15. D. G. Hangauer, A. F. Monzingo, B. W. Matthews, *Biochemistry* **23**, 5730 (1984).
16. O. Dideberg *et al.*, *Nature* **299**, 469 (1982); J. M. Ghuysen, J. Lamotte-Brasseur, B. Joris, G. D. Shockman, *FEBS Lett.* **342**, 23 (1994).
17. W. Stocker and W. Bode, *Curr. Opin. Struct. Biol.* **5**, 383 (1995).
18. K. L. Constantine *et al.*, *J. Mol. Biol.* **223**, 281 (1992); A. R. Pickford, J. R. Potts, J. R. Bright, I. Phan, I. D. Campbell, *Structure* **5**, 359 (1997).
19. I. E. Collier, P. A. Krasnov, A. Y. Strongin, H. Birkedal-Hansen, G. I. Goldberg, *J. Biol. Chem.* **267**, 6776 (1992).
20. L. Bányai, H. Tordai, L. Patthy, *ibid.* **271**, 12003 (1996).
21. F. Willenbrock *et al.*, *Biochemistry* **32**, 4330 (1993); M. W. Olson, D. C. Gervasi, S. Mobashery, R. Fridman, *J. Biol. Chem.* **272**, 29975 (1997); C. M. Overall *et al.*, *ibid.* **274**, 4421 (1999).
22. J. Hodgson, *Biotechnology* **13**, 554 (1995); A. E. Yu, R. E. Hewitt, E. W. Connor, W. G. Stetler-Stevenson, *Drugs Aging* **11**, 229 (1997); S. A. Watson and G. Tierney, *Biodrugs* **9**, 325 (1998).
23. XDS [W. Kabsch, *J. Appl. Crystallogr.* **21**, 916 (1988)]; CCP4 programs [CCP4, Collaborative Computational Project No. 4, Daresbury, UK, *Acta Crystallogr.* D **50**, 760 (1994)]; SFCHECK [A. A. Vagin, J. Richelle, S. J. Wodak, *ibid.* **55**, 191 (1999)]; AMORE [J. Navaza, *Acta Crystallogr.* A **50**, 157 (1994)]; O [T. A. Jones, J.-Y. Zou, S. W. Cowan, M. Kjeldgaard, *ibid.* **47**, 110 (1991)]; and X-PLOR [A. T. Brünger, *X-PLOR, Version 3.1: A System for X-Ray Crystallography and NMR* (Yale Univ. Press, New Haven, CT, 1992)].
24. Figures 1, 2A, 3, A and B, and 4A were made with MOLSCRIPT [P. J. Kraulis, *J. Appl. Crystallogr.* **24**, 946 (1991)] and RASTER3D [E. A. Merrit and M. E. P. Murphy, *Acta Crystallogr.* D **50**, 869 (1994)]. Figures 3, C and D, and 4B were made with GRASP [A. Nichols, K. A. Sharp, B. Honig, *Proteins* **11**, 281 (1991)].
25. Supported by grants from the Swedish Cancer Foundation, EC project BMH4-CT 96-0012, Novo Nordisk Foundation, and Hedlund's Foundation. We thank Tiina Berg, Ilkka Miinalainen, and Kristian Tryggvason for technical assistance with insect cell cultures. We are also grateful to Richard Kahn and Tatjana Sandalova for assistance with the data collection. Beam time was provided by the ESRF.

21 December 1998; accepted 28 April 1999

Genetics of Mouse Behavior: Interactions with Laboratory Environment

John C. Crabbe,[1]* Douglas Wahlsten,[2] Bruce C. Dudek[3]

Strains of mice that show characteristic patterns of behavior are critical for research in neurobehavioral genetics. Possible confounding influences of the laboratory environment were studied in several inbred strains and one null mutant by simultaneous testing in three laboratories on a battery of six behaviors. Apparatus, test protocols, and many environmental variables were rigorously equated. Strains differed markedly in all behaviors, and despite standardization, there were systematic differences in behavior across labs. For some tests, the magnitude of genetic differences depended upon the specific testing lab. Thus, experiments characterizing mutants may yield results that are idiosyncratic to a particular laboratory.

Targeted and chemically induced mutations in mice are valuable tools in biomedical research, especially in the neurosciences and psychopharmacology. Phenotypic effects of a knockout often depend on the genetic background of the mouse strain carrying the mutation (*1*), but the effects of environmental background are not generally known.

