Assessing the Genetic Structure of Microbial Populations

Richard E. Lenski

Proceedings of the National Academy of Sciences of the United States of America, Volume 90, Issue 10 (May 15, 1993), 4334-4336.

Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 4334–4336, May 1993

Commentary

Assessing the genetic structure of microbial populations

Richard E. Lenski

Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824-1325

A strenuous debate continues to rage over the use of DNA typing as forensic evidence (1–5). One of the most contentious issues has been how to calculate the probability of a coincidental match between the DNA of a suspect and DNA taken from the scene of a crime (and thought to belong to the perpetrator). In principle, this probability depends upon the genetic structure of the human population, including the extent to which alleles are associated statistically due to limited recombination among genetically distinct subpopulations.

More quietly, the genetic structure of microbial populations has been the subject of growing interest for more than a decade. No microbe faces conviction in a court of law for its offenses, although many have caused harm that would be the envy of any mass murderer. But an understanding of the genetic structure of microbial populations is relevant to tracking down the sources of epidemic outbreaks of pathogens, to managing the spread of resistance to drugs, and to improving the efficacy and safety of genetically engineered microorganisms intended for environmental applications.

The basic issue is whether bacteria and certain other microorganisms, including parasitic protozoa, exist as a series of asexual clones or as promiscuous, freely intermixing populations. The rapid spread of antibiotic-resistance genes led to a widespread view that genetic exchange was prevalent among bacteria in nature (6) and that clonal growth was an artifact of laboratory studies. But antibiotic-resistance genes are typically encoded by horizontally transmissible plasmids, and these genes have come under extraordinarily intense selection (7). Hence, antibiotic-resistance genes may not give a complete picture of the population structures of bacteria and other microorganisms (8).

In this issue of the *Proceedings*, Maynard Smith *et al.* (9) shed some new light on the analysis of population genetic structure in bacteria and parasitic protozoa. At first glance, their results are discomforting, because they show that conclusions regarding population structure sometimes depend on details of how one analyzes the genetic data. Fortunately, however, the differences that emerge depending on the method of analysis are illuminating and may reveal important biological processes.

The data sets analyzed by Maynard Smith *et al.* (9) are based on multilocus enzyme electrophoresis, whereby alleles at different loci are distinguished on the basis of differences in the mobility of their resulting proteins in an electrical field. Though the resolution of the method is now regarded as crude relative to DNA sequencing or fingerprinting, it has the virtue that hundreds of isolates can be screened for differences at tens of loci quickly and inexpensively (10). In 1973, Milkman (11) performed the first systematic survey of electrophoretic variation in a microorganism, showing that *Escherichia coli* harbored a tremendous wealth of molecular genetic variation. In 1980, Selander and Levin (12) showed that, despite this tremendous allelic variation, there was much less genotypic variation in *E. coli* than one might have expected, because certain alleles were almost invariably found in association with one another—e.g., the two-locus haploid genotypes *AB* and *ab* might both be common, whereas *aB* and *Ab* are absent or extremely rare. The scarcity of certain multilocus genotypes relative to expectations under free recombination is an indication of *linkage disequilibrium*. Later studies strengthened Selander and Levin's inference that linkage disequilibrium in *E. coli* was due primarily to infrequent recombination, rather than population subdivision or other factors that could create disequilibrium. In particular, geographical variation accounts for only a small fraction of genetic variation in *E. coli* (13), and the same multilocus electrophoretic genotypes are found to persist stably over several decades (14).

Since Selander and Levin's benchmark paper (12), clonal population structures have been reported for many other bacterial species, including numerous pathogens as well as environmentally important bacteria (15, 16). It has also been reported that some parasitic protozoa have clonal population structures, raising the possibility that clonality may be common in this group of organisms as well (17). But even as the clonal structure of bacterial populations was being elevated to the status of a paradigm, there were also some indications that things might not be quite so simple. Dykhuizen and Green (18) showed that phylogenies constructed for different chromosomal genes (using DNA sequences) from the same set of *E. coli* isolates were significantly different from one another. This result makes sense only if recombination between lineages has been sufficiently common to disrupt the associations between loci, but not so common as to preclude constructing gene phylogenies in the first place. And two studies published last year suggest that recombination among chromosomal genes may be much more frequent in some other bacterial species. Istock *et al.* (19) found only slight linkage disequilibrium in a population of *Bacillus subtilis* taken from one site in an Arizona desert; what disequilibrium did exist was much less extreme than in a correspondingly localized population of *E. coli* (collected from a single host, ref. 20). They concluded that recombination must be much more frequent relative to binary fission in *B. subtilis* than in *E. coli*. Souza *et al.* (21) investigated *Rhizobium leguminosarum* nodulating wild and cultivated beans at several different geographical scales, ranging from individual host plants to throughout the Western Hemisphere. Although they observed extremely strong linkage disequilibrium in the data set as a whole, most of this disequilibrium could be attributed to geographical subdivision. Souza *et al.* (21) also found that a local population of *R. leguminosarum* showed much less extreme disequilibrium than a local population of *E. coli*, suggesting that recombination may be much more frequent in *Rhizobium* than in *E. coli*.

