Estimating and interpreting heterozygosity and effective population size of parasites and vectors: caveat emptor!

Tiago Antao[*][†], Andrés Pérez-Figueroa[‡],

Ian M. Hastings[†], Martin J. Donnelly[†], Gordon Luikart[§]

Abstract

Expected heterozygosity and effective population size are crucial parameters related to parasite and vector population viability and are increasingly used in parasitology, with potential applications to monitor and evaluate control and elimination policies. Here we make the argument that long-term N_e and heterozygosity estimation do not provide a reliable monitoring strategy for evaluating the short-term efficacy of control and elimination measures and that current research might be incorrectly using such estimators. We suggest that contemporary based N_E estimation might be used to evaluate control and elimination measures as long as the genetic sampling is sizable enough. We provide suggestions on how to obtain a more precise and unbiased estimate of effective population size. We argument that *Plasmodium falciparum* cannot use Linkage Disequilibrium based estimators due to its mating strategy which includes selfing.

[*] Corresponding author. E-mail: Tiago.Antao@liverpool.ac.uk

[†] Address: Liverpool School of Tropical Medicine, Liverpool, L3 5QA, UK.

[‡] Departamento de Bioquímica, Genética e Inmunología. Facultad de Biología. Universidad de Vigo. 36310 Vigo, Spain.

[§] Fish and Wildlife Genomics Group, Flathead Lake Biological Station and Division of Biological Sciences, University of Montana, Polson, MT 59860-6815, USA

1 Introduction

Effective population size (N_e) determines the rate of genetic change in the composition of a population caused by genetic drift (Charlesworth, 2009), and the relative efficiency and natural selection in the face of drift. N_e has many biological applications, being a fundamental factor in determining population viability as larger N_e entails larger population genetic variability which is paramount for species survival and adaptation. In conservation genetics it is a fundamental and widely used parameter to define and implement management actions in order to protect threatened species (Luikart et al., 2010). N_e is also a commonly reported parameter in population genetic studies of *Plasmodium falciparum* at least since Anderson et al. (2000a) and of *Anopheles Gambiae* starting from Lehmann et al. (1998).

Estimating N_e is mainly done through two different approaches: Long-term estimation based on silent nucleotide site diversities (Hughes and Verra, 2001) or heterozygosity (**?**) assuming mutation-drift equilibrium and that mutation rate can be estimated. If populations are small (typically in conservation genetics scenarios, but also for many parasites and vectors with N_e estimates below 1,000) N_e can also be estimated for contemporary values from observed changes in allele frequencies between generations (Wang, 2001), linkage disequilibrium patterns (Hill, 1981; Waples, 2006) or heterozygosity loss over generations (Miller and Waits, 2003).

Estimations of N_e in *P. falciparum* are done, using long-term estimation approaches as N_E is quite high and contemporary estimators of N_e are not precise with large N_e values (Waples and Do, 2009). Malaria effective population size and thus its population genetic diversity increases with transmission intensity with estimations varying between 1,000 for low transmission areas in South America and 20,000 in high transmission areas like Central Africa (Anderson et al., 2000a). Estimation is commonly done using an heterozygosity based estimator (GOOD CITE), therefore the proprieties of heterozygosity have direct consequences on this kind of N_E estimation. As far as we know contemporary based estimation was never attempted with *P. falciparum*, though in areas of low transmission (expected low N_E) might allow a precise estimate with contemporary estimators especially if the sampling strategy is large enough.

For vector species both contemporary and long-term estimators have been routinely used (see, e.g., **???**). Indeed the most used contemporary N_E estimator (**?**) was developed to study the impact of insecticide resistance. This method requires two temporal samples and works by computing the variance in allele frequencies over time.

HETEROZYGOSITY TESTS (GORDON!)

PINTO2002 problem long term.

Estimating heterozygosity, short- and long-term N_e has many potential applications from simple estimation of parasite genetic variability on parasites (Anderson et al., 2000a; Iwagami et al., 2009; Susomboon et al., 2008) or vectors (Lehmann et al., 1998), impact of control interventions on parasites (Gatei et al., 2010) and vectors (**???**) or vector seasonality (**?**)

We provide arguments supporting the following key conclusions: 1) We cannot use this estimator in particular, and heterozygosity in general, to detect the effectiveness of control and elimination interventions and 2) Heterozygosity based N_e is not informative of contemporary demographic processes and changes in heterozygosity over time are also not instructive. We also provide guidelines on how to more precisely estimate long-term N_e. Most of our conclusions are generally applicable to all species where N_e is high (i.e. above 1,000).