Different laboratories commonly employ their own idiosyncratic versions of behavioral test apparatus and protocols, and any laboratory environment also has many unique features. These variations have sometimes led to discrepancies in the outcomes reported by different labs testing the same genotypes for ostensibly the same behaviors (*2*). Previous studies could not distinguish between interactions arising from variations in the test situation itself and those arising from subtle environmental differences among labs. Usually, such differences are eventually resolved by repetition of tests in multiple labs. However, null mutants and transgenic mice are often scarce and tend to be behaviorally characterized in a single laboratory with a limited array of available tests.

We addressed this problem by testing six mouse behaviors simultaneously in three laboratories (Albany, New York; Edmonton, Alberta, Canada; and Portland, Oregon) using exactly the same inbred strains and one null mutant strain (*3*). We went to extraordinary lengths to equate test apparatus, testing protocols, and all possible features of animal husbandry (*4*). One potentially important feature was varied systematically. Because many believe that mice tested after shipping from a supplier behave differently from those reared in-house, we compared mice either shipped or bred locally at the same age (77 days) starting at the same time (0830 to 0900 hours local time on 20 April 1998) in all three labs (*5*). Each mouse was given the same order of tests [Day 1: locomotor activity in an open field; Day 2: an anxiety test, exploration of two enclosed and two open arms of an elevated plus maze; Day 3: walking and balancing on a rotating rod; Day 4: learning to swim to a visible platform; Day 5: locomotor activation after cocaine injection; Days 6 to 11: preference for drinking ethanol versus tap water (*6*)].

Despite our efforts to equate laboratory environments, significant and, in some cases, large effects of site were found for nearly all variables (Table 1). Furthermore, the pattern of strain differences varied substantially among the sites for several tests. Sex differ-

Table 1. Statistical significance and effect sizes for selected variables in the multisite trial. Color of cell depicts Type I error probability or significance of main effects and two-way interactions from 8 × 2 × 3 × 2 analyses of variance: blue, $P < 0.00001$; purple, $P < 0.001$; gold, $P < 0.01$; dashes with no shading, $P > 0.01$. Cell entries are effect sizes, expressed as partial omega squared, the proportion of variance accounted for by the factor or interaction if only that factor were in the experimental design (range = 0 to 1.0). Multiple R^2 (unbiased estimate) gives the proportion of the variance accounted for by all factors. For the water escape task, results are based on only seven strains because most A/J mice never escaped because of wall-hugging. We recognize that the issue of appropriate alpha level correction for multiple comparisons is contentious. Details of the statistical analyses are available on the Web site (*4*), including a discussion of our rationale for presenting uncorrected values in this table.

Task	Measure	Eight Genotypes	Three Sites	Two Sexes	Local vs Shipped	Genotype x Site	Genotype x Sex	Genotype x Ship	Multiple R^2
Open field	Distance in 15 min	.600	.157	----	----	.059	.045	----	.604
Open field	# vertical movements	.788	.281	.039	----	----	----	----	.772
Cocaine	Difference from Day 1	.338	.053	----	----	.086	----	----	.342
Plus maze	Total arm entries	.385	.327	----	----	.210	----	----	.660
Plus maze	Time in open arms	.082	.212	----	----	.066	----	----	.266
Water maze	Mean escape latency	.221	----	----	.026	----	----	----	.177
Alcohol preference	Alcohol consumed (g/kg)	.483	----	.043	----	----	----	----	.451
Body size	Weight (g)	.408	.204	.637	----	.071	.070	----	.698

[1]Portland Alcohol Research Center, Department of Veterans Affairs Medical Center and Department of Behavioral Neuroscience, Oregon Health Sciences University, Portland, OR 97201, USA. [2]Department of Psychology, University of Alberta, Edmonton, Alberta, Canada T6G 2E9. [3]Department of Psychology, State University of New York at Albany, Albany, NY 12222, USA.

