Genet. Res., Camb. (2001), **78**, *pp.* 209–212. © 2001 Cambridge University Press
DOI: 10.1017/S0016672301005444 Printed in the United Kingdom

Review

Genealogical inference of closely related species based on microsatellites

CHRISTIAN SCHLÖTTERER*
Institut für Tierzucht und Genetik, Veterinärmedizinische Universität Wien, 1210 Wien, Austria

Summary

Despite their unmatched popularity in many research areas, microsatellites have not yet become a major tool for the inference of genealogical relationships of closely related species. Recent studies have successfully extended the repertoire of microsatellite analysis beyond population genetics and demonstrate that phylogenetic relationships of closely related species can be inferred accurately with fewer loci than previously assumed.

Even though well-founded phylogenies of closely related species and populations play a central role in the understanding of evolutionary processes, the inference of genealogical relationships remains a significant challenge. Even under the simple model of an instantaneous speciation event, ancestral polymorphisms are shared between the two separate groups. For large population sizes, substructured populations, and recent speciation events, this problem is even more pronounced (Edwards *et al*., 2000). Hence, allelic variants within and between taxonomic groups are very likely to predate the species split. Genealogical relationships inferred from a single gene may or may not coincide with the genealogy of the species (Pamilo *et al*., 1988). Consequently, multiple genes need to be analyzed, preferably for multiple individuals from each species in order to estimate the species phylogeny.

A different approach to the genealogical inference of closely related species takes advantage of variation shared across groups. The process of differentiation of the frequencies of non-selected ancestral alleles in different taxonomic groups due to genetic drift begins immediately after a speciation event. Initially, allozymes were used for genealogical inference based on differences in allele frequency, but the frequent non-neutral behavior of allozymes and the limited number of informative marker systems often severely compromised their use. About a decade ago, microsatellites were introduced as a novel molecular marker (Litt *et al*., 1989; Tautz, 1989; Weber *et al*., 1989). This class of DNA evolves mostly under neutrality, and several thousand loci are typically present in eukaryotes (Schlötterer, 2000). In contrast to allozymes, microsatellites have an exceptionally high mutation rate. Therefore, not only genetic drift, but also mutation pressure affects allele frequencies. To account for this, several microsatellite specific distance measurements have been introduced, such as R_{ST} (Slatkin, 1995) and $(\delta\mu)^2$ (Goldstein et al., 1995). While these distance measurements have the desired property of linearity with time, they also have a large variance. Computer simulations suggest that 50–100 microsatellite loci may be required for reliable phylogenetic reconstruction (Takezaki *et al*., 1996). Further problems for microsatellite based genealogical inference are imposed by a presumed length constraint, which keeps the repeat number at a locus below a certain threshold (Bowcock *et al*., 1991). Modifications of microsatellite-based distance measurements have been proposed to account for this (Feldman *et al*., 1997; Pollock *et al*., 1998; Zhivotovsky, 1999), more recent studies on the mutation process of microsatellites suggest, however, that mutational behavior of microsatellites is specific to the repeat number (Ellegren, 2000; Harr *et al*., 2000; Xu *et al*., 2000). While deviations from the strict stepwise mutation model do not affect phylogenetic reconstruction, variation in mutation rates among species remains a major concern (Goldstein *et al*., 1997). Due to these anticipated complications, only a very limited number of experimental studies have been performed to test the applicability of microsatellites for genealogical inference.

* Tel: +43-1-25077-5603. Fax: +43-1-25077-5693. e-mail: christian.schloetterer@vu-wien.ac.at

1. The Drosophila model

In comparison to many other experimental systems, the *Drosophila melanogaster* complex offers a wealth of information relating to genealogy, including data on allozyme and inversion polymorphism, numerous gene trees, and a full body of speciation studies (Powell, 1997). As a result, a good *a priori* hypothesis exists, against which the genealogical relationships inferred from microsatellites can be tested.

