L. Bish, S. Chipera, *Los Alamos Nat. Lab. Rep. LA-11289-MS* (1988).
6. J. Quade and T. E. Cerling, *Science* **250**, 1549 (1990).
7. J. F. Whelan and J. S. Stuckless, *Proceedings of the American Nuclear Society International Meeting on High-Level Radioactive Waste Management*, Las Vegas, NV, 8 to 12 April 1990 (American Nuclear Society, La Grange Park, IL, 1990), p. 930; J. F. Whelan and J. S. Stuckless, unpublished manuscript.
8. J. S. Stuckless and Z. E. Peterman, unpublished manuscript; J. L. Banner, G. J. Wasserburg, P. F. Dobson, A. B. Carpenter, C. H. Moore, *Geochim. Cosmochim. Acta* **53**, 383 (1989); N. Clauer and Y. Tardy, *C. R. Acad. Sci. Ser. D* **273**, 2191 (1971); K. D. Collerson, W. J. Ullman, T. Torgersen, *Geology* **16**, 59 (1988); A. Starinsky, M. Bielski, A. Ecker, G. Steinitz, *Isot. Geosci.* **1**, 257 (1983).
9. R. H. McNutt, S. K. Frape, P. Fritz, M. G. Jones, I. M. MacDonald, *Geochim. Cosmochim. Acta* **54**, 205 (1990); P. C. Smalley, R. Blomqvist, A. Raheim, *Geology* **16**, 354 (1988).
10. P. J. Patchett, *Earth Planet. Sci. Lett.* **50**, 181 (1980).
11. L. V. Benson and P. W. McKinley, *U.S. Geol. Surv. Open-File Rep. 85-484* (1985).
12. D. R. Muhs *et al.*, unpublished manuscript; D. R. Muhs, unpublished data.
13. B. J. Szabo and I. J. Winograd, unpublished manuscript.
14. I. J. Winograd, B. J. Szabo, T. B. Coplen, A. C. Riggs, *Science* **242**, 1275 (1988).
15. I. J. Winograd and W. Thordarson, *U.S. Geol. Surv. Prof. Paper 712-C* (1975).
16. D. B. Marshall *et al.*, *Proceedings of the American Nuclear Society International Meeting on High-Level Radioactive Waste Management*, Las Vegas, NV, 8 to 12 April 1990 (American Nuclear Society, La Grange Park, IL, 1990), p. 921.
17. R. W. Spengler and Z. E. Peterman, *Proceedings of the American Nuclear Society Meeting on High-Level Radioactive Waste Management*, Las Vegas, NV, 28 April to 3 May 1991 (American Nuclear Society, La Grange Park, IL, 1991), p. 1416.
18. B. D. Marshall, Z. E. Peterman, J. S. Stuckless, K. Futa, S. A. Mahan, unpublished manuscript.
19. Z. E. Peterman, J. S. Stuckless, J. S. Downey, E. D. Gutentag, *Geol. Soc. Am. Abs. Programs* **22**, A295 (1990); U.S. Geol. Surv., unpublished data.
20. B. D. Marshall, Z. E. Peterman, K. Futa, J. S. Stuckless, *Proceedings American Nuclear Society International Meeting on High-Level Radioactive Waste Management*, Las Vegas, NV, 28 April to 3 May 1991 (American Nuclear Society, La Grange Park, IL, 1991), p. 1423.
21. R. E. Zartman and L. M. Kwak, unpublished manuscript.
22. J. K. Osmond and J. B. Cowart, in *Uranium Series Disequilibrium—Applications to Environmental Problems*, M. Ivanovich and R. S. Harmon, Eds. (Clarendon, Oxford, England, 1982), p. 202.
23. J. N. Rosholt *et al.*, *U.S. Geol. Surv. Open-File Rep. 85-540* (1985).
24. K. R. Ludwig, K. R. Simmons, B. J. Szabo, A. C. Riggs, I. J. Winograd, *Geol. Soc. Am. Abstr. Programs* **22**, A310 (1990).
25. Environmental Protection Agency, *U.S. Environ. Prot. Agency Rep. EMSL-LV-539-4* (1976); R. A. Zielinski and J. N. Rosholt, *U.S. Geol. Surv. J. Res.* **6**, 489 (1978); J. K. Osmond and J. B. Cowart, unpublished data from B. J. Szabo, personal communication.
26. The data summarized in this report are the results of efforts by a number of people at the U.S. Geological Survey and Lawrence Livermore National Laboratory including B. D. Marshall, D. T. Vaniman, J. F. Whelan, and R. E. Zartman. E. M. Taylor provided instructive comments and assisted in developing the final pedogenic model for the veins exposed at Trench 14. Unpublished manuscripts (*7, 8, 12, 13, 19, 21*) have been submitted as parts of a USGS Bulletin. Supported by the U.S. Department of Energy, Yucca Mountain Site Characterization Project.

