14

Tools for monitoring the genetic structure and stability of mosquito populations

Gregory C. Lanzaro[#]*, Sergey Nuzhdin*[##] *and Frederic Tripet*[#]

Abstract

Interest in mosquito population genetics has risen dramatically over the past decade, driven mainly by renewed interest in vector control as a means of controlling malaria and dengue fever. Earlier work in mosquito population genetics focused on resolving taxonomic issues, especially in distinguishing and defining the geographic distributions of cryptic taxa that are common in mosquitoes, especially in the genus *Anopheles*. The lessons learned from this early work include the realization that our concept of vector species is often incorrect and that even at the within-species level substantial genetic divergence among local populations exists. The explosion of research into the molecular genetics of mosquito vectors has dramatically altered the direction of, and interest in, mosquito population genetics. The most obvious difference is in the nature of the genetic markers that lie at the heart of studies aimed at describing the genetics of mosquito populations. Perhaps even more significantly, the study of mosquito molecular biology has led to changes in the questions being asked and in our ability to provide answers. The tools (e.g. markers) and questions are intimately related because the availability of new methodologies has allowed us to seek the answers to questions that seemed intractable in the past.

In this paper we discuss several tools that are currently available, but not yet widely used in population-genetics studies. We do not, however, provide a comprehensive review of those tools that are currently used in mosquito population genetics (with the exception of microsatellite DNA), as these have been reviewed elsewhere (Norris 2002).

Keywords: population genetics; microsatellites; single-nucleotide polymorphisms; sequence-specific amplification polymorphisms; microarrays

The raw material: Procuring specimens from natural populations

Unquestionably the most important tools for anyone interested in studying the biology of mosquito populations remain a pair of sturdy boots, a passport and the willingness to travel and spend time in the field. The point here is that the development of a sampling strategy that satisfies the question(s) being asked is the most important component of work aimed at understanding the genetics of natural populations. The sampling plan must include temporal and/or spatial components and

[#] Department of Entomology, University of California, One Shields Avenue, Davis, CA 95616, USA. E-mail: gclanzaro@ucdavis.edu and ftripet@ucdavis.edu

[##] Section of Evolution and Ecology, University of California, Davis, CA USA. E-mail: svnuzhdin@ucdavis.edu

provide for the procurement of an adequate sample size from each population under study, regardless of scale, which may vary from a single household within a village, to broad ecological areas spanning multiple national boundaries. The best way to obtain the necessary samples is to participate directly in collecting expeditions, which also provides the important benefit of learning something of the natural history of the species under study. Thus, an important role to be filled by the population geneticist is bridging the gap between the molecular genetics and field ecology.

Although there is little doubt that obtaining material for population studies by collecting it oneself is the best approach, doing so is expensive, both in time and in money, and often the questions being asked involve populations that may have already been extensively sampled. Therefore it is reasonable that organized specimen archives be created and maintained. Informal ‘specimen sharing’ among mosquito population geneticists is part of our tradition, however in the past, studies based on chromosome or isozyme markers usually led to the destruction of samples, so that only samples in excess of what was needed for a study were available. Most of the markers being used today are PCR-based, requiring only a small portion of the total amount of single mosquito-genomic DNA, so single specimens should be available for a relatively large number of assays. In addition, a method known as Multiple Displacement Amplification exists which allows a 100–400-fold amplification of whole mosquito genomes (Gorrochotegui-Escalante and Black IV 2003). This essentially makes individual mosquito DNA samples available for a very large number of studies. A network of collaborating investigators could be established and their contact information and description of material available posted on an appropriate website available to the community at large. Beyond this, arrangements can be made by individual investigators with respect to sharing material.

Microsatellite DNA

The discovery of hyper-variable microsatellite-DNA sequences undoubtedly revolutionized the fields of population genetics and ecology (Zhang and Hewitt 2003). For population geneticists, having at hand a marker evolving much faster than mitochondrial genes or genes coding for isozymes, equated to being able to resolve the structure of populations at a much finer geographical and evolutionary scale. For behavioural ecologists it translated into being able to establish kin relationships using DNA from smaller and smaller organisms, samples from live organisms, or even their gametes. Because microsatellites provide higher resolution for estimating genetic differentiation between populations within taxa, they allowed population biologists to make better inferences about population structure, and in some cases, about the movement of individuals between populations.

