Imperial College London

MSc COMPUTATIONAL METHODS IN ECOLOGY AND EVOLUTION

Project Proposal: Joint Estimation of Population Size and Migration Rate for Mosquito Population Control

---

Author:
Sam TURNER
CID:
01804283

Primary Supervisor:
Austin BURT
Department of Life Sciences
a.burt@imperial.ac.uk

March 30, 2020

Key Words

Population genetics; Joint parameter estimation; Genome simulation; Mosquito population structure; Migration rates; Sampling strategies

Project Idea and Proposed Questions

Female Anopheles mosquitos transmit malaria, causing around 400,000 deaths globally each year, 70% of which are children under 5 [WHO, 2018]. Whilst population control of Anopheline mosquitos has been effective in reducing this burden [Mabaso ML, 2004], widespread elimination using current strategies (primarily bednetting and IRS) is unlikely. The desired scaling up of this strategy, requiring an increase in funding from \$2.7bn to \$9bn per year, is still predicted leave malaria in 62 countries by 2030 [Griffin, 2016]. As such, there is renewed interest in genetic methods for population control: a promising approach is the use of homing endonuclease genes (HEGs), which use gene drive to allow a lethal or sterilising genetic construct to spread through a population [Burt, 2003].

The evaluation of genetic control strategies is requires an understanding of the population genetics of the target species. In this project, we aim to develop a method for jointly estimating effective population size and migration rates. These parameters are of central interest to HEG work, allowing modelling of construct spread and evolution of host resistance.

Methods for estimating recent population size of closed populations exist, often based on allele frequency fluctuations [Waples, 1989], or linkage disequilibrium [Hill, 1981]. However, mosquitos are known to exist in structured population [Miles et al., 2017], posing a challenge for parameter inference: migration has been shown bias population size estimates from both allelic fluctuation [Wang and Whitlock, 2003] and linkage disequilibrium [Waples and England, 2011] based methods.

We aim to study the effect of population structure on the outcome on a number of population size estimators. Further, differences in the effect of migration on these estimators may allow us to construct a method for joint estimation of population size and migration rate. By analysing the properties of the joint estimator, we aim to determine how sampling strategy affects the confidence of parameter estimates, informing decisions on the sampling strategy used to collect mosquito genomes.

Proposed Methods

We will examine the properties of preexisting population size estimators, both analytically and with simulation, which may lead to an analytically derived joint estimator. Further, the simulated genomes can directly be used for joint estimation. The properties of the joint estimator can then be studied both analytically and with simulation

Genome Simulation Simulation of genomes using the coalescent has become an important part of population genetics [Kelleher and Lohse, 2020], and will be used to test the behaviour of estimators. Here, we will produce coalescent trees using msprime [Kelleher et al., 2016], which represent samples from an idealised population under a range of demographic scenarios: this includes varying population sizes, migration rates, and migration patterns.

Examining Estimator Behaviour We will examine the behaviour of estimators both analytically and via simulation. In the analytic approach, the properties of estimators are expressed as a function of the parameters of interest. The validity of the derived relationships can then be tested using genomes simulated under known demographic scenarios. However, quantities such as the variance or distribution of estimators may be difficult to determine analytically, so may be approximated by applying the