24 April 1991; accepted 1 August 1991

Gene Trees and the Origins of Inbred Strains of Mice

WILLIAM R. ATCHLEY AND WALTER M. FITCH

Extensive data on genetic divergence among 24 inbred strains of mice provide an opportunity to examine the concordance of gene trees and species trees, especially whether structured subsamples of loci give congruent estimates of phylogenetic relationships. Phylogenetic analyses of 144 separate loci reproduce almost exactly the known genealogical relationships among these 24 strains. Partitioning these loci into structured subsets representing loci coding for proteins, the immune system and endogeneous viruses give incongruent phylogenetic results. The gene tree based on protein loci provides an accurate picture of the genealogical relationships among strains; however, gene trees based upon immune and viral data show significant deviations from known genealogical affinities.

ESTIMATING PHYLOGENETIC RELATIONSHIPS among groups of organisms is complicated by the existence of different types of phylogenetic trees. For example, species trees represent the putative evolutionary pathways of a group of species or populations whereas gene trees represent putative evolutionary pathways depicting changes in homologous genes sampled from different taxa (*1–3*). Gene trees can be constructed from DNA sequences, protein sequences, electrophoretic data, and antigenic data. Trees can also be constructed from genetic data obtained from DNA-DNA hybridization, chromosome structure, and morphological traits. However, different data often produce quite different trees and, as a result, provide qualitatively different estimates of phylogenetic relationships (*1–9*).

A major hindrance to evaluating concordance among different estimates of evolutionary relationships is a lack of examples where extensive data on genetic divergence exist for groups of organisms with well-documented phylogenies. For inbred strains of mice (*Mus musculus*), there is a wealth of genetic data from various levels of organization for a large number of genetically distinct strains having reasonably well-known genealogies. Thus, they can provide a powerful source of data to test important evolutionary hypotheses including those about concordance between species and gene trees when the latter are based upon different types of genetic data (*4–8*).

Herein, we examine genetic divergence among 24 well-known inbred strains of mice using 144 distinct gene loci. These data permit examination of two important questions: First, do patterns of genetic divergence among these 144 loci accurately reflect known relationships among strains? Second, will different structured subsamples of loci give equivalent phylogenetic conclusions, that is, are gene trees based upon different types of loci concordant?

Data and analyses. The 24 inbred strains included in these analyses are (in alphabetic order): 129, A, AKR, BALB/c, BDP, BUB, CBA, CE, C3H, C57BL, C57BR, C58, DBA/1, DBA/2, I, LP, NZB, P, RF, SEA, SEC, SJL, ST, and SWR. The actual genotypes for the 144 homozygous loci employed in these analyses for each of the 24 inbred strains are listed by Lyon and Searle (*10*). Only cladistically informative loci are included (those with at least two alleles, each present in two or more strains). A list of the actual loci included in these analyses is available from the authors. When genetic differences exist among sub-lines of any of these strains, the Jackson Laboratory sub-line was chosen as standard. Figure 1A summarizes the known genealogy of these 24 strains (*10–12*). These 24 strains were deliberately chosen to include some of the most widely used inbred mouse strains including commonly used stocks of uncertain origin (*13*). For purposes of this discussion, "genealogical relationships" refers to affinities among strains known from the original crossing experiments (Fig. 1A). "Phylogenetic relationships" refers to estimations of affinities inferred from analyses of the genetic data. Phylogenetic trees presented here are gene trees because their structure reflects patterns in reduction of residual heterozygosity (*1*).

Phylogenetic analyses of all loci. A matrix of pairwise genetic distance estimates (D) was generated for the 24 strains based upon all 144 loci. The genetic distance value (D) represents the percentages of the fixed alleles that differ between a pair of strains (*5*). All loci are homozygous within a strain.

As points of reference, the following are strains of known relationships together with their corresponding D values and an esti-

W. R. Atchley, Department of Genetics, North Carolina State University, Raleigh, NC 27695.
W. M. Fitch, Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92717.

mate of genealogical relationship derived from the original crosses and computed as (c = 1 − coefficient of kinship) (*14*): DBA/1 and DBA/2 (D = 11.8, c = 0.0); CBA and C3H (D = 20.1, c = 0.38); C57BL and C57BR (D = 23.8, c = 0.75); C57BR and C58 (D = 23.1, c = 0.88); and C57BL and C58 (D = 30.4, c = 0.88). By comparison, pairs of strains with independent origins, where c = 1.0 (Fig. 1A), have much larger D values: for example, for A and 129, D = 48.8; for BALB/c and AKR, D = 48.1; and for C58 and CBA, D = 58.4.

Figure 1B shows the phylogenetic relationships among taxa based upon all 144 loci when the parsimony algorithm is used (*15*). Bootstrap procedures (*16*) give measures of support for individual groupings given by the phylogenetic analysis. Comparing the phylogenetic analysis with the known genealogy shows close agreement. Indeed, the phylogenetic analysis provides correct estimates of evolutionary relationships for almost every instance where genealogical relationships are known among strains. Minor discrepancies sometimes arise when crosses among strains were originally used to generate mouse strains leading to a pattern of reticulated evolution. The computer algorithm assumes a conventional bifurcating phylogeny and cannot accurately reproduce reticulated patterns.