Not surprisingly, the number of studies taking advantage of their versatility has grown exponentially and the advances made possible by microsatellites render them indispensable in many fields (Zhang and Hewitt 2003). Today, an equally important body of literature points out the limitations of microsatellites for some applications (e.g. Chambers and MacAvoy 2000; Balloux and Lugon-Moulin 2002; Zhang and Hewitt 2003). The stepwise mutation process that adds or subtracts repeats to existing alleles (Armour et al. 1999; Eisen 1999) results in alleles of identical size having different mutational histories, a phenomenon known as allele size homoplasy (Estoup and Cornuet 1999; Estoup, Jarne and Cornuet 2002). Homoplasy is made more likely if the range of possible allele sizes itself is constrained (Garza, Slatkin and Freimer 1995; Lehmann, Hawley and Collins 1996; Estoup, Jarne and Cornuet 2002). There is evidence that some microsatellite tracts located in promoter regions may affect gene expression and protein binding (Kashi and Soller 1999; Rothenburg et al. 2001). These loci are clearly not neutral because selection may favour an optimal size range of repeat tracts (Estoup, Jarne and Cornuet 2002; Zhang and Hewitt 2003). Generally speaking, microsatellite loci that are near genes that are themselves under selection will be subject to hitchhiking and background selection and this may be another source of deviation from neutrality (Charlesworth, Nordborg and Charlesworth 1997; Barton 2000). Homoplasy is a concern when estimating genetic divergence between taxa (Garza, Slatkin and Freimer 1995; Estoup, Jarne and Cornuet 2002). Genetic distances such as Wright's F_{ST} estimate (1931), Weir and Cockerham's θ F_{ST} estimate (1984), or Nei's *Ds* and *Da* distances (Nei 1972; Takezaki and Nei 1996) are based on the assumption that mutations generate only new alleles (infinite-allele model). Homoplasy will therefore result in an underestimation of genetic distance. As a result a number of new genetic distance statistics that assume stepwise or mixed stepwise and non-stepwise mutational models have been proposed. These include R_{ST} (Slatkin 1995), $(\delta\mu)^2$ (Goldstein et al. 1995) and D_{SW} (Shriver et al. 1995). Simulations presented along with these distances show that they out-perform classic distances for phylogenetic inferences (Slatkin 1995; Goldstein et al. 1995; Shriver et al. 1995). However, assessing which is better suited for use on real data remains difficult and depends on the evolutionary scale and the organisms considered. For population-genetic studies, F_{ST}'s (Wright 1931; Weir and Cockerham 1984) are still widely used and it is generally accepted that they perform better in studies of populations that exchange migrants – e.g. subdivided populations or hybrid zones (Rousset 1996; Estoup, Jarne and Cornuet 2002).

Another major concern for using microsatellites was the practice of directly translating F_{ST}'s into *Nm*, the number of migrants per generation, using the simple relationship $F_{ST} \approx 1/(4Nm+1)$ (Slatkin 1985; 1987). As emphasized by Whitlock and McCauley (1999), the temptation of translating F_{ST} estimates into units that make immediate ecological sense is understandable but treacherous. This is because few populations meet the assumptions required for translating F_{ST}'s into *Nm*'s, and F_{ST}'s may be biased by homoplasy (Bossart and Prowell 1998; Whitlock and McCauley 1999). Here again, it is generally recognized that *Nm* estimates have to be interpreted with caution. As a result, they are best suited for qualitative comparisons except when their reliability has been assessed using direct measures of dispersal or by comparing them to estimates of hybridization rates (Taylor et al. 2001; Tripet, Dolo and Lanzaro 2005).

Sequence-specific amplification polymorphisms

Sometimes, it is necessary to develop a set of markers private for one genotype and absent in the remaining individuals within a population. One of the promising and inexpensive ways is Sequence-specific amplification polymorphism (SSAP) (Waugh et al. 1997). SSAP analysis has been designed to resolve genetic distances of very closely related crop-plant varieties. It relies on the presence of multiple transposable elements – that are similar to retroviruses – frequently inserting into new genomic positions. SSAP is similar to AFLP except that only those bands are visualized that are tagged into highly polymorphic transposable element sites. Specifically, one ligates ~20bp adapter to the ends of restricted DNA (usually with a four-cutter enzyme), and PCR-amplified DNA using a labelled primer homologous to the transposable element sequence and an unlabelled primer homologous to the adapter.

Bands of different size correspond to transposable elements from different occupation positions. An advantage of these markers is that they are as polymorphic as microsatellites. As the mosquito genome is sequenced and its transposable element population becomes known, thousands of SSAP markers can be developed within a week. But unlike microsatellite markers, SSAP bands are frequently population- and individual-specific (Yang and Nuzhdin 2003).

