates of historical and contemporary effective population size (N_e)

Isabelle Rajendiran

Acknowledgements

My sincere thanks to Dr Vassiliki Koufopanou and Dr Tin-Yu Hui for their guidance throughout the project and also Dr Jon Brenas and Josh Reynolds for their coding-related assistance.

Table of Contents

1. **Abstract**
2. **Introduction**
3. **Methods and Materials**
   - **3.1. Terminology**
   - **3.2. Data source and filtering**
   - **3.3. Estimating historical N_e using nucleotide diversity (π)**
   - **3.4. Estimating contemporary N_e using genetic drift**
4. **Results**
   - **4.1. Estimation from nucleotide diversity (π)**
   - **4.2. Estimation of contemporary N_e**
5. **Discussion**
   - **5.1. Comparing the nucleotide diversity (π) estimates**
   - **5.2. Comparing the temporal F_a estimates**
   - **5.3. Comparing historical and contemporary N_e estimates**
   - **5.4. Conclusions and future directions**
6. **References**

1. Abstract

Mosquitos of the *Anopheles gambiae species complex* are vectors of diseases like malaria, a life-threatening disease with over millions of annual cases globally. Estimates of *Anopheles spp.* population sizes is key for vector control strategies e.g., to monitor the success of an intervention such as insectides. Here, using genomic data from the *Anopheles gambiae* 1000 Genome Project (Ag1000G) collected from Niono, Mali in 2013 and 2015, we calculate and compare two estimates of effective population size (N_e): historical and contemporary N_e which are derived from the principles of diversity and genetic drift, respectively. We also compare the data types used in these estimates and how they influence N_e estimates e.g., chromosome 3 vs X and neutral (no selection) genomic sites (intergenic vs 4-fold coding degenerate sites (4-CDS)). We found that the historical N_e was larger than contemporary $N_{e,}$ likely reflecting the differences in their spatial-temporal scale of population demographics. Additionally, our findings reveal unexpected differences between the 4-CDS and intergenic sites, and suggests they have differential influence on diversity measures between chromosome X and 3R. Similar conclusions were found in previous studies including Ag1000G (2017), but the underlying reasons for these observations remain a question of interest.

2. Introduction

Commonly, population size estimates have known conservation purposes e.g., to maintain a certain size to avoid inbreeding depression or measure population decline of endangered species (Antao, Pérez-Figueroa & Luikart, 2011; Wang et al., 1999). However, population size estimation is also used for vector control strategies against diseases like Malaria, which had a reported 247 million cases globally in 2021 (World Health Organisation, 2022). Mosquitoes of the *Anopheles gambiae species complex* including *An. coluzzii* and *An. gambiae* are important vectors of human malaria found in Africa, and thus one of the main targets for such strategies (Ag1000G, 2017). This includes using population size to measure the success of an intervention i.e., decreased population post-intervention such as insecticides, but also to determine the necessary amount of intervention required. An example is for gene-drive technology, in which knowing the wild mosquito population size is essential to calculate and release enough genetically modified mosquitoes to ensure spread of the gene in the wild population and thus suppress the wild population (Willis & Burt, 2021). Release of too few mosquitoes will lead to loss of the gene by genetic drift i.e., the change in allele frequencies due to random sampling of gametes (Charlesworth, 2009; Willis & Burt, 2021).

Here, genomic data was used to estimate the effective population size (N_e) which is the size of an ideal population which loses heterozygosity at the same rate as the observed population (Luikart et al., 2010). This differs from the census population size (N_c) which is the observed number of adults in a population (Luikart et al., 2010). A popular method of estimating N_c is using mark-release-recapture in which the ratio of the re-captured: total captured individuals is assumed to be equivalent to the ratio of marked individuals: total population size (N_c); and this has been used previously to study seasonal variation in *Anopheles spp.* population size as part of vector control strategies (Epopa et al., 2017). While both useful, N_c and N_e are used for different applications e.g., N_c would be preferred to deduce mosquito biting rate which depends on the adult population size (e.g., for epidemiological purposes); but N_e is more suited if you are interested in the genetic health and adaptive potential of the population for purposes like the spread of gene-drive technology.

There are multiple estimators of N_e, each giving information about different questions of interest/reflecting different spatial-temporal scales of population dynamics; in this case, historical and contemporary N_e was estimated. Historical N_e measures the population size from over a large time-frame and a wide geographical region as it is based on diversity which accumulates from over ten-to-thousands of generations, and through gene flow from neighbouring regions (Luikart et al., 2010). Contemporary N_e, estimated from drift, is a more local measure taken over recent generations and restricted to a smaller region from which the samples are obtained from (Luikart et al., 2010). Therefore, while historical N_e gives insight into the long-term population demographic e.g., bottleneck, expansions and migration, contemporary N_e is more useful to study current genetic health and population size (Charlesworth, 2009; Luikart et al., 2010).

Assuming the standard neutral model (SNM) (i.e., a panmictic population of constant size with no migration and under no selection (Kimura, 1983)), historical and contemporary N_e

can be calculated using nucleotide diversity (π) (i.e. the average number of pairwise differences in DNA nucleotides per site calculated for a cohort (Nei & Li, 1979) and genetic drift, respectively. The principle is that a smaller effective population size experiences stronger genetic drift as chance events have a greater influence on allele frequencies; consequently this leads to a loss of genetic diversity over time due to loss and fixation of alleles (figure 1) (Wright, 1931).

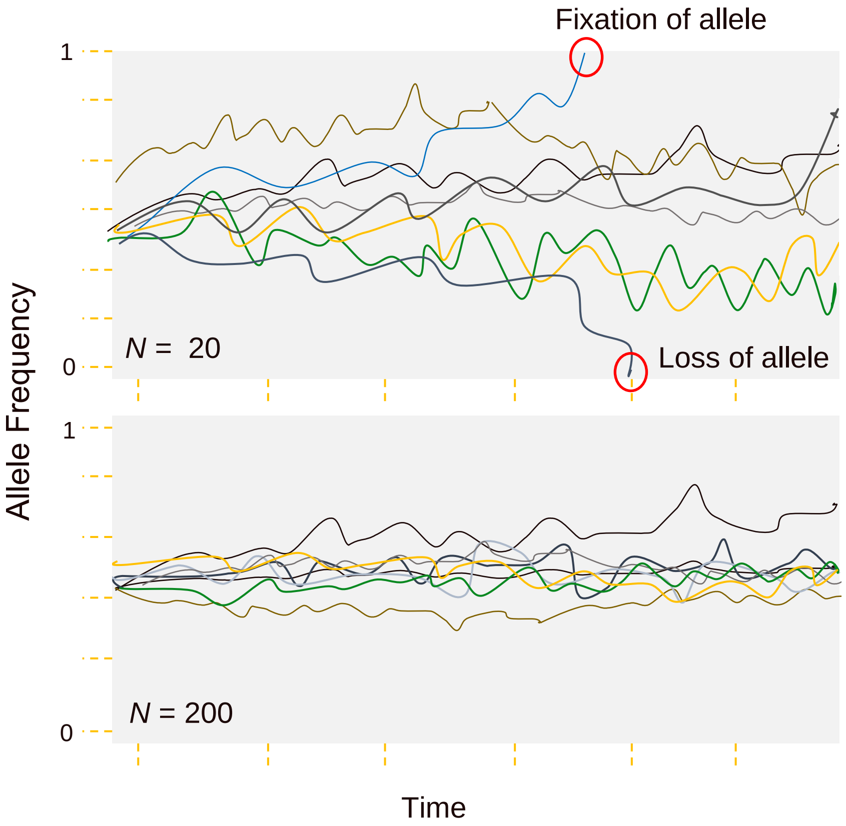


Figure 1. Wright-Fisher model simulation of genetic drift which demonstrates that the strength of genetic drift is stronger in a smaller population. Each line represents an allele of a site, and its change in frequency over time. The smaller population (*N* = 20) experiences more noise in its change in allele frequencies i.e., greater changes due to genetic drift compared to the larger population (*N* = 200) (Wright, 1931). In all populations, the effect of drift leads to eventual fixation/loss of alleles (seen in red) i.e., loss of polymorphisms, but the rate at which loss of polymorphism occurs is faster in smaller populations (Wright, 1931).