*To whom correspondence should be addressed. E-mail: crabbe@ohsu.edu

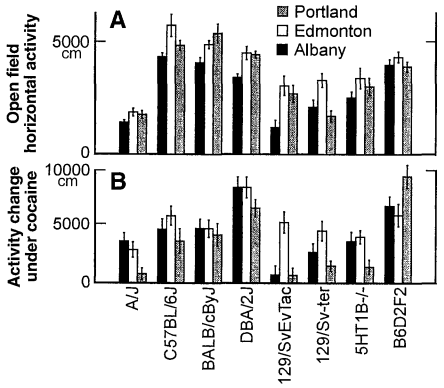

Fig. 1. Group means (±SEM for $n = 16$ mice) for activity in a 40 cm by 40 cm open field for eight strains tested at the same time of day in identical apparatus in three laboratories. (**A**) Horizontal distance (centimeters) traveled in 15 min on the first test on Day 1. (**B**) Cocaine-induced activation, expressed as the difference between horizontal activity (centimeters in 15 min) after cocaine (20 mg/kg) on Day 5 minus the score on Day 1.

ences were only occasionally detected, and, much to our surprise, there were almost no effects of shipping animals before testing. Large genetic effects on all behaviors were confirmed, which is not surprising because we chose strains known to differ markedly on these tasks.

Results for locomotor activity and the effect of a subsequent cocaine injection on locomotion are shown in Fig. 1. Expected strain differences in undrugged activity were found: A/J mice were relatively inactive at all three sites, whereas C57BL/6J mice were very active. An effect of laboratory was also found: mice tested in Edmonton were, on average, more active than those tested in Albany or Portland. In addition, the pattern of genetic differences depended on site. For example, 129/SvEvTac mice tested in Albany were very inactive compared to their counterparts in other labs. Similar results were seen for sensitivity to cocaine stimulation. For example, B6D2F2 mice were very responsive (and A/J mice quite insensitive) to cocaine in Portland, but not at other sites.

In the elevated plus maze, a very similar pattern was seen: strong effects of genotype, site, and their interaction. This was true both for activity measures and for time spent in open arms, the putative index of anxiety (Fig. 2). For total arm entries, the testing laboratory was particularly important for the 5-HT_{1B} knockout mice versus their wild-type 129/Sv-ter background controls. Knockout mice had greater activity than wild types in Portland and tended to have less activity in Albany, while not differing in Edmonton. Edmonton mice of all strains spent more time in open arms (lower anxiety). Portland mice also spent less time in open arms, but this was especially true for strains A/J, BALB/cByJ, and the B6D2F2 mice.

Although the testing laboratory was an important variable, there was a good deal of consistency to the genetic results as well. For example, comparison of the genotype means (averaged over sites) for the initial 5 min of the activity test on Day 1 with the total arm entry scores from the plus maze yielded a high correlation between strains ($r = 0.91$, $P < 0.002$). This indicates that a strain's characteristic activity in novel apparatus is robust and occurs in different apparatus as well as different labs (*7*).

For some behaviors, laboratory environment was not critical. For example, ethanol drinking scores were closely comparable across all three labs, and genotypes alone accounted for 48% of the variance (Table 1 and Fig. 3). The genetic differences showed the well-known pattern of C57BL/6J mice strongly preferring and DBA/2J mice avoiding ethanol (*8*). Females drank more, as is also well known (*8*), but there were no significant effects of site, shipping, or any other interactions. Unlike the other five tests, ethanol preference testing extended over 6 days in the home cage and involved a bare minimum of handling mice by the experimenter.

For some measures, the difference between 5-HT_{1B} null mutant and wild-type mice depended on the specific laboratory environment. In Edmonton, for example, no difference was observed between +/+ and –/– mice in distance traveled in the activity monitor, whereas there was greater activity in the knockouts at the other two sites, especially Portland ($P = 0.002$). In the elevated plus maze, knockouts were considerably more active than wild types only in Portland (Fig. 2A; $P = 0.02$).

The numbers of mice we tested made formal statistical assessment of reliability infeasible, but it would be important to know whether each laboratory would obtain essentially the same strain-specific results if this experiment were repeated. Because our experiment included an internal replication, we estimated the lower bounds of reliability for each site separately by correlating the mean scores for each strain (collapsed over sex and shipping group) obtained during the two replicates of the experiment. These correlations differed depending upon the behavior, and were consonant with the relative importance of genotype in the overall analysis. For example, for locomotor activity, the correlations were 0.97, 0.74, and 0.87 for the three sites. For open-arm time on the plus maze, possibly the most intrinsically unstable task we employed, the correlations were lower (0.32, 0.52, and 0.26). These can be compared to correlations for body weight, which can serve as a type of control variable not influenced by idiosyncratic dynamics of the test situation (0.83, 0.74, and 0.90). No site had generally higher or lower reliability than the others, and formal analyses of replication in analyses of variance indicated no strong interactions of strain by replication. We conclude that reasonable estimates of strain-specific scores are highly dependent on behavioral endpoint, and that some behaviors are highly stable.