Single-nucleotide polymorphisms

As the amount of sequence data available from many organisms increases and entire genomes are being assembled, more and more researchers have the possibility to use yet another type of marker. Single-nucleotide polymorphisms (SNPs) have much lower mutation rates than microsatellites and provide an alternative tool for pedigree analyses (Blouin et al. 1996; Glaubitz, Rhodes and Dewoody 2003). Their increasing popularity is driven in large part by advances in biomedicine where genomic studies have linked SNPs with phenotypic characteristics of diseases and hosts, and increasingly powerful methods are used for screening variation at multiple loci (Vignal et al. 2002; Hirschhorn et al. 2002). Here we discuss their potential use as an alternative to microsatellites for population-genetic studies as proposed elsewhere (Brumfield et al. 2003; Morin et al. 2004). We are currently involved in studies of the population structure of *Anopheles gambiae* Giles, the main vector of malaria in Africa. Since the entire genome has been sequenced (Holt et al. 2002), SNPs are a real alternative to microsatellites in this organism.

Microsatellites versus SNPs

There are a number of aspects that need to be taken into account when comparing the two types of markers. One of the big advantages of SNPs, when whole genome sequences are available, is their abundance. Users can decide the polymorphism they prefer (transition, transversion or both) and pick loci away from coding regions to insure that they are not influenced by selection on nearby genes. In theory, when genomes are available, microsatellite loci could also be selected away from genes but their lower density may prevent that luxury in many study organisms. With regard to mutational processes, SNPs with their low mutational rate, i.e. ~10^{-9} compared to ~10^{-4} to 10^{-6} for microsatellites, are expected to feature few alleles per locus (Hancock 1999; Zhang and Hewitt 2003). In fact, in most cases, SNPs will be equivalent to di-allelic markers (Vignal et al. 2002; Morin et al. 2004). It is expected that multi-allelic microsatellites should have higher power – per locus – over di-allelic SNPs for estimating genetic divergence or gene flow using *F*-statistics or assignment tests (Vignal et al. 2002; Brumfield et al. 2003; Morin et al. 2004). Mariette et al. (2002) estimated that four to ten times more di-allelic markers (dominant markers were simulated in this study) were necessary for reliably estimating genome-wide levels of variation. Studies by Blouin et al. (1996) and Glaubitz, Rhodes and Dewoody (2003) suggest that measures of pair-wise genetic relationships using SNPs would require analysis of more than 5 times more loci.

Comparing the resolution of SNPs to that of the faster evolving microsatellites is critical to assess their potential for population genetics. Despite extensive discussion of the potential for SNPs in population-genetic studies (Brumfield et al. 2003; Morin et al. 2004), there are, as yet, no qualitative comparisons of the two types of markers available.

Practical implications for population studies

Assessing what is the optimal marker for the organism and populations under study is a fundamental step in designing and planning population genetic studies. In many instances, however, time and cost optimization is just as important for the success of a project. The costs and technical aspects relating to the use of either marker have been adequately evaluated and discussed elsewhere (reviewed in Kwok 2001; Chen and Sullivan 2003; Zhang and Hewitt 2003). Despite important developments in methods of SNP allele discrimination and detection, all techniques rely on PCR amplification. Given the much higher number of SNP loci required, the costs in reagents and manpower will be multiplied 6-10-fold. Choosing SNPs over microsatellites also involves considerable investment in equipment that often has higher operating costs than the sequencers used for typing microsatellites. Nowadays, microsatellite libraries can be ordered from companies at a reasonable cost. More importantly they can be ordered with pre-evaluated primer pairs thus significantly cutting down the costs of manpower required for these steps.

In conclusion, SNPs should theoretically generate better estimates than microsatellites when the populations under study are fully isolated either reproductively or geographically. Switching to SNPs does not, however, prevent biases due to co-ancestry and it should also be noted that the impact of ascertainment biases or problems of null alleles remains to be adequately evaluated. In animal systems or populations with either known ongoing gene flow or low microsatellite mutation rates, e.g. *Drosophila* (Zhang and Hewitt 2003), the benefit of using SNPs is questionable. For the reasons discussed above and because advantages and potential flaws of microsatellites are so well documented, we predict that they will remain essential tools in population genetics. The popularity of SNPs will strongly depend on the number of whole genomes available, the development of simpler protocols for their design in other organisms, and the availability of affordable automated PCR procedures for processing large number of loci.