The *D. melanogaster* complex consists of four species. *D. melanogaster* and *D. simulans* are two cosmopolitan species, and *D. sechellia* and *D. mauritiana* are endemic to the islands of the Seychelles and Mauritius. The basal position of *D. melanogaster* was established a long time ago, but the grouping of the remaining three species has only recently been established. 'Speciation genes', which contribute directly to some aspects of biological divergence, are expected to reflect the species phylogeny more accurately than other genes. Use of the *Odysseus* (*OdsH*) locus, a homeobox gene involved in hybrid male sterility in the *D. simulans* clade, implies that *D. simulans* and *D. mauritiana* cluster together (Ting *et al.*, 2000). This grouping is also supported by a joint sequence analysis of 14 different genes (Kliman *et al.*, 2000).

Approximately 30 individuals from each species were analyzed for 39 microsatellite loci (Harr *et al.*, 1998). In a tree of individuals, in which each individual is treated as a separate OTU, all four species are well separated and each species node was supported by a high bootstrap value. Most importantly, the genealogical relationships were identical to those based on the 'speciation gene' *OdsH*. Hence, this Drosophila model has clearly demonstrated that microsatellite analysis can be used for phylogenetic reconstruction.

2. Other studies

A substantial number of studies using a small number of loci (< 10) to infer genealogical relationships have been published, but they are difficult to evaluate in the context of usefulness of microsatellites for phylogenetic reconstruction. The first study to systematically evaluate the usefulness of microsatellites for phylogenetic inference was based on samples from 13 geographic areas representing three different bear species (Paetkau *et al.*, 1997). Using 10 loci the authors were not able to resolve the close sister relationship of polar bears and brown bears from more distantly species, irrespective of the distance measurement used. Despite the undoubted importance of the study at the time published, in the light of other studies (see below), it appears likely that a larger number of loci would have resulted in different outcome. Similarly, a study based on 10 microsatellites failed to differentiate humans and three primate species (Bowcock *et al.*, 1994).

The mammalian family Bovidae includes a variety of globally distributed ungulates. One tribe, the Bovini, contains many important domestic animals, such as cattle, Yak and Buffalo. Ritz *et al.* (2000) used 20 microsatellites to infer the phylogenetic relationship of the tribe Bovini. Two different distance measurements, $(\delta\mu)^2$ (Goldstein *et al.*, 1995) and Cavalli-Sforza and Edwards' chord distance (Cavalli-Sforza *et al.*, 1967) gave a consistent grouping of species. As expected from the results of the computer simulation study of Takezaki and Nei (1996), the chord distance resulted in higher bootstrap support values (Ritz *et al.*, 2000).

The study by Goldstein *et al.* (1999) on the island foxes, a diminutive form of the mainland gray fox *Urocyon cinereoargenteus*, provides a nice example of genealogical inference at the population level based on microsatellites. Island foxes currently occupy six of California's Channel islands. With a dataset of 19 microsatellite loci and using a distance measure based on the proportion of shared alleles, 181 out of 183 foxes were assigned correctly to their geographic origin. While the topology of population trees based on $(\delta\mu)^2$ matched the presumed colonization history of the islands, the trees derived with the proportion of shared alleles did not place the gray fox as an outgroup (Goldstein *et al.*, 1999).

The last example for microsatellite based phylogeny reconstruction addressed the old question of whether the two oak species, *Q. robur* and *Q. petraea*, are two separate taxonomic units (Muir *et al.*, 2000). The authors analyzed five populations from each of the two species with 20 microsatellites. A grouping of the populations according to geography would be expected if *Q. robur* and *Q. petraea* are not distinct taxonomic units. However, consistent with morphological data, Muir *et al.* (2000) found that the two species were well separated with high bootstrap values supporting the grouping of populations from the same species.

3. Indels in flanking sequence

Microsatellite variability is predominantly scored as PCR product length variation, hence microsatellite repeat number changes are measured jointly with indels in the flanking sequence. While in *D. melanogaster* flanking sequence variation seems to be a frequent phenomenon Colson *et al.*, 1999; Harr & Schlötterer, unpublished observations, for many species the length of PCR product usually varies by multiples of the repeat unit size, suggesting that mutations of the microsatellite account for most of the variation in PCR product size. Between species, however, flanking sequences accumulate indel muta-

https://doi.org/10.1017/S0016672301005444 Published online by Cambridge University Press

tions with increasing phylogenetic distances. A laborious but feasible approach to circumvent this problem would be to determine the exact repeat counts for each species and locus by DNA sequencing. Interestingly, data from the *D. melanogaster* complex suggest that indels in the flanking sequence are not compromising the recovery of the correct topology. Using PCR product lengths rather than repeat number inferred by sequencing of a single representative individual, Harr *et al.* (1998) recovered the identical topology, but the bootstrap support was slightly lower for the data set based on PCR-product lengths.