Phylogenetic analyses with different types of genetic data. An important question underlying comparative phylogenetic analyses is whether independent subsamples of the genome give equivalent estimates of phylogenetic relationships. Several previous reports indicate a lack of concordance in estimates of genetic similarity on the basis of morphological and single locus data (*4, 6, 9*). Incongruity in phylogenetic analyses is expected when data represent quite different levels of developmental organization and complexity. However, unanswered is whether structured subsamples of loci at the same level of organization and complexity give congruent estimates of phylogenetic relationships.

To test the null hypothesis that structured subsamples give equivalent estimates of phylogenetic relationships, the data were subdivided into protein (including enzyme) coding loci (70 loci), loci coding for the immune system (41 loci), and endogenous viral loci (22 loci). (Eleven loci could not be easily placed in these categories and were not included in these separate analyses.) The phylogenetic analyses and bootstrap procedures were repeated on these three subsamples of loci.

Protein loci. The relationships shown in Fig. 1 are preserved by the protein-coding loci, except for the placement of strains A, 129, and LP (Fig. 2A). Strain A is placed in the C3H, CBA, and CE lineage. Strain LP and 129 are genetically similar but no longer form a clade; rather, they appear as separate lineages. A seeming deviation involving NZB is resolved in the alternative equally parsimonious tree. Likewise, a seeming deviation in the relationship of BUB with SWR is also resolved by the alternative tree. Thus, with minor exceptions, protein data provide an accurate reflection of the relationships shown in Fig. 1.

Immune loci. Phylogenetic results for the immune loci (Fig. 2B) deviate significantly from the genealogical relationships shown in Fig. 1A. Some known genealogical relationships are reproduced by the immune data (for example, C57BL and C58; P and BDP; CBA and C3H; DBA/1 and DBA/2; and SEC and BALB/c), many other known genealogical relationships are not correctly described by the immune data. Among the latter are the placement of C57BR, SJL, LP, A, and 129. Further evidence for lack of robustness of the immune data is found in the much lower levels of statistical confidence for the various branching sequences given by the bootstrap procedures.

Endogenous viral loci. Data exist for approximately 40 loci that reflect integration sites of proviral genomes of endogenous ecotropic and nonecotropic murine leukemia viruses (*10*). DNA copies of retroviral genomes integrate into mouse chromosomes and persist as stable genetic elements (*17*). These proviruses replicate with the