It is important to keep in mind that drift is an accumulative process (accumulation of the successive random changes in allele frequency over time) and cannot be directly observed, only through samples (see figure 2).

$$\begin{tikzcd}
p_0 \arrow[r] \arrow[d] \arrow[rrrr, bend left] & p_1 \arrow[r] \arrow[d] & \cdots \arrow[r] & p_{t-1} \arrow[r] \arrow[d] & p_t \arrow[d] \\
S_0 & S_1 & \cdots & S_{t-1} & S_t
\end{tikzcd}$$

Figure 2. A model describing the accumulative process of genetic drift over time. The population at different timepoints is represented as *p_0, …, p_t* and contains the true allele frequencies which follow the Wright-Fischer model i.e., changes by the process of genetic drift (shown as black horizontal arrows →). The measure of drift between p_0 and p_t (horizontal black arrow →) is therefore an accumulation of the successive changes in the true allele frequencies between those populations. The true allele frequencies cannot be directly observed, and instead represented by the sampled allele frequencies obtained from population samples (*S_0, …, S_t*). Note that the drift estimate from the sampled allele frequencies also contains sampling error (vertical black arrows ↓) which needs to be accounted for. The resulting drift estimate can inform us about the effective population size (N_e) between p_0 and p_t, which has governed the process. Adapted from Hui & Burt (2015).

Therefore, drift and diversity can be used to inform our N_e estimates.

The methods of using drift and diversity to estimate contemporary and historical N_e of *Anopheles spp.* have been well established in previous literature including the *Anopheles gambiae* 1000 Genome Consortium (Ag1000G)'s publications in 2017 and 2020, and Hui, Brenas & Burt (2021). However, these studies focus solely on either contemporary or historical N_e. In this project, we aimed to compare both measures, to test the expectation that under the SNM both historical and contemporary N_e would be the same. We expect this because usually, historical N_e is a long term population estimate from diversity which accumulates over many generations and thus reflects the diversity from the influx of migrant mosquitoes as well as the long-term demographic history e.g., bottlenecks and expansions (Charlesworth, 2009; Luikart et al., 2010). Contemporary N_e only reflects the events which occur between two temporally spaced samples and do not consider any past changes in population size (Luikart et al., 2010). As the SNM assumes a constant and closed population size, the diversity of a population under SNM is not affected by migrant mosquitoes or any bottlenecks/expansion events. Therefore, the historical N_e should not differ from the contemporary N_e, as long as the assumptions hold.

Here, we aimed to provide N_e estimates for a town in the Ségou Region of Mali known as Niono, which has yet to be studied. Additionally, we analyse the data types utilised in the calculations to see how they influence the N_e estimates. In particular, we compare chromosome 3R vs chromosome X, chromosome positions and intergenic vs 4-fold degenerate sites (4-CDS) which are two types of presumably neutral sites (i.e. sites that are not under selection). We hypothesised that results from both intergenic and 4-CDS sites would be same, as under the SNM, they would both be equally (and completely) neutral. Also, we expected the 3R:X diversity estimation ratio to be 1:0.75 to reflect the chromosomes in a male-female pair (3 Xs, 1 Y and 4 chromosome 3s so a 4:3 ratio which simplifies to 1:0.75) (Hammer et al., 2008). Therefore, in a panmictic population with an equal male-female ratio, the X chromosome should have ~75% of the genetic diversity of chromosome 3 (and other autosomes) (Hammer et al., 2008).

3. Methods and Materials

3.1 Terminology

To clarify, from here onwards, a sample refers to an individual mosquito. Samples from a particular year e.g., 2013 samples refers to all the individual mosquitoes taken in 2013, unless otherwise specified. A nucleotide site, hereafter called a site, is a physical location on a genome, and can be fixed (only one allele) or polymorphic (more than one allele). Polymorphic sites are also referred to as single nucleotide polymorphisms (SNPs).

The designation of an intergenic site (sites in-between genes) here are sites that are more than 10 kbp from a gene. A 4-fold coding degenerate site is a site which can have any nucleotide substitutions at the third codon position and not substitute the amino acid.

3.2 Data source and filtering

All genomic data was provided by the Anopheles gambiae 1000 Genomes Project (Ag1000G) which was established in collaboration with MalariaGEN with the aim of studying mosquito genome variation and evolution (Ag1000G, 2017). Here, the mosquito sequences were obtained from the Ag1000G project phase 3 data who have collected samples from 2,784 individual mosquitoes belonging to *Anopheles gambiae, Anopheles coluzzii* and *Anopheles arabiensis* species from across 19 countries in sub-Saharan Africa (Ag1000G, 2021). The details related to the sampling, whole-genome sequencing, accessibility, SNP calling and quality control can be found on Ag1000G (2020) and their website (https://www.malariagen.net/data/ag1000g-phase3-snp).

For this project, a total of 177 *An. coluzzii* collected from Niono, Mali in 2013 (73 individuals) and 2015 (104 individuals) were used. Niono is a commune (third-level administrative unit) located in the Ségou Region of Mali. The last census (taken in 2009) reported a population size of 365,443 in Niono (INSTAT, n.d.)

All code was run on R (version 3.6.1) or Python (version 3.9.13). The malariagen_data (version 3.0.0) (https://github.com/malariagen/malariagen-data-python) and scikit-allel (version 1.3.5) (https://github.com/cggh/scikit-allel) Python packages were used primarily. Using these packages, the samples metadata and genomic data (containing the genotypes at each site and the site position information) was retrieved for the 2013 and 2015 samples. The sites have previously undergone quality control measures (see Ag1000G, 2020) e.g., to remove non-accessible sites (only accessible sites can be used in analyses). Further filtering to exclude missing sites i.e., sites which have missing genotype information for at least one individual.

3.3 Estimating historical *N_e* using nucleotide diversity (*π*)

To calculate nucleotide diversity (π), both the polymorphic (biallelic and multiallelic) and fixed sites from the 2013 and 2015 samples are necessary. To allow a fair comparison between the 2013 and 2015 samples, only a sample size of 70 individuals from each temporal sample was used in the calculation as the number of individuals affect the diversity calculated from them. The equation to estimate π is below (Nei & Li, 1979):

$$\pi = \frac{\sum_{i=1}^{K} 2 \times p_i \times q_i}{K} \tag{1}$$

where p_i and q_i are the allele frequencies of major (p) and alternate ($q = 1 - p$) alleles of site i and K is the number of sites.

A 95% confidence interval for the π estimate is calculated using the malariagen_data package by using a block jackknife procedure of 200 resamples (Ag1000G, 2021). To summarise, the total number of sites (intergenic/4-CDS) across the chromosome (X/3R) are divided into 200 blocks and used to generate 200 jackknife resamples in which each resample has removed one of the blocks. π is calculated for each jackknife resample and used to estimate the standard deviation of the π values which is used to construct the confidence intervals.

N_e can be estimated using the following equation (Charlesworth, 2009):

$$N_e = \frac{\pi}{4 \times \mu} \qquad \text{(Re-arranged from } \pi = 4N_e\mu\text{)} \qquad (2)$$

where μ is the mutation rate per-site-per-generation. Currently, there is no estimated mutation rate for *An. coluzzii* so an estimate of 2.8×10^9 from *Drosophila* is used instead (Keightley et al, 2014).

To investigate the heterogeneity of nucleotide diversity across the entire chromosome, a sliding window analysis of π (calculated using equation (1)) was conducted using non-overlapping windows of 10,000 loci. For this analysis, the 2013 and 2015 samples were pooled together. Using the positions of the regions of maintained diversity across the chromosomes, nucleotide diversity (π) and N_e were re-calculated using equations (1) and (2).