estimator to simulated genomes. In particular, the response of bias and variance to sampling strategy will be of interest.
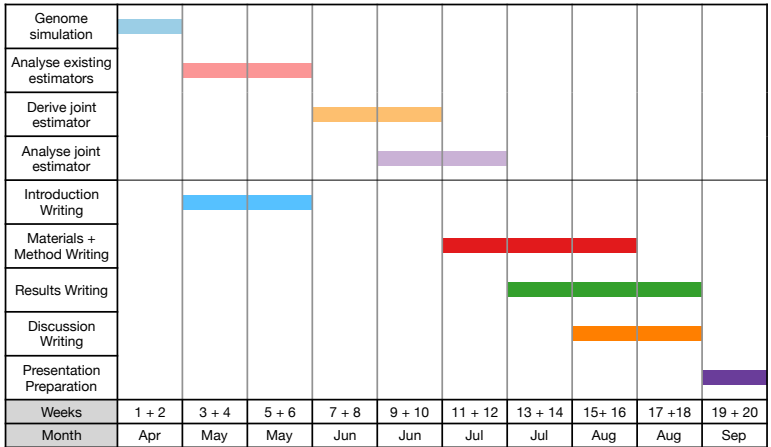
Joint Estimation   The difference between preexisting estimators may be a useful statistic for estimating migration rate and population size. If so, we will construct a joint estimator using either analytical (maximum likelihood, method of moments [Wang and Whitlock, 2003]) or algorithmic (Approxmiate Bayesian Computation [Sunnåker et al., 2013]) approaches. Similarly, the ratio of a single estimator taken in different demes may also hold information that allows a joint estimator to be constructed.

Outputs and outcomes

Estimates of population size and migration rate are critical for HEG modelling. This project aims both to produce an estimator, and to determine how sequencing resources can best be allocated to provide precise parameter estimates. In particular, the group is involved with the Ag1000 project, which has sequenced 1142 *A. gambiae* genomes to date [Clarkson et al., 2020]. These have uncovered structured local populations with gene flow and varying effective size, which must be quantified for genetic control mechanisms to be evaluated [Clarkson et al., 2020]. To this end, Ag1000 will be sequencing many more genomes, the allocation and analysis of which can be informed by this work.

Timeline and Feasibility

The simulation and mathematical techniques which we propose to use are well established, and have been used in similar estimation work before, including in this group [Hui and Burt, 2015]. It is highly likely that a set of estimators can be found that jointly give information on population size and migration rate, although it is possible that analytical estimator construction will be harder. However, numerical alternatives using simulation are available, so it is very feasible that a sound estimator is produced by the end of the project. Also, the current requirement for remote working will not be problematic as the work is computational, with communication possible over email, telephone, or video call.

	Weeks	1 + 2	3 + 4	5 + 6	7 + 8	9 + 10	11 + 12	13 + 14	15+ 16	17 +18	19 + 20
Genome simulation											
Analyse existing estimators											
Derive joint estimator											
Analyse joint estimator											
Introduction Writing											
Materials + Method Writing											
Results Writing											
Discussion Writing											
Presentation Preparation											
	Month	Apr	May	May	Jun	Jun	Jul	Jul	Aug	Aug	Sep

Figure 1: Gantt chart showing timeline for project completion.

Budget

High performance computing time: £500   Extensive use of Imperial's HPC services will be required as coalescent simulation is computationally expensive, with many simulations required to densely cover the parameter space for Approximate Bayesian Computation.

2

References

[Burt, 2003] Burt, A. (2003). Site-specific selfish genes as tools for the control and genetic engineering of natural populations. *Proc. Biol. Sci.*, 270(1518):921–928.

[Clarkson et al., 2020] Clarkson, C. S., Miles, A., Harding, N. J., Lucas, E. R., Battey, C. J., Amaya-Romero, J. E., Cano, J., Diabate, A., Constant, E., Nwakanma, D. C., Jawara, M., Essandoh, J., Dinis, J., Le Goff, G., Robert, V., Troco, A. D., Costantini, C., Rohatgi, K. R., Elissa, N., Coulibaly, B., Midega, J., Mbogo, C., Mawejje, H. D., Stalker, J., Rockett, K. A., Drury, E., Mead, D., Jeffreys, A. E., Hubbart, C., Rowlands, K., Isaacs, A. T., Jyothi, D., Malangone, C., Kamali, M., Henrichs, C., Simpson, V., Ayala, D., Besansky, N. J., Burt, A., Caputo, B., della Torre, A., Fontaine, M., Godfray, H. C. J., Hahn, M. W., Kern, A. D., Lawniczak, M. K. N., O'Loughlin, S., Pinto, J., Riehle, M. M., Sharakhov, I., Schrider, D. R., Vernick, K. D., White, B. J., Donnelly, M. J., and Kwiatkowski, D. P. (2020). Genome variation and population structure among 1,142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii. *bioRxiv*.

