

Different Data from Different Labs: Lessons from Studies of Gene–Environment Interaction

Douglas Wahlsten,¹ Pamela Metten,² Tamara J. Phillips,² Stephen L. Boehm II,² Sue Burkhart-Kasch,² Janet Dorow,² Sharon Doerksen,¹ Chris Downing,³ Jennifer Fogarty,³ Kristina Rodd-Henricks,³ René Hen,⁴ Carrie S. McKinnon,² Catherine M. Merrill,² Cedar Nolte,² Melike Schalomon,¹ Jason P. Schlumbohm,² Jason R. Sibert,² Charlotte D. Wenger,² Bruce C. Dudek,³ John C. Crabbe²

¹ Centre for Neuroscience and Department of Psychology, University of Alberta, Edmonton, AB T6G 2E9, Canada

² Portland Alcohol Research Center, Department of Veterans Affairs Medical Center and Department of Behavioral Neuroscience, Oregon Health & Science University, Portland, Oregon 97201

³ Department of Psychology, State University of New York at Albany, Albany, New York 12222

⁴ Center for Neurobiology and Behavior, 722 West 168th Street, New York, New York 10032

ABSTRACT: It is sometimes supposed that standardizing tests of mouse behavior will ensure similar results in different laboratories. We evaluated this supposition by conducting behavioral tests with identical apparatus and test protocols in independent laboratories. Eight genetic groups of mice, including equal numbers of males and females, were either bred locally or shipped from the supplier and then tested on six behaviors simultaneously in three laboratories (Albany, NY; Edmonton, AB; Portland, OR). The behaviors included locomotor activity in a small box, the elevated plus maze, accelerating rotarod, visible platform water escape, cocaine activation of locomotor activity, and ethanol preference in a two-bottle test. A preliminary report of this study presented a conventional analysis of conventional measures that revealed strong effects of both genotype and laboratory as well as noteworthy interactions between genotype and laboratory. We now report a more detailed analysis of additional measures and view the data for each test in different ways. Whether mice were shipped from a supplier or bred locally had negligible effects for almost every measure in the six tests, and sex differences were also absent or very small for most behaviors, whereas genetic effects were almost always large. For locomotor activity, cocaine activation, and elevated plus maze, the analysis demonstrated the strong dependence of genetic differences in behavior on the laboratory giving the tests. For ethanol preference and

water escape learning, on the other hand, the three labs obtained essentially the same results for key indicators of behavior. Thus, it is clear that the strong dependence of results on the specific laboratory is itself dependent on the task in question. Our results suggest that there may be advantages of test standardization, but laboratory environments probably can never be made sufficiently similar to guarantee identical results on a wide range of tests in a wide range of labs. Interpretations of our results by colleagues in neuroscience as well as the mass media are reviewed. Pessimistic views, prevalent in the media but relatively uncommon among neuroscientists, of mouse behavioral tests as being highly unreliable are contradicted by our data. Despite the presence of noteworthy interactions between genotype and lab environment, most of the larger differences between inbred strains were replicated across the three labs. Strain differences of moderate effects size, on the other hand, often differed markedly among labs, especially those involving three 129-derived strains. Implications for behavioral screening of targeted and induced mutations in mice are discussed. © 2003 Wiley Periodicals, Inc. *J Neurobiol* 54: 283–311, 2003

Keywords: inbred strain; mouse; knockout; serotonin 1B receptor; elevated plus maze; anxiety; locomotor activity; cocaine; accelerating rotarod; water escape learning; ethanol preference; gene–environment interaction; test reliability

Correspondence to: D. Wahlsten (wahlsten@ualberta.ca).

Contract grant sponsors: the Natural Sciences and Engineering Research Council of Canada, the Office of Behavioral and Social Science Research at NIH, the National Institute on Alcoholism and Alcohol Abuse and the National Institute of Drug Abuse at NIH, and the Department of Veterans Affairs.

© 2003 Wiley Periodicals, Inc.

Published online in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/neu.10173

INTRODUCTION

Many behaviors of great importance to society and public health are influenced substantially by genes, and, furthermore, are genetically complex (Wahlsten, 1999). These behaviors include abusive patterns of self-administration of alcohol and other drugs (Crabbe, 2002) as well as other psychiatric disorders such as depression, impulsivity, and schizophrenia (Moldin and Gottesman, 1997; Phillips et al., 2002). Behavior is a property of a whole organism and functions at the level of the individual. Genes code for proteins and function at the molecular level. A gene does not code for a single behavioral phenotype in the whole organism, or even for a component of a behavioral phenotype or specific psychologic process. Rather, most genes have multiple phenotypic effects (pleiotropy) and influence behavior via diverse physiologic and developmental pathways.

The unraveling of specific gene effects to understand influences on the integrated whole of an organism's behavior has been termed behavioral genomics (Plomin and Crabbe, 2000). Because the molecular activities of many genes are regulated by features of the external environment or by behavior itself (Gottlieb, 1998), it is expected that behavioral effects of a mutation or genetic polymorphism will depend to some extent on the animal's environment (gene \times environment interaction). The specific features of the environment that are most effective in modulating gene activity will, of course, depend on the specific gene, and in most cases are presently unknown (Sokolowski and Wahlsten, 2001). Gene products are part of an integrated metabolic system. Therefore, the effects of polymorphism are also expected to depend on genotype at other loci (epistasis) and on the genetic background of the host strain (Gerlai, 2001).

Recent Interest in Mouse Phenotyping

The rapid growth of mammalian genetics has focused on the mouse as a model organism having well-documented genetic homologies with humans (Moldin et al., 2001). Numerous targeted mutations in mice have been created to serve as models for hereditary human diseases (Bolivar et al., 2000; Anagnostopoulos et al., 2001), and random mutagenesis is also now being used to generate many new mutants. Techniques have been devised to map genes with relatively modest effects on complex behaviors (quantitative trait loci or QTLs), offering the promise that the specific genes underlying complex traits can be identified and serve as targets for innovative therapies.

The replicability of these mapping findings, as well as the analysis of the more subtle effects of transgenic manipulations such as gene knockouts, requires the use of closely comparable behavioral tests in different laboratories. However, widely accepted, standard procedures for testing mouse behaviors are not currently available (see Würbel, 2000, 2002; Wahlsten, 2001; van der Staay and Steckler, 2002). Instead, behavioral assessment protocols tend to be unique to each laboratory testing mice.

The need for improved behavioral testing protocols in the neuroscience community became the main topic discussed at a meeting convened by the MacArthur Foundation in 1995, "Animal models of psychiatric diseases: To man from mouse." Participants recognized that a rich diversity of behaviors expressed by mice make it an ideal model mammal for genetic and pharmacologic research on behavior. It was also agreed that there was an important need for stable behavioral norms to aid comparison with genetically and pharmacologically altered animals. As studies with targeted mutant mice proliferated, it became clear that the behavioral "footprint" of the knockout often differed dramatically, depending upon which genetic background (strain) the mutant was placed (see, e.g., Bowers et al., 2000). Thus, normative data on a wide range of standard mouse strains would help to choose the most appropriate genetic background strains on which to place the next generation of targeted mutants (e.g., tissue-specific and conditional knockouts).

Several scientific meetings in 1996 (e.g., an NIMH Workshop on Behavioral Phenotypes of Inbred Strains, a Society for Neuroscience short course "What's Wrong with My Mouse?"; see Takahashi, 1996) subsequently highlighted the growing interest in testing mouse behavior as well as the lack of a common and convenient methodology. A review of some available tests (Crawley et al., 1997) documented the need for careful standardization and the paucity of systematic information available for more than one or two standard inbred mouse strains. Two subsequent meetings were convened by the Office of Behavioral and Social Sciences Research of the NIH. The first of these resulted in partial funding for the experiment discussed in this article. A second meeting of a much larger group discussed tasks in the general behavioral domains surrounding activity and anxiety. In 1999, The Jackson Laboratory convened a "Strain Characteristics Database Summit" to discuss mouse phenotyping. This group determined that a large-scale, multiyear effort to collect standard phenotypic information about certain inbred mouse strains was an area that should be pursued. After identifying several

sources of funding, this has been instantiated as the Mouse Phenome Project (Paigen and Eppig, 2000; <http://www.jax.org/phenome>).

Following these initial efforts, the NIH issued numerous Requests for Applications (RFAs) to stimulate phenotyping of genetically defined mice, including behavioral phenotyping. Several institutes, led by NIMH, convened a meeting in 2000, which led to a report urging funding for multiple studies (http://www.nih.gov/science/models/mouse/genomics/priority_setting_genomics.pdf). For the most part, the motivation for this increased attention has been the recent NIH undertaking of large-scale mutagenesis projects (Moldin et al., 2001), all of which include behavioral screens. It is unfortunate that the systematic phenotyping currently underway did not precede the mutagenesis screens, but at least as the screens mature they can take advantage of the phenotypic information as it develops.

Stability of Genetic Influences on Mouse Behavior

With so many projects proceeding in parallel and dividing the labor by examining different arrays of phenotypes, there appears to be a presumption in the field that the site where a study is done is not a very important factor, provided that the work is done with sufficient expertise. At the same time, failures to replicate effects of certain genetic knockouts on behavior in different laboratories has led to growing concern about the replicability of genetic influences on behavior across labs (Wahlsten, 2001). That such concerns are well grounded is shown by many studies of inbred strains tested after rearing in different conditions within a laboratory (Henderson, 1970, 1976; Erlenmeyer-Kimling, 1972; Wahlsten and Gottlieb, 1997).

We decided to address the question of reproducibility of results in different labs systematically by studying standard mouse strains that show characteristic patterns of behavior (Crawley et al., 1997; Crawley, 2000). To test the resilience of such genotypic influences, we tested mice of eight genotypes simultaneously in three laboratories on a battery of six simple behaviors (Crabbe et al., 1999). We standardized apparatus, test protocols, and other environmental variables to the best of our ability to minimize nongenetic sources of variability. We also asked whether behavioral genetic differences would be affected by shipping the animals from a supplier versus breeding them in house. The results were striking in several ways, and the initial report of the results (Crabbe et al., 1999) has been fairly widely cited and

discussed. The purpose of this article is to present additional primary data from the study, discuss in more depth than previously possible the interpretational nuances of those data, present some of our thoughts about subsequent reactions to the article, and consider the implications of available data for standardization of laboratory testing.

The 1998 Study Revisited

The primary goal of the original study was to determine whether different labs would obtain essentially the same results when testing the same strains of mice on the same behavioral tests. The alternative possibility was that strain differences might be lab specific. In terms of statistical analysis (Wahlsten, 1990; Sokolowski and Wahlsten, 2001), the question we asked was best answered by the strain \times lab interaction term in an analysis of variance (ANOVA). If average test scores were generally higher in one lab than the others but the strain differences were nevertheless similar, this would amount to a lab environment main effect in the ANOVA but it would not undermine conclusions about genetic effects. Lab differences in average test scores are interesting environmental effects that warrant serious investigation, but they were not the focus of our experiment. Indeed, we deliberately sought to minimize environmental differences between our three labs.

We recognized from the outset that equating all relevant variables across labs was futile, and we never hoped to achieve such perfection. We were incapable of making even our three labs do things in exactly the same way, and we were aware that the rest of the mouse testing world would never agree to a single standard (Wahlsten, 2001). Instead, we sought to (a) equate the physical test apparatus and test protocols, and (b) minimize differences in the lab environments where this could be done with little inconvenience in all three labs. Noteworthy differences remained among the lab environments in many respects, and we wanted to know whether these variables would have a substantial impact on test results.

Every test of behavior involves a physical apparatus with which the mouse interacts, and stimulus conditions that impinge on the apparatus as well as a specific protocol of things that are to be done for every animal that is tested on separate occasions. Collectively, these things may be referred to as the *test situation*. The *phenotype* of an animal is then the measured aspects of its behavior when observed in the specific test situation. If two test situations are substantially different, such as a small, square box in the dark and a large, round open field under bright lights,

activity in the two situations may be thought of as two different phenotypes. No behavioral phenotype exists separately from a test situation, because behavior is a reaction to something. It is this reaction that we seek to measure.

The *laboratory environment* is then defined by exclusion as everything that impinges on the animal outside and prior to the test situation. Thus, different lighting conditions during a trial on the elevated plus maze would constitute different test situations, whereas lighting in the colony room involves the lab environment. Whenever one thing is defined as the complement of the other, there are always things at the border of the two that cannot readily be assigned to a category. In our previous report (Crabbe et al., 1999), this intersection of categories was not addressed. Instead, we regarded everything that was not equated as being part of the lab environment. When matters are examined more closely, however, a zone of ambiguity is seen. For this reason, we present our methods in greater detail.