Brown *et al.* (22, 23) proposed a useful index to describe linkage disequilibrium, which is based on the distribution of allelic mismatches between pairs of organisms over several loci. For example, if one organism has haploid genotype $A_1B_1C_1D_1E_1$ and another has genotype $A_2B_1C_2D_2E_1$, then the number of mismatches between them is three. The mean number of pairwise mismatches among a set of organisms is a measure of genetic distance. The variance in the number of pairwise mismatches, relative to that expected under the hypothesis of panmixia (i.e., random association of alleles), provides an index that allows clonality to be evaluated statistically. This

index has been used in several studies of the genetic structure of bacterial populations, but it was unclear whether the most appropriate units to compare in a pairwise fashion were all individual isolates or the subset of distinct genotypes. Some authors favored the former, on the grounds that the relative abundance of different genotypes provides important information about population structure; others favored the latter, because it was more conservative with respect to rejecting the null hypothesis of panmixia.

Maynard Smith *et al.* (9) show that this decision can dramatically affect one's interpretation of population structure. In *Neisseria meningitidis*, for example, there is a highly significant clonality when the 688 isolates are compared in a pairwise fashion; but when the 331 electrophoretically distinct genotypes are compared, they appear to be almost indistinguishable from panmixia. What accounts for this effect? It does not appear to be simply a reduction in statistical power associated with the loss of information: when the 331 genotypes were used instead of the 688 isolates, the index of multilocus linkage disequilibrium was reduced from 1.96 to 0.21, whereas its associated standard error increased only slightly, from 0.05 to 0.08. Rather, Maynard Smith *et al.* (9) suggest that such a discrepancy indicates an epidemic population structure, which arises "because of the recent, explosive increase in particular electrophoretic types" in species that nonetheless undergo rather frequent recombination. They predict that, with the passage of time, the association between a particular electrophoretic type and its evolutionary success will be eroded by recombination (see also ref. 24), and they offer evidence to suggest that this erosion is indeed occurring. So it seems reasonable to conclude that the relatively recent proliferation of a few clones in an epidemic fashion is largely responsible for the linkage disequilibrium that is observed in *N. meningitidis*.

By contrast, *Neisseria gonorrhoeae* gives no indication of deviation from panmixia even when all isolates are compared in a pairwise fashion (9). Evidently, there can be considerable diversity in population genetic structure even among closely related species. The difference between *N. gonorrhoeae* and *N. meningitidis* could be due to intrinsic differences in their capacity for gene transfer, but that is not the only possible explanation. The genetic diversity of pathogens within an individual host, which determines the likelihood that genetic exchange will produce an electrophoretically distinguishable recombinant genotype, depends critically on the multiplicity of infection. As Maynard Smith *et al.* (9) say, "It must also be the case that human behavior ensures frequent opportunities for gonococci of different genotypes to meet." Further information concerning the multiplicity of infections by distinct genotypes, as well as more direct estimates of rates of genetic exchange, would be necessary to explain the difference in population structure between these two species of *Neisseria*. But it is clear from the results of Maynard Smith *et al.* (9) that simultaneous analyses of linkage disequilibrium on both isolates and genotypes can reveal potentially important insights into the causes of population structure.

Going from the full set of all isolates to the subset of electrophoretically distinct genotypes is only one of several potentially meaningful ways to subdivide data on population structure. Whittam *et al.* (13) and Souza *et al.* (21) have previously subdivided data obtained for *E. coli* and for *R. leguminosarum* on geographic bases. In *E. coli*, geographic subdivision appears to contribute very little to either allelic diversity or linkage disequilibrium (13), whereas in *R. leguminosarum*, both allelic diversity and linkage disequilibrium increase markedly with geographic distance (21). Maynard Smith *et al.* (9) also consider the effects of subdividing sets of data on phylogenetic grounds. In *Rhizobium meliloti*, the average genetic distance between electrophoretic genotypes could be reduced by about half by splitting this species into two taxonomic divisions, which also apparently differ in their geographic distribution. When the two divisions were analyzed separately, there was little or no linkage disequilibrium in either group. The reasonable conclusion is that recombination between genotypes within either division is rather frequent, but there is little or no recombination between genotypes belonging to the different divisions. Without further information, however, one cannot determine whether the lack of genetic exchange between the two divisions is due to an intrinsic inability to exchange genes or whether representatives of the two divisions are rarely found in close proximity because of differences in their geographic range or habitat specificity.