3.4 Estimating contemporary N_e using genetic drift

The 2013 and 2015 samples were filtered to contain only biallelic SNPs. Only SNPs with MAF ≥10% in the first temporal sample i.e., 2013 were included in the 2013 sample and the corresponding SNPs in the 2015 samples. This is to remove rare SNPs close to loss/fixation across the two time points as rare SNPs are more likely to be lost during the timepoints or have less potential to increase/decrease in frequency and thus provide less information. Also, to avoid linked SNPs (due to linkage disequilibrium), 1 SNP per 1000bp was randomly selected from the 2013 sample, and again, the corresponding SNPs were selected from the

2015 samples. This aims to remove replicate signals as two linked SNPs will have the same drift signal/pattern and needs to be interpreted as one signal (rather than 2 separate ones). Note that all the samples (i.e., individual mosquitoes) from the 2013 (73) and 2015 (104) samples were used for drift calculation, unlike the diversity calculations.

A measure of genetic drift per site called the temporal F_a statistic is calculated here. F_a is the sum of the squared differences in the normalised allele frequencies, average to per SNP i.e., calculates the accumulative drift across all the SNPs. The equation is shown below (Hui, Brenas & Burt, 2021; Waples, 1989):

$$F_a = \frac{1}{K} \sum_{i=1}^{K} \frac{(X_{it} - X_{i0})^2}{X_{i0}(1 - X_{i0})} \tag{3}$$

where K is the number of SNPs and X_{it} and X_{i0} represent the allele frequencies of the major (or can use the minor) allele at the first (X_{i0}) and second (X_{it}) temporal time points for every site (i). $(X_{it}-X_{i0})^2$ is the squared differences in allele frequencies and it is normalised to account for different starting allele frequencies using $X_{i0}(1 - X_{i0})$.

95% confidence intervals (CI) for F_a are computed according to Hui, Brenas & Burt, (2021) which takes into account the correlation among the changes in allele frequency at each SNP, which is mainly a function of the starting linkage disequilibrium and recombination rates. Note that despite picking 1 SNP per 1000bp, some linkage between SNPs remain. Here is a simple summary of the procedure. A $K \times K$ pairwise correlation matrix is computed which measures the covariance of the changes in allele frequency between pairs of SNPs i.e., the strength of linkage between pairs of SNPs. The eigenvalues of the matrix have an approximate chi-squared distribution (Q^2) which represents the variance of F_a. The 95 CI for F_a can then be obtained from the 2.5 and 97.5 percentiles of the Q^2 distribution. For the correlation matrix, here, a recombination rate of 1.1 cM/Mb and 3.2 cM/Mb was used for chromosome X and 3R, respectively (Zheng et al., 1996) instead of the genome-wide average of 1.6 cM/Mb (Pombi et al., 2006) as this more accurately represents the difference in recombination rates between chromosomes.

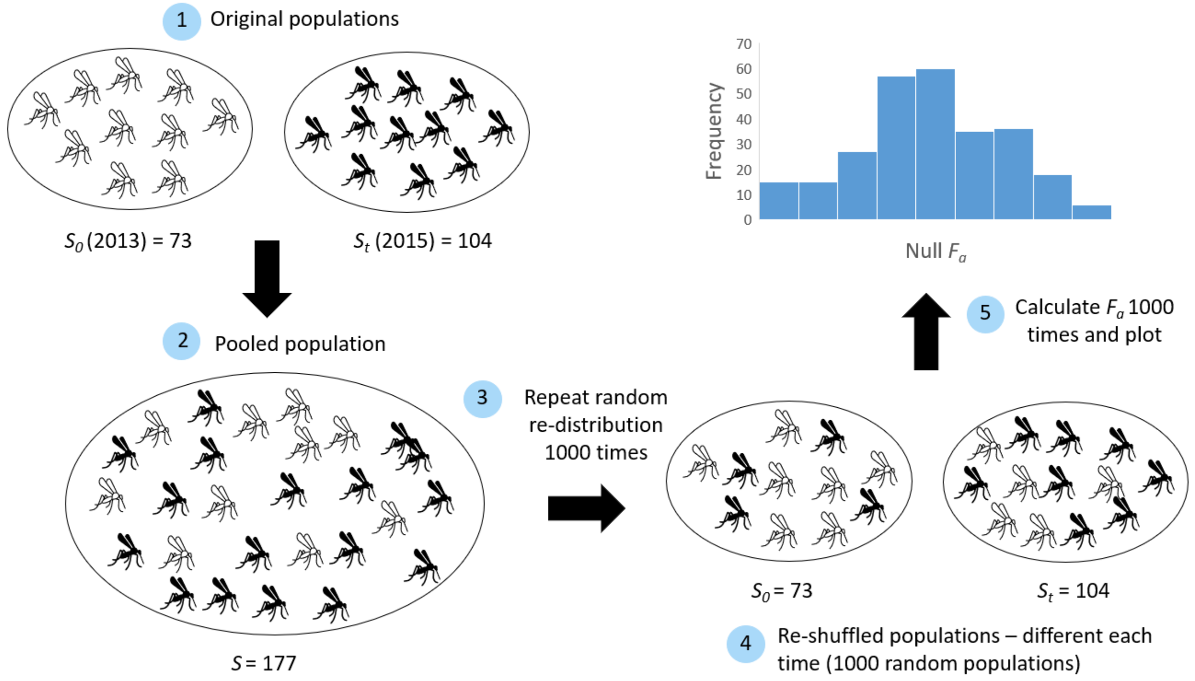
Using the F_a statistics calculated, one can estimate N_e using the equation below (Hui, Brenas & Burt, 2021; Waples, 1989):

$$N_e = \frac{t}{2\left(F_a - \frac{1}{2S_0} - \frac{1}{2S_t}\right)} \qquad (4)$$

where *t* is the number of generations (= 20 generations for a temporal difference of 2 years (O'Loughlin et al., 2016)) and S_0 and S_t are the sample sizes at the first and second timepoints, respectively. $F_a - \frac{1}{2S_0} - \frac{1}{2S_t}$ represents the F_a post sampling error in our samples.

The 95% CI for N_e are calculated from the 95% CI of the F_a values calculated above.

A randomisation test was conducted to test if the F_a calculated for our samples differ from that of no drift i.e., if the two samples in fact belong to the same population (initially assumed they belong to two temporally separate populations). The process of this method is displayed in figure 3 below.



1 Original populations
S_0 (2013) = 73
S_t (2015) = 104
2 Pooled population
S = 177
3 Repeat random re-distribution 1000 times
S_0 = 73
S_t = 104
4 Re-shuffled populations – different each time (1000 random populations)
5 Calculate F_a 1000 times and plot
Frequency
0
10
20
30
40
50
60
70
Null F_a

Figure 3. Schematic diagram of the process of a randomisation test of the F_a statistic. The two original temporal mosquito population samples taken in 2013 and 2015 have a size of 73 (S_0) and 104 (S_t) individuals, respectively. For the randomisation test, the mosquitoes from both populations are pooled into one, shuffled and separated into two populations of size 73 and 104. The F_a statistic for that population is calculated according to equation (3) (this will be called the null F_a). This is repeated 1000 times (calculating F_a for each random shuffled population) and a frequency

histogram of the null F_a is plotted to observe the distribution. Refer to Good (1994) for detail on randomisation test.

4. Results

4.1 Estimation from nucleotide diversity (π)

Once the sites for the 2013 and 2015 samples of Niono, Mali were filtered for non-missing fixed and polymorphic sites, nucleotide diversity (π) was calculated for the whole of chromosome X and 3R and from either intergenic or 4-fold degenerate coding sites (4-CDS) according to equation (1). Note that a sample size of 70 from each year was used to allow a fair comparison of diversity. Using the π values, N_e was estimated using equation (2) (table 1a).

Table.1a Summary of the historical N_e estimates from nucleotide diversity (π).