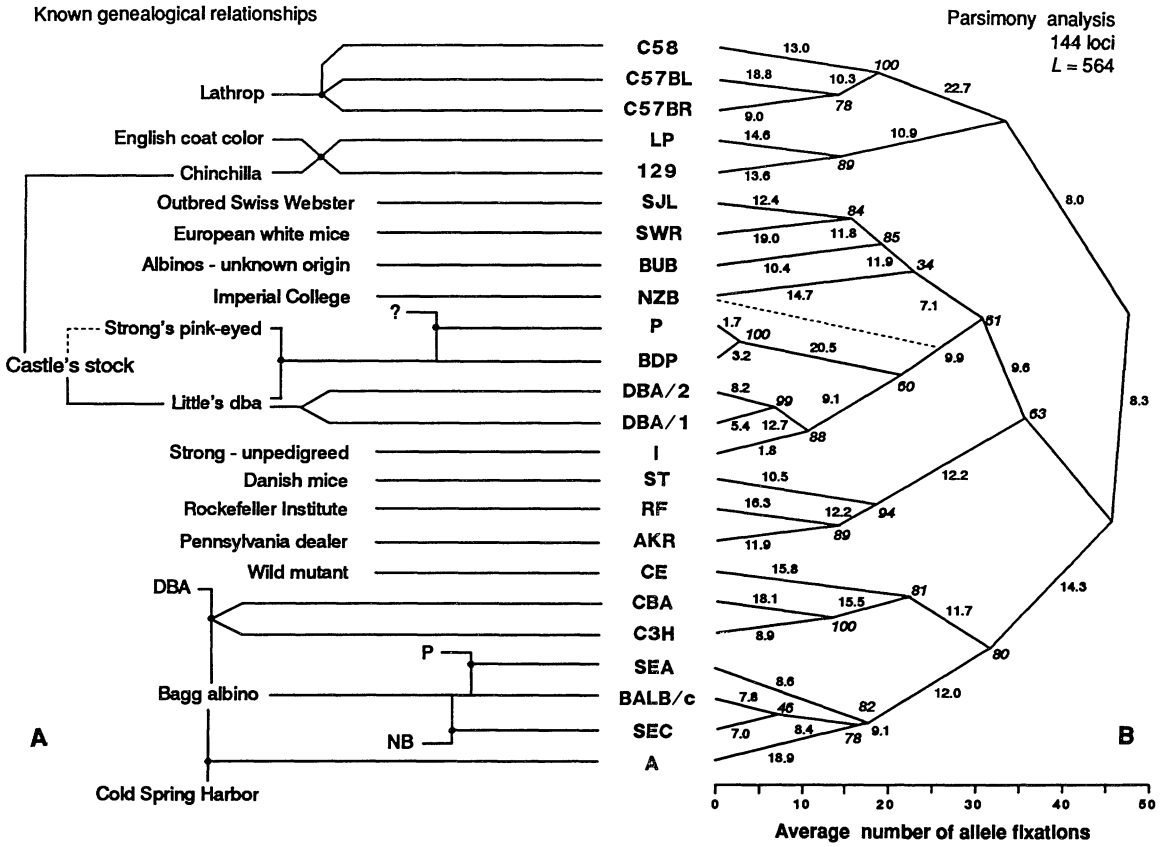

Fig. 1. (**A**) Genealogy of 24 inbred strains of mice (*10–12*). Black dots refer to crosses made in production of a particular strain. (**B**) Parsimony analysis of 144 single locus genotypes (rather than a genetic distance matrix) for 24 inbred strains of mice. Phylogenetic trees representing estimated relationships among strains were produced by using the parsimony algorithm (*15*). The length of the tree (L) reflects a total of 565 allele fixations. The branch lengths are in allele fixations (arabic numbers). The numbers in italics reflect the number (out of 100 resamplings of the data) that give the same clade (*16*). Nonintegral numbers of substitutions (allele fixations) arise from averaging over all possible parsimonious solutions for the tree. Dotted lines represent alternative but equally parsimonious joinings of taxa.

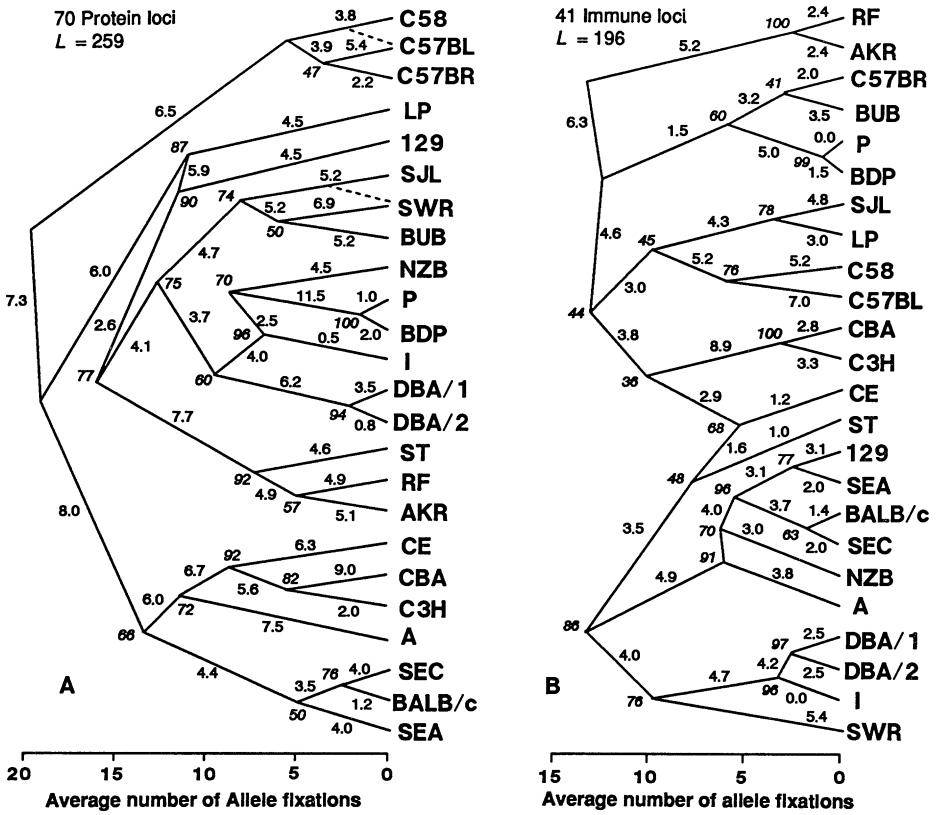

Fig. 2. (**A**) Parsimony analysis of 70 protein loci for 24 inbred strains of mice. (**B**) Parsimony analysis of 41 immune-coding loci for 24 inbred strains of mice. See the legend to Fig. 1B for additional details.

cellular genome and can be inherited through the germline as classical Mendelian genes (*18–20*). In addition, inbred mouse strains can be distinguished by their patterns of spontaneous and induced virus expression (*20, 21*). Some authors suggest that viral DNA integration occurred during speciation of the various murine lineages preceding establishment of inbred mouse lines (*22*). Furthermore, these integrations are relatively stable (*19*). Thus, these proviral genes contain information that is potentially useful in phylogenetic analyses.

Twenty-two viral loci exhibit genetic variation among these 24 strains and are included in these subsequent phylogenetic analyses (Fig. 3). These 22 loci include 18 used in the analyses described in Fig. 1B and four additional loci where only a single strain exhibits a polymorphism (that is, cladistically uninformative loci). Of these 22 viral loci, two are endogenous ecotropic murine leukemia viruses and 20 are endogenous nonecotropic viruses (*10*).

