***Appendix. Assumptions made in MalHaploFreq analyses and clarifications on it application.***

MalHaploFreq necessarily makes some assumptions in the calculations; these are described elsewhere[11] but are re-iterated here for transparency. A general introduction to the principles of ML and its application in parasitology can be found elsewhere.[13] It will be desirable to study the impact of these assumptions in future work but it is likely to become statistically complex. Meanwhile it makes sense to establish the general principles of prevalence vs frequency, identify best methods of analysis, consider potential sources of bias and the effect of GSL as matter of some urgency before proceeding to these more subtle aspects of analysis.

*Assumptions.*

(1) ML analysis assumes MOI is measured correctly whereas this may not be the case for two main reasons. Firstly, identical marker alleles (e.g. in MSP1, MSP2, glurp) may occur in different clones infecting the same human purely by chance; if two clones have the same marker allele then only a single clone will be counted and the true MOI will be underestimated. A more detailed description of this problem, although in a different context, can be found elsewhere[25]. Secondly, MOI estimation also has a detection limit so, in principle, minority marker alleles may be missed if the other marker alleles are identical and form a 'dominant' genotyping signal. However, highly variable marker genes makes this statistically unlikely; for example, even if 2 clones are identical purely by chance in a sample with MOI=5 then their ratios of genotyping signals will be 2:1:1:1 and none will fall below the GSL of 0.3. The chance of three clones all sharing an identical marker allele depends on the sum of the marker allele frequencies cubed so are very unlikely; consequently, it is likely that the first factor will be the major reason for underestimating MOI. It is difficult to anticipate the consequences of underestimating MOI in our analyses. We cannot see how this would systematically bias the estimates (e.g. by underestimating the true frequency of low-frequency alleles or haplotypes). Preliminary analyses suggest the magnitude of MOI underestimation will be small (A. Ross, personal communication).

(2) It is assumed, in common with others[14, 26, 27] that clones within hosts are acquired at random and are genetically unrelated. Genetic recombination often occurs in the oocysts of the mosquito and it is generally assumed that this gives rise to 4 distinct, but genetically related, haplotypes per oocyst. It therefore seems reasonable to conclude that several genetically related haplotypes may commonly be inoculated into humans by the same mosquito bite. Whether this leads to genetically-related clones establishing simultaneous patent infections in the same host is not clear. It has been suggested to occur in low transmission areas[28] but these regions of low MOI are not particularly problematic for analysis (see main text). It is less plausible in high-transmission areas because inoculation rates greatly exceeds MOI so it seems reasonable to assume the vast majority of inoculations fail and that consequently the chance of two-genetically related parasites both establishing successful infections from the same bite may be negligible We explicitly searched for genetic relatedness in parasites co-infecting the same human in an area of moderately high transmission of Malawi by analysing oocysts obtained from local mosquitoes and found no evidence of genetic relatedness. [29]

(3) MalHaploFreq necessarily implements GSL in a rather crude manner. It assumes that all clones are equally 'detectable' so that if one clone is mutant in a sample with MOI=4 then the mutation will be present in 25% of the parasites. In reality, the relative numbers of parasites in different clones varies enormously so it is probably that in at least some cases the mutant clone will have a large number of parasites (i.e. >>25%) and be detected. This means that the strict GSL cut-off should in principle be replaced by a more probabilistic criterion but this will be extremely complex; for example is the probability of detecting one mutant clone when MOI=4 the same as detecting two mutant clones in a MOI=8, and so on? This is an area that would reward further work although we suspect it may rapidly become statistically very complex and impossible to calibrate.

*Clarifications*

(1) Maximum likelihood is only one of several methods for statistical inference of allele and haplotype frequencies. The most obvious alternative is a Bayesian analysis based on Markov Chain Monte Carlo (MCMC) methods as described previously[27] but all reasonable methods of statistical inference should give identical (or extremely similar) estimates. We

used the MCMC method to analyses the single SNP frequencies ignoring GSL and confirmed near-identical estimates to that obtained by ML (results not shown).