Microarrays for population genetics

In addition to PCR-based techniques, SNP-scoring techniques relying on ligation and hybridization are being developed and evaluated. Because of the large sample sizes required for studies of population-genetic structure, ligation-based techniques are currently the only real options for typing large numbers of SNPs at an affordable cost. For ligation-based approaches, two primers are designed to ligate if they perfectly hybridize to a PCR-amplified genomic template. The specificity of the ligation reaction is much higher than that of polymerization, thus the typing error is much smaller. The use of a bar-code system of ligation detection greatly reduces labour. It enables typing at the cost of about US $0.05 per SNP. Millions of SNPs are rapidly and reliably typed with this approach (Genissel et al. 2004). However, current applications focus on scoring large numbers of SNPs within a few large amplicons. For population-genetic studies, the reverse should be achieved, namely scoring fewer loci but from a large number of amplicons spread out across the genome, but here again the number of PCR reactions required can be very large. Assuming that a genome is available and money would not be a limiting factor, SNP typing can be made at a much larger scale with microarrays. Currently available Affymetrix CustomSeqTM re-sequencing arrays enable the analysis of up to 30,000 bases of double-stranded sequence. These arrays carry in excess of 240,000 features, each feature being 20x25 micrometer of glass surface covered by millions of copies of a 25-mer oligonucleotide. To identify the nucleotide at a given position, the Affymetrix platform compares the levels of template hybridization to four oligonucleotides that match the reference genome sequence and are identical except at the position that is being analysed. This position (the exact middle of the oligo) contains either A, T, C or G. The strongest hybridization indicates which of the four oligonucleotides represents a perfect match, as opposed to a mismatch. This inference is confirmed if the sequencing of the two opposite strands produces concordant results. To identify the next nucleotide, the analysis is repeated with all oligos shifted by one base. Affymetrix arrays are intended for hybridizations with templates amplified by LPCR as described above. This technology provides base calls at >99.99% accuracy and 90% calling rate. This compares favourably with the accuracy and calling rate achievable with direct ABI sequencing. Using this platform also generates data on small indels.

References

Armour, J.A.L., Alegre, S.A., Miles, S., et al., 1999. Minisatellites and mutation processes in tandemly repetitive DNA. *In:* Goldstein, D.B. and Schlötterer, C. eds. *Microsatellites: evolution and applications*. Oxford University Press, Oxford, 24-33.

Balloux, F. and Lugon-Moulin, N., 2002. The estimation of population differentiation with microsatellite markers. *Molecular Ecology,* 11 (2), 155-165.

Barton, N.H., 2000. Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences,* 355 (1403), 1553-1562.

Blouin, M.S., Parsons, M., Lacaille, V., et al., 1996. Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology,* 5 (3), 393-401.

Bossart, J.L. and Prowell, D.P., 1998. Genetic estimates of population structure and gene flow: limitations, lessons and new directions. *Trends in Ecology and Evolution,* 13 (5), 202-206.

Brumfield, R.T., Beerli, P., Nickerson, D.A., et al., 2003. The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution,* 18 (5), 249-256.

Chambers, G.K. and MacAvoy, E.S., 2000. Microsatellites: consensus and controversy. *Comparative Biochemistry and Physiology. Part B. Biochemistry and Molecular Biology,* 126 (4), 455-476.

Charlesworth, B., Nordborg, M. and Charlesworth, D., 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research,* 70 (2), 155-174.

Chen, X. and Sullivan, P.F., 2003. Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. *The Pharmacogenomics Journal,* 3 (2), 77-96.

Eisen, J.A., 1999. Mechanistic basis for microsatellite instability. *In:* Goldstein, D.B. and Schlötterer, C. eds. *Microsatellites: evolution and applications*. Oxford University Press, Oxford, 34-48.

Estoup, A. and Cornuet, J-M., 1999. Microsatellite evolution: inferences from population data. *In:* Goldstein, D.B. and Schlötterer, C. eds. *Microsatellites: evolution and applications*. Oxford University Press, Oxford, 49-65.

Estoup, A., Jarne, P. and Cornuet, J.M., 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology,* 11 (9), 1591-1604.

Garza, J.C., Slatkin, M. and Freimer, N.B., 1995. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Molecular Biology and Evolution,* 12 (4), 594-603.

Genissel, A., Pastinen, T., Dowell, A., et al., 2004. No evidence for an association between common nonsynonymous polymorphisms in delta and bristle number variation in natural and laboratory populations of *Drosophila melanogaster*. *Genetics,* 166 (1), 291-306.