Several sources of these laboratory-specific behavioral differences could be ruled out by the rigor of the experimental design. For example, Edmonton mice might have been

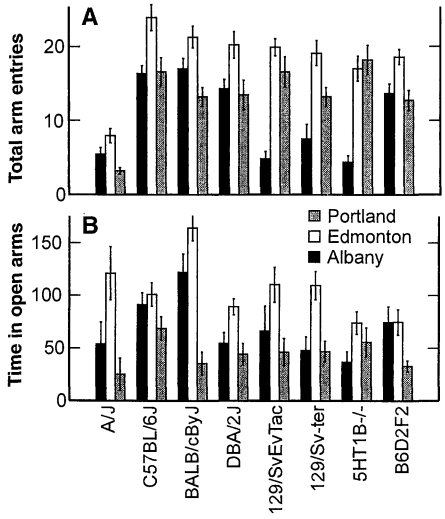

Fig. 2. Group means (±SEM for $n = 16$ mice) for behavior videotaped for 5 min on elevated plus mazes having two open and two enclosed arms. (**A**) Total number of entries into any arm (defined as all four limbs in the arm). (**B**) Time (seconds) spent in the two open arms during the 300-s test. Smaller amounts of time indicate higher levels of anxiety.

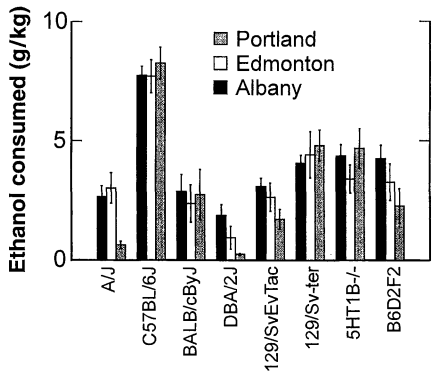

Fig. 3. Mean (±SEM) ethanol consumed per day, expressed as grams per kilogram body weight, over 4 days of an ethanol preference test where each mouse had free access to two drinking bottles, one with local tap water and the other with 6% ethanol in tap water.

more sensitive to cocaine-induced locomotion because the source of cocaine differed from the other two sites (*4*), but this could not explain the relatively marked response of the three 129-derived strains in Edmonton only. However, specific experimenters performing the testing were unique to each laboratory and could have influenced behavior of the mice. The experimenter in Edmonton, for example, was highly allergic to mice and performed all tests while wearing a respirator—a laboratory-specific (and uncontrolled) variable.

Whether animals were bred in each laboratory or shipped as adults 5 weeks before testing had no consistent influence on results in this experiment. Shipped animals took routes of varying duration and difficulty. For example, some Taconic mice were trucked to Albany from nearby Germantown, New York, whereas others spent 2 days in transit during a flight in midwinter to Edmonton. At least in this experiment, allowing animals a lengthy period of acclimation to new quarters was sufficient to overcome any strong effects of putative shipping stress on subsequent behavior.

These results support both optimistic and pessimistic interpretations. Seen optimistically, genotype was highly significant for all behaviors studied, accounting for 30 to 80% of the total variability, and several historically documented strain differences were also seen here. In general, we conclude that very large strain differences are robust and are unlikely to be influenced in a major way by site-specific interactions. However, a more cautious reading suggests that for behaviors with smaller genetic effects (such as those likely to characterize most effects of a gene knockout), there can be important influences of environmental conditions specific to individual laboratories, and specific behavioral effects should not be uncritically attributed to genetic manipulations such as targeted gene deletions.

When studying mutant mice, relatively small genetic effects should first be replicated locally before drawing conclusions (*9*). We further recommend that, if possible, genotypes should be tested in multiple labs and evaluated with multiple tests of a single behavioral domain (such as several tests of anxiety-related behavior) before concluding that a specific gene influences a specific behavioral domain. We also suggest the possibility that laboratory-specific effects on genetic differences will affect phenotypes other than behaviors to an extent similar to that we report.