Sample Year	Number of individuals	Genomic Region	No. of polymorphic sites used		No. of fixed sites		Polymorphic/ fixed ratio		Nucleotide diversity (π) (per bp)					Historical N_e	
			Intergenic	4-CDS	Intergenic	4-CDS	Intergenic	4-CDS	Intergenic (95% CI)	4-CDS (95% CI)	Intergenic/ 4-CDS ratio	X/3R ratio		Intergenic (95% CI)	4-CDS (95% CI)
												Intergenic	4-CDS		
2013	70	3R	1,817,338	139,780	8,274,762	425,694	0.2203	0.3284	0.0136 (0.0132 - 0.0141)	0.0240 (0.0229 - 0.0250)	0.5667	0.59	0.33	1,214,286 (1,178,571 - 1,258,929)	2,142,857 (2,044,643 - 2,232,143)
		X	494,335	37,452	3,332,169	211,839	0.1484	0.1768	0.0080 (0.0075 - 0.0086)	0.0078 (0.0074 - 0.0083)	1.0256			714,286 (669,643 - 767,857)	696,429 (660,714 - 741,071)
2015	70	3R	1,810,391	137,811	8,283,562	427,972	0.2186	0.3220	0.0136 (0.0131 - 0.0140)	0.0239 (0.0229 - 0.0250)	0.5690	0.59	0.33	1,214,286 (1,169,643 - 1,250,000)	2,133,929 (2,044,643 - 2,232,143)
		X	485,906	36,436	3,343,024	213,063	0.1453	0.1710	0.0080 (0.0074 - 0.0086)	0.0078 (0.0074 - 0.0082)	1.0256			714,286 (660,714 - 767,857)	696,429 (660,714 - 732,143)

Note: The calculations are computed for *Anopheles coluzzii* samples taken in 2013 and 2015 in Niono, Mali. Results are computed for combinations of intergenic and 4-CDS sites with chromosomes X and 3R. A sample size of 70 individuals from each year were used for the calculations to allow fair comparison between the years. The number of fixed and polymorphic (biallelic and multiallelic) sites provided do not include missing or inaccessible sites. The 95% confidence interval (CI) for the π and N_e values (given in the brackets) were calculated using a block jackknife resampling procedure of 200 re-samples (see further in methods).

Table 1a shows no significant difference in π between the years (e.g., intergenic sites in 3R between 2013 and 2015) (overlap in the CIs). However, we observe a significant difference between the 3R and X chromosomes. For the same type of sites (e.g., intergenic), the X chromosome has a significantly lower genetic diversity than 3R as we expected. Whilst we predicted a 3R:X ratio of 1:0.75 (Hammer et al., 2008), our observed ratios was far lower - intergenic and 4-CDS sites gave a 3R:X ratio of 1:0.59 and 1:0.33, respectively (for both years). Lastly, we note a significant difference in the results using intergenic or 4-CDS sites. π ratio between intergenic and 4-CDS for chromosome 3R shows an almost half the diversity using intergenic compared to 4-CDS. Interestingly, this does not apply for chromosome X (the ratio is close to 1).

A sliding window analysis was performed to test for heterogeneity in diversity of the chromosomes, this can be seen in figure 4 below. As the 2013 and 2015 sample showed no significant differences in the table 1a, the samples were pooled together for this analysis as it allows us to increase our sample size.

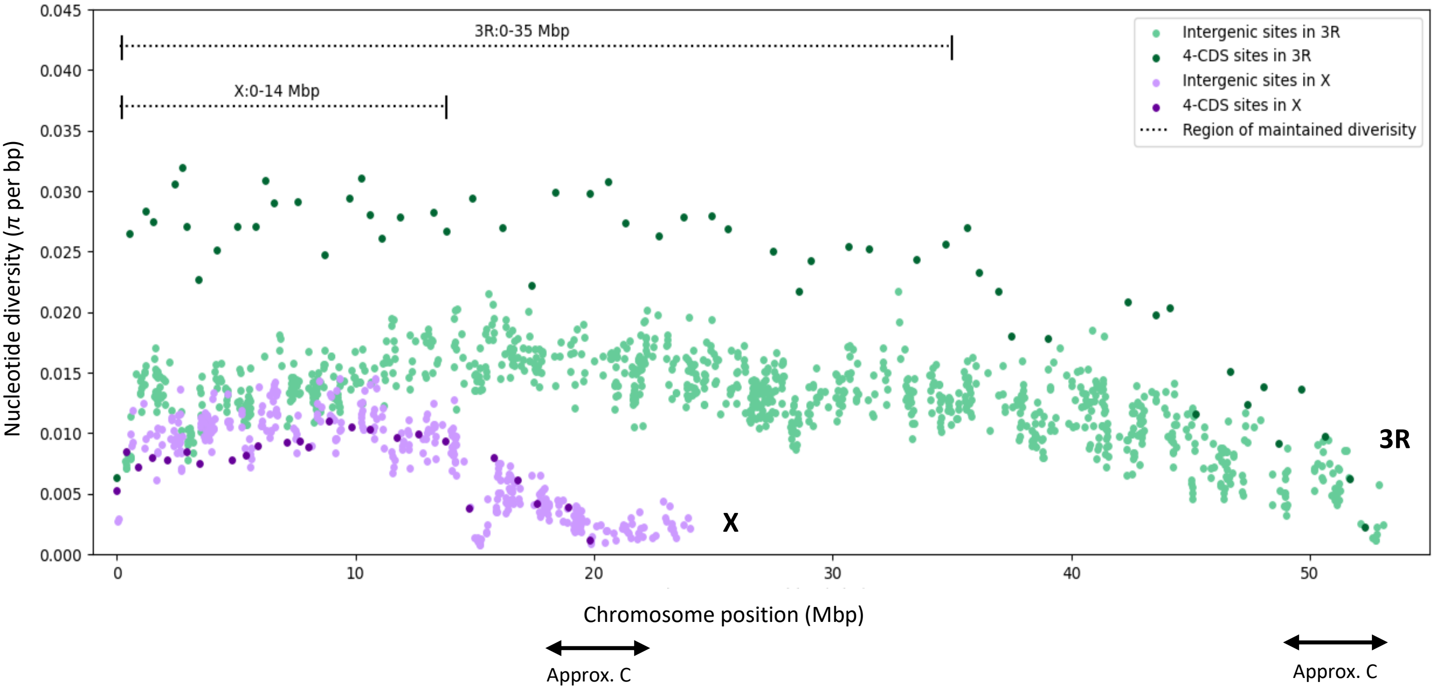


Figure 4. The heterogeneity of nucleotide diversity (*π*) of chromosomes 3R and X calculated using intergenic and 4-CDS sites. Nucleotide diversity values (per bp) were calculated using non-overlapping sliding windows of 10,000 sites. The start position of each window has been plotted. A total of 177 individuals were used, pooled from the 2013 (73 individuals) and 2015 (104 individuals) samples (see table 1a). Diversity starts to drop after 14 Mbp for chromosome X and after 35 Mbp for chromosome 3R (for both site types). The approximate centromeric regions (Sharakhova et al., 2010) is shown as black arrows for both chromosomes.

The sliding window analysis shows the same pattern as for the overall *π* from table 1 – the 4-CDS sites on chromosome 3R shows almost double the diversity compared to intergenic sites, but no such difference for chromosome X. Furthermore, we detect a decrease in diversity towards the centromere for each chromosome.

Due to the heterogeneity in diversity across the chromosome, we decided to re-calculate *π* and N_e (as for table 1a) but only using the chromosome positions of constant diversity (shown in figure 4) to obtain a more accurate estimate not skewed by the drop in diversity towards the centromeres. The results are displayed in table 1b below.

Table.1b Summary of the historical *N_e* estimates from nucleotide diversity (π) using regions 0-14 Mbp for chromosome X and 0-35 Mbp for chromosome 3R.