Strains BUB, DBA/1, and DBA/2 have identical viral genotypes, as do P and BDP, SEC, and BALB/c, and C3H and CBA (Fig. 3). In all these instances (except for BUB), identical viral genotypes are observed between closely related strains. An argument for chance infection by BUB for the viruses integrated into the DBA/1 and DBA/2 genomes can probably be discounted because these three strains share possession of eight viral integrations. Strains BUB, DBA/1, and DBA/2 differ at one virus locus from strain I and by two changes from P and BDP. Some clades in Fig. 1B are reasonably preserved including P, BDP, I, BUB, DBA/1, and DBA/2. However, most other clades in Fig. 1B are not preserved in Fig. 3. Bootstrap results indicate considerable support (>90%) for all branch points. Such high levels of statistical support are the consequence of the original sample being so small that it contains, by chance, no incongruent characters.

Incongruity in estimates of genetic divergence. The elements of the protein distance matrix are plotted as a function of those of the immune and viral distance matrices in Fig. 4. A matrix permutation test (*23*) was carried out to assess the extent of association among the three possible comparisons between genetic distance matrices. Ten thousand permutations were employed and the Kendall correlation statistic (r_τ) was used as a measure of association. Pairwise matrix correlations and their associated probability levels are as follows: protein versus immune, $r_\tau = 0.18, P = <0.001$; protein versus viral, $r_\tau = 0.25, P = <0.001$; and immune versus viral, $r_\tau = 0.07$, $P = 0.065$. In spite of having statistically significant Kendall statistics, Fig. 4 shows that the various sets of pairwise distance estimates are not strongly correlated. Indeed, it is clear from Figs. 2 and 3 that protein, immune, and viral data provide different estimates of the patterns of genetic divergence and, as a result, give different phylogeny estimates. Immune and viral data often deviate significantly from the known genealogy. Similar discrepancies are reported for electrophoretic, HLA, and red cell antigenic loci in humans (*24*).

Protein loci provide a much more accurate picture of genealogical relationships among these 24 strains than immune or endogeneous viral loci. There are several possible reasons.

1) Sampling variation: There are two aspects of sampling variation. First, different numbers of loci are included in these three data sets. Logically, one might expect a better fit to the known genealogy when significantly more data are available. Indeed, theoretical studies (*3*) suggest that a large number of independent loci are needed to determine which competing estimate of phylogeny is correct. However, the effects of differing numbers of loci on tree structure are difficult to rigorously evaluate. Conventionally, one could carry out a statistical resampling study to formally test the hypothesis that different sample sizes of loci give equivalent phylogenetic trees. For example, one would repeatedly obtain random subsamples of protein and immune loci from the full data, construct phylogenetic trees for each random sample, and analyze the resultant trees for equivalency. However, unlike conventional resampling studies, no statistical procedures exist to test whether the hundreds of trees obtained in this resampling procedure are equivalent. Lack of rigorous statistical procedures to evaluate the extent of homogeneity among estimated phylogenetic trees remains a serious problem in evolutionary studies.

Second, there are differing proportions of missing data per locus in the protein and immune data. Missing data arise when certain strains have not been characterized for a particular gene. In several instances, genetic distance estimates for immune data are based on as few as eight or nine loci, which represent only about 20% of the complete immune data (for example, comparisons involving SEC and SEA), whereas the data are complete in comparisons involving CBA and C3H. Thus, one might conclude that aberrant assignment of some strains results from stochastic effects of sampling variation.

However, sampling variation involving missing data is not the entire story. In some instances where the fewest loci are involved, the resultant phylogenetic arrangement is reasonably accurate. For example, 12 to 13 immune loci are used in estimating genetic

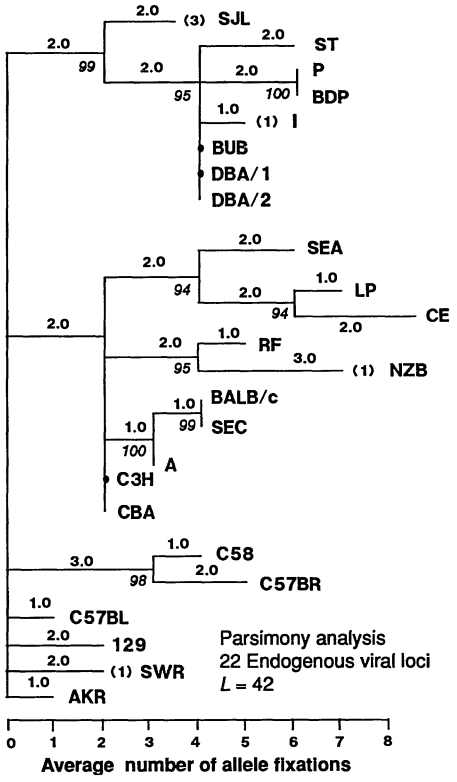

Fig. 3. Parsimony analysis of 22 endogenous viral loci for 24 inbred strains of mice. Numbers in parentheses after some strain names refer to the number of autapomorphies. This figure differs from previous ones in that branch lengths are proportional to the number of alleles that have been fixed (excluding autapomorphies) in the descent from node to node. Confidence levels for each node from the bootstrap procedure are given in italics.

