Comparative Analysis of Population Genetic Parameters at Different Sample Sizes

Ying Chen, Kevin McDermott, Peiwen Li and Alyson Van Natto
Team Name: Wenzi Instructor: Dr. Rob Colautti

Introduction

Problem/Background:
High-throughput Next Generation Sequencing (NGS) facilitates population genomic analysis using Single Nucleotide Polymorphisms (SNP) in model and non-model organisms[5]. Sample collection in the field is often ad hoc and obtaining a large sample size can be unrealistic. However, little is known of how sampling and analytical designs will influence SNP calling and downstream population genomic analyses.

Question

Does sample size affect the estimation of population genetic parameters?

Methods and Materials

Data:
A total of 98 mosquitoes (97 adults and 1 larva) were collected in August and November of 2013 from Olama and Nyabessan, West Africa[6] (Figure 2).

Methods:
- *Resampling:* We resampled 10 (small), 25 (medium) and 37 (large) individuals from each site with 10 replicates respectively (30 datasets total).
- *STACKS:* We performed *de novo* SNP calling for each dataset[2] (Figure 1).
- *NMDS:* Analysis was used to determine SNP genetic distances from STACKS outputs[8].
- *STRUCTURE:* We tested for the existence of distinct genetic clustering using cluster (K) values 1 to 3[9,10].
- *GenoDive:* We calculated observed and expected heterozygosity and pairwise F_{ST} with 999 permutations[7].
- *Statistical Analysis:* We generated pairwise NMDS distances for each dataset and performed a bootstrap analysis[1]. Additionally, we performed analysis of variance (ANOVA) tests (Figure 5a-f).

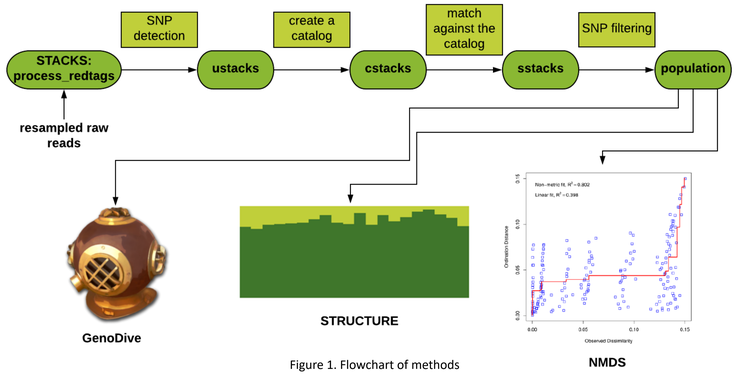

Figure 1. Flowchart of methods

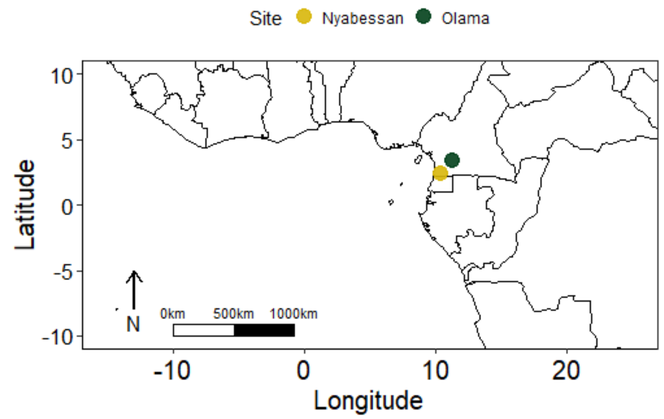

Figure 2. Location of sampling sites

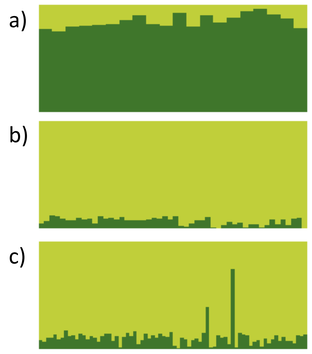

Figure 4. STRUCTURE plots of a) small sample sizes, b) medium sample sizes, and c) large sample sizes at a cluster (K) number of 2. Each colour represents one genetic cluster, and each bar represent one individual.

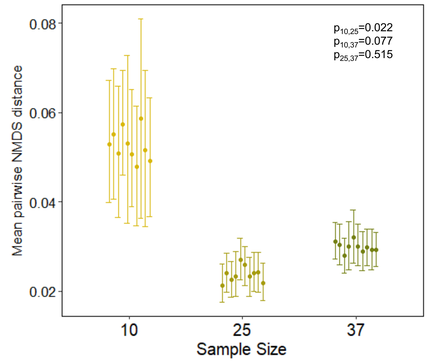

Figure 5. Mean and 95% confidence interval of mean pairwise NMDS distance at each replicate of sample size 10, 25 and 37 after 1000 bootstrap runs.