Sample Year	Number of individuals	Genomic Region	No. of polymorphic sites used		No. of fixed sites		Polymorphic /fixed ratio		Nucleotide Diversity (π)					Historical N_e	
			Intergenic	4-CDS	Intergenic	4-CDS	Intergenic	4-CDS	Intergenic (95% CI)	4-CDS (95% CI)	Intergenic /4-CDS ratio	X/3R ratio Intergenic	 4-CDS	Intergenic (95% CI)	4-CDS (95% CI)
2013	70	3R: 0-35 Mbp	1,390,264	111,111	5,536,574	284,842	0.2511	0.3901	0.0151 (0.0148 - 0.0154)	0.0271 (0.0263 - 0.0279)	0.5572	0.73	0.32	1,348,214 (1,321,429 - 1,375,000)	2,419,643 (2,348,214 - 2,491,071)
		X: 0-14 Mbp	374,052	32,204	1,695,956	152,688	0.2206	0.2109	0.0110 (0.0108 - 0.0113)	0.0088 (0.0085 - 0.0091)	1.2500			982,143 (964,286 - 1,008,929)	785,714 (758,929 - 812,500)
2015	70	3R: 0-35 Mbp	1,385,308	109,401	5,547,244	286,749	0.2497	0.3815	0.0150 (0.0147 - 0.0153)	0.0270 (0.0262 - 0.0278)	0.5556	0.73	0.32	1,339,286 (1,312,500 - 1,366,071)	2,410,714 (2,339,286 - 2,482,143)
		X: 0-14 Mbp	369,607	31,536	1,700,039	153,494	0.2174	0.2055	0.0110 (0.0107 - 0.0113)	0.0087 (0.0084 - 0.0091)	1.2644			982,143 (955,357 - 1,008,929)	776,789 (750,000 - 812,500)

Note: The calculations are computed for *Anopheles coluzzii* samples taken in 2013 and 2015 in Niono, Mali. Results are computed for combinations of intergenic and 4-CDS sites with the regions of maintained diversity in chromosomes X and 3R, observed from the sliding window analysis (figure 4). A sample size of 70 individuals from each year were used for the calculations to allow fair comparison between the years. The number of fixed and polymorphic (biallelic and multiallelic) sites provided do not include missing or inaccessible sites. The 95% confidence interval for the π and N_e values (given in the brackets) were calculated using a block jackknife resampling procedure of 200 re-samples (see further in methods).

Here, we observe a slight increase in the nucleotide diversities and N_e for all calculations as one would expect; and similar conclusions deduced from table 1 and the sliding window analysis are reached in table 1b too. However, the observed 3R:X ratio using intergenic sites increased to 1:0.73 for both samples, bringing it closer to the theoretical expectation of 1:0.75. No difference was observed for the 4-CDS 3R:X ratio.

4.2 Estimation of contemporary N_e

Compared to the diversity sites, only biallelic SNPs and SNPs with MAF ≥10% in the 2013 samples and the corresponding SNPs in the 2015 samples were included. Additionally, 1 SNP per 1000bp was randomly selected from the 2013 sample, and again, the corresponding SNPs were selected from the 2015 samples to avoid linked SNPs. A measure of genetic drift, temporal F_a statistic, was estimated from the allele frequencies of the major allele of the remaining SNPs using equation (3). Contemporary N_e was estimated from this using equation (4). The results are shown in table 2 below.

Table.2 Summary of the contemporary N_e estimates from temporal F_a statistic.

Genomic Region	Number of individuals		Number of sites *(K)*		F_a		Sampling error $(\frac{1}{2S_0}+\frac{1}{2S_t})$	Contemporary N_e	
	S_0	S_t	Intergenic	4-CDS	Intergenic (95% CI)	4-CDS (95% CI)		Intergenic (95% CI)	4-CDS (95% CI)
3R	73	104	13,580	5,170	0.01126 (0.01048 - 0.01206)	0.01124 (0.01023 - 0.01234)	0.01166	Inf (24,666 - Inf)	Inf (14,648 - Inf)
X	73	104	4,677	1,238	0.01189 (0.01064 - 0.01326)	0.01114 (0.00957 - 0.01301)	0.01166	43,193 (6,241 - Inf)	Inf (7,386 - Inf)

Note: The calculations are computed for *Anopheles coluzzii* samples taken in 2013 and 2015 in Niono, Mali (which corresponds to *t* = 20 generations ((O'Loughlin et al., 2016)). The sample size for 2013 and 2015 are represented by S_0 and S_t, respectively. Results are computed for combinations of intergenic and 4-CDS sites with chromosomes X and 3R. The number of sites (*K*) provided do not include missing or inaccessible sites and have been filtered to be biallelic and exclude sites with a minor allele frequency (MAF) <10% in the 2013 sample from both the 2013 and 2015 samples. The 95% confidence interval for the F_a and N_e values (given in the brackets) were calculated according to Hui, Brenas & Burt, (2021) (refer to methods) and using the published recombination rates of 1.1 cM/Mb and 3.2 cM/Mb for chromosome X and 3R, respectively (Zheng et al., 1996). The sampling error is calculated according to the equation shown and part of equation (3) to calculate N_e. Infinity values have been shortened to Inf.

Point F_a estimates were calculated from the intergenic and 4-CDS SNPs for each of the chromosomes. As you can see in table 2, however, the sampling error (based on the sample sizes of both temporal samples) was larger than the F_a for all combinations except for intergenic sites within the X chromosome. This means that the F_a minus sampling error (as in equation 4) results in a negative F_a, which suggests that the drift was too weak to be detected and thus gives an approximation of infinite population size. Note that when the difference between the F_a and the sampling error is close to zero, N_e becomes very sensitive, and it is not uncommon to have wide confidence intervals. The N_e estimated from the intergenic SNPs in X was 43,193.

The 95% confidence intervals estimated for F_a was able to provide the upper bound values of F_a which was larger than the sampling error for all combinations. Using this, non-infinity lower bound estimations of N_e was calculated for the other three combinations and this ranged from 7,386 – 24,666. Though the confidence intervals overlap (due to infinity values), looking at only the lower bound values for all four, the N_e estimates from chromosome X are lower than 3R as expected (and as observed with the diversity results).

We suspected that one of the reasons for the negative F_a detected was due to the 2013 and 2015 samples belonging to one population, not two temporally separate populations as we had imagined. This would mean that the measure of genetic drift is zero as drift is an accumulation of the changes of allele frequencies over time (and generations) but here, we would only have one population taken at a single timepoint. Therefore, there was no change in time for there to be a change in allele frequencies i.e., no drift to be detected. To test this, we conducted randomisation test (as described in figure 3). The resulting histograms are displayed in figure 5 below.