It is not clear whether standardization of behavioral assays would markedly improve future replicability of results across laboratories. Standardization will be difficult to achieve because most behaviorists seem to have differing opinions about the "best" way to assay a behavioral domain. For example, two of us typically test behavior during the light phase of the animals' cycle, whereas the third typically tests during the dark phase (but switched to the light phase for this study). Which apparatus specifications or test protocol to employ is also a subject of differing opinion. There is a risk of prematurely limiting the "recommended" tests in a domain to those deemed "industry standard," because this may constrain the intrinsic richness of a domain and obscure interesting interactions. On the other hand, increased communication and collaboration between the molecular biologists creating mutations and behavioral scientists interested in the psychological aspects of behavioral testing will benefit both groups.

References and Notes

1. M. Sibilia and E. F. Wagner, *Science* **269**, 234 (1995); R. Gerlai, *Trends Neurosci.* **19**, 177 (1996); M. Nguyen *et al.*, *Nature* **390**, 78 (1997).
2. It has been known for some time that comparisons of multiple genotypes on learning-related tests do not always yield consistent results across laboratories [D. Wahlsten, in *Psychopharmacology of Aversively Motivated Behavior*, H. Anisman and G. Bignami, Eds. (Plenum, New York, 1978), pp. 63–118]. For another example, the Crabbe laboratory has reported that C57BL/6 mice show a small enhancement of locomotor activity after low doses of ethanol, while the Dudek laboratory finds no such stimulant response [J. C. Crabbe *et al.*, *J. Comp. Physiol. Psychol.* **96**, 440 (1982); B. C. Dudek and T. J. Phillips, *Psychopharmacology* **101**, 93 (1990)]. Similar variation has been reported in other measures of activity in various laboratories and apparatus [J. M. LaSalle and D. Wahlsten, in *Techniques for the Genetic Analysis of Brain and Behavior: Focus on the Mouse*, D. Goldowitz, D. Wahlsten, R. E. Wimer, Eds. (Elsevier, Amsterdam, 1992), pp. 391–406].
3. We tested males and females from the inbred strains: A/J, BALB/cByJ, C57BL/6J, DBA/2J, 129/Sv-ter, and 129/SvEvTac; the F_2 hybrid cross of C57BL/6J and DBA/2J (B6D2F2); and the serotonin receptor subtype null mutant, 5-$HT_{1B}^{-/-}$, which is maintained on the 129/Sv-ter background. Mice were obtained from the Jackson Laboratory (Bar Harbor, ME), Taconic Farms (Germantown, NY), or the colonies of R. Hen (Columbia University, New York, NY). Because many targeted deletions are placed on the 129/SvEvTac background, we included this close relative of 129/Sv-ter. The genealogy of many 129 substrains has recently been discussed [E. M. Simpson *et al.*, *Nature Genet.* **16**, 19 (1997); D. W. Threadgill, D. Yee, A. Matin, J. H. Nadeau, T. Magnuson, *Mamm. Genome* **8**, 390 (1997)].
4. Details of procedures and test protocols are given in the Web site for this study (www.albany.edu/psy/obssr). Variables explicitly equated across laboratories included apparatus, exact testing protocols, age of shipped and laboratory-reared mice, method and time of marking before testing, food (Purina 5001; Purina 5000 for breeders), bedding (Bed-o-cob, 1/4 inch; Animal Specialties, Inc., Hubbard, OR), stainless steel cage tops, four to five mice per cage, light/dark cycle, cage changing frequency and specific days, male left in cage after births, culling only of obvious runts, postpartum pregnancy allowed, weaned at 21 days, specific days of body weight recording, and gloved handling without use of forceps. Unmatched variables included local tap water, requirement of filters over cage tops in Portland only, variation of physical arrangement of colonies and testing rooms across sites, different air handling and humidity, and different sources of batches of cocaine and alcohol.