METHODS

Mice

The same eight genetic groups of mice were assessed in all three labs. Identity in most instances was guaranteed by procuring highly inbred strains from the same source and shipping to the three labs on the same day. Some mice were used to constitute breeding pairs, whereas others were housed for later testing. The inbred strains A/J (A), BALB/cByJ (cBy), C57BL/6J (B6), and DBA/2J (D2) along with the F₁ hybrid B6D2F1/J were obtained from the Jackson Laboratory, Bar Harbor, ME, and the inbred strain 129/SvEvTac (129) was obtained from Taconic Farms, Germantown, NY. B6D2F1/J mice were cross bred in each lab to obtain B6D2F2/J (F2) offspring for testing. To see whether a null mutant would respond similarly in different laboratories, we included a strain in which the serotonin 1B receptor gene had been deleted by homologous recombination. The 5-HT_{1B} +/+ and -/- strains were obtained by each lab from the colony of R. Hen at Columbia University in New York. These two strains comprised a mixture of three different 129 substrains, 129/SvPas, 129/SvEvTac, and 129/Sv-ter (Phillips et al., 1999) and were still segregating for genetic differences among the three 129 substrains (Simpson et al., 1997), but care was taken to ensure that the three labs received animals with similar backgrounds. Thus, we are confident that the genetic variable was effectively equated among the three labs.

The eight genotypes were chosen to span a wide range of genetic variation and include strains that were already known to differ substantially on several of the behavioral tests, such as locomotor activity (Thompson, 1953; South-

wick and Clark, 1968) and ethanol preference (McClearn and Rodgers, 1959; Fuller, 1964). It should be emphasized that we designed the study to make the strain main effect in our data analyses large, reasoning that an interaction with the genetic variable would be easier to detect when strains differ substantially within a lab. We also sought to minimize environmental variation.

Shipping

Some labs have facilities to maintain their own breeding colony, whereas many others purchase the mice and test them not long after arriving at the lab. This environmental difference could exert a major influence on tests of anxiety, for example, but no systematic study of the topic had been published. Thus, we compared mice that were bred in our own colonies with those shipped directly from the supplier. Timing of shipping and mating was rigorously controlled to insure all mice were close to the same age on the day when behavioral testing began. The breeding stock was shipped from all three suppliers to all three labs on December 2 or 3, 1997, when mice were about 6 weeks of age, and these animals were then mated January 13, 1998, in all three labs. A second batch of mice from the suppliers was shipped March 15 to 17, 1998, at about 6 weeks of age, so that they would be in the same age range as those bred in our own colonies. Logistics of equating apparatus and protocols dictated that testing began 5 weeks after arrival of shipped mice. Because the period for acclimatizing to the lab environments was longer than occurs in many studies, our experiment is not a definitive examination of this issue.

The actual manner of shipping was not identical for all three sites. For mice sent to Portland and Edmonton from the eastern United States, they of course went by air freight, but those for Albany traveled by truck. Shipping to Edmonton always required a trip of 2 to 3 full days, versus 1 day for Albany. At the last minute when the second shipment in March was to occur, we learned that one supplier could not deliver mice of one group (B6D2F2/J). Fortunately, surplus animals had been bred in Edmonton and Portland. Edmonton then shipped B6D2F2/J mice to Albany. Edmonton also shipped mice to itself, flying them to the Toronto airport, where Dr. Barbara Bulman-Fleming of the University of Waterloo received them and promptly sent them back to Edmonton on a separate waybill. Portland shipped mice it had bred on a flight to Los Angeles, with instructions on the waybill that they be returned to Portland, but for some reason they ended up sitting for 2 days at the Toledo, OH, airport before being returned to Portland. Thus, the shipping treatment was not identical for all mice. We are confident, however, that the experiences of all mice in the shipped condition were very different and more stressful than the lives of genetically identical mice bred in our colonies.

Breeding and Housing

Mice bred locally were treated in a very similar manner in each lab. One female was housed with one male, and cages

were inspected daily for birth of a litter. Litters were not culled or handled prior to weaning, but cage bedding was replaced weekly without removing the nest. Weaning occurred within one day of 21 days of age, and mice were then housed with same-sex littermates. We generally did not retain or test any animal that was housed alone. Breeding mice were fed the high fat Purina 5020 chow, whereas weaned animals were maintained with free access to local tap water and Purina 5001 chow. In each case the supplier of the chow was the local Purina dealer, and there could have been local differences in the exact nutritional formulation (Tordoff et al., 1999). Mice were all housed in plastic shoebox cages with 1/4" Bed-o-cob bedding obtained locally. Cage tops had stainless steel bars, but in Portland filter tops were also used to comply with colony regulations. All mice were maintained on the same light–dark schedule with colony lights on at 0600 and off at 1800. Lighting levels in the colony rooms were not identical.

Mice shipped to the labs were housed four of the same sex per cage until shortly before testing began. They were kept in the same colony room as the weaned mice bred locally, and all aspects of husbandry were the same, including caging, bedding, food, and water.

Testing

Mice were tested in two replications 1 week apart, so that the animals in the second replication were about 1 week older than the first replication and, for shipped mice, had been in the lab an additional week. In Edmonton it was necessary to run an additional squad of only eight mice to fill certain groups where breeding had been slow. Behaviors of these animals appeared to be similar to those in the second replication, and they were pooled with the second replication for the purpose of analysis.

One week before testing began, mice were housed two per cage. The two mice in a cage were littermates if bred locally but may not have been littermates in the shipped condition. The order of the 32 cages on the rack for the first week of testing was determined by random numbers, and a different random order was used for the second replication in each lab. Mice were always tested in the same order on different days. The order of placement on the rack determined the order of testing during a day, and the two cage mates were always tested at the same time. Strain identification information was removed at the start of the testing day, so that only the cage number was visible to the experimenter. Thus, mice of different strains, sexes, and shipping condition were well mixed during the test day, and testing was blind with respect to shipping condition. Testing was to a lesser extent blind with respect to strain. D2 mice have a unique coat color (dilute brown), while there were two albino strains (A, cBy), three black agouti (129 and the two 5-HT_{1B} groups), and one group with mixed coat colors (F2). B6 mice are black, but so were some of the F2 mice. Sex was apparent.

Testing was done over a period of 11 days for one replication. On Monday morning from 0830 to 0900 local

time, all mice for that replication were tail marked with a black Sharpie pen (red Sharpie pen for black mice) and then weighed to the nearest 0.1 g. The rack with all 32 cages was kept close to the testing room, and cages were brought to the vicinity of the apparatus shortly before the test for each pair of mice. The order of tests for the first week was (a) locomotor activity on Monday, (b) elevated plus maze on Tuesday, (c) accelerating rotarod on Wednesday, (d) visible platform water escape on Thursday, and (e) locomotor activity after a cocaine injection on Friday. After a break from testing on Saturday, mice were given ethanol preference tests the next week. After the completion of testing, mice were euthanized and their brains were removed for histologic analysis. Data on brain anatomy in relation to behavior have been published separately (Wahlsten et al., 2001). Analysis of those data demonstrated that absence of the corpus callosum in the three 129-derived strains was unrelated to individual differences in measures of behavior reported here.

Apparatus

Locomotor Activity. The test chamber was a clear plastic box 40 × 40 × 30-cm high with a plastic floor and removable lid. Animal movement was monitored by the AccuScan (formerly Omnitech) Digiscan system that used a grid of photocell beams 2 cm above floor level; another grid 6 cm above floor level detected rearing (vertical movements). Each system consisted of four activity monitors in separate cubicles that allowed four mice to be tested at one time. The Albany and Portland labs already had this apparatus, and new apparatus was generously provided on loan to Edmonton from Dr. R.H. Kant of AccuScan. Data collected by computer included distance traveled in each time period, number of horizontal movements, number of vertical movements, and time spent moving. The system also recorded the time spent in the center 25 × 25-cm zone. Each activity monitor in Albany and Portland was enclosed in a small cubicle with an exhaust fan, but the cubicles were somewhat different and external light sources were different. Edmonton built new enclosures that were about the same dimensions as the other two labs and included an exhaust fan. We opted to test mice in total darkness because that was the one thing we could definitely make the same in all three labs. Thus, in the locomotor activity test, the recording system, the physical apparatus in contact with the mouse and lighting were identical, but the surrounding cubicles themselves were similar but not strictly the same. The same apparatus was used for cocaine activation testing.

Elevated Plus Maze. Apparatus were constructed in the Department of Psychology shop in Edmonton, and two copies were provided for each lab. The black plastic floor of the maze consisted of four arms 5-cm wide and 30-cm long that met at a 5 × 5-cm center zone, and the apparatus was mounted on a clear plastic pedestal that placed the arms 50 cm above the floor. Bedding was placed below the maze should the mice fall. The two enclosed arms had clear

plastic walls 15 cm high, whereas the open arms had a low rim 5 mm high. The end of each arm and the wall itself was rounded, and the walls on both enclosed and open arms could be removed for cleaning. Lights were procured in Edmonton and shipped to the other two labs. The light consisted of a 15-watt frosted tungsten bulb mounted in an aluminum reflector and suspended 1.0 m above the arms of the maze to create a light intensity of about 100 Lux. The two mazes in each lab were separated by a white partition so that mice could not see each other during testing but their behaviors could be viewed by the same video camera. We tried to make the general arrangement of the two mazes in the test room similar, especially the distances of mazes from walls and objects, but the stimulus surroundings were not identical. Behavior during plus maze testing was recorded on video tape in each lab, although cameras were different.

Accelerating Rotarod. Each lab received two new copies of the AccuRod apparatus on loan from AccuScan. Two rods were run by one computer, but they were independent and could be started and stopped at different times. The rod itself was 11 cm long and 2.5 cm diameter, and it was suspended 30 cm above a trough filled with bedding. The fall of a mouse from the rod was detected by photocells about 1 cm above the bedding. If a mouse happened to miss the photocells, the trial could be ended by a pushbutton. In an attempt to make the all-important rod surface identical, we glued a strip of 320 grit emery paper to the surface of each rod with rubber cement. In an extreme effort to ensure identical surfaces, sandpaper from a single source in Portland was sent to the other labs, much to the amusement of office staff in Edmonton who had never seen four sheets of sandpaper delivered by courier. In the end, this effort was for naught, and there was reason to believe the rod surfaces were not identical in the three labs. The seam where edges of the paper met turned out to be crucial because mice could gain a toe hold on a seam that was not perfectly formed. Lighting was provided by the room fluorescent lights.

Water Escape Tank. Six identical, seamless polyethylene tanks were molded in Edmonton and two were shipped to each of the other labs. The circular tank was 70-cm diameter and 30-cm deep, and it was filled with clear, fresh tap water 25 cm deep at the start of the test day. Water was maintained at 25 to 26°C. The rim of the tank was covered with a 1.5-cm thick black rubber tube to make it clearly visible. A visible platform was located at the center of each tank and consisted of a 10-cm diameter metal mesh painted black and protruding 5 mm above the surface, with a 6.4-cm diameter black ball located 7 cm above the platform. Lighting was provided by the same light as used for the plus maze and was about 100 Lux at the water surface. As for the plus mazes, the two water tanks were separated by a white partition and behavior was recorded by a video tape recorder.

Ethanol Preference. During preference testing, mice were housed singly and allowed free access to food, but the single

large water bottle was replaced by two 25-mL graduated cylinders placed symmetrically in the cage top 4.5 cm from the two sides of the lid. Each cylinder had a size 14.5 rubber stopper with a stainless steel drinking spout that extended about 6 cm into the cage at an angle of about 30 degrees below horizontal. Cylinders, stoppers, and spouts were cleaned with a dilute bleach solution (0.03% sodium hypochlorite) prior to the start of testing. Ethanol solution was 6% v/v of absolute ethanol in distilled water. Food was freely available and dispersed around and between the cylinders in the cage lid. Control cages (without mice) were placed on the racks to correct for spillage and evaporation.

Procedures

Locomotor Activity. The trial was 15 min long, and data were recorded separately for three 5-min samples. Four mice were tested at the same time, consisting of two cages of two mice each. The home cage was brought close to the apparatus and the lid was removed. The first mouse was grasped by the tail and placed into the center of the box facing away from the experimenter. The plastic lid was put on the box, the cubicle was closed, and a button was pressed to start the trial for that mouse. The same procedure was followed for the remaining three mice. Mice were removed and returned to their home cages, fecal boli were counted and removed, and excess urine was removed with a clean paper cloth. Finally, the inside surfaces of the chamber were wiped clean with 70% isopropyl alcohol and allowed to dry before the start of the next batch of mice.

Elevated Plus Maze. The trial of 5-min duration was video taped for later scoring of behaviors. Two mice from one cage were run at the same time. The first animal was picked up from the cage by its tail and placed gently onto the center of the maze facing the opposite open arm. Care was taken to perform this operation quickly but smoothly to avoid any abrupt tug on the tail that would induce the mouse to dart into the open arm. As soon as the tail was released, a stop watch was started, and then the procedure was repeated for the other mouse. The experimenter moved out of sight behind a partition and observed the mice on a TV monitor. At the end of the trial, neither mouse was disturbed until the 5 min had elapsed for both mice. The animals were returned to their home cage, walls of the maze were removed, boli were counted, and then the surfaces of the maze and the walls were cleaned with 70% isopropyl alcohol. After the study was completed, data were scored from video tapes at each site. Number of entries into each arm was scored on one pass through the tape, defining an arm entry as an occasion when a mouse placed all four feet within the arm after having all four feet outside the arm. On a second pass through the tape, time spent in each arm was recorded with stopwatches.