Results

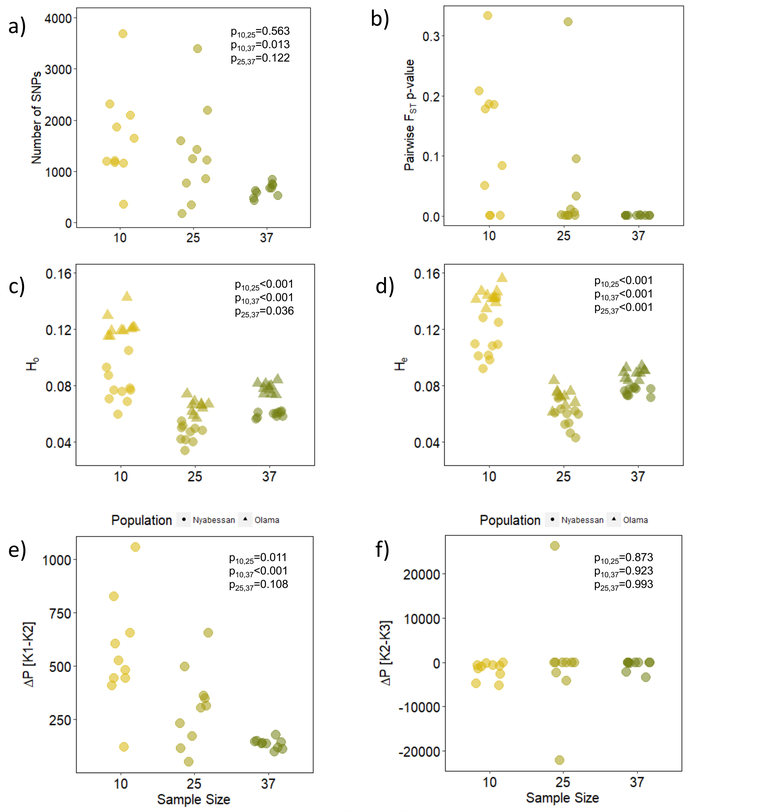

Figure 3. a) Number of SNPs for each replicate, which are significantly different among sample sizes (p=0.0154). b) Pairwise F_{ST} analysis p-values. Pairwise F_{ST} values were not significantly different (p=0.18, df=2) among datasets, with values ranging from 0.001 to 0.017 across sample sizes. c,d) Observed and expected heterozygosity for each site in each replicate, which are significantly different among sample sizes (p<0.005, df=38). e) Differences in probabilities (ΔP[K1-K2]) between K=1 and K=2 in STRUCTURE analysis. ΔP[K1-K2] is significantly different between small and medium sample sizes (p=0.011, df=18), and between small and large sample sizes (p<0.001, df=18), but not significantly different between medium and large sample sizes (p=0.108, df=18). f) Differences in probabilities (ΔP[K2-K3]) between K=2 and K=3 in STRUCTURE analysis. ΔP[K2-K3] is not significantly different among sample sizes (p>0.05, df=18).

Discussion

The large sample size called a more consistent and smaller number of SNPs compared to the medium and small sample sizes (Figure 3a). We have three possible explanations for this:
- Greater sample size may allow the calling of the SNPs with more confidence, eliminating false positive calls.
- Given the same filtering parameters, outlier individuals may have a greater impact in smaller sample sizes, which can cause the greater variation and larger number of SNPs in smaller sample size.
- At the Nyabessan site only 37 individuals were sampled and therefore, all 37 sample size replicates had the same individuals.

A difference in SNP calling results also affected the estimation of population genetic parameters (Figure 3c-f; Figure 5). This may be because individuals from the small and medium sample size datasets may not be representative of the true population, in terms of observed and expected heterozygosity, pairwise $F_{ST,}$ genetic clustering, and NMDS pairwise distance.

Conclusions and Future Directions

Conclusion:
Sample size did affect the estimation of population genetic parameters in the mosquito *(Anopheles moucheti*) dataset. Using a larger sample size gave more consistent results. We recommend using the largest sample size possible.

Future Directions:
- For the dataset we used, we will decrease the largest sample size to ensure resampling obtains a variety of individuals or assign ambiguous ID's so resampling individuals with replacement is possible
- Repeat the experiment using multiple datasets of a variety of species to determine a minimum sample size for an accurate estimation of population genetic parameters.

Contact

Ying Chen, Kevin McDermott, Peiwen Li and Alyson Van Natto
Queen's University
116 Barrie Street, Kingston, Ontario, Canada, K7L 3J9
GitHub: https://github.com/kevinmcdermott062/812Final_Assignment/

References

1. Canty, A., and Ripley, B. (2017). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-20.
2. Catchen, J., Hohenlohe, P., Bassham, S., Amores, A., and W. Cresko, W. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124-3140.
3. Davison, A. C., and Hinkley, D. V. (1997). Bootstrap Methods and Their Applications. *Cambridge University Press, Cambridge*. ISBN 0-521-57391-2.
4. Earl, D.A,. and vonHoldt, B.M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* ,**4**(2), 359-361.
5. Ekblom, R., and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity,* **107**, 1-15.
6. Fouet, C., Kamdem, C., Gamez, S., and White, B.J. (2017). Extensive genetic diversity among populations of the malaria mosquito *Anopheles moucheti* revealed by population genomics. *Infection, Genetics and Evolution,* **48**, 27-33.
7. Meirmans, P.G., and Van Tienderen, P.H. (2004), GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* **4,** 792-794.
8. Oksanen, J., Guillaume Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Henry, M., Stevens, H., Szoecs, H., and Wagner, H. (2018). vegan: Community Ecology Package. R package version 2.4-6.
9. Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype. *Genetics,* **155**, 945-959.
10. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution,* **38**, 1358-1370.
11. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.