4. How many loci?

While computer simulations suggest that approaching 100 microsatellite loci may be required for accurate phylogenetic reconstruction, the genealogy of the *D. melanogaster* species group could be inferred reliably with just 39 microsatellites. Even a smaller number of loci resulted in the same topology (Harr *et al.*, 1998). The authors attributed the accuracy of their phylogenetic reconstruction to the low mutation rate of Drosophila microsatellites (Schlötterer *et al.*, 1998; Schug *et al.*, 1997). An alternative explanation would be that some taxa require fewer loci for accurate phylogenetic reconstruction. The studies by Ritz *et al.* (2000) and Muir *et al.* (2000), however, were based on only 20 microsatellites, and also provided a strong phylogenetic signal despite the observation that microsatellites in their taxonomic groups were not described to have low mutation rates. Future studies will confirm whether or not a moderate number (20–40) of microsatellite loci is sufficient to infer reliable microsatellite based phylogenies.

5. Which distance measurement?

Takezaki and Nei (1996) proposed to use D_A (Nei *et al.*, 1983) or Cavalli-Sforza and Edwards' chord distance (Cavalli-Sforza *et al.*, 1967) for the inference of the topology. $(\delta\mu)^2$ on the other hand should be used to estimate branch lengths. While the proportion of shared alleles (Bowcock *et al.*, 1994) was not evaluated by the authors, the experimental data by Harr *et al.* (1998) and Muir *et al.* (2000) suggest that this distance measurement is also effective in obtaining the correct genealogical relationship.

Given that distance estimators based on allele frequencies rely on genetic drift, they are sensitive to changes in population size. It has been known for some time that bottlenecks inflate genetic distances (Chakraborty *et al.*, 1977), an effect that has been confirmed by computer simulation studies using microsatellites (Takezaki *et al.*, 1996). A recent study demonstrated, however, that in cases of low migration rates the $(\delta\mu)^2$ estimator could be corrected for demographic effects such as bottlenecks and population expansions (Zhivotovsky, 2001). The only caveat of this method is that the variability in the ancestral population is required for this correction. Using an extant population with levels of variability similar to the presumed ancestral one Zhivotovsky (2001) improved the microsatellite based estimate for the out of Africa colonization of humans from 34,000 to 57,000 years.

While the approach by Zhivotovsky (2001) could potentially be applied to genealogical inference, the variability of the ancestral populations is required for each node of the tree (including internal nodes). Whether or not estimates based on extant populations and additional information about the history will be sufficient to improve genealogical inference based on $(\delta\mu)^2$ requires further investigation.

6. Final remark

Given that several studies have demonstrated that a moderate number of loci permit a microsatellite based phylogenetic inference of closely related species, I anticipate that more studies with a larger number of loci will verify the usefulness of microsatellites for phylogenetic reconstruction. Based on larger data sets, it will be possible to decide whether loci with a low mutation rate are more effective for phylogenetic reconstruction and how many loci are required for a sufficient statistical support. Whether or not divergence times are best estimated by Baysian or likelihood based approaches, which require a specific demographic model, or by model free algorithms remains to be clarified (Stumpf *et al.*, 2001).

I am grateful to R. Achmann, B. Harr, M. Kauer, and L. Zhivotovsky for helpful discussions. Many thanks to G. Gibson for his encouragement. C.S. is supported by grants of the Fonds zur Förderung der wissenschaftlichen Forschung.

References

Bowcock, A. M., Kidd, J. R., Mountain, J. L., Hebert, J. M., Carotenuto, L., Kidd, K. K. & Cavalli-Sforza, L. L. (1991). Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. *Proceedings of the National Academy of Sciences of the USA* **88**, 839–843.

Bowcock, A. M., Ruiz-Lineares, A., Tonfohrde, J., Minch, E., Kidd, J. R. & Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–457.

Cavalli-Sforza, L. L. & Edwards, A. W. F. (1967). Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics* **19**, 233–257.

Chakraborty, R. & Nei, M. (1977). Bottleneck effects on average heterozygosity and genetic distances with the stepwise mutation model. *Evolution* **31**, 347–356.

Colson, I. & Goldstein, D. B. (1999). Evidence for complex mutations at microsatellite loci in Drosophila. *Genetics* **152**, 617–627.

Edwards, S. V. & Beerli, P. (2000). Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* **54**, 1839–1854.

Ellegren, H. (2000). Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genetics* **24**, 400–402.

Feldman, M. W., Bergman, A., Pollock, D. D. & Goldstein, D. B. (1997). Microsatellite genetic distances with range constraints: analytic description and problems of estimation. *Genetics* **145**, 207–216.

Goldstein, D. B. & Pollock, D. D. (1997). Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. *Journal of Heredity* **88**, 335–342.

Goldstein, D. B., Ruiz Lineares, A., Cavalli-Sforza, L. L. & Feldman, M. W. (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences of the USA* **92**, 6723–6727.

Goldstein, D. B., Roemer, G. W., Smith, D. A., Reich, D. E., Bergman, A. & Wayne, R. K. (1999). The use of microsatellite variation to infer population structure and demographic history in a natural model system. *Genetics* **151**, 797–801.

Harr, B. & Schlötterer, C. (2000). Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. *Genetics* **155**, 1213–1220.

Harr, B., Weiss, S., David, J. R., Brem, G. & Schlötterer, C. (1998). A microsatellite-based multilocus phylogeny of the *Drosophila melanogaster* species complex. *Current Biology* **8**, 1183–1186.

Kliman, R. M., Andolfatto, P., Coyne, J. A., Depaulis, F., Kreitman, M., Berry, A. J., McCarter, J., Wakeley, J. & Hey, J. (2000). The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156**, 1913–1931.

Litt, M. & Luty, J. A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics* **44**, 397–401.

Muir, G., Fleming, C. C. & Schlötterer, C. (2000). Species status of hybridizing oaks. *Nature* **405**, 1016.

Nei, M., Tajima, F. & Tateno, Y. (1983). Accuracy of estimated phylogenetic trees from molecular data. *Journal of Molecular Evolution* **19**, 153–170.

Paetkau, D., Waits, L. P., Clarkson, P. L., Craighead, L. & Strobeck, C. (1997). An evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. *Genetics* **147**, 1943–1957.

Pamilo, P. & Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5**, 568–583.

Pollock, D. D., Bergman, A., Feldman, M. W. & Goldstein, D. B. (1998). Microsatellite behavior with range constraints: parameter estimation and improved distances for use in phylogenetic reconstruction. *Theoretical Population Biology* **53**, 256–271.

Powell, J. R. (1997). *Progress and prospects in evolutionary biology: The Drosophila model*. Oxford University Press, Oxford.

Ritz, L. R., Glowatzki-Mullis, M. L., MacHugh, D. E. & Gaillard, C. (2000). Phylogenetic analysis of the tribe Bovini using microsatellites. *Animal Genetics* **31**, 178–185.

Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**, 365–371.

Schlötterer, C., Ritter, R., Harr, B. & Brem, G. (1998). High mutation rates of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. *Molecular Biology and Evolution* **15**, 1269–1274.

Schug, M. D., Mackay, T. F. C. & Aquadro, C. F. (1997). Low mutation rates of microsatellite loci in *Drosophila melanogaster*. *Nature Genetics* **15**, 99–102.

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

Stumpf, M. P. & Goldstein, D. B. (2001). Genealogical and evolutionary inference with the human Y chromosome. *Science* **291**, 1738–1742.

Takezaki, N. & Nei, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite data. *Genetics* **144**, 389–399.

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research* **17**, 6463–6471.

Ting, C. T., Tsaur, S. C. & Wu., C. I. (2000). The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. *Proceedings of the National Academy of Sciences of the USA* **97**, 5313–5316.

Weber, J. L. & May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* **44**, 388–396.

Xu, X., Peng, M. & Fang, Z. (2000). The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics* **24**, 396–399.

Zhivotovsky, L. A. (1999). A new genetic distance with application to constrained variation at microsatellite loci. *Molecular Biology and Evolution* **16**, 467–471.

Zhivotovsky, L. A. (2001). Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. *Molecular Biology and Evolution* **18**, 700–709.