5. All breeding stock was shipped on 2 or 3 December 1997, and mating pairs were set simultaneously on 13 January 1998 in all labs to provide "unshipped" mice for testing. On 15 to 17 March 1998, a second batch of mice from each genotype was shipped to each laboratory. These "shipped" mice, age matched with the unshipped cohort already in place, were allowed to acclimate to the laboratory for 5 weeks before testing commenced. We tested 128 mice in each lab, in two groups of 64 separated by 1 week. With an $n = 4$ mice in each genotype/shipping condition/sex/laboratory condition, we had 16 mice per group for the crucial genotype × laboratory comparisons. This sample size gave us statistical power of 90% to detect modest interactions of genotype × laboratory when Type I error probability was set at 0.01 [J. Cohen, *Statistical Power Analysis* (Erlbaum, Hillsdale, NJ, 1988); D. Wahlsten, *Behav. Brain. Sci.* **13**, 109 (1990)]. For results of analysis of variance, we report only effects significant at $P < 0.01$. The Web site in (*4*) provides detailed protocols used for each test, descriptions of the laboratory conditions rigorously equated across labs, and raw data that may be examined for other interesting patterns.
6. AccuScan Digiscan monitors (AccuScan Instruments, Columbus, OH) were generously loaned to D. Wahlsten by R. H. Kant to match those available in the other two laboratories. AccuScan also provided all sites with rotarod apparatus. Mouse-scaled water mazes and elevated plus mazes were constructed by D. Wahlsten and shipped to the other two labs. On the first test day, each mouse was tested for 15 min in a Digiscan open-field monitor in a dark, sound-attenuated chamber. On Day 2, each mouse was videotaped for 5 min in an elevated plus maze. On Day 3, mice were given 10 trials on a rotarod set to accelerate from 0 to 100 rpm in 75 s. After all mice had been tested on the rotarod, mice were pretrained briefly to escape from the water maze. On Day 4, mice were given eight massed trials of escape learning to a visible platform in the water maze. On Day 5, the activity test was repeated immediately following an ip injection of 20 mg of cocaine per kilogram. After 2 days of rest, mice were individually housed, given only tap water for 2 days, and then tested for 4 days for drinking of 6% ethanol in tap water versus tap water alone.
7. J. Flint *et al.*, *Science* **269**, 1432 (1995); S. R. Mitchell, J. K. Belknap, J. C. Crabbe, unpublished observations.
8. G. E. McClearn and D. A. Rodgers, *Q. J. Stud. Alcohol* **20**, 691 (1959); J. K. Belknap, J. C. Crabbe, E. R. Young, *Psychopharmacology* **112**, 503 (1993); L. A. Rodriguez *et al.*, *Alcohol. Clin. Exp. Res.* **19**, 367 (1995).
9. It was previously reported that 5-HT_{1B} null mutant mice drank much more alcohol than the 129/Sv-ter wild-type strain [J. C. Crabbe *et al.*, *Nature Genet.* **14**, 98 (1996)]. In the experiments here, no site detected this difference (Fig. 3 and Table 1). The original outcome was replicated four times (J. C. Crabbe *et al.*, unpublished data). It is possible that residual polymorphisms for genes segregating in the 129/SvPas substrain that served as the original source of the embryonic stem cell line and in the 129/Sv-ter substrain to which the null mutant was crossed have subsequently been fixed differentially in the 5-HT_{1B} +/+ and –/– strains maintained at Columbia University (*3*). If so, these genes must exert very large epistatic effects on the 1B gene deletion's phenotypic effects on drinking (*1*). Alternatively, some undetected variable (for example, a change in animal care personnel) may have occurred specifically at the Portland site between the original (1995–96) observations and the current experiments.

10. Supported by the Office of Behavioral and Social Sciences Research, NIH, as supplements to grants AA10760 (J.C.C.) and DA10731 (J. Marley and B.C.D., co–principal investigators), and by the Natural Sciences and Engineering Research Council of Canada Grant # 45825 (D.W.), the Department of Veterans Affairs (J.C.C.), and a K02 Award to B.C.D. AA00170. We thank R. H. Kant at AccuScan for the generous loan of equipment and R. Hen for providing the serotonin receptor mutants. We appreciate the comments of C. Cunningham, R. A. Harris, J. Janowsky, and G. Westbrook on a draft of this manuscript. We also thank S. Boehm II, S. Burkhart-Kasch, J. Dorow, S. Doerksen, C. Downing, J. Fogarty, K. Henricks, C. McKinnon, C. Merrill, P. Metten, C. Nolte, T. Phillips, M. Schalomon, J. Schlumbohm, J. Sibert, J. Singh, and C. Wenger for valuable assistance.

1 February 1999; accepted 7 May 1999