In order to substantiate and better quantify some of the arguments presented here, we conducted some forward-time individual based population genetics simulations using simuPOP (Peng and Kimmel, 2005). These results, while not novel, provide detailed quantitative evidence for the arguments presented here.

2 The problems with heterozygosity based N_e estimation

The heterozygosity based N_e estimator assumes that in a population in mutation-drift equilibrium, heterozygosity is a function of the product of N_e and the mutation rate μ. In the event of a bottleneck, there will be heterozygosity loss at a rate of $\frac{1}{2N_{E2}}$ (CITE HARTL MAYBE? CHECK), where N_{E2} is the effective population size after the bottleneck event. Heterozygosity loss is a very slow process, indeed assuming from data presented Anderson et al. (2000a) that a high-transmission zone has an N_e of around 18,000 (H_e of 0.8) and a low-transmission zone has an N_e of 1,400 (H_e of 0.4) then a control intervention that would cause a reduction from 18,000 to 1,400 would take 600 generations (more than 100 years) just to approximate an H_E of 0.5. Almost 1,000 generations would be needed to approximate a H_e of 0.4. To put this in perspective if the usage of Chloroquine (introduced around 1947) were to impose a continuous bottleneck of the intensity described above, we would get a resonably precise estimation of the bottleneck by sampling around the year 2020 (this making the benign, but unrealistic, assumption of no mutation). This illustrates that there is no relationship between this estimator of N_e and contemporary demographic processes. In the supplement we present several examples of the impact of bottlenecks on H_E.

These results have direct application on the interpretation of the results presented in (Gatei et al., 2010). These authors compared parasite genetic diversity before and after the introduction of ITNs; the results show that there was no significant difference in expected heterozygosity led the authors to conclude that the population maintained “overall stability in genetic diversity”. (Gatei et al., 2010) computed the expected heterozygosity using eight “neutral” microsatellites before the introduction of ITNs and compared the result with a sampling done 5 years after. This was done in western Kenya, a high-transmission area. ITN use causes a substantial reduction in malaria cases and human deaths (health organization WHO, 2008), thus potentially reducing the size of the parasite population. The authors report an increase in heterozygosity from 0.75 to 0.79. With regards to heterozygosity and the prevalence of mixed infections the authors state: "The stable overall genetic diversity after dramatic reduction in transmission intensity observed in the current study was unexpected by the initial prediction. The counter-intuitive results suggest that other factors may be involved in offsetting the effect of transmission reduction on parasite genetic diversity and/or stabilization of the overall genetic diversity of malaria parasite." It is quite clear that the results are neither "counter-intuitive" nor "unexpected": expected heterozygosity is a slow moving indicator especially in high-transmission areas even when efficient control measures (i.e. imposing strong bottlenecks) are in place. What (Gatei et al., 2010) is observing is expected "artifact noise." In this case the study conclusions are probably over-pessimistic: the ITN intervention might be having a impact on parasite diversity. Similar arguments could be raised for a study (**?**) on *A. gambiae* about the impact of indoor spraying with DDT or several other vector studies which make contemporary inferences using H_e.

Long-term N_E estimation is also influenced by the sample size (i.e. the number of individuals and loci sampled) as it influences precision and bias, but this problem (further detailed in the supplement) is of considerable less importance than contemporary interpretations of long-term N_E estimators. Loci under selection will also bias the estimator and, for instance, in Susomboon et al. (2008) it is suggested that 3 of the 12 microsatellite loci used to estimate N_e had strong genetic differentiation between samples taken in severe and uncomplicated malaria patients, though no formal test for selection was conducted. No known study of long-term estimation, in both parasites and vectors, included any test for selection. Tests to detect these loci should thus be performed and repeated for all datasets: just because a set of loci was documented as neutral in the past, there is no assurance neutrality is maitained.

If large number of alleles are detected at each locus then a larger sample size is required in order to obtain an accurate heterozyogisty estimate. The minimum possible heterozygosity estimate is a function of the number of alleles detected and the sample size and its value can be easily calculated by noting that the minimum heterozygosity value happens when for n alleles and r samples, $n-1$ alleles have a single sample and one single allele has $r-n+1$ samples. We plot, on figure 1 the minimum possible heterozygosity as a function of number of alleles detected and sample size. Having a sample size that is close to the number of observed alleles will thus bias the heterozigosity estimate up.