Accelerating Rotarod. Ten trials were given with an inter-trial interval of 30 s. Each mouse was placed onto the center of the motionless rod and a button was pressed to start the

rotation after both mice were in position. Acceleration rate was 80 rpm/min. The experimenter stepped back to a position about 1.2 m from the apparatus and watched the mice closely. Time to fall from the rod was usually detected by the photocells, but it was sometimes necessary for the experimenter to record the latency by hand. After both mice had fallen, a stop watch was started for the intertrial interval. Mice remained in the bedding trough of the apparatus during this period. The experimenter made notes on the animals' behavior at this time, making special mention of occasions when a mouse jumped rather than fell from the rod or flattened itself on the rod by holding onto the surface and rotating passively with the rod rather than walking on it.

Water Escape Tank. Pretraining was given at the end of the day on Wednesday and consisted of three trials. The first was a 60-s period of swimming with no platform present. Next the mouse was placed onto the visible platform and allowed to remain there 15 s. Finally, the mouse was placed into the water facing the platform and allowed to climb onto it and remain there 15 s. The next day, eight training trials were given with an intertrial interval of 30 s and a trial duration limit of 40 s. Two mice from one cage were run in close succession in adjacent tanks. Each animal was kept in a separate holding cage with three paper towels on the floor during training. The first animal was picked up by the tail and placed facing the wall of the tank at one of four randomly chosen compass positions (N, S, E, W). Each block of four trials included one instance of each position. As soon as the tail was released, the experimenter started a stop watch and stepped back at least 1 m from the tank. When the mouse had climbed onto the platform with all four feet, the latency was recorded and the mouse was allowed to remain on the platform for 10 s before being returned to the holding cage. As soon as one animal had been returned to its holding cage, a timer was started for the intertrial interval and the trial for the other mouse was started. In an instance when one mouse required more than 30 s to escape, the intertrial interval for its cage mate was extended beyond 30 s.

Cocaine Activation. Mice were weighed at the start of the session on Friday. Cocaine hydrochloride in a dose of 20 mg/kg was administered by intraperitoneal injection shortly before the 15-min trial in the activity box. The four mice in one squad were injected quickly in succession and then the trial was conducted in the same way as the locomotor activity test on Monday. The cocaine was obtained from a different source in each lab (Albany, National Institute of Drug Abuse Clearing House; Edmonton, BDH lot #113428/2411 via Health Canada; Portland, Sigma lot #34H0200).

Ethanol Preference. On Sunday morning, mice were housed individually in a clean cage and given two graduated cylinders containing fresh tap water placed symmetrically; water levels were read from the tubes. On Monday at the same time, water levels were again read without disturbing the apparatus. On Tuesday, water levels were read and then

two clean cylinders were put in place, one containing fresh tap water and the other containing 6% ethanol. For Tuesday and Wednesday, the ethanol solution was on the left side of the cage for all mice. The tubes were not disturbed on Wednesday when readings were taken. On Thursday, levels were read and two clean bottles were put in place—one on the left with fresh tap water, and one on the right with 6% ethanol. Readings were taken at the same time on Friday and Saturday.

Data Analysis

Data from all tests at the three sites were collated and entered into a spreadsheet at Albany and then imported into SPSS and saved as a SAV type of file that could be analyzed with either SPSS or SYSTAT. Data were scrutinized for errors and outliers using a variety of methods. The principal technique for the final analysis was factorial analysis of variance (ANOVA). Although the mice bred locally were tested in littermate pairs, litter membership of most of the shipped mice was not known. Consequently, the individual mouse, not the litter, was adopted as the unit of analysis. The original design called for four mice of each strain–sex–shipping condition to be tested at each lab for a total of 128 mice per lab. This sample size was sufficient to detect a moderate strain \times lab interaction with power of 90%, and power to detect moderate main effects was much higher (Cohen, 1988; Wahlsten, 1990). Because so many tests of significance were done, we decided to adopt $\alpha = .01$ as the criterion for significance, and we gave serious attention only to effects that met a more stringent criterion of $\alpha = .001$. Our simple procedure yielded conclusions that were very similar to those based on the more sophisticated significance criterion adjustment recommended by Benjamini et al. (2001). Because the F -ratio in the ANOVA is so closely related to the p -value for the test of significance, we chose to report p -values and an indicator of effect size. For effects with greater than one degree of freedom in the numerator, a convenient indicator is partial ω^2 , an estimate of the proportion of variance attributable to the between-group effect when only that one effect is compared with variance within a group.

RESULTS

The design entailed $8 \times 3 \times 2 \times 2 = 96$ independent groups with four per cell for a total of 384 mice, whereas we obtained valid data for 378 or 379 mice on most measures. Every cell had at least one mouse at each site, and the shortage of mice, due to poor breeding, was confined to the 5-HT_{1B} +/+ animals in Edmonton. Thus, the problem of unequal sample sizes in the ANOVA was negligible.

Significance and Effect Size

The overall pattern of results from this study is shown in Figure 1 where color expresses the level of statistical significance and numerical values express effect size. As expected from our choice of a wide range of strains, the genotype factor was in every instance significant and for most measures a very large effect. A sex difference was seen primarily for drinking in the ethanol preference test. Shipping effects were generally not apparent, with one small exception. Site effects, on the other hand, were significant and large for measures of locomotor activity and anxiety as well as response to cocaine, and several significant and very substantial interactions between genotype and site were observed for those measures. On the locomotor activity, elevated plus maze and cocaine activation tests, site differences were as large or even larger than genotypic effects for several measures, whereas genotypic differences were much larger than site variation on the visible platform water escape and ethanol preference tests.

Although the global effects in an ANOVA reveal the general pattern of results, detailed examination of data for each kind of test is needed to learn which genotypes were most sensitive to differences in the laboratory environment. For several tests, there was more than one way to express the data or there were animals for whom certain measures were effectively meaningless, and we also wanted to know how results would change when these more subtle things were taken into account. Furthermore, we were concerned about the reliability of data and wanted to know whether the lack of significant site or interaction effects for certain variables was attributable to a high noise level or low test reliability. It was possible to assess reliability in terms of interanimal consistency across time periods or days for each test except elevated plus maze where there was only one brief trial.

Basal Locomotor Activity and Response to Cocaine

Reliability of measures of locomotor activity was evaluated by consistency of performance across 5-min blocks within a trial when all animals in all conditions were pooled for analysis. Correlations of scores in the second and third blocks were 0.79 or greater for all five measures listed in Table 1. Correlations of scores in the first and second blocks were less than 0.5 for center time and speed of movement but 0.77 or greater for the other measures. No test-retest reliability could be calculated across trials.

Activity testing on Monday and the cocaine acti-

vation test on Friday were done with the same mice tested in the same order in the same Accuscan apparatus for 15 min, and data were collected in blocks of 5 min each day. Each of the various measures of behavior was therefore combined into one large analysis of variance with four between-subject factors (genotype, sex, site, shipping) and two within-subject factors (naive vs. cocaine, time block within a session). In almost every analysis, the sex and shipping factors as well as interactions with sex were not significant at $p < .01$, and these two factors are not discussed further. Results of the ANOVAs are summarized in Table 1, and patterns of strain, lab, and cocaine effects are shown in Figure 2.

The dynamics of the five measures of activity-related behavior in the activity box averaged over all mice are expressed in Figure 3 in a way that emphasizes relative change across time and between naive and cocaine conditions. Distance traveled, percent of the time spent moving, and speed of movement declined within a trial in naive mice and increased greatly after cocaine administration. Rearing increased over time for naive mice and was reduced after cocaine, whereas time spent near the walls of the chamber increased over time in naive mice and was increased further on the cocaine day. Generally speaking, cocaine greatly increased amount and speed of movement while reducing rearing. The increase in time near walls after cocaine was similar to what might be expected on a second day of testing in the activity box, and was not necessarily a result of cocaine, whereas the other four measures changed substantially in a direction opposite what would be expected on a second test and therefore were clearly influenced by cocaine.

The correlation across all 379 mice between distance traveled and percent time moving was very high ($r = .90$) in naive mice. Distance was also highly correlated with speed ($r = .68$) and rearing ($r = .70$) but not with time in the center ($r = .26$). Both distance and rearing were positively related to activity and tended to be high in the most active mice. Nevertheless, cocaine effects on distance traveled and rearing were opposite, which reveals that the two measures are not simply reflections of the same trait.

For each of the five measures, the genotype \times site interaction was significant and was also substantial except for speed. The magnitudes of cocaine effects were genotype-dependent for every measure, and they were site-dependent for all but center time. The cocaine effects even showed higher order dependence on the joint genotype \times site condition for all measures except margin time. This more complete analysis reveals that genotype \times lab site interactions were not

Results of Univariate ANOVAs, showing partial ω^2 values

Measure	Geno (df = 7)	Sex (df=1)	Site (df=2)	Ship (df=1)	Geno x Site (df=14)	Mult. R ²
Open field (Monday) - one 15 min trial						
Distance traveled (cm)	.606		.158		.062	.703
Number of vertical movements	.792	.039	.282			.829
Time moving	.384		.147		.104	.590
Time in center	.076		.211			.394
Speed when moving	.741		.058			.777
Elevated plus maze (Tuesday) - one 5 min trial						
Time spent in 5 x 5 cm center	.303		.180		.133	.523
Number of arm entries	.390		.328		.217	.660
Percent time in open arms	.051		.265			.445
Water escape to visible platform (Thursday) - 8 trials with 40 sec limit						
Average latency on trials 1 to 4	.672		.034			.716
Trials to 1st success (< 5 sec)	.420			.022		.542
Inconsistency score over 8 trials	.317				.053	.490
Open field after 20 mg/Kg cocaine (Friday) - one 15 min trial						
Distance traveled (cm)	.503		.086		.107	.619
Number of vertical movements	.641		.050		.121	.704
Time moving	.186		.117		.191	.510
Time in center	.198		.144		.164	.551
Speed when moving	.748		.054		.067	.783
Ethanol preference (Monday to Thursday, week 2) - 4 days of 2 bottle choice						
Total volume of liquid consumed	.307	.039	.053			.496
Total volume per g body weight	.124	.129				.418
Ethanol consumed (g/Kg/day)	.489	.044				.590
Ethanol preference ratio	.447					.545
Preference ratio on days 2 and 4	.463					.557

Figure 1 Results of univariate analysis of variance (ANOVA) for several measures of behavior on five tests. Each ANOVA was done as a complete factorial design with the four between-groups factors shown in the table. All interactions were assessed, but only the results for the genotype \times site interaction are shown. Multiple R^2 is the proportion of total variance attributable to all main effects and interactions in the ANOVA, whereas the values shown for each specific effect are partial ω^2 , an estimate of effect size that compares the specific effect to variation within groups. For each ANOVA, degrees of freedom within groups are 282 or 283 for a total sample of $N = 378$ or 379 . Significance is shown by the color of each cell: Blue, $p < .00001$; Pink, $p < .0001$; Green, $p < .001$; Gold, $p < .01$; Gray, $p < .1$. Only effects significant at $p < .01$ are taken seriously in this study.

Table 1 Results of Repeated-Measures Analysis of Variance for Measures of Locomotor Activity

Measure	Between-Subject Effects			Within-subject effects			
	Genotype	Site	Genotype × Site	Cocaine	Cocaine × Genotype	Cocaine × Site	Cocaine × Genotype × Site
Distance traveled	0.590	0.119	0.109	0.718	0.343	0.053	0.090
Number of rearings	0.803	0.209	0.109	0.489	0.122	0.039	0.074
Percent time moving	0.302	0.138	0.188	0.656	0.061	0.094	0.152
Percent time in center	0.218	0.284	0.165	0.073	0.082	N.S.	N.S.
Speed when moving	0.343	0.058	0.062	0.629	0.178	0.057	0.058

Values in the table are partial omega-squared, the proportion of variance accounted for by the effect in question when only that effect is compared with variation within groups. Values are similar but not identical to those shown in Figure 1, which involved separate ANOVAs for activity on day 1 and then under the influence of cocaine on day 5. Degrees of freedom for the residual variance were 282 for all effects shown in the table.

N.S. indicates an effect not significant at $p < .01$. The repeated-measures analysis also involved terms for change within a session, but these are not shown in the table. Change across samples within a session was unquestionably significant and the rate of change depended on genotype for every measure.

limited to the two measures reported previously. On the contrary, the interaction effects were pervasive for all measures of activity-related behavior.

Abstract statistics in Table 1 cannot distinguish between interactions where the same quality of effect is observed for every genotype but the effects differ substantially in degree, versus effects that may be fundamentally different for certain genotypes. The group means shown in Figure 2 reveal that several effects were indeed very different in the three labs for the three main variables that describe activity.

For distance traveled, cocaine activation in the 129-derived strains was strong in Edmonton but almost absent in Portland. In Albany, cocaine activation was absent in 129/SvEvTac but present in 5-HT_{1B}+/+ and -/-. Portland also observed minimal activation for the A strain, contrary to effects in the other two labs. This large difference between labs could not be attributed to differential activity of the cocaine, because the three labs found almost identical effects for four of the groups (B6, cBy, D2, F2).

Rearing behavior differed greatly among genotypes when naive, and it was generally lower in Portland. Cocaine effects were strongly genotype dependent, to the extent that rearing was almost totally suppressed in the four strains that had lower levels of distance traveled and rearing when naive. The suppressive effect was seen in all three labs, but because the level of rearing differed between labs in naive mice of these strains, the floor effect also contributed to the cocaine × genotype × site interaction. Rearing was not suppressed by cocaine at all in B6 or F2 mice in Edmonton.