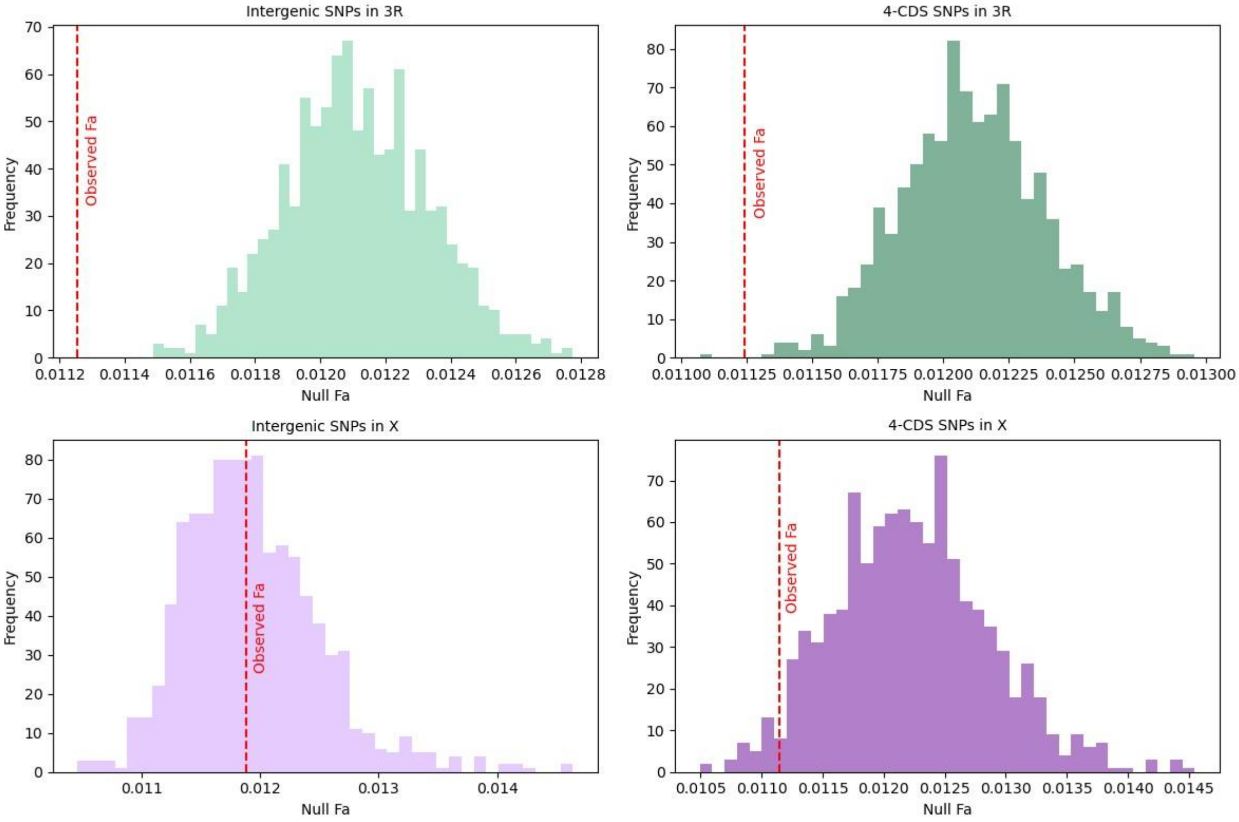


Figure 5. Frequency histograms of the null *F_a* generated from the randomisation test. The aim of the randomisation test is to test whether the observed F_a (calculated from our samples) differ from a random distribution of null F_a calculated from pooling our 2013 and 2015 samples into one population, randomly re-distributing into two samples (of size 73 and 104 to match the size of the original temporal samples) and calculating F_a from them. The frequency histogram was built from repeating this procedure 1000 times and calculating F_a for each random shuffled population. The observed F_a are labelled on the distributions as a vertical dotted red line for each.

All the observed F_a except for the intergenic SNPs in chromosome 3R fall within the null distribution (figure 5), implying that the 2013 and 2015 populations are one population. While the observed F_a for 4-CDS sites in 3R is within the distribution, it is still towards the left of the distribution and a rarer F_a compared to the others. Interestingly, despite the observed F_a for intergenic sites in X being positive post sampling error (table 3) (implying observable drift), here we see that it falls within the null distribution and implying there is no drift as the temporal samples are in fact one population.

5. Discussion

5.1 Comparing the nucleotide diversity (π) estimates

In section 4.1, we estimated diversity and historical N_e for a 2013 and a 2015 sample of *Anopheles coluzzii* from Niono, Mali. We found no significant difference between the years and this is likely due to 2 years not being a large enough timescale for differences in diversity to accumulate (only 20 generations), so this was expected.

Also, we observed a X/3R ratio of less than the expected 0.75 (Hammer, 2008) for 4-CDS samples (table 1b). This has been reported in previous literature too including Ag1000G (2017) but remains unexplained. Summary of some past literature results have been included in table 3.

Table 3. Summary of the nucleotide diversity (π) and historical N_e estimates from existing literature.

Species	Location	Genomic Region	SNP site type	Number of individuals	Number of sites	Nucleotide Diversity (π) (per bp)	X/3 ratio of π	Historical N_e	Reference
An. coluzzii	Burkina Faso	3R: 0-3.7 Mbp	Intergenic	82	19,587	0.01041	0.87	929,170	Bort (2022)
An. coluzzii	Burkina Faso	X: 0-1.25 Mbp	Intergenic	82	6,218	0.01074		808,107	
An. coluzzii and An. gambiae	Multiple	3	Intergenic	N/A	N/A	0.015	0.80	N/A	Ag1000G (2017)
An. coluzzii and An. gambiae	Multiple	X	Intergenic	N/A	N/A	0.012		N/A	
An. coluzzii and An. gambiae	Multiple	3	4-CDS	N/A	N/A	0.032	0.31	N/A	
An. coluzzii and An. gambiae	Multiple	X	4-CDS	N/A	N/A	0.010		N/A	

Note that Bort (2022) used the same method as this work with the adjustment of different regions of the chromosome of maintained diversity and using fixed and biallelic sites. The Ag1000G (2017) authors calculated only π (using multiallelic and fixed sites). They combined both *An. coluzzii* and *An. gambiae* species, and multiple locations which can be found on Ag1000G (2017).

The diversity (π) of our *An. coluzzii* 2013 and 2015 samples ranges from 0.0110 – 0.0271 per base pair (see table 1b). Previous studies for *Anopheles* (see table 3) show similar π values (apart from a π of 0.032 per base pair from Ag1000G (2017)). However, caution is needed when directly comparing these values as there are differences in the spatial scale (town vs multiple countries) and types of species used to calculate them. Nonetheless, we can see a trend in which the X/3 ratio for diversity calculated using intergenic sites meets the theoretical expectation of 0.75 (apart from the 0.87 which exceeds it) while those calculated using 4-CDS sites are below the expected. A potential reason for the lower than anticipated diversity in X could be due to sex-biased evolutionary forces which act on allosomes e.g., unequal sex ratios or sex-biased migration (Hammer et al., 2008). For example, if the sex ratios are male-biased (more males than females, then the number of X chromosomes in the population would be lower, resulting in decreased X-linked diversity (Hammer et al., 2008). Contrastingly, a female-biased population will exhibit an increased X-linked diversity (Hammer et al., 2008), as could be the case with Bort (2022)'s π ratio of 0.87 per base pair.

Another question of interest is why this below than expected X/3 ratio is only observed for 4-CDS sites. This is likely due to an almost double in diversity measured from 4-CDS sites compared to intergenic, but only for chromosome 3/3R, as seen according to our results and Ag1000G (2017). Doubling the denominator results in a reduction of the X/3 ratio by half, thus showing a ratio closer to 0.375 (half of the expected 0.75). Nonetheless, we need to investigate the underlying cause for this increased diversity using 4-CDS in chromosome 3. Though intergenic sites are non-coding sites between genes and 4-CDS are coding sites within the genes, they are both known to be neutral sites. While under the SNM, they should be completely neutral, this assumption is less likely to be upheld in a real population. Other literature including Lynch et al. (2017) and Martin et al. (2016) observed a similar discrepancy in *Daphnia pulex* and *Heliconius melpomene*, respectively. The intergenic/4-CDS diversity ratios were 0.49 and 0.8, respectively, indicating a higher nucleotide diversity in 4-CDS. There is evidence in *Drosophila* that shows non-coding sequences including intergenic regions are more constrained than 4-CDS due to containing functional elements like regulatory and non-coding RNA genes which experience selection and thus less diverse (Casillas, Barbadilla & Bergman, 2007). However, this should apply to both chromosomes and does not explain the higher diversity only in chromosome 3 and not X.

Alternatively, this could be due to recombination differences between 4-CDS and intergenic sites. It is known that recombination decreases towards the centromere, and lower recombination is linked with decreased diversity (Langley et al., 2012), as seen in the sliding window analysis (see figure 4). As the deviation between diversity measured from 4-CDS and intergenic sites decreases towards the centromere for both chromosomes, this leads us to suspect the difference is related to recombination differences. Pombi et al., (2006) states a lower recombination rate in chromosome X is found compared to chromosome 3 (1.1cM/Mb vs 3.2cM/Mb). It is possible that there is additionally differing recombination rate between 4-CDS and intergenic sites. If intergenic sites have lower recombination, it would have lower diversity. Given more time, this could have been tested by plotting a measure of recombination rate/linkage disequilibrium across the chromosomes for different site types.