3 Contemporary estimation of N_E

Contemporary estimation of N_E is mostly used with vectors **???**Lehmann et al. (see e.g. 1998) and no study with malaria parasites is known. Most contemporary studies of N_e use temporal estimators which require two samples over time. Precision of temporal based N_E estimators is lost with large N_E, small temporal spacing between samples and low number of alleles and sample size. Figure XXX shows the impact of these factors in precision. Very high N_E (above 2,000) and low temporal spacing (below 12 generations) are fundamental factors accruing loss of precision. **?** shows that commonly used temporal based methods might not be appriate to detect bottlenecks (normally resulting from control interventions) or seasonality patterns. Temporal based methods are more appropriate to detect averages over a period of time. For early detection of intervetions, the usage of methods based on linkage disequilibrium (LD) was recommended instead.

4 Linkage Disequilibrium and biological assumptions

Malaria biology is known to diverge from standard population genetics' models. For instance, selfing is common especially in low transmission areas (Arnot, 1998). While malaria is not clonal, selfing will impact LD in similar ways. (de Meeűs and Balloux, 2004) shows associations between loci are maintained for several generations in clonal populations. This has two important consequences: contemporary estimators of N_e based on LD (Hill, 1981; Waples, 2006; Waples and Do, 2008) will probably produce errouneous results as LD patterns are maintained over time. Also, any study in the change of multi-locus LD will probably also be slow moving. Therefore it is probable that LD is also of little usage to evaluate the impact of malaria control measures (though, for other parasites without clonal reproduction or vectors, LD might be useful).

XXXX

The bias and accuracy issues that we point out are due to demographic events and sampling strategies. These effects will have to be compounded with the already well known impact of varying mutation rates. Indeed, we took a very benevolent approach to both mutation rate and mutation model as we assumed a constant mutation rate for all loci based on the value commonly used for N_e estimations with *P. falciparum* malaria and a Stepwise Mutation Model (SMM) whereas some loci show patterns of variation which are inconsistent with the SMM (Anderson et al., 2000b). Even with such assumptions, several limitations of this estimator of effective population size became clear.

5 Conclusions and guidelines

Over-interpretation of data is being made with regards to the impact of control and elimination measures. Heterozygosity is not appropriate to detect recent, sudden changes in population size. Heterozigosity based longitudinal analysis will be overly pessimistic as to the impact of control and elimination methods and cannot be used to evaluate contemporary measures of control, elimination and eradication. It is widely known that heterozygosity is a slow moving indicator and theoretical estimates put the loss of heterozygosity at $\frac{1}{2N}$ per generation. This means that the larger the N_C value, the slower heterozygosity is expected to change. Even for values typical in conservation genetics scenarios (i.e. less than 1,000), heterozygosity based estimates of N_e have been shown to be slow moving (**?**). We illustrated, in the supplment, that theoretical expectations hold for typical population sizes of malaria. Even for extreme bottlenecks the N_e estimated after 100 generations (i.e. around 20 years, assuming 5 malaria generations per year) is still closer to the original value than to the post-bottleneck value. Indeed for areas that have had continuous high-transmission, it is not clear at all that estimates of N_e based on heterozygosity can even detect any effect from the introduction of treatment drugs like Chloroquine decades ago. While theoretical predictions and simulation studies mainly done in conservation genetics would easily predict this estimator to be slow moving, it is staggering that some interventions in the distant past, especially in high transmission areas, cannot be detected even today using this estimator.

Contemporary estimators of N_E cannot be used with very high N_E, but for malaria parasites in low transmission areas (where N_E is expected to be below 2,000) or vectors, contemporary N_E estimation might be feasible. Due to reproduction strategies of malaria (which include selfing), methods based on LD are not usable.

Here we provide a set of suggestions in order to help future studies of effective population sizel, heterozygosity and LD for *P. falciparum* malaria or insect vectors:

1. Long-term N_E estimators should never be used to infer contemporary demographic processes. There is little information that can be inferred for the long-term N_E estimator as to the impact of control and elimination measures.

2. Contemporary estimators of N_E cannot be readily applied to *P. falciparum* malaria or any other organisms if N_E is expected to be above 2,000.