Center time was highest in all eight genotypes in Edmonton and lowest in Portland. Center time was

decreased in large measure by cocaine for many mice in all three labs, but the effect for the 129-derived strains was minimal in Albany and Portland. Center time was actually increased for the 5-HT_{1B}+/+ and -/- groups in Edmonton. This genotype-dependent effect on center time was particularly striking because the low activity A and 129-derived strains showed similar cocaine effects on other measures of cocaine-related behavior change but radically opposite effects on center time in two of the three labs.

Thus, the patterns of group means in the three labs reveal that the substantial interaction effects seen in the ANOVAs represented different directions of cocaine effects or complete absence of cocaine effects for certain measures in specific genotype-lab combinations.

Elevated Plus Maze

The plus maze is widely employed as a test of anxiety in mice and rats (Crawley, 2000; Hogg, 1996; R.J. Rogers and Dalvi, 1997), and factor analysis of multiple measures of behavior on the plus maze as well as several other tests suggest the presence of an anxiety factor. At the same time, it is evident that different assays of mouse anxiety are not tapping the same underlying neurobiologic substrates (see Clément et al., 2002), because strain rank orders using different tests of anxiety or modifications of the same basic test are not always the same, nor are the genomic regions implicated in genetic mapping studies (Toye and Cox, 2001; Turri et al., 2001).

In our simple plus maze, each mouse was started at the center, and there were several animals that remained there for most or all of the 5-min trial. Several

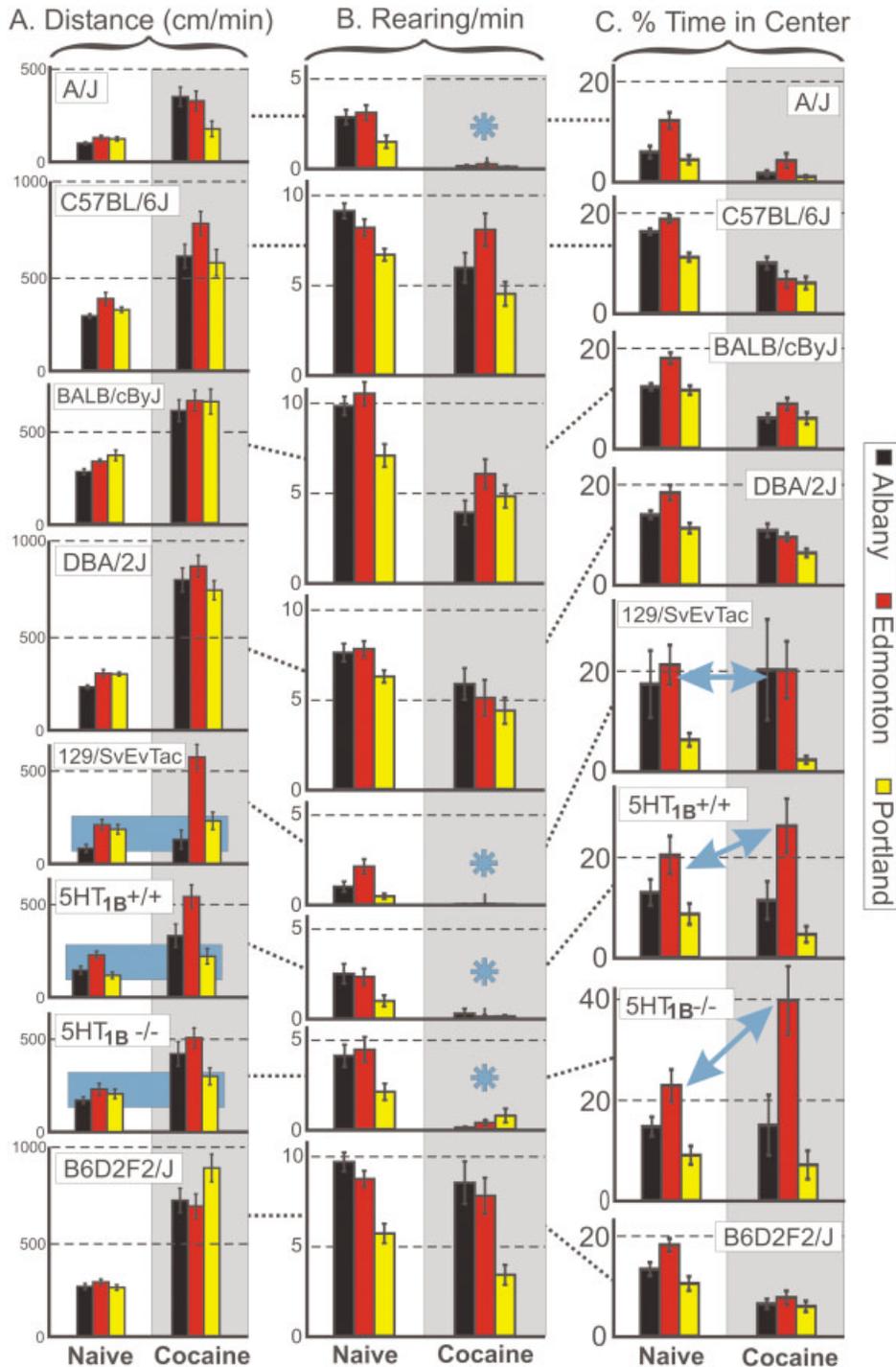


Figure 2 Three measures of locomotor activity in naive mice and the same animals under the influence of 20 mg/kg cocaine, showing group means \pm 1 standard error of the mean. (A) For distance traveled, cocaine generally increased activity, but this effect was either absent or quite small for the three 129-derived strains (highlighted in blue) when tested in Portland and to a lesser extent in Albany. All three sites detected substantial cocaine activation for the mice obtained from the Jackson Labs. Distance traveled was greatest on average in Edmonton. (B) Rearing was considerably lower in Portland. Cocaine reduced rearing in most cases, although not for B6 and F2 mice in Edmonton, and it virtually eliminated rearing for the A strain and the three 129-derived strains at all sites (blue asterisks). (C) Time spent in the center of the box was considerably greater in Edmonton. Cocaine reduced center time for almost all genotypes in Albany and Portland, but it had no effect on 129 mice in Albany and Edmonton while increasing center time for the $+/+$ and $-/-$ mice in Edmonton (blue arrows).

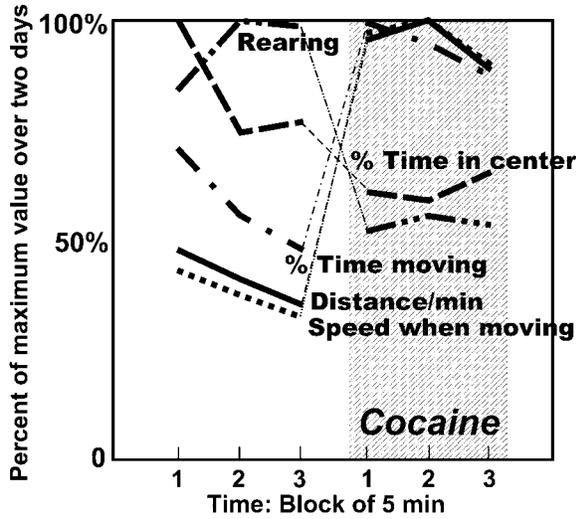


Figure 3 Change across time within a trial in naive mice and under the influence of 20 mg/kg cocaine, averaged over all genotypes and labs. In naive animals, rearing increased during a trial while the other measures declined, but there was little change across time under cocaine. Cocaine greatly increased distance traveled, percent time moving, and speed of movement, while it reduced rearing.

other mice moved from the center into one of the arms and then remained there for the remainder of the trial. None of these problems involved any error in measuring behavior; instead, the behavior itself posed a challenge. There was doubt about whether data from these animals should be included in the assessment of

mouse anxiety, because low motor activity on the plus maze, by itself, is not necessarily an indicator of high anxiety, and it subverts the more widely accepted indicators of anxiety, especially the relative amount of time spent in open versus closed arms.

The magnitude of the challenge may be seen in Figure 4, where values for every mouse in all three labs are plotted. We believe it is very important at this phase of the analysis that the researchers not consider the strain or environmental background of the mice; rather, the goal was to establish criteria for excluding certain data from the final analysis without introducing bias for or against certain genotypes or treatments. In Figure 4(A) it is apparent that animals remaining at the center for most of the trial also made very few arm entries (identified as Exclusion Zone 1). One purpose of using a 5-min trial on the elevated plus maze is to obtain a sufficiently large sample of an animal's behavior in the two kinds of arms. If it remains at the center for more than 240 s, then in effect it has experienced a trial of less than 1-min duration in the arms, and so short a trial does not provide a suitably reliable indicator of anxiety. Thus, we consider it reasonable to exclude mice that spent more than 4 min of the 5-min trial at the center of the maze. Figure 4(C) shows that there was no preponderance of time spent in the open versus closed arms for these animals, and it illustrates how this cluster of data points was far from the main group of mice.

Figure 4(A) also shows that several mice that left the center made very few arm entries, and Figure 4(B)

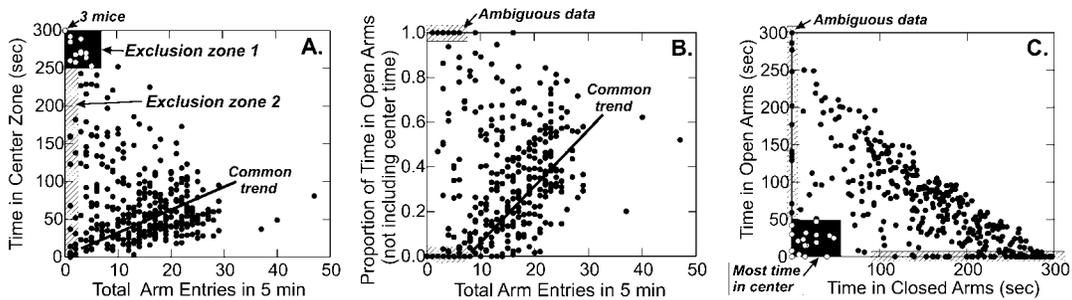


Figure 4 Scatterplots for measures of behavior on the elevated plus maze when all mice are pooled. (A) For most animals, time in the center 5×5 -cm zone increased as mice entered more arms, because they necessarily spent little time in any one arm before returning to the center. Several mice with very low motor activity spent most of their time in the center, however (Exclusion zone 1). Animals with fewer than three arm entries yielded data that were difficult to interpret (Exclusion zone 2). Mice in the two exclusion zones were eliminated from the data set for certain analyses. (B) Animals with very few arm entries tended to spend all of their time in either an open or a closed arm, mainly because of hypoactivity, whereas those with higher numbers of arm entries tended to spend relatively more time in the open arms. (C) Animals that never entered the closed arms spent widely varying amounts of time in the open arms and in some instances remained in the center most of the trial. For animals that did not enter both kinds of arms several times, the measure of preference for open arms is ambiguous.

Table 2 Results of ANOVA for the Elevated Plus Maze with Full and Reduced Data Sets

Variable	<i>N</i>	Genotype (<i>df</i> = 7)	Site (<i>df</i> = 2)	Genotype × Site (<i>df</i> = 14)	Multiple <i>R</i> ²
Time in center	379	0.302	0.180	0.134	0.523
	341	0.368	0.252	0.131	0.576
Total arm entries	379	0.389	0.327	0.217	0.660
	341	0.308	0.300	0.143	0.623
Percent time in open arms	379	0.050 ^a	0.265	N.S.	0.445
	341	0.095	0.310	N.S.	0.527

^a $p < .01$; all other effects significant at $p < .001$. N.S. indicates $p > .01$.

Values for specific effects are partial omega-squared coefficients.

shows that several of these mice then spent all the remaining time in either a closed or an open arm (shaded as ambiguous data). There was a general trend apparent in the main bulk of the data suggesting that mice entering more arms also spent a higher proportion of time actively exploring the open arms. Thus, low numbers of arm entries may be important facts when considering anxiety. Nevertheless, if there are too few entries, data become highly unstable, as shown by several animals with few arm entries that spent all of the time in an open arm. For example, several mice were observed to move quickly into the opposite open arm after being released at the center and then remain there for most of the 5-min trial. Others entered a closed arm and then remained there. Because of the obvious relation between few arm entries and percent time on the open arms, we adopted an exclusion criterion of zero, one or two arm entries [Exclusion Zone 2, Fig. 4(A)]. This criterion left mice in the sample that had very few entries but spent all their time in the open arms. Although mice with very few (but more than 3) entries that spent most of their time in the closed arms were clearly on the same continuum as the bulk of the sample, we considered excluding their data and data from mice with low numbers of entries that favored open arms. However, it seemed unwise to exclude the latter animals because they did something we could not understand—remaining out in the open the entire time. As shown in the zones identified as ambiguous data in Figure 4(B) and (C), there were also several mice that had a substantial number of arm entries but nevertheless experienced only one kind of arm. Because they returned to the center zone so often, we conclude they had ample opportunity to enter both kinds of arms but chose not to do so. Hence, data from all mice except those in Exclusion Zones 1 and 2 remained in the sample.