5.2 Comparing the temporal F_a estimates

In section 4.2, we estimated point F_a for each combination of chromosomes and site types. Surprisingly, we found that only intergenic SNPs within chromosome X gave us a positive F_a, meaning genetic drift was too weak to be detected for the others. One potential cause for this is our sample sizes for the 2013 and 2015 samples were too small. For the same F_a, the smaller the sample size, the more likely F_a is negative as the sampling error becomes larger than the F_a (refer to equation 4). However, Bort (2022) and Hui, Brenas & Burt, (2021) had smaller samples sizes than here ($S_0 = 73$, $S_t = 104$) and calculated some positive F_a values which can be found in table 4.

Table 4. Summary of the temporal *F_a* statistic and contemporary *N_e* estimates from existing literature.

Species	Location	Genomic Region	Site type	S_0	S_t	t	K	F_a	Sampling error$^{\$}$	Contemporary N_e	X/3R ratio	Reference
An. coluzzii	Burkina Faso	3R: 0-3.7 Mbp	Intergenic	82	53	20	17,970	0.0141 (0.0133 - 0.0149)	0.0155	Inf* (Inf* - Inf*)	N/A	Bort (2022)
An. coluzzii	Burkina Faso	X: 0-1.25 Mbp	Intergenic	82	53	20	5,561	0.0169 (0.0154 - 0.0185)	0.0155	7,302 (3,408 - 50,094)		Bort (2022)
An. coluzzii	Burkina Faso	3R	Intergenic	82	53	20	17,837	0.0166038	0.0155	9,325 (5,314 - 36,748)	N/A	Hui, Brenas & Burt (2021)

Note that Bort (2022) used the same method as this work with the adjustment of different chromosome regions of constant recombination, a different filtering of MAF <5% at both timepoints as with Hui, Brenas & Burt (2021), and calculating the 95% CI with the recombination rates of 1.0 cM/Mb and 1.6 cM/Mb for chromosome X and 3R, respectively. Hui, Brenas & Burt (2021) used a recombination rate of 1.4 cM/Mb. $^{\$}$Sampling error was calculated according to $\frac{1}{2S_0} + \frac{1}{2S_t}$ (refer to equation 4). *Infinity values have been used where a negative *N_e* value was reported (calculated from a negative *F_a* post sampling error).

Interestingly, though Bort (2022) and Hui, Brenas & Burt (2021) calculated F_a for the same populations (*An. coluzzii* 2012 and 2014 samples taken in Burkina Faso), had the same sample sizes and similar number of SNPs, only Bort (2022) reported a negative F_a from intergenic SNPs in chromosome 3R. One reason for this discrepancy could be due to changes to the designation of intergenic SNPs in the Ag1000G dataset (Ag1000G, 2021), post Hui, Brenas & Burt's publication in (2021), leading to Bort (2022) using a potentially different set of SNPs. Alternatively, the particular set of SNPs chosen in Bort (2022), which was selected in part by randomly choosing 1 SNP per 1000bp to avoid linkage disequilibrium (as was done in this work), had a lower F_a estimate (less difference in the allele frequencies between SNPs at the two temporal timepoints). This could be investigated by re-selecting and re-computing F_a values numerous times (e.g., 1000 repeats) and viewing the distribution of in a F_a histogram. If there is a lot of variation e.g., ranges from positive and negative F_a values post sampling error, it can be concluded the particular result found in Bort (2022) was due to that particular set of SNPs. F_a was briefly re-computed in this manner a few times for our samples except intergenic SNPs in chromosome X (i.e., for the negative F_a values) and we found repeated negative F_a values each time (results not shown). Another reason for low F_a values could be due to the way in which both Bort (2022) and Hui, Brenas & Burt (2021) filter to remove SNPs with rare alleles. They both excluded SNPs which had a minor allele frequency (MAF) <5% at either timepoints. However, for those SNPs which had a starting allele frequency of 5%, this filtering will in effect remove all SNPs which decreased in frequency by the second timepoint, but keep those which would have increased. Therefore, we would detect an overall increase in allele frequency and a reduced variance in the change in allele frequencies, leading to a smaller F_a. Alternatives to this is to have a stronger restriction on the first temporal sample e.g., exclude SNPs with a MAF of <10% and no restrictions on the second temporal sample (as was done here); or to use a weighted average MAF of <5% across both timepoints.

Our negative F_a results prompted us to ask whether the 2013 and 2015 samples in fact belong to a single population i.e., not two temporally separate populations, by conducting a randomisation test. We found that all the observed F_a except for the intergenic SNPs in chromosome 3R fall within the null distribution (though 4-CDS SNPs in chromosome 3R is still rarer than most null F_a); even for intergenic SNPs in X despite having a positive point F_a

post sampling size originally. This is likely because the upper bound F_a for intergenic SNPs in X is negative, and thus overlaps with the null F_a distribution (which indicates zero drift). This result suggests that the 2013 and 2015 samples are from the same population and that drift is insignificant. However, we were able to obtain finite N_e values through the lower bound CI for each combination (highest possibly drift), and thus we cannot fully rule out the possibility that they are from separate populations. When the difference between the F_a and the sampling error is close to zero, N_e becomes very sensitive, and it is not uncommon to have wide confidence intervals as was observed here. Furthermore, looking at the histograms (in figure 5), the observed F_a from chromosome 3R lies much further to the left than chromosome X and even outside the null distribution for intergenic SNPs. This would alter the interpretation/conclusion on the significance of drift compared to the X chromosome, despite expecting the same result regardless of chromosome and site types. Possibly, given the sample size (same for both), it can be that we can finite N_e for X but not autosomes, and this is why they give different interpretations. Finally, despite our efforts, we could not find any potential explanations for why the observed F_a for intergenic SNPs in chromosome 3R lies to the left of the null F_a i.e., implying a lower drift than that of no drift. This does not seem right and is of interest to be studied in the future.

5.3 Comparing historical and contemporary N_e estimates

According to our historical N_e estimates, the effective population size of *An. coluzzii* for Niono, Mali in 2013 and 2015 ranged from 7.5×10^5 - 1.0×10^6 for X chromosome and 1.3×10^6 to 2.5×10^6 for chromosome 3R (including the 95% CI for both) across both site types (table 2). In comparison, the contemporary effective population size ranged from 6.2×10^3 – 4.3×10^4 and 1.4×10^4 – 2.5×10^4 for chromosome X and 3R, respectively (omitting the infinity values i.e., using the 95% CI lower bound values and point N_e estimate for intergenic SNPs in X) (table 3). Our expectation was that under the SNM, both estimates would be the same. Clearly, this was not the case and it seems contemporary N_e estimates are smaller than historical N_e estimates. This implies that our populations are not 'ideal' and break at least one of the assumptions of the SNM model – a constant sized panmictic population with no selection nor migration (Kimura, 1983). It is likely that most of these assumptions are broken. For example, there is evidence that *Anopheles* mosquitoes take part in migration (Ag1000G (2017); Ag1000G (2020)) and thus the nucleotide diversity calculated would reflect this (increased diversity from gene flow of neighbouring populations) and give a higher historical N_e. Wang & Whitlock (2003) also state that migration can greatly bias contemporary N_e estimate as in the short-term, migration will show a signal similar to high drift (alters allele frequencies more) and thus leads to an underestimation of N_e. Moreover, it is unlikely for a population to have remained a constant size throughout history e.g., population bottlenecks. Therefore, diversity-based methods may underestimate the current N_e if there has been a population bottleneck in the past (reduces diversity severely even following recovery of population size/population expansion) (Charlesworth, 2009; Khatri & Burt, 2019). There are other methods of estimating effective population sizes which do not use the same assumptions e.g., jointly inferring N_e and migration rates (Wang & Whitlock, 2003) or inferring N_e from the number of independent lineages following a partial soft sweep (rise in beneficial mutations across multiple but not all independent lineages) as seen in Khatri & Burt (2019) which can give a more accurate measure of the current day N_e. Khatri & Burt (2019) estimated N_e using the Ag1000G dataset too, and calculated an estimate of 6.2×10^7 (95% CI 2.7×10^7, 1.2×10^8). This is larger than the estimates calculated here but comparisons should be approached with caution as there are differences in the temporal and spatial context between our methods, and each method has different assumptions. For example, Khatri and Burt (2019) provide a continent-wide N_e estimate, but state that under the assumption of the a panmictic continent, regional estimates should match the continental estimate. Nonetheless, here, we aimed to estimate effective population sizes under the SNM (i.e., purely theoretical). The differences between the contemporary and historical N_e thus allows us to infer how our populations deviate from the standard neutral model.