3. The number of loci and/or number of individuals sampled should be enough (Details can be seen in e.g. **?**Antao et al. (2010); **?**); **?**); England et al. (2010)). If the number of individuals sampled cannot be increased then sampling more loci can increase estimation accuracy of long-term N_E as long as the ratio between number of different alleles detected and number of samples is enough. This issue is much more serious with contemporary estimators both because precision is more influenced but also sampling more loci is normally less informative than sampling more individuals. With temporal based estimators spacing between samples should be as large as possible. Further research is needed to understand if genotyping thousands of SNPs

(as it is possible with next generation sequencing) can increase precision, especially of contemporary estimators.

4. Loci should be tested for selection as these can bias the estimation. While testing for selection is commonly made for different geographical areas, in the specific case of malaria selection over time for the same population is also relevant due to the impact of human interventions (e.g., drug deployment and subsequent selection for resistance). Many methods developed for geographical selection detection can be easily used to detect selection over time.

5. In the specific case of malaria and due to reproduction with selfing, inferences based on LD (including contemporary N_E estimators based on LD) should be used with great care, most probably avoided.

6. The appropriate contemporary estimator should be correctly choosen for different research problem. The LD estimator is probably better suited for detecting the impact of intervetions and maybe seasonlity, whereas temporal based approaches can be used to have an average over a period of time (though temporal estimators have been developed to detect population fluctuations (**?**), but were never used or evaluated in parasite or vector contexts)

While it would be important to have an accurate and precise estimate of contemporary effective population size for malaria and its vectors, especially in the context of control and elimination measures, it is not clear that that objective is always feasible: methods explicitly targeted for contemporary N_E are only applicable for small populations and the heterozygosity based method cannot reliably estimate contemporary size. We recommend caution in interpreting N_E estimates and suggest that other strategies should be used to assess the impact of control and elimination measures on parasite genetic diversity.

6 Acknowledgements

TA was supported by research grant SFRH/BD/30834/2006 from Fundacao para a Ciencia e Tecnologia, Portugal.

References

T. J. Anderson, B. Haubold, J. T. Williams, J. G. Estrada-F ranco, L. Richardson, R. Mollinedo, M. Bockarie, J. Mokili, S. Mharakurwa, N. French, J. Whitworth, I. D. Velez, A. H. Brockman, F. Nosten, M. U. Ferreira, and K. P. Day. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol Biol Evol*, 17(10):1467–1482, Oct 2000a.

T. J. Anderson, X. Z. Su, A. Roddam, and K. P. Day. Complex mutations in a high proportion of microsatellite loci from the protozoan parasite plasmodium falciparum. *Mol Ecol*, 9(10):1599–1608, Oct 2000b.

Tiago Antao, Andrs Prez-Figueroa, and Gordon Luikart. Early detection of population declines: high power of genetic monitoring using effective population size estimators. *Evolutionary Applications*, pages no–no, 2010. ISSN 1752-4571. URL http://dx.doi.org/10.1111/j.1752-4571.2010.00150.x.

D. Arnot. Unstable malaria in sudan: the influence of the dry season. clone multiplicity of *Plasmodium falciparum* infections in individuals exposed to variable levels of disease transmission. *Trans R Soc Trop Med Hyg*, 92(6):580–585, 1998.

Brian Charlesworth. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205, Mar 2009.

T. de Meeűs and F. Balloux. Clonal reproduction and linkage disequilibrium in diploids:

a simulation study. *Infection, Genetics and Evolution*, 4(4):345–351, 2004. ISSN 1567-1348.

P.R. England, G. Luikart, and R. Waples. Early detection of population fragmentation using linkage disequilibrium estimation of effective population size. *Conservation Genetics*, 11(6), 2010.

W. Gatei, S. Kariuki, W. Hawley, F. ter Kuile, D. Terlouw, P. Phillips-Howard, B. Nahlen, J. Gimnig, K. Lindblade, E. Walker, et al. Effects of transmission reduction by insecticide-treated bed nets (ITNs) on parasite genetics population structure: I. The genetic diversity of Plasmodium falciparum parasites by microsatellite markers in western Kenya. *Malaria Journal*, 9(1):353, 2010.

World health organization WHO. *World malaria report 2008*. WHO, 2008. ISBN 9241563699.

William G. Hill. Estimation of effective population size from data on linkage disequilibrium. *Genetics Research*, 38(03):209–216, 1981.

A. L. Hughes and F. Verra. Very large long-term effective population size in the virulent human malaria parasite plasmodium falciparum. *Proc Biol Sci*, 268(1478):1855–1860, Sep 2001. doi: 10.1098/rspb.2001.1759. URL http://dx.doi.org/10.1098/rspb.2001.1759.