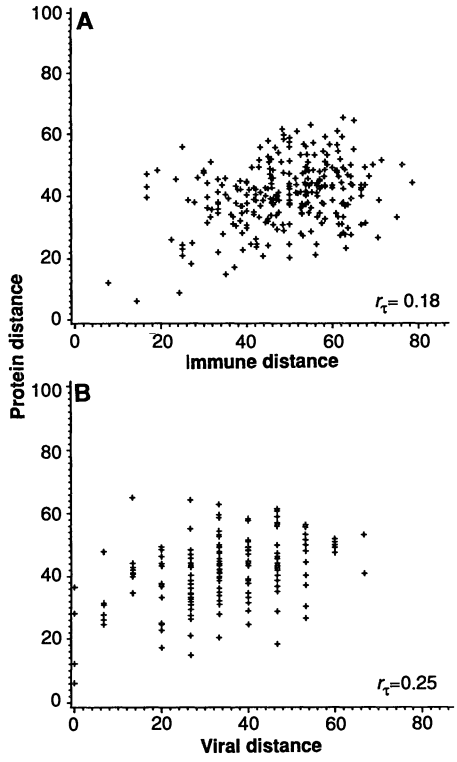

Fig. 4. Bivariate plots of the genetic distances for (**A**) protein loci as a function of immune loci and (**B**) protein loci as a function of viral loci.

distances among BALB/c, SEC, SEA, and 129. The first three taxa are correctly placed but 129 is not. Only 14 immune loci are involved in the correct placement of P with BDP, whereas 19 to 20 are involved with the accurate placement of I, DBA/1, and DBA/2. On the other hand, 36 to 37 immune loci are used to estimate genetic distances of C57BR to C57BL and C58, 26 to 27 are available between BDP and the two DBA strains, 26 loci are used for NZB and A, and 24 for SJL and LP. In each of these latter instances, calculated phylogenetic affinities do not agree with the known genealogy or analyses of the complete data.

2) Number of alleles: Allied to problems of sampling variability are complications arising from allelic diversity. Many more different alleles at several immune loci have become fixed by inbreeding compared to protein and viral loci. For example, eight different alleles are represented at the *H-2* and *Igh-1* loci and six alleles at the *H-1* and *Igh-2* loci. Accurate phylogenetic reconstructions can be more difficult when genes have a large number of alleles that may become independently fixed by random drift during inbreeding. Alleles are more likely to become uniquely fixed in some strains, making the position less informative cladistically, with the result that the effective number of loci is smaller than locus numbers alone would suggest.

3) Linkage disequilibrium: Tight linkage can affect the information content of loci in phylogenetic analyses because genes do not segregate independently. Thus, the effective information content for phylogenetic analyses might be significantly reduced if extensive linkage is present. Nine loci in these analyses (*C-4, H-2, H-2s, Qa-1, Qa-2, Qa-3, Qa-4, Slp*, and *Tla*) are part of the major histocompatibility complex (MHC) and map to the same region of chromosome 17. Similarly, five loci involved with immunoglobulin heavy chain complex map to the same region of chromosome 12, including *Igh-1, Igh-2, Igh-4, Igh-Src*, and *Igh-V*. Thus, rather than 40 independent loci, the immune data may reflect as few as about 28 independent loci.

4) Selection: One assumes with inbred strains that, apart from linkage to recessive lethals, gene fixation is the result of random drift and that overt selection for combinations of alleles has not occurred. However, existence of maternal-fetal histocompatibility interactions (*25*) suggests that inadvertent selection may have occurred for combinations of maternal-fetal genotypes during establishment of these inbred lines. If such selection occurred, it may have distorted estimates of phylogenetic relationships obtained with only immune loci.

These analyses suggest that not all genetic data have equivalent information content for phylogenetic reconstructions, in spite of the large numbers of gene loci used and statistically significant correlations among genetic distance estimates for these three data sets. Complications arising from sampling variability, allelic diversity, linkage, and possible historical selection, may affect the information content of these various subsets of loci. Phylogenetic analyses based on 41 separate immune loci give results that deviate significantly from the known evolutionary history of the organisms. Analyses of the 22 viral loci similarly deviate from the known genealogy.

Because of the magnitude of genetic data used in these analyses, these results provide new information about the ancestors and genealogy of inbred strains whose origins have previously been unclear. These analyses show that the diversity of genetical and genealogical information available for the inbred mouse strains makes them an excellent source of material for testing certain evolutionary hypotheses about the concordance among different types of genetic data, rates of evolutionary change, and similar problems.