5.4 Conclusions and future directions

To conclude, in this work, we have estimated contemporary and historical N_e using genomic data from genetic drift and nucleotide diversity, respectively from *An. coluzzii* samples taken in Niono, Mali in 2013 and 2015. We have also investigated the differences in the estimates between the X and 3R chromosome as well as using intergenic or 4-fold degenerate coding sites (4-CDS). Despite our expectations, we found our population to deviate from the standard neutral model (SNM). It could be of interest to study in the future how our population here deviates from the assumptions of the SNM. Moreover, we were unable to explain the differences in diversity measures caused by using either 4-CDS or intergenic sites (despite both being considered neutral sites). We suspect recombination differences between the sites causing a decrease in estimates from intergenic sites (lower diversity caused by lower recombination) but we were unable to determine this given the time constraints on this project. Lastly, we found surprising results during the randomisation test in which the drift detected for the intergenic SNPs in chromosome 3R is below that of the drift if the 2013 and 2015 samples belonged to the same population (i.e., equivalent to no drift). This did not seem correct to us and should be studied in the future.

6. References

Antao, T., Pérez-Figueroa, A. & Luikart, G. (2011) Early detection of population declines: high power of genetic monitoring using effective population size estimators. *Evolutionary Applications.* 4 (1), 144-154. 10.1111/j.1752-4571.2010.00150.x.

Bort, J. L. (2022) *Comparing contemporary and historical effective population size in Anopheles coluzzii and Anopheles gambiae using autosomes and sex chromosomes*. Undergraduate final year project thesis. Imperial College London

Casillas, S., Barbadilla, A. & Bergman, C. M. (2007) Purifying selection maintains highly conserved noncoding sequences in Drosophila. *Molecular Biology and Evolution.* 24 (10), 2222-2234. 10.1093/molbev/msm150.

Charlesworth, B. (2009) Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics.* 10 (3), 195-205. 10.1038/nrg2526.

Epopa, P. S., Millogo, A. A., Collins, C. M., North, A., Tripet, F., Benedict, M. Q. & Diabate, A. (2017) The use of sequential mark-release-recapture experiments to estimate population size, survival and dispersal of male mosquitoes of the Anopheles gambiae complex in Bana, a west African humid savannah village. *Parasites & Vectors.* 10 (1), 376. 10.1186/s13071-017-2310-6.

Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer.

Hammer, M. F., Mendez, F. L., Cox, M. P., Woerner, A. E. & Wall, J. D. (2008) Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genetics.* 4 (9), e1000202. 10.1371/journal.pgen.1000202.

Hui, T. J., Brenas, J. H. & Burt, A. (2021) Contemporary Ne estimation using temporally spaced data with linked loci. *Molecular Ecology Resources.* 21 (7), 2221-2230. 10.1111/1755-0998.13412.

Hui, T. J. & Burt, A. (2015) Estimating effective population size from temporally spaced samples with a novel, efficient maximum-likelihood algorithm. *Genetics.* 200 (1), 285-293. 10.1534/genetics.115.174904.

Institut National de la Statistique du Mali (INSTAT). (n.d.) *Region de Segou*. https://web.archive.org/web/20110722215805/http://instat.gov.ml/documentation/segou.pdf [Accessed 3rd June 2023].

Keightley, P. D., Ness, R. W., Halligan, D. L. & Haddrill, P. R. (2014) Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family. *Genetics.* 196 (1), 313-320. 10.1534/genetics.113.158758.

Khatri, B. S. & Burt, A. (2019) Robust Estimation of Recent Effective Population Size from Number of Independent Origins in Soft Sweeps. *Molecular Biology and Evolution.* 36 (9), 2040-2052. 10.1093/molbev/msz081.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press

Langley, C. H., Stevens, K., Cardeno, C., Lee, Y. C. G., Schrider, D. R., Pool, J. E., Langley, S. A., Suarez, C., Corbett-Detig, R. B., Kolaczkowski, B., Fang, S., Nista, P. M., Holloway, A. K., Kern, A. D., Dewey, C. N., Song, Y. S., Hahn, M. W. & Begun, D. J. (2012) Genomic variation in natural populations of Drosophila melanogaster. *Genetics.* 192 (2), 533-598. 10.1534/genetics.112.142018.

Luikart, G., Ryman, N., Tallmon, D. A., Schwartz, M. K. & Allendorf, F. W. (2010) Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conservation Genetics.* 11 355-373.

Lynch, M., Gutenkunst, R., Ackerman, M., Spitze, K., Ye, Z., Maruki, T. & Jia, Z. (2017) Population Genomics of Daphnia pulex. *Genetics.* 206 (1), 315-332. 10.1534/genetics.116.190611.

Martin, S. H., Möst, M., Palmer, W. J., Salazar, C., McMillan, W. O., Jiggins, F. M. & Jiggins, C. D. (2016) Natural Selection and Genetic Diversity in the Butterfly Heliconius melpomene. *Genetics.* 203 (1), 525-541. 10.1534/genetics.115.183285.

Nei, M. & Li, W. H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America.* 76 (10), 5269-5273. 10.1073/pnas.76.10.5269.

Pombi, M., Stump, A. D., Della Torre, A. & Besansky, N. J. (2006) Variation in recombination rate across the X chromosome of Anopheles gambiae. *The American Journal of Tropical Medicine and Hygiene.* 75 (5), 901-903.

Sharakhova, M. V., George, P., Brusentsova, I. V., Leman, S. C., Bailey, J. A., Smith, C. D. & Sharakhov, I. V. (2010) Genome mapping and characterization of the Anopheles gambiae heterochromatin. *BMC Genomics.* 11 459. 10.1186/1471-2164-11-459.

The Anopheles gambiae 1000 Genomes Consortium (Ag1000G). (2021) Ag1000G phase 3 SNP data release. MalariaGEN. https://www.malariagen.net/data/ag1000g-phase3-snp.

The Anopheles gambiae 1000 Genomes Consortium (Ag1000G). (2020) Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii. *Genome Research.* 30 (10), 1533-1546. 10.1101/gr.262790.120.

The Anopheles gambiae 1000 Genomes Consortium (Ag1000G). (2017) Genetic diversity of the African malaria vector Anopheles gambiae. *Nature.* 552 (7683), 96-100. 10.1038/nature24995.

Wang, J., Hill, W. G., Charlesworth, D. & Charlesworth, B. (1999) Dynamics of inbreeding depression due to deleterious mutations in small populations: mutation parameters and inbreeding rate. *Genetical Research.* 74 (2), 165-178. 10.1017/s0016672399003900.

Wang, J. & Whitlock, M. C. (2003) Estimating effective population size and migration rates from genetic samples over space and time. *Genetics.* 163 (1), 429-446. 10.1093/genetics/163.1.429.

Waples, R. S. (1989) A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics.* 121 (2), 379-391. 10.1093/genetics/121.2.379.

Willis, K. & Burt, A. (2021) Double drives and private alleles for localised population genetic control. *PLoS Genetics.* 17 (3), e1009333. 10.1371/journal.pgen.1009333.

World Health Organisation. (2022) *World malaria report 2022.* https://www.who.int/publications/i/item/9789240064898 [Accessed 3rd June 2023].

Wright, S. (1931) Evolution in Mendelian Populations. *Genetics.* 16 (2), 97-159. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1201091/.

Zheng, L., Benedict, M. Q., Cornel, A. J., Collins, F. H. & Kafatos, F. C. (1996) An integrated genetic map of the African human malaria vector mosquito, Anopheles gambiae. *Genetics.* 143 (2), 941-952. 10.1093/genetics/143.2.941.