Moritoshi Iwagami, Pilarita T Rivera, Elena A Villacorte, Aleyla D Escueta, Toshimitsu Hatabu, Shin ichiro Kawazu, Toshiyuki Hayakawa, Kazuyuki Tanabe, and Shigeyuki Kano. Genetic diversity and population structure of plasmodium falciparum in the philippines. *Malar J*, 8:96, 2009. doi: 10.1186/1475-2875-8-96. URL http://dx.doi.org/10.1186/1475-2875-8-96.

T. Lehmann, W. A. Hawley, H. Grebert, and F. H. Collins. The effective population size of anopheles gambiae in kenya: implications for population structure. *Mol Biol Evol*, 15(3):264–276, Mar 1998.

G. Luikart, N. Ryman, D.A. Tallmon, M.K. Schwartz, and F.W. Allendorf. Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conservation Genetics*, 11(2):355–373, 2010. ISSN 1566-0621.

C.R. Miller and L.P. Waits. The history of effective population size and genetic diversity in the Yellowstone grizzly (Ursus arctos): implications for conservation. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):4334, 2003.

B. Peng and M. Kimmel. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687, 2005. doi: 10.1093/bioinformatics/bti584.

Pannapa Susomboon, Moritoshi Iwagami, Noppadon Tangpukdee, Srivicha Krusood, Sornchai Looareesuwan, and Shigeyuki Kano. Differences in genetic population structures of plasmodium falciparum isolates from patients along thai-myanmar border with severe or uncomplicated malaria. *Malar J*, 7:212, 2008. doi: 10.1186/1475-2875-7-212. URL `http://dx.doi.org/10.1186/1475-2875-7-212`.

J. Wang. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetics Research*, 78(3):243–257, Dec 2001.

R. S. Waples and Chi Do. Linkage disequilibrium estimates of contemporary n_e using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, 2009. URL `http://dx.doi.org/10.1111/j.1752-4571.2009.00104.x`.

R.S. Waples. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci*. *Conservation Genetics*, 7(2):167–184, 2006. ISSN 1566-0621.

R.S. Waples and C. Do. LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, 8(4):753–756, 2008. ISSN 1755-0998.

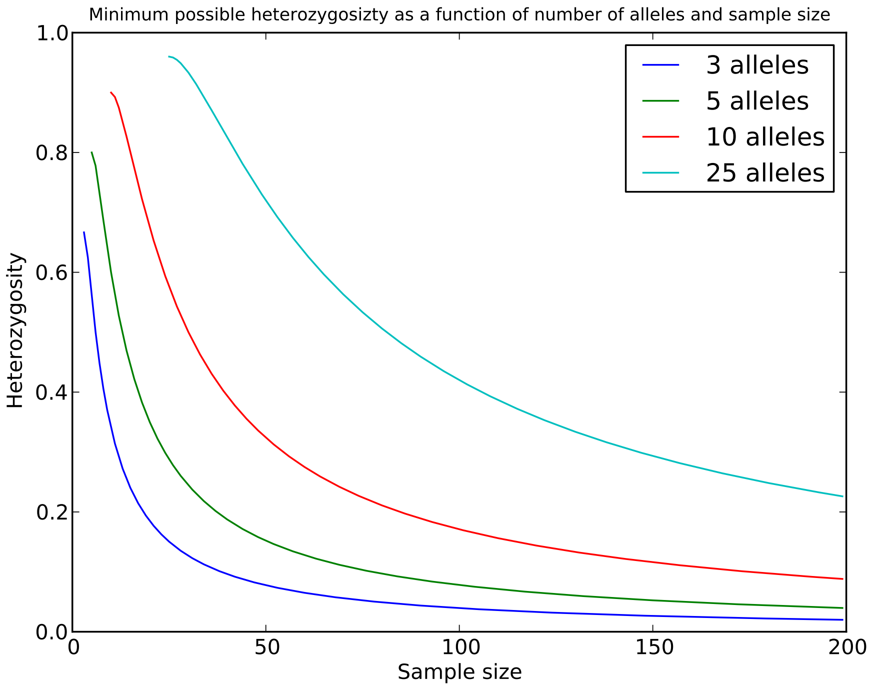

Figure 1: Minimum possible estimated heterozygosity assuming a number of different alleles is detected. If the number of alleles detected is equal to the sample size, then heterozygosity is maximized as all alleles exist in equal proportion. In order to have an accurate estimate of heterozygosity the number of samples has to be several times higher than the number of alleles.