REFERENCES AND NOTES

1. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
2. P. Pamilo and N. Nei, *Mol. Biol. Evol.* **5**, 568 (1988); N. Takahata, *Genetics* **122**, 957 (1989).
3. C.-I. Wu, *Genetics* **127**, 429 (1991).
4. W. R. Atchley, S. Newman, D. E. Cowley, *ibid.* **120**, 239 (1988).
5. W. M. Fitch and W. R. Atchley, *Science* **228**, 1169 (1985). The genetic distance matrix for the 144 loci is available from the authors.
6. _____, in *Molecules and Morphology in Evolution, Conflict or Compromise?*, C. Patterson, Ed. (Cambridge Univ. Press, London, 1987), pp. 203–216.
7. S. D. Ferris, R. D. Sage, E. M. Prager, U. Ritte, A. C. Wilson, *Genetics* **105**, 681 (1983).
8. R. K. Wayne and S. J. O'Brien, *J. Mammal.* **67**, 441 (1986); M. F. W. Festing and T. H. Roderick, *Genet. Res.* **53**, 45 (1989).
9. M.-C. King and A. C. Wilson, *Science* **188**, 107 (1975); J. M. Lowenstein, *Am. Sci.* **73**, 541, 1985; C. Patterson, Ed., *Molecules and Morphology in Evolution, Conflict or Compromise?* (Cambridge Univ. Press, London, 1987); C. G. Sibley and J. Ahlquist, *J. Mol. Evol.* **20**, 2 (1984); A. H. Bledsoe and R. J. Raikow, *ibid.* **30**, 247 (1990); D. M. Hillis, *Annu. Rev. Ecol. Syst.* **18**, 23 (1987); K. J. Sytsma, *Trends Ecol. Evol.* **5**, 104 (1990); A. R. Wyss, M. J. Novacek, M. C. McKenna, *Mol. Biol. Evol.* **4**, 99 (1987).
10. M. F. Lyon and A. G. Searle, Eds., *Genetic Variants and Strains of the Laboratory Mouse* (Oxford Univ. Press, Oxford, 1989).
11. M. F. W. Festing, *Inbred Strains in Biomedical Research* (Oxford Univ. Press, Oxford, 1979).
12. J. Staats, *Cancer Res.* **45**, 945 (1985).
13. Six of these inbred strains (C57BL, C3H, BALB/c, DBA/2, CBA, and A) made up about 70% of the mice used in 1600 publications surveyed by Festing (*12*). Figure 1 shows that these six and a few additional widely used strains (for example, BDP and 129) have complicated and often interrelated origins whereas other commonly used strains (for example, AKR, CE, I, NZB, RF, SJL, ST, and SWR) have independent or uncertain origins and relationships with other strains.

14. c_{ij} = 0.75 when the initial mice used to found an inbred strain were related as full sibs, (for example, C57BL and C57BR) because their gametes had 25% of their alleles identical by descent. c_{ij} = 0.375 for C3H and CBA because these strains arose from a single cross of the Bagg albino and DBA lines. DBA had been inbred for about 11 years so C3H and CBA should be identical by descent for all alleles from DBA and as full sibs for those alleles from Bagg albino. See (*5*) for more details and examples.
15. W. M. Fitch, *Syst. Zool.* **20**, 406 (1973).
16. The bootstrap procedure sampled with replacement the loci as many times as they were present in the original data set. The ANCESTOR program of Fitch (*15*) was used to find a most parsimonious tree and each clade (the group of taxa comprising all descendants of a given ancestral node) of that tree was given a unit of support. If there was more than one most parsimonious tree, say t of them, each clade of each most parsimonious tree received $1/t$ units of support. By repeating this procedure 100 times, the units of support could range from zero to 100 and hence can be treated as a measure of data's support of the tree. However, under circumstances where not all most parsimonious trees are guaranteed to be found (and that is the case here), or those trees found are not a random sample of all of them, the measure of support is biased in a way that would slightly inflate that estimate (W. M. Fitch, unpublished work). Nevertheless, because clades of trees with no measures of support give the appearance of being more certain than trees with them, we think that presenting those values is more likely to reduce rather than inflate confidence in the results. The boot strap values indicate the robustness of the data for a given result, not the representativeness of the data nor the probability that the correct tree has been found.
17. Weiss *et al.* (*19*) provide the following characterization of ecotropic and nonecotropic viruses: Ecotropic viruses will grow in cells of the species from which they were isolated—for example, mouse virus propagates best in mouse cells, to a much lesser extent in other rodent cells, and not at all in higher primates. Nonecotropic or xenotropic viruses, on the other hand, are endogenous to one species but cannot replicate in that species. Xenotropic viruses are not pathogenic in any animal. Endogenous ecotropic viruses, on the other hand may be pathogenic. Transmission may be from one host animal to another by contact; however, a frequent mode of transmission is from parent to offspring. Vertical transmission may be by contact infection or by genetic transmission. (Inbred mice have a single genetic locus, *Fv-1*, that controls susceptibility or resistance to different virus strains.)
18. R. Weiss, N. Teich, H. Varmus, J. Coffin, *RNA Tumor Viruses* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, ed. 2, 1982).
19. N. A. Jenkins, N. G. Copeland, B. A. Taylor, B. K. Lee, *J. Virol.* **43**, 26 (1982).
20. C. A. Kozak, *Adv. Cancer Res.* **44**, 295 (1985).
21. S. F. Chen, D. Struuck, M. L. Duran-Reynals, F. Lilly, *Cell* **21**, 849 (1980).
22. C. A. Kozak and R. R. O'Neill, *Curr. Top. Microbiol. Immunol.* **127**, 349 (1986).
23. E. J. Dietz, *Syst. Zool.* **32**, 21 (1983). Standard bivariate product moment correlation coefficients are inappropriate for assessing the strength of association between pairwise distance statistics because of problems with independence.
24. T. McLellan, L. B. Jorde, M. H. Skolnick, *Am. J. Hum. Genet.* **36**, 836 (1984); N. Ryman, R. Chakraborty, M. Nei, *Hum. Hered.* **333**, 93 (1983).
25. I. Hings and R. E. Billingham, *J. Reprod. Immunol.* 7, 337 (1985); C. B. Coulam, S. B. Moore, W. M. O'Fallon, *Am. J. Reprod. Immunol. Microbiol.* **14**, 54 (1987); C. M. Warner, M. S. Brownell, M. Ewoldsen, *Biol. Reprod.* **38**, 17 (1988); P. W. Hedrick, *Genetics* **119**, 199 (1988).
26. We thank D. E. Cowley, J. F. Crow, M. M. Goodman, E. L. Green, M. McClure, H. Temin, R. Swanstrom, and B. Weir for their helpful advice and comments on the manuscript and E. Owens for technical assistance. Supported by NSF grants BSR-8507855 and BSR-8605518 and NIH grant GM-45344 to W.R.A. and NSF grant BSR-9096152 to W.M.F.