Glaubitz, J.C., Rhodes, O.E. and Dewoody, J.A., 2003. Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology,* 12 (4), 1039-1047.

Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L., et al., 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences of the United States of America,* 92 (15), 6723-6727.

Gorrochotegui-Escalante, N. and Black IV, W.C., 2003. Amplifying whole insect genomes with multiple displacement amplification. *Insect Molecular Biology,* 12 (2), 195-200.

Hancock, J.M., 1999. Microsatellites and other simple sequences: genomic context and mutational mechanisms. *In:* Goldstein, D.B. and Schlötterer, C. eds. *Microsatellites: evolution and applications*. Oxford University Press, Oxford, 1-9.

Hirschhorn, J.N., Lohmueller, K., Byrne, E., et al., 2002. A comprehensive review of genetic association studies. *Genetics in Medicine,* 4 (2), 45-61.

Holt, R.A., Subramanian, G.M., Halpern, A., et al., 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science,* 298 (5591), 129-130,141-149.

Kashi, Y. and Soller, M., 1999. Functional roles of microsatellites and minisatellites. *In:* Goldstein, D.B. and Schlötterer, C. eds. *Microsatellites: evolution and applications*. Oxford University Press, Oxford, 10-23.

Kwok, P.Y., 2001. Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics,* 2, 235-258.

Lehmann, T., Hawley, W.A. and Collins, F.H., 1996. An evaluation of evolutionary constraints on microsatellite loci using null alleles. *Genetics,* 144 (3), 1155-1163.

Mariette, S., Le Corre, V., Austerlitz, F., et al., 2002. Sampling within the genome for measuring within-population diversity: trade-offs between markers. *Molecular Ecology,* 11 (7), 1145-1156.

Morin, P.A., Luikart, G., Wayne, R.K., et al., 2004. SNPs in ecology, evolution and conservation. *Trends in Ecology and Evolution,* 19 (4), 208-216.

Nei, M., 1972. Genetic distance between populations. *American Naturalist,* 106 (949), 283-292.

Norris, D.E., 2002. Genetic markers for study of the anopheline vectors of human malaria. *International Journal of Parasitology,* 32 (13), 1607-1615.

Rothenburg, S., Koch-Nolte, F., Rich, A., et al., 2001. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proceedings of the National Academy of Sciences of the United States of America,* 98 (16), 8985-8990.

Rousset, F., 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics,* 142 (4), 1357-1362.

Shriver, M.D., Jin, L., Boerwinkle, E., et al., 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Molecular Biology and Evolution,* 12 (5), 914-920.

Slatkin, M., 1985. Gene flow in natural populations. *Annual Review of Ecology and Systematics,* 16, 393-430.

Slatkin, M., 1987. Gene flow and the geographic structure of natural populations. *Science,* 236 (4803), 787-792.

Slatkin, M., 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics,* 139 (1), 457-462.

Takezaki, N. and Nei, M., 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics,* 144 (1), 389-399.

Taylor, C., Touré, Y.T., Carnahan, J., et al., 2001. Gene flow among populations of the malaria vector, *Anopheles gambiae,* in Mali, west Africa. *Genetics,* 157 (2), 743-750.

Tripet, F., Dolo, G. and Lanzaro, G.C., 2005. Multi-level analyses of genetic differentiation in *Anopheles gambiae s.s.* reveal patterns of gene flow important for malaria-fighting mosquito projects. *Genetics*, 169 (1), 313-324.

Vignal, A., Milan, D., SanCristobal, M., et al., 2002. A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, 34 (3), 275-305.

Waugh, R., McLean, K., Flavell, A.J., et al., 1997. Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). *Molecular and General Genetics,* 253 (6), 687-694.

Weir, B.S. and Cockerham, C.C., 1984. Estimating F-Statistics for the analysis of population structure. *Evolution,* 38, 1358-1370.

Whitlock, M. C. and McCauley, D. E., 1999. Indirect measures of gene flow and migration: F-ST not equal 1/(4Nm+1). *Heredity,* 82 Part 2, 117-125.

Wright, S., 1931. Evolution in Mendelian populations. *Genetics*, 16, 97-159.

Yang, H.P. and Nuzhdin, S.V., 2003. Fitness costs of Doc expression are insufficient to stabilize its copy number in *Drosophila melanogaster*. *Molecular Biology and Evolution,* 20 (5), 800-804.

Zhang, D.X. and Hewitt, G.M., 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology,* 12 (3), 563-584.