[Griffin, 2016] Griffin, J. (2016). Potential for reduction of burden and local elimination of malaria by reducing plasmodium falciparum malaria transmission: a mathematical modelling study. *Lancet Infect Dis.*, 16(4):465–472.

[Hill, 1981] Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetical Research*, 38(3):209–216.

[Hui and Burt, 2015] Hui, T.-Y. and Burt, A. (2015). Estimating effective population size from temporally spaced samples with a novel, efficient maximum-likelihood algorithm. *Genetics*, 200:285–293.

[Kelleher et al., 2016] Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*, 12(5):1–22.

[Kelleher and Lohse, 2020] Kelleher, J. and Lohse, K. (2020). *Coalescent Simulation with msprime*, pages 191–230. Springer US, New York, NY.

[Mabaso ML, 2004] Mabaso ML, Sharp B, L. C. (2004). Historical review of malarial control in southern african with emphasis on the use of indoor residual house-spraying. *Trop Med Int Health.*, 9(8):846–856.

[Miles et al., 2017] Miles, A., Harding, N. J., Bott?, G., Clarkson, C. S., Ant?o, T., Kozak, K., Schrider, D. R., Kern, A. D., Redmond, S., Sharakhov, I., Pearson, R. D., Bergey, C., Fontaine, M. C., Donnelly, M. J., Lawniczak, M. K. N., Kwiatkowski, D. P., Donnelly, M. J., Ayala, D., Besansky, N. J., Burt, A., Caputo, B., Della Torre, A., Fontaine, M. C., Godfray, H. C. J., Hahn, M. W., Kern, A. D., Kwiatkowski, D. P., Lawniczak, M. K. N., Midega, J., Neafsey, D. E., O'Loughlin, S., Pinto, J., Riehle, M. M., Sharakhov, I., Vernick, K. D., Weetman, D., Wilding, C. S., White, B. J., Troco, A. D., Pinto, J., Diabat, A., O'Loughlin, S., Burt, A., Costantini, C., Rohatgi, K. R., Besansky, N. J., Elissa, N., Pinto, J., Coulibaly, B., Riehle, M. M., Vernick, K. D., Pinto, J., Dinis, J., Midega, J., Mbogo, C., Bejon, P., Wilding, C. S., Weetman, D., Mawejje, H. D., Donnelly, M. J., Weetman, D., Wilding, C. S., Donnelly, M. J., Stalker, J., Rockett, K., Drury, E., Mead, D., Jeffreys, A., Hubbart, C., Rowlands, K., Isaacs, A. T., Jyothi, D., Malangone, C., Vauterin, P., Jeffery, B., Wright, I., Hart, L., Kluczy?ski, K., Cornelius, V., MacInnis, B., Henrichs, C., Giacomantonio, R., Kwiatkowski, D. P., Cornelius, V.,

MacInnis, B., Henrichs, C., Giacomantonio, R., and Kwiatkowski, D. P. (2017). Genetic diversity of the African malaria vector Anopheles gambiae. *Nature*, 552(7683):96–100.

[Sunnåker et al., 2013] Sunnåker, M., Busetto, A., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate bayesian computation. *PLoS Comput Biol*, 9(1).

[Wang and Whitlock, 2003] Wang, J. and Whitlock, M. C. (2003). Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*, 163(1):429–446.

[Waples, 1989] Waples, R. S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, 121(2):379–391.

[Waples and England, 2011] Waples, R. S. and England, P. R. (2011). Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics*, 189(2):633–644.

[WHO, 2018] WHO (2018). *World malaria report 2018*. World Health Organization.

Declaration

"I have seen and approved the proposal and the budget"

Approved: Austin Burt

Austin BURT

March 30, 2020