We prefer to rely on our findings for the cleansed sample of 341 mice because we have greater confidence in the proper interpretation of one of the mea-

asures, percent time on the open arms, when mice with lowest activity levels on the elevated plus maze are excluded. For the number of arm entries, on the other hand, the full data set is meaningful, and there is no reason to exclude any mouse. Applying the two exclusion criteria eliminated 38 of the 379 mice, but it did not eliminate any single group from the sample, and an ANOVA with four factors (strain, sex, lab, shipping) was possible for the cleansed data set. The same analysis was also done for the full sample in order to assess possible consequences of data exclusion.

Table 2 summarizes the results of the ANOVAs for three major variables. There were no effects of sex, shipping, or interactions with sex or shipping significant at $\alpha = .01$, but strain differences in the level of activity or inactivity were very large indeed. Of particular importance were the dramatic differences among the three labs in the overall level of activity on the maze (number of arm entries) and the large interaction between strain and lab, as shown in Figure 5(A). Albany and Portland obtained almost identical results for four of the five groups from the Jackson Labs (B6, cBy, D2, F2), whereas Edmonton was consistently and considerably higher for these four groups. On the other hand, for the 129-derived strains, Edmonton and Portland found very similar levels of activity, whereas values in Albany were notably low. Albany found low activity levels for both A and the 129-derived strains, whereas strain A, well known for hypoactivity in previous research (Xu and Domino, 1994), clearly was least active of the eight groups in both Edmonton and Portland. Not surprisingly, excluding mice with low numbers of arm entries tended to reduce strain and site effects on number of arm entries as well as the strain by site interaction, and for this measure the full sample is clearly preferable.

Percent time in open arms differed remarkably among the three labs, whereas the strain difference on this classical indicator of anxiety was barely significant. As shown in Figure 5(B) and Table 2, the strain

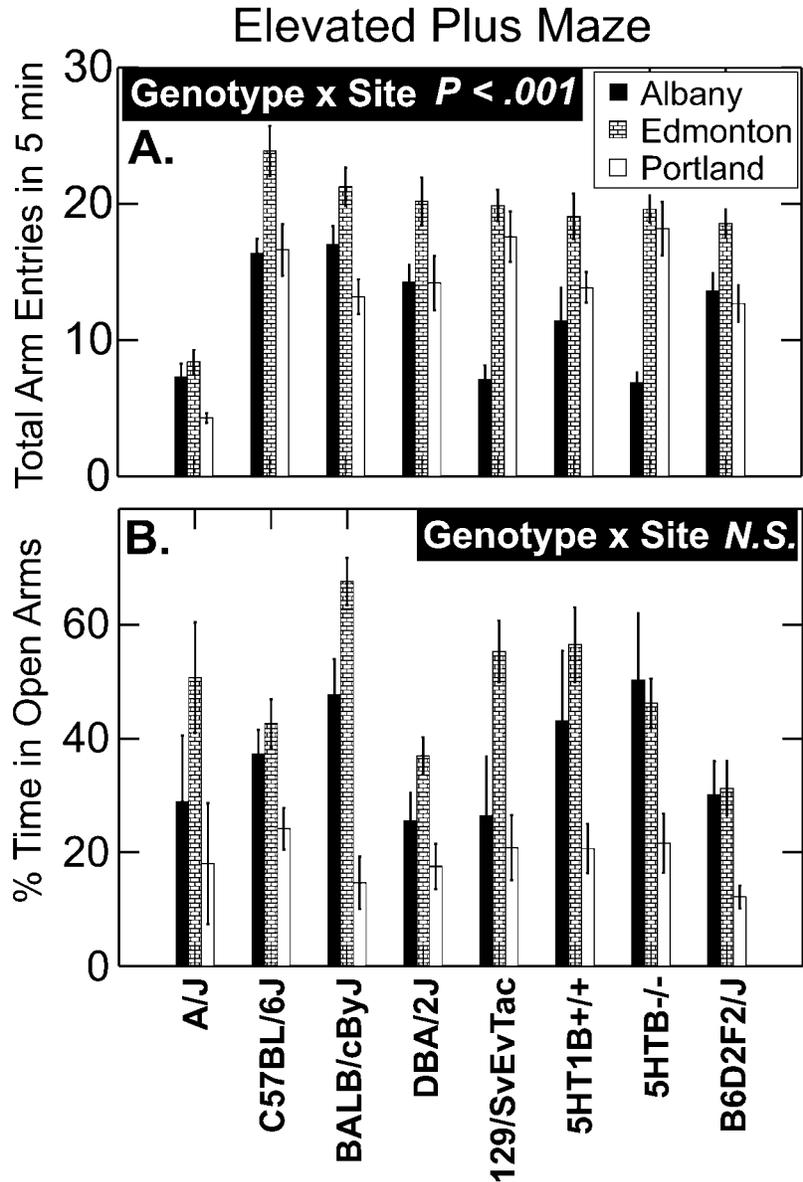


Figure 5 Group means \pm standard error of the mean for eight strains tested in three labs on the elevated plus maze. (A) Total number of arm entries, indicative of the level of locomotor activity, was generally highest in Edmonton, but results were very similar in Edmonton and Portland for the three 129-derived strains, whereas results were very similar in Albany and Portland for the mice obtained from the Jackson Labs. (B) Percent time spent in the open arms was generally high in Edmonton and low in Portland with Albany being intermediate. Strain differences were not nearly as large as for number of arm entries.

difference was not small merely because the test was insensitive or unreliable; after all, the difference between labs was very large. At the same time, the standard error bars for percent time in the open arms were relatively larger than for total arm entries [Fig. 5(B)] and the multiple R^2 was smaller (Table 2). The multiple R^2 and the strain differences were substantially larger for the cleansed sample lacking hypoac-

tive mice. Mice of almost all strains spent considerably more time in the open arms in Edmonton, whereas they evidenced strong preference for the closed arms of identical mazes in Portland, with Albany being intermediate for almost every strain. Although Albany and Edmonton obtained very similar results for three groups (B6, -/-, F2), the ANOVA failed to detect a significant interaction.

The relation between activity and anxiety was clearly quite complex, especially when different labs were introduced into the equations. Caution is warranted when a psychologic relation between variables is based solely on data from a single lab. By involving three labs in the study, the range of variation among and within strains was increased considerably. For certain strains, the extended range seemed to clarify the relation between activity and anxiety (data not shown), but this was certainly not true for all strains. The results for motor activity levels are perhaps easier to grasp and also more in line with the existing literature, whereas the vast range of individual scores on the anxiety indicator and occurrence of many mice, especially cBy in Edmonton, with a clear preference for the open arms, prompts us to wonder how the results would appear with other tasks believed to sense anxiety. Our results cannot be well understood unless a wider range of tasks is examined. By testing mice simultaneously in three labs, we have identified an important issue for mouse behavioral genetics, but the scope of this study was not sufficient to identify the fundamental psychologic processes responsible for our results.

Accelerating Rotarod

In this task, the mice extracted their revenge! For the reasons detailed below, we believe that the rotarod data we collected are essentially uninterpretable, and we therefore did not report them earlier (Crabbe et al., 1999), nor do we present any ANOVAs here. As the deadline for starting the experiment neared, we noticed that the surfaces of the rotarods were not identical. Even though the genotypes and conditions were randomized across rods in each site, we were concerned that this could distort the estimated performance of some groups. One of us (J.C.) made the last-minute and very ill-advised decision to cover all rotarod surfaces with 320 grit wet/dry sandpaper to achieve uniformity. This surface works well on other rotarods he uses in his laboratory, but those rods have substantially larger diameters.

The accelerating rotarod task is designed to place increasing demand on the mouse until it is no longer able to stay on top and falls. The behavioral strategy it seeks to measure is a constant shifting of position as the rod rotates beneath the mouse, akin to a log-rolling contest for humans. However, the combination of relatively small-diameter rotarod with sandpaper offered the mice a second (and superior!) behavioral strategy, which was to “flatten” themselves against the rod and essentially wrap themselves around it. On trials when a mouse adopted this flattened posture and

grip, its latency to fall was dramatically elevated (latencies of 35–70 s vs. latencies of 5–25 s for most mice early in training). There were many long latencies on the very first trial, and behavior remained highly variable throughout training. The correlation for all 378 mice between latencies on two successive trials was close to $r = .4$ across all 10 trials, and this low reliability was primarily caused by inconsistent flattening from trial to trial.

Some mice clearly learned to flatten, based on increasing numbers of flattened trials in the second five versus the first five trials, while others learned to walk skillfully atop the rod, and many others remained highly variable. Strains clearly differed in the extent to which they engaged in the flattening strategy. For example, a large proportion of B6 mice engaged in this strategy across sites, while D2 mice rarely did. Unfortunately, sites differed greatly in the proportion of trials they declared as “flattening” (Fig. 6), and we concluded that one contributor to this dramatic difference was different levels of skill on the part of the human experimenters in applying the sandpaper to the rod. The seam at the intersection of the edges of the sandpaper tended to be wider in Edmonton, and many mice were able to get a claw or two into the seam and cling to the rod, even on the first trial. We tried to cleanse the data of trials with flattening, but this proved futile because some cells in the design became empty. We reluctantly decided that the results of this test could not be interpreted. The raw data from the Portland site are available from J.C. for the interested peruser.

Water Escape

Perhaps the most striking observation in the water escape task was the abysmal failure of the A strain to escape on the first trial or improve over trials. This was a consequence of frequent wall hugging by these mice; they swam constantly and never floated, but they seldom left the wall. Only five individuals of the 48 A mice (three in Edmonton, two in Portland) showed clear evidence of improvement across trials that lead to proficient performance. Many mice that escaped quickly on one trial nevertheless reverted to wall hugging on subsequent trials. Thus, the failure to improve over trials was not caused by a lack of experience with successful escape. Instead, performance of the A strain both early and late in training showed strong interference from wall hugging in the water tank. No mouse in any other strain showed wall hugging on more than two of the eight trials. For this reason, data for A were not included in our original statistical analyses (Crabbe et al., 1999).

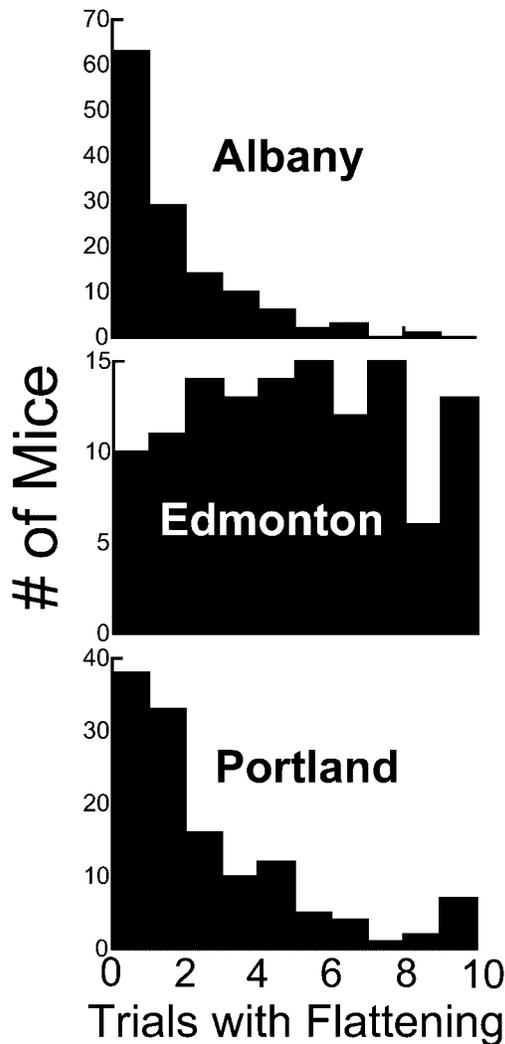


Figure 6 Number of trials out of 10 when a mouse showed a flattening behavior in which it clung to the rotating rod and was carried passively for at least one full rotation. The excess of flattening in Edmonton arose from a defect in attaching the 320 grit sandpaper to the rod.

Reliability, as indicated by correlations of escape latencies on the first four trials, was strongly affected by the presence of the A mice. With all eight strains in the data set, the correlations were substantial (from .40 to .62), whereas they were much lower (from .18 to .36) when A mice were omitted.

For the other seven strains, there were highly skewed distributions of latencies, and after trial four there was no further improvement in any strain. To address the skewness, a square root transformation was employed for data subjected to significance tests. Repeated measures ANOVA on the first four trials revealed an obvious strain effect, $F(6, 246) = 26.0$, $p < .000001$, and a significant shipping effect, $F(1,$

$246) = 7.7$, $p = .006$, but no sex effect or interactions among any group variables.

Excluding one entire strain and half the trials gave what appeared to be a satisfactory portrayal of our results, but we were uneasy about eliminating so much valid information, and therefore, sought other useful indicators of performance. The central question was whether lab site effects would appear and perhaps interact with strain after the initial analysis with data exclusion pointed to an absence of site effects.

For most mice, learning to find the visible platform was very rapid. Expressing data as mean escape latency gave a fairly good portrayal of behavior, but it conflated rapidity of learning with variability in performance, especially when considering the merits of individual animals. We devised new indicators that more clearly distinguished between these two aspects of behavior. With respect to learning, proficient performance was evident when the mouse swam directly to the platform, perhaps deviating from a straight line but never taking a circuit around the entire tank. We observed that direct approach and successful climbing onto the platform generally occurred in about 5 s or less, so 5 s was adopted as the criterion for a success. Whether the one success would indeed lead to competence or was only a brief episode in the trials of a klutz depended on performance in subsequent trials. In this respect, we also observed numerous mice that acquired the task rapidly and performed consistently well for a few trials but then had a relapse or two in which they swam around the tank several times before approaching the platform.