3 April 1991; accepted 10 July 1991

Activation of Early Gene Expression in T Lymphocytes by Oct-1 and an Inducible Protein, OAP^{40}

KATHARINE S. ULLMAN, W. MICHAEL FLANAGAN, CYNTHIA A. EDWARDS, GERALD R. CRABTREE*

After antigenic stimulation of T lymphocytes, genes essential for proliferation and immune function, such as the interleukin-2 (IL-2) gene, are transcriptionally activated. In both transient transfections and T lymphocyte–specific in vitro transcription, the homeodomain-containing protein Oct-1 participated in the inducible regulation of transcription of the IL-2 gene. Oct-1 functioned in this context with a 40-kilodalton protein called Oct-1–associated protein (OAP^{40}). In addition to interacting specifically with DNA, OAP^{40} reduced the rate of dissociation of Oct-1 from its cognate DNA-binding site, suggesting that a direct interaction exists between Oct-1 and OAP^{40}.

T LYMPHOCYTES CAPABLE OF RESPONDING to virtually any antigen are produced in the thymus as a result of intrathymic differentiation and selection. Mature T cells migrate to peripheral lymphoid organs and remain in a quiescent state until the presentation of a specific antigen. Antigen binding produces an orderly and sequential series of gene activations that result in proliferation and immunologic function. Once activated, T cells help to coordinate the immune response through the production of cytokines necessary for the function of B cells, macrophages, and other cell types (*1*). One of the first cytokines to be produced in this process is IL-2, which induces synthesis of its own receptor (*2*) and autoregulates T cell proliferation (*3*). The minimal IL-2 enhancer contains recognition sites for several DNA binding proteins, including two octamer motifs (Fig. 1A) (*4, 5*); however, it is not known which octamer protein or proteins function at these sites. We now demonstrate that Oct-1 (OTF-1, OBP100, NF-III) and an inducible auxiliary protein participate in activating transcription of the IL-2 gene. The association of Oct-1 with a second transcriptional regulatory protein is reminiscent of the VP16–Oct-1 interaction that reprograms gene expression after infection by herpes simplex virus (*6*).

The proximal octamer motif is found in an antigen-receptor response element (ARRE-1) that confers transcriptional activation on a heterologous promoter in response to signals initiated at the antigen receptor (*4*). To assess the function of an octamer-binding protein in the inducible regulation of ARRE-1, we introduced site-directed mutations into the IL-2 enhancer (Fig. 1, A and B), which was ligated upstream of the chloramphenicol acetyltransferase (CAT) reporter gene. Transcriptional activity of these constructs was measured by transient transfection in Jurkat cells, a human T cell lymphoma line that mimics early events of T cell activation when stimulated through the antigen receptor (*7*). A 2-bp mutation that changed the proximal octamer sequence of the IL-2 enhancer to a related octamer motif found in the histone H2b promoter increased activity to 134% (mean) of the wild-type enhancer after stimulation with calcium ionophore and phorbol 12-myristate 13-acetate (PMA) (Fig. 1C). In contrast, when we mutated two additional base pairs to render the histone octamer nonfunctional (*8*), inducible activity dropped to 31% of wild-type (Fig. 1C). On the basis of methylation interference assays that identified the contact guanosines in the octamer motif (Fig. 1B), we mutated two noncontact guanosines immediately 5′ to the octamer site. These changes reduced activation to 28% of wild-type activity, indicating that the octamer motif is important for ARRE-1 induction but is not the sole sequence needed for inducible activity.

To further test the function of the sequences within ARRE-1, we developed an in vitro transcription system. The IL-2 enhancer is induced four- to fivefold by in vitro transcription, as compared to the 100-fold activation measured in transient transfections. Although not as sensitive as in vivo assays, in vitro transcription qualitatively reflects the complex requirements for activating the IL-2 gene (*9*). Using the internal

K. S. Ullman, Department of Microbiology and Immunology, Stanford University, Stanford, CA 94305.
W. M. Flanagan and G. R. Crabtree, Howard Hughes Medical Institute, Beckman Center, Stanford University, Stanford, CA 94305.
C. A. Edwards, Genelabs, Redwood City, CA 94063.

*To whom correspondence should be addressed.