(2) There is no formal 'denominator' in ML analyses. Simple counting or prevalence calculations divide one quantity by another so a denominator is required. In contrast, ML simply varies a set of parameters (haplotype frequencies in this case) until a best match to the data is obtained[13]; no division is entailed.

(3) We make no assumptions about the underlying distribution of MOI within patients and simply use the observed MOI in each of the samples[11]; hence our requirement that each sample in our datasets have estimate of MOI. This contrast to the ML analysis of Hill and Babiker[14] and Rannala and colleagues [26] neither of whom knew MOI for their samples and were consequently forced to assume that MOI followed a Poisson[14] or negative binomial distribution.[26]

(4) It is likely that not all malaria clones in a human will enter the blood sample because sequestration means that many will be present at very low numbers in venous blood. This does not affect our analysis because even if only a subset of clones are sampled and counted in the MOI, this set simply represents the number of clones entering our analyses. Note that exactly this same effect occurs in prevalence estimates and counting methods.

(5) MalHaploFreq has an inherent barrier to increasing our assumed level of GSL because 'mixed' wildtype+mutant samples are not compatible with high GSL. In the *dhfr* analyses the assumed GSL could not be increased beyond 0.33 because at this point it becomes impossible to have a mixed sample if MOI=3: there must be 1 clone of one type and 2 clones of the other so the proportion of the minor genotype is 1/3=0.33. GSL values higher than 0.33 this are incompatible with mixed samples of MOI=3 because the minor clone should not be detectable. For *dhps* analyses the limit occurred at 40% which corresponds to mixed infection when MOI=5 (because minor clones would be present at a proportion 2/5 =0.4). MalHaploFreq contains a routine to detect these 'impossible' samples and terminate the analysis; the 'impossibility' would result in a likelihood of zero.

(6) The ML analysis cannot rectify biases present in sampling schemes; it simply estimates the frequencies of alleles and/or haplotypes in the samples. It is extremely difficult to avoid sampling bias. Sampling blood after drug treatment will results in over-representation of resistant alleles, as will taking blood from people who have residual levels of drug from previous treatments[20, 21, 30]; blood taken from asymptomatic people may have fewer resistance alleles if these alleles are associated with a fitness penalty. [31, 32, 33] Many of these problems disappear if selectively neutral markers are investigated but many genetic studies explicitly study selected markers such as those encoding drug resistance or vaccine insensitivity.

This leads to several points of clarification

- There is no implicit mathematical requirement that the loci being analysed should be selectively neutral.
- It is the responsibility of investigators to design appropriate sampling scheme; blaming ML estimation for reflecting shortcomings in the data is untenable.
- The problems of sampling bias are not specific to ML frequency estimate but apply equally to all summary measurement, including prevalence estimates and direct counting.
- We make no claim that the PNG and Malawian datasets are directly comparable in any clinical or epidemiological sense but do assert that they are methodologically comparably because they were genotyped by the same person in the same laboratory.

It is important to make these points because readers from a background in fieldwork (at whom this paper is aimed) tend to appraise the data and then judge the analysis in this context. This approach is appropriate for epidemiological studies comparing regions, but this is a methodological paper comparing methods whose aim was to demonstrate and discuss the properties of different types of analysis. We therefore urge readers to take the datasets at face values and concentrate on understanding how different methods are affected by, and become biased by, factors such as MOI and GSL.

(7) Statisticians prefer artificial, simulated dataset because it is easier to quantify, rather than simply illustrate, effects such as bias. However we chose to analyze real datasets to illustrate the results because experience has shown they are more compelling to our intended audience of field workers who we wish to encourage to report frequencies and to

incorporate GSLs. MalHaploFreq does has a facility to simulate datasets and we intend to extend this facility in future work to allow us to examine some of the assumptions detailed elsewhere i.e. possible relatedness between clone, varying clonal densities, errors in assigning MOI, and so forth. This will require some rather detailed statistical analysis that would constitute a distraction in the current manuscript.