We adopted the following two indicators of performance. Rapidity of initial acquisition was expressed in the number of trials required to achieve a success (escape latency < 6 s). This index revealed that many animals were successful on the very first trial, which showed that they had learned about the visible platform during pretraining the previous day. We assigned these mice a score of 0, denoting that no training trials were needed to be successful. The second indicator was average inconsistency of performance across trials. This was calculated as the sum of absolute values of the difference between adjacent trial latencies, divided by 7. For an animal that began with a very short latency of 2 s and then reached the platform in 1 s thereafter, this index would be $(1 + 0 + 0 + 0 + 0 + 0)/7$. An animal that went the full 40 s on every trial would have an inconsistency score of 0. Figure 7(A) shows examples of three mice that achieved successful performance on trial 3 but differed greatly in the progress of performance. Figure 7(B) shows that inconsistency scores were on average greatest for animals having intermediate values for

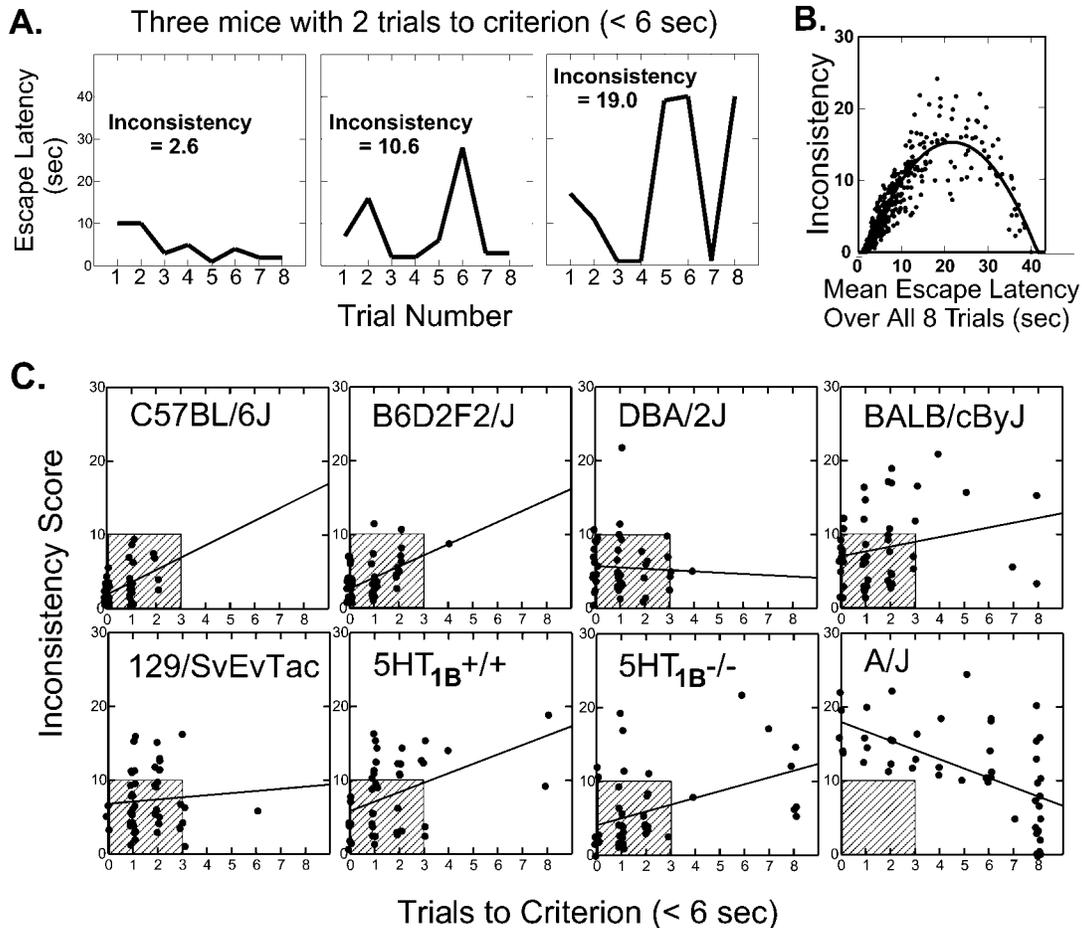


Figure 7 Inconsistency scores (see text) and learning on the visible platform water escape task. (A) Examples of three mice that met the proficiency criterion of an escape in less than 6 s after two trials but had very different inconsistency scores. (B) Inconsistency score versus mean escape latency across all eight trials, showing that the least proficient mice were consistently poor. (C) Strain differences in the pattern of trials to criterion (successful escape in <6 s) versus inconsistency. Even though most mice of the cBy and 129-derived strains experienced success early in training, many of them did not then embark on a series of proficient escapes. No A strain mouse ever showed consistent success on this task.

mean latency over the eight trials, but variability in inconsistency was also greatest at intermediate mean latencies.

The patterns of trials to the first success versus inconsistency were very different for the eight genetic groups [Fig. 7(C)]. Many B6 and F2 mice were successful on the first or second trial and never erred thereafter. Three F2 mice did very well initially but later had a relapse when they took a lengthy excursion around the tank that placed their inconsistency scores near 10. D2 mice were a little slower to achieve good performance but usually were consistently good after experiencing a success, although there were a few exceptions. Most A mice had high inconsistency scores, even if they had a success early in training,

which suggested they did not benefit much from the success. Some were truly horrid and always reached the 40-s limit. For cBy and the 129-derived strains, most experienced a success within the first two trials but showed considerable variability on subsequent trials, although a few mice in these four groups performed exceedingly well. Hence, cBy and the 129-derived strains were best characterized by the high degree of variability both between mice and within a mouse. Comparing D2 and cBy, similar numbers of animals encountered success within the first four training trials, but D2 mice were much more consistent after a success and rarely relapsed.

Results of ANOVAs done on all eight strains are summarized in Figure 1. For no measure was a sex

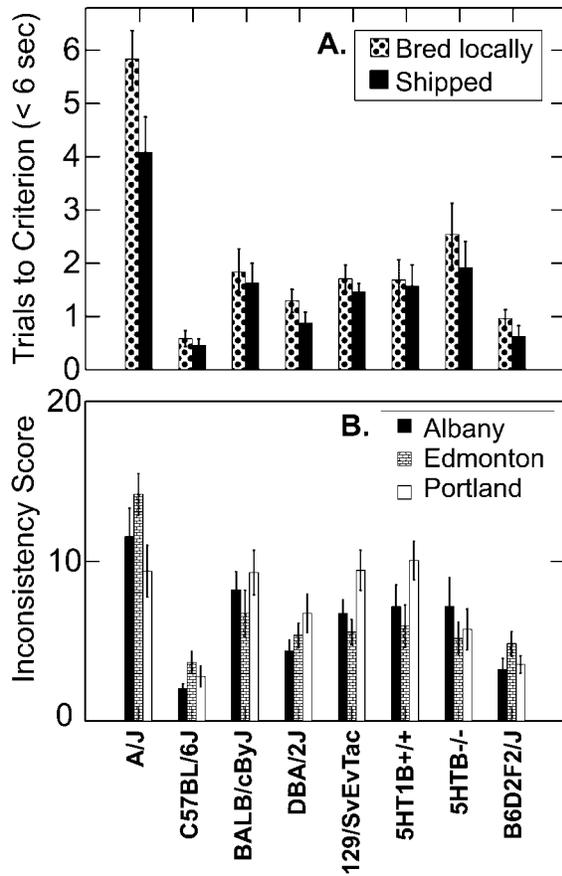


Figure 8 (A) Trials to criterion (escape in <math>< 6\text{ s}</math>) on the visible platform water escape task were slightly lower in animals that were shipped, but the difference in escape latency appeared only on the first two trials. (B) Inconsistency scores showed a moderate genotype \times site interaction (see Fig. 1) but no main effect of site.

difference or interaction with sex detected. Strain differences were obviously very large, as expected from Figure 7. For trials to the first success, there was a small superiority of mice that had been shipped to our labs [Fig. 8(A)], whereas for inconsistency score there was a modest interaction between strain and lab [Fig. 8(B)] but no shipping effect. Analyzing the latency data trial by trial (using a square root transform to reduce heteroscedasticity), the shipping effect appeared only on the first two trials, whereas a lab effect emerged after the first two trials. Thus, the shipping effect was transitory. The lab effect on speed was quite small, amounting to only a 1-to 2-s difference between the labs on the first four trials, and this arose mainly from a more pronounced tendency in Edmonton to record a latency of 1 s. The lab difference in inconsistency score could not be attributed entirely to the abundance of 1-s latencies in Edmonton, because the effect was largest for the slowest

strains. The lab effect might reflect a difference in the personal definitions of a platform escape by technicians at the three labs, combined with a tendency to assign lower scores to strains that were expected to perform best.

Ethanol Preference

One of the greatest challenges to obtaining a valid indicator of ethanol preference is a strong left or right bottle preference in individual mice. A mouse that begins to drink from a particular water bottle tends to consume most of its daily liquid from that bottle when another, identical bottle is available. We controlled or compensated for this positional bias by placing the ethanol on one side for 2 days and then the opposite side for another 2 days. As shown in Figure 9, the correlations between preference ratios on days 1 and 2 and between days 3 and 4 were very high, partly because of the positional bias, whereas the correlation between days 2 and 3 when the ethanol was on opposite sides was substantially lower. The day 2 versus day 3 pattern involved some mice that stubbornly resisted the change of ethanol bottle and instead remained with the preferred side, as well as mice that shifted sides to stay with or away from the ethanol. Thus, the cluster of mice with preference scores near 1.0 on both days 1 and 2 was a heterogeneous group of animals, some of which strongly preferred ethanol and others that had little preference for ethanol but a strong side preference. A similar phenomenon occurred for mice with both days 1 and 2 near 0 preference; some detested ethanol, whereas others had a strong side preference. When the ethanol side was reversed, those with a strong preference or aversion for ethanol switched sides, whereas those with no ethanol preference remained on the preferred side. This pattern of weaker and stronger individual preferences for bottle position was similar to observations by Collins (1975) and Biddle and Eales (1999) for paw preference when mice are required to reach for food into asymmetrical tubes opposite their bias. One consequence of these strong individual differences was that the center zone where individuals had no strong preference on either trial was almost vacant. Only when preference ratio was averaged over the 4 days did many intermediate scores appear, those being mice that stayed with their position preference.

Analysis of variance was conducted for five measures of drinking behavior (Fig. 1). Total volume of tap water and 6% ethanol consumed over the 4 days differed substantially among strains, labs, and sexes, but no shipping effect or interaction was apparent. When volume consumed was divided by body weight,

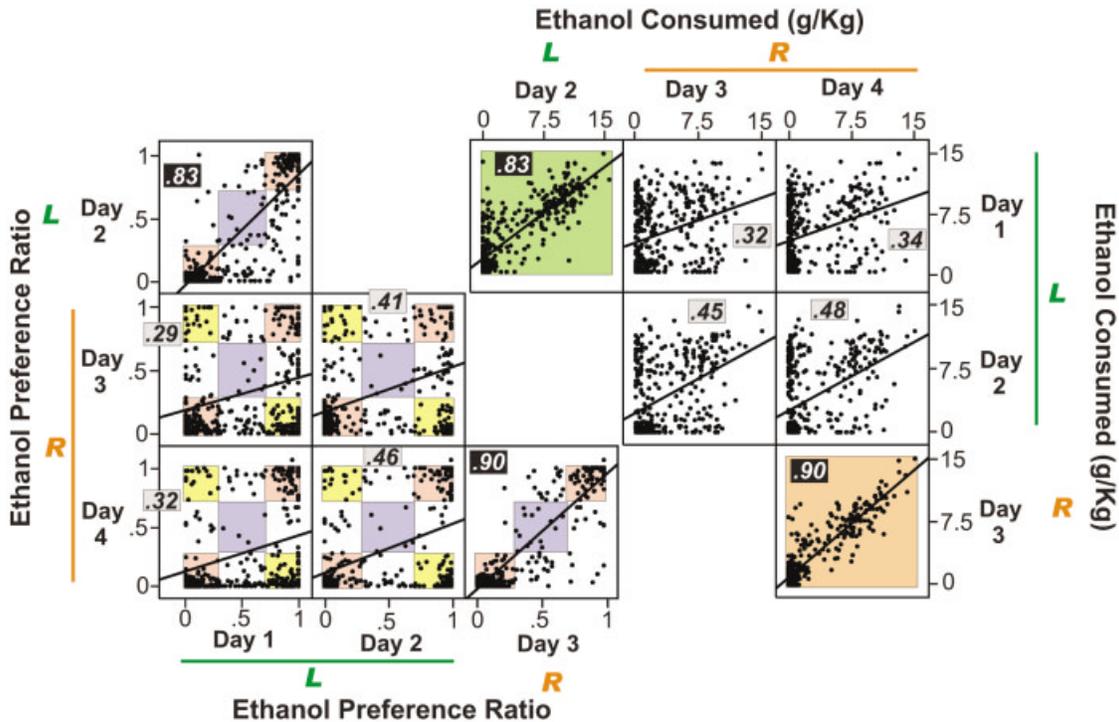


Figure 9 Scatterplots of scores on the 4 days of ethanol preference testing, showing the correlation of values on each pair of days. The bottle containing ethanol was on the left side of the cage for the first 2 days and the right side on the last 2 days. (A) For preference ratio, the correlations were much higher between days with the ethanol bottle on the same side, and there were very few mice with an intermediate preference ratio on any 1 day (blue region), which showed that preference ratio was strongly influenced by stubborn positional preferences. Pink regions show mice that had strong preference or aversion for ethanol that overcame positional preferences, while yellow regions show animals that remained with the preferred side and had abrupt changes of preference ratio. (B) For amount of ethanol consumed per g body weight, correlations were also highest when bottle position was not changed, but many more mice with intermediate values were seen than for preference ratio. The greater clustering of scores near 0 on days 3 and 4 reflected a general downward trend in the amount of ethanol consumed (see Fig. 10).

the strain difference was reduced, the lab effect was eliminated, but the sex difference increased. Thus, the lab difference in volume consumed was simply a consequence of body size difference, whereas females drank considerably more liquid relative to their body weights.

Appetite for ethanol was quantified in two ways. The average amount of ethanol consumed over the 4 days relative to body weight revealed a very large strain difference and a moderate sex difference but no lab or shipping effect. Ethanol preference ratio, on the other hand, indicated only a very large genotype effect. As shown in Figure 10, most strains gradually reduced the amount of ethanol consumed over the 4 days, whereas B6 mice remained consistently high. D2 and to a lesser extent A and cBy mice showed aversion over all 4 days. As expected, the F2 hybrid between B6 and D2 averaged close to 50% preference

owing to genetic variation in ethanol preference. That is, some F2 mice showed low and some high preference, presumably because individuals had inherited different proportions of B6 and D2 alleles at genes relevant for ethanol drinking. The three 129-derived strains had intermediate average scores, but scatter plots revealed that there was an extraordinary degree of variation among mice within each of these groups, including animals with strong aversion to ethanol, others with a strong preference for ethanol, and still others with strong positional preferences. It was also noteworthy that there were no differences in preference drinking between 5-HT_{1B} *+/+* and *-/-* mice.

Because preference ratio on the day when bottle positions were changed was so strongly influenced by positional biases, an alternative indicator of preference was considered: the average preference ratio on days 2 and 4, each being the day following a bottle

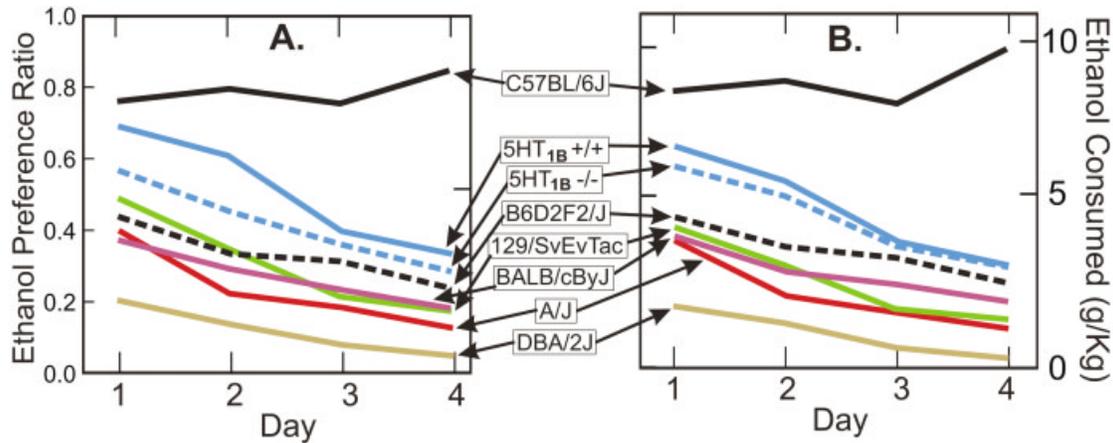


Figure 10 Mean ethanol preference ratio and amount of ethanol consumed over the 4 testing days for eight genotypes. A general downward trend was apparent for all but B6 mice, and the strain distribution patterns were very similar for the two measures.

position change. We have used this index previously to characterize strain differences (Phillips et al., 1994). The correlation between days 2 and 4 ($r = .455$) was substantially higher than the correlation between days 1 and 3 ($r = .288$), and the correlation between the average of days 2 and 4 versus the average of all 4 days was nearly perfect ($r = 0.977$); apparently data from days 1 and 3 were so strongly influenced by positional effects that they contributed little to the overall average. Consequently, the magnitude of the strain difference (Fig. 1) was somewhat greater for the average of days 2 and 4 than for the average of all 4 days.

For no indicator of ethanol consumption or preference was any difference among sites detected. Only the total volume of fluid consumed showed a site effect, and that effect evidently arose from a site difference in body size.

Correlations across Tests

In view of the diverse measures obtained on several of the tests, it is interesting to examine patterns of correlated measures across tasks, especially correlations that arose from hereditary differences. These correlations are shown in Table 3 for the principal measures from each task. As we might expect, groups that traveled longer distances in the activity box on Monday made more arm entries in the elevated plus maze on Tuesday and spent less time at the center hub of the plus maze. Groups that reared more on Monday in the activity box spent less time in the center of the plus maze and made more plus maze arm entries, a relation that may be related to overall locomotor activity levels. Center time in the activity box on Monday was

strongly related to number of arm entries and weakly correlated with percent time in the open arms in the elevated plus maze on Tuesday but not to plus maze center time.

The more active groups in the activity box and on the plus maze also required fewer trials to achieve a success on the water maze on Thursday. Furthermore, groups that spent more time in the center zone of the activity box improved more quickly on the water escape task. It would be hazardous to conclude learning ability and activity level are genetically related, given the relatively small sample of genotypes observed in this study. As shown in Figure 11, correlations with center time in the activity box were strongly influenced by the presence of A strain mice that hugged the walls in both the water tank and the activity box. Interfering response tendencies of certain strains can thus substantially alter the relations between variables in a genetic correlation or factor analytic study and yield misleading results if not detected.

We found that the higher correlations among variables were detected in all three labs, with certain exceptions involving center time and arm entries on the elevated plus maze. Likewise, the generally low correlations of measures in the ethanol consumption test with other tests was confirmed in all three labs. Beyond these general observations, we are reluctant to draw conclusions from what proved to be a very complex pattern of moderate correlations among many other variables, some of which were positive in one lab but negative in another. It is hazardous to interpret moderate strain correlations as having a purely genetic origin when they are based on repeated measurements of the same animals and thus may be

influenced by prior testing on a different kind of apparatus and well as common genetic variance.

DISCUSSION

In the following sections, we first compare results and interpretations from our initial report of this study with those from the expanded analysis presented here, and we discuss our findings in the context of reactions to the earlier publication. Then we offer more global interpretations of the study’s implications for test standardization and the study of mutations.

Comparison of Current and Earlier Results and Interpretations

For the two locomotor activity tests, we initially presented separate analyses in naive mice and following cocaine administration for two measures of activity (Crabbe et al., 1999). We have now analyzed three additional measures of activity and have examined cocaine effects on each variable in one large analysis that provides formal tests of cocaine effects and interactions involving cocaine. Cocaine generally increased distance traveled, percent time moving, and locomotion speed, while reducing rearing and time in the center of the apparatus, but this pattern did not occur in all three labs for all genotypes. The genotype × site interactions were clearly significant for all variables, as were the genotype × cocaine interactions (see Fig. 1). Furthermore, the current analyses reveal clearly that cocaine had genotype-dependent effects on activity variables (with the exception of center time) that differed among sites (Table 1). Thus, our earlier conclusions about the importance of genotype × environment interactions have been amply confirmed and expanded for locomotor activity and cocaine effects on activity.

For the elevated plus maze, we reported that total entries and time in open arms were significantly affected by genotype, site, and their interaction (Crabbe et al, 1999). Our expanded analysis explored the effects of eliminating noncompliant mice that spent most of the time in the center of the maze or had very few arm entries. The current analyses agree well with those reported earlier. For total arm entries, site and genotype were important, and their interaction was pronounced. Percent time in open arms yielded results very similar to the earlier-reported variable, time in open arms (uncorrected for time spent in the center); that is, the effect of site was quite pronounced, and genotype had a lesser effect. Site and genotype did not interact significantly using this better-articulated vari-

Table 3 Correlations of Eight Strain Means for Principal Measures of Behavior on Five Tests

	Locomotor Activity			Activity under Cocaine			Elevated Plus Maze			Water Escape		Ethanol Preference	
	Rearing	Center Time	Center Time	Dist.	Rearing	Center Time	Center Time	# Entries	% Open Time	Trials to Criterion	Fluid/g	Pref24	
Loco-motor activity	0.946	0.550	0.304	0.871	0.917	-0.202	-0.781	0.877	0.230	-0.699	0.475	0.308	
				0.908	0.951	-0.413	-0.845	0.694	-0.035	-0.515	0.599	0.113	
				0.417	0.342	0.635	-0.121	0.843	0.606	-0.873	-0.151	0.356	
Activity under Cocaine					0.949	-0.302	-0.841	0.702	-0.186	-0.657	0.536	0.000	
						-0.449	-0.916	0.716	-0.164	-0.633	0.399	0.188	
							0.679	0.153	0.628	-0.247	-0.051	0.129	
Elev. plus maze								-0.583	0.300	0.502	-0.181	-0.046	
									0.477	-0.915	0.216	0.429	
Water escape										-0.282	0.014	0.467	
Ethanol pref											-0.043	-0.407	
												-0.369	

Each strain mean is based on n = 48 scores. Correlations of 0.7 or more are shown as **bold** type. No formal tests of significance.

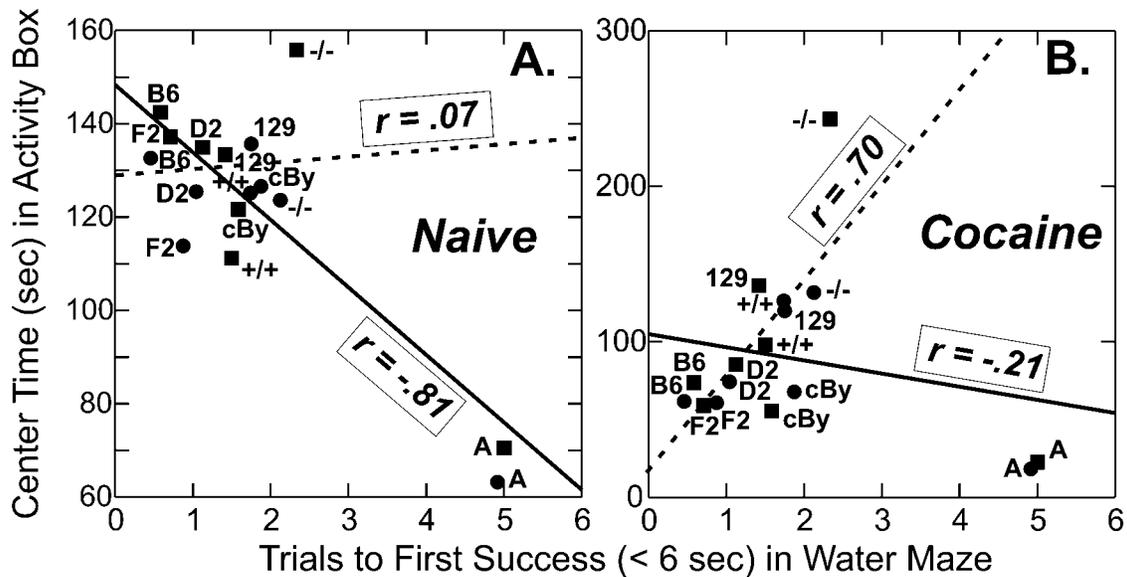


Figure 11 Associations between trials to achieve the first success (escape in <6 s) on the water maze on Thursday and time spent in the center zone of the activity box (A) in naive mice on Monday and (B) under the influence of 20 mg/kg cocaine on Friday. Means for males are shown by \blacksquare and females by \bullet . The linear relation with all 16 strain/sex groups is shown by a solid line, whereas the relation when the A strain is removed from the data set is shown by a dashed line. Pearson correlations of group means were markedly altered by excluding A mice.

able. The genetic effects on open arm exploration were statistically larger for the cleansed sample, but eliminating inactive mice did not change the pattern of results.

For the accelerating rotarod, we originally eschewed any report at all, given the ability of the mice to subvert the validity of the measures. The current results do, however, demonstrate well that one needs to carefully examine the behavior as it is being performed by the mouse. If we had simply recorded latencies on the rotarod, we would have quite erroneously concluded that the task was highly variable genetically, when in fact the data were dominated by an artifact of gluing sandpaper that differed between labs and allowed many mice to grab a free ride on the rotarod without walking or running.

Others have noted the tendency of mice to clasp a rotarod rather than try to balance on it. Tarantino et al. (2000) tested four inbred strains of mice on a 3-cm rotarod where the rod accelerated at a constant rate from 4–36 Rpm. They reported passive rotations but did not count them. They also reported strain differences in the latency to make the first passive rotation on each of the six trials. Because they did not report the proportion of mice making the passive rotation response, it is difficult to offer any interpretation of the strain differences in latency to make this response. Another group (Hyde et al., 2001) tested Ts54Dn

(chromosome 16 trisomic) mice using a 6-cm diameter rod accelerating from 0.8–30 rpm and noted that no mice could passively rotate. In an additional experiment, they used a rotarod with a 3-cm diameter and tested mice in 60-s trials at fixed speeds, increasing the speed of rotation between trials. Under these conditions, many mice adopted the passive rotation strategy on one or more trials.

We believe that when some animals adopt the passive rotation strategy while others do not, and when use of this alternative strategy varies from trial to trial within an animal, it is no longer possible to interpret results based on the rotation speed when the mice eventually fall. Apples are being compared to oranges here. Although it is not impossible that gradual increases in usage of this strategy could serve as an index of learning, the studies that have reported it have not presented their data in this way.

To the measure of water escape learning originally reported, we added an assessment of the variability of performance during acquisition, the inconsistency score. Results for escape learning overall remained as originally reported—a large strain difference, plus a small and transitory effect of shipping status. There was a small effect of site on acquisition. Only the inconsistency measure showed a modest interaction between genotype and site.

Generally speaking, when the lab effect is statistically significant but numerically small, as occurred for water escape learning over the first few trials, concerns about the exact definitions of behavior at each lab are legitimate, especially when the technician must make a decision about when to stop the clock. This problem can best be addressed in the future with automated, computer-based video tracking or photocell detection in each lab.

Finally, for ethanol preference drinking, we include here reports of preference ratio and total fluid consumption. We also report on a measure of preference we have used elsewhere to characterize genotypes that eliminates some of the bottle position bias (Phillips et al., 1994). Basically, we saw for the additional variables what we had earlier reported. Females drink more than males, and the genotypes differ drastically. There were no site \times genotype interactions in either analysis for any alcohol variables. Furthermore, measures in the ethanol preference tests showed low genotypic correlations with measures on other tests.

Analyses of genotype-specific effects cannot be extricated from the situation in which they are assessed. This fundamental conclusion from the earlier report (Crabbe et al., 1999) remains appropriate, even when the data collected are subjected to more extensive scrutiny. The additional variables we report here also showed significant evidence for genotype \times site interaction. In many cases, those interactions were substantial, and in some cases, a given genotypic comparison would have led to one conclusion in one laboratory but an opposite conclusion in another.

In both the original report and after assessing additional measures, effects of shipping condition and sex were generally small and nonsignificant. Only the greater ethanol consumption by females stands as a noteworthy sex difference in our data. These results do not justify the study of only one sex in behavioral research, but they do suggest that power and sample size calculations used to plan a study can often be done for the combined number of male and female mice, provided equal numbers of each sex are employed. Exceptions to this rule need to be made for tests where sex differences are expected or may be of central interest in the research. The general absence of noteworthy shipping effects on commonly studied behaviors may provide comfort to labs that use different approaches to obtaining mice, but our data cannot exclude the possibility of shipping effects when mice are tested soon after arrival in the lab.

Response to the Original Report

Reactions to the 1999 report in *Science* covered the spectrum of opinion from an indignant sigh that we all knew this already to hysterical outbursts that our findings invalidate the entire field of behavioral genetics. Most comments, of course, were more moderate and thoughtful. Some of the interpretations and misinterpretations have been discussed well in a feature article by a neuroscientist in *The Sciences* (Sapolsky, 2000). Given our interpretation of the results restated above, we were puzzled when discussing these data publicly to find that many apparently interpreted the results pessimistically to indicate that behavioral tasks were intrinsically unreliable in mice, and that strain differences were unstable across environmental conditions. Thus, we were curious about what those who chose to cite the article would have to say.

ISI Web of Science listed 133 papers (excluding books and book chapters) in December, 2001, that cited Crabbe et al. (1999). The fairly widespread citation of this article is congruent with our personal experience; we have been approached many times about the study by other neuroscientists as well as journalists. The interpretation of the results depended to some extent on the eye of the beholder. In an attempt to characterize the nature of the citations, we evaluated a selection of the 133 citing papers. We excluded seven that were authored by one of us. Using a random number table, we selected 25 papers of the remaining 126 for analysis. We eliminated six papers because the journal could not be located easily in Portland. We eliminated three more papers because the author or the author(s) group was already represented by a previous selection. These nine deletions were replaced by nominations from the next nine random numbers.

Using professional judgment, tempered with a standard dose of ox-gored authority, one of us (J.C.) categorized the nature of the citation with one or more descriptors, as accurate (i.e., sharing the authors' biases), inaccurate (by reason of a demonstrable, non-trivial error in the content of the citation), overly negative (i.e., accurate, but stating the pessimistic interpretation cited in the preceding paragraph), misleading (e.g., citing the article as evidence for an author-favored position not addressed by the article), or irrelevant (i.e., to the goal here: the typical irrelevant citation merely cited the article as showing strain differences in ethanol preference drinking). A final category, bizarre, was reserved for one unparseable citation. The citations were apportioned as follows: 15 accurate, 3 overly negative, 3 inaccurate and mislead-

ing, 3 irrelevant, and 1 bizarre. Another author (T.P.) reviewed these same 25 papers, and her opinion, unguided by the initial criteria for evaluation, agreed essentially with J.C.

We found the results of this survey rather encouraging, but also sobering. On the good side, 60% of citations were in our view apposite and reflected our own cautionary interpretation of the data. We viewed only three of the 25 as overly pessimistic. On the discouraging side, though, were the four articles that were inaccurate, misleading, or both. Nonetheless, this informal survey says to us that the articles' main points were generally grasped by those who cited it.

On the other hand, our contacts with scientists outside the field of mouse neural and behavioral genetics have generally shown their perception of the article's message to be the pessimistic interpretation—behavior is just not reliable. To some degree, we expect that this is because many did not read the article closely, but relied on summary reports of it or, even worse, article headlines that were not even written by the author of the news article. Media coverage of this article ranged widely from very cogent (Sapolsky, 2000) to appalling (Immen, 1999). A reasonably accurate news report of our study by Enserink (1999) in the same issue of *Science* was accompanied by a pejorative headline proclaiming fickle mice, while a report in *The Globe and Mail* sensationalized our findings, focusing on laid-back Edmonton mice, and claiming that lab differences invalidate findings about behavior (Immen, 1999). The study continues to be the subject of reanalysis and commentary in the media, exemplified by a recent proclamation that results were caused by fluoridation of drinking water in Edmonton (Darmouth professor, 2002).

Given the large body of research in mouse behavioral genetics dating from Yerkes (1907), no informed scientist should be shocked by a report that environment can alter mouse behavior. Quantitative genetic studies of mouse behavior have virtually unanimously found that nongenetic sources of individual differences are very important (Wehner et al. 2001). Decades of research have taught us that complex behaviors are always multifactorial and are influenced substantially by both genetic and environmental sources of variation. Within the realm of the genes, we also know from QTL mapping studies that the genetic part of the equation is itself complex and strain differences are always products of polymorphisms at several loci. Likewise, the environment itself is complex and multidimensional. The lab environments in our studies differed in several ways, and it is most unlikely that the dramatic interaction effects we observed can be attributed to differences in a

single environmental variable. Nonetheless, we received numerous e-mail suggestions nominating a hidden environmental variable as the sole or major cause of the lab effects (e.g., composition of the local tap water, variations in emissions from the fluorescent light sources and how they were filtered by different kinds of plastics, etc.).

To Standardize, or Not to Standardize?

One inference that could be drawn from the laboratory–genotype interactions revealed by this study is that standardization of behavioral tests is desirable. The results of our study, however, suggest that standardization of the test situation would not guarantee identical results in different labs because of large effects of laboratory environments. At the same time, test standardization allows us to discriminate between effects of interlab differences in the test situation versus the lab environment. Our findings may inspire others to pursue test standardization, but our experience also highlights the daunting nature of this challenge. We were able to equate several of the tests only by adopting features of the apparatus and procedures that were probably not optimal; that is, convenient compromises were needed to meet the stringent constraints of our study. If there is to be a widely accepted standard, there will need to be agreement that the test parameters are the best available.

There are nearly as many ways of testing a particular behavioral construct as there are experimenters, and few articles offer sufficient detail to enable the naive reader to replicate the test situation. Information germane to the laboratory environment is almost never presented. This has led some to urge creation of a database comprising very detailed, behaviorally relevant information for each lab, which could be useful (Surjo and Arndt, 2001). There are many issues to be considered before undertaking such a venture (Wahlsten, 2001). A thoughtful discussion of the pros and cons of such standardization has been presented elsewhere (van der Staay & Steckler, 2001).

The idea of establishing a standardized battery to characterize basic sensorimotor, behavioral and developmental functions in mice has a long history (Yerkes, 1907; Fox, 1965; Irwin, 1968). Several good batteries have been proposed recently (Crawley & Paylor, 1997; D.C. Rogers et al., 1997, 1999; Gold, 1999; D.C. Rogers et al., 1999), but there is currently no widely accepted battery of behavioral tests covering the range of behavioral endpoints important for the understanding of human disease.

For any given behavioral domain, multiple tests exist, but these differ in many details among labora-

tories. Whereas establishing that the forelimb placing reflex is intact is fairly straightforward (Fox, 1965), more complex behavioral assays are far less straightforward. One might contemplate an efficient behavioral screen with a single test of anxiety, one for hyperactivity, and another for learning deficits, but this would be an exercise in wishful thinking. Standardization within any broad behavioral domain presumes that a single test of the psychologic construct indeed captures the essence of the modeled behavior. Unfortunately, this situation generally does not prevail. Instead, it is likely the case that no single test in any one behavioral domain holds enough construct validity to be nominated as a standard (Boehm et al., 2000). In addition, there are issues surrounding test order effects (McIlwain et al., 2001), which were not addressed by our study but are probably important. Despite the complexity of the undertaking, there remains a possibility that a well-conceived, thorough approach to capturing the range of complex behavioral domains could be achieved through well-chosen, multiple tests (Brown et al., 2000), but this has not yet been achieved in any mouse testing laboratory.

Implications for Studies with Mutants

We included a null mutant and its wild-type in this study to see whether a manipulation targeting a single gene would result in differential sensitivity to laboratory environment. A strain of mice null mutant for serotonin 1B receptors (5-HT_{1B}^{-/-}) had been shown to have twofold increased ethanol consumption and reduced sensitivity to ethanol-induced motor incoordination in the grid test of ambulatory ataxia (Crabbe et al., 1996). Because many different substrains of the 129 inbred strain have been used as a background for targeted mutagenesis (Simpson, et al., 1997), we also included in our study one of the popular 129 strains, 129/SvEvTac, as well as the wild-type strain from the colonies of RH (5-HT_{1B}^{+/+}). When preference scores for these three 129-derived strains were calculated, we were greatly surprised to see no difference in ethanol preference between the null mutants and their wild types. The original drinking difference had been replicated four times (Crabbe et al., 1996), and the current results showed no difference three times—once in each laboratory. This finding led to an intensive analysis of the reason for the loss of the original phenotypic difference over generations. Examination of the breeding scheme used to maintain the animals suggests strongly that the gradual introduction of more 129/SvEvTac alleles into the background strain maintaining the null mutant diluted the original genetic difference through an epistatic interaction with

other genes in the background (Phillips et al., 1999). The 129/SvEvTac strain showed lower preference overall than the other two 129 strains, and 129 strains are known to differ substantially genetically (Simpson et al., 1997).

Whether the serotonin receptor gene itself affects ethanol drinking remains difficult to determine for certain. Another population of the 5-HT_{1B}^{-/-} null mutants and wild types, also derived from the colonies of RH, did not show the difference in preference drinking (Bouwknicht et al., 2000). However, a subsequently obtained group of animals, also from the colonies of RH, shows the original, twofold preference drinking difference. The choice of a 4-day test for preference drinking may have attenuated the differences in preference between null mutant and wild type in the current studies, as we have seen that the preference difference emerges most clearly after about 7–8 days of testing (Phillips and Crabbe, unpublished findings). In addition, as we have selectively bred animals for preference drinking or aversion starting with an F2 intercross of 5-HT_{1B}^{+/+} × 5-HT_{1B}^{-/-}, we have found that the frequency of the 5-HT_{1B} gene cosegregates with drinking phenotype in the predicted direction, even though there was no preference difference between ^{-/-} and ^{+/+} genotypes in the F2 population (Phillips and Belknap, 2002).

On the other hand, another trait differentially expressed by the 5-HT_{1B} null mutants, increased activity when first exposed to an apparatus (Castanon et al., 2000; Wahlsten et al., 2001), was seen in all three laboratories (Fig. 2), although it was expressed more strikingly in Portland than the other two laboratories. Interestingly, the pattern of activity differences in the elevated plus maze showed lower activity of the mutants in Albany, no difference in Edmonton, and higher activity in Portland [Fig. 5(A)]. The locomotor response to cocaine had also previously been reported to be enhanced in the null mutants (Castanon et al., 2000), but this difference was seen only in Edmonton in the current studies.

We remain convinced that the likely reason for the replication of only some of the behavioral differences between 5-HT