It is clear that subdividing data on population structure into meaningful subsets can reveal useful information. I would suggest, however, that such procedures also require a degree of caution. One can easily imagine that a data set that has been sufficiently subdivided may lose the statistical power to reject the null hypothesis of panmixia, so that one must be cautious in accepting the joint inference that subsets are in linkage equilibrium whereas data sets as a whole exhibit clonality. One can see a possible illustration of this type of problem in the case of *Bordetella bronchiseptica*, which shows highly significant linkage disequilibrium when all 304 isolates are analyzed but which shows only a marginally significant deviation from panmixia when the 21 distinct electrophoretic genotypes are analyzed (9). Should we conclude that this species, like *N. meningitidis*, has an epidemic population structure? In the case of *N. meningitidis*, the index of disequilibrium dropped dramatically, and the standard error increased only slightly, when distinct genotypes were analyzed instead of all isolates. For *B. bronchiseptica*, however, the index of disequilibrium dropped only slightly (from 1.29 to 0.99), whereas the standard error increased to a much greater extent (from 0.08 to 0.30) when genotypes were analyzed instead of isolates. Maynard Smith *et al.* (9) suggest that one should "distinguish only between values of [the index] that differ significantly from zero and those that do not." But this recommendation risks, in my opinion, confusing statistical significance with biological importance. I see no simple resolution to this dilemma and suggest that both statistical significance and the magnitude of any change in the index should be respected when drawing inferences on population structure from subdivided data.

Leaving aside troubling statistical issues that may arise in borderline cases, Maynard Smith *et al.* (9) have demonstrated how one can get much useful information about the genetic structure of microbial populations by subdividing the data along biologically meaningful lines. Their results support the growing suspicion that all bacteria are not equally clonal.

1. National Research Council (1992) *DNA Technology in Forensic Science* (Natl. Acad. Press, Washington, DC).
2. Lewontin, R. C. & Hartl, D. L. (1991) *Science* **254,** 1745–1750.
3. Chakraborty, R. & Kidd, K. (1991) *Science* **254,** 1735–1739.
4. Risch, N. & Devlin, B. (1992) *Science* **255,** 717–720.
5. Weir, B. S. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 11654–11659.
6. Falkow, S. (1975) *Infectious Multiple Drug Resistance* (Pion, London).
7. Davey, R. P. & Reanney, D. C. (1980) *Evol. Biol.* **13,** 113–147.
8. Maynard Smith, J., Dowson, C. G. & Spratt, B. G. (1991) *Nature (London)* **349,** 29–31.
9. Maynard Smith, J., Smith, N. H., O'Rourke, M. & Spratt, B. G. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 4384–4388.
10. Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N. & Whittam, T. S. (1986) *Appl. Environ. Microbiol.* **51,** 873–884.
11. Milkman, R. (1973) *Science* **182,** 1024–1026.
12. Selander, R. K. & Levin, B. R. (1980) *Science* **210,** 545–547.
13. Whittam, T. S., Ochman, H. & Selander, R. K. (1983) *Mol. Biol. Evol.* **1,** 67–83.

14. Ochman, H. & Selander, R. K. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 198–201.
15. Selander, R. K., Musser, J. M., Caugant, D. A., Gilmour, M. M. & Whittam, T. S. (1987) *Microb. Pathog.* **3,** 1–17.
16. Piñero, D., Martinez, E. & Selander, R. K. (1988) *Appl. Environ. Microbiol.* **54,** 2825–2832.
17. Tibayrenc, M., Kjellberg, F., Arnaud, J., Oury, B., Breniere, S. F., Darde, M.-L. & Ayala, F. J. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 5129–5133.
18. Dykhuizen, D. E. & Green, L. (1991) *J. Bacteriol.* **173,** 7257–7268.
19. Istock, C. A., Duncan, K. E., Ferguson, N. & Zhou, X. (1992) *Mol. Ecol.* **1,** 95–103.
20. Caugant, D. A., Levin, B. R. & Selander, R. K. (1981) *Genetics* **98,** 467–490.
21. Souza, V., Nguyen, T. T., Hudson, R. R., Piñero, D. & Lenski, R. E. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 8389–8393.
22. Brown, A. H. D., Feldman, M. W. & Nevo, E. (1980) *Genetics* **96,** 523–536.
23. Brown, A. H. D. & Feldman, M. W. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 5913–5916.
24. Milkman, R. & Bridges, M. M. (1990) *Genetics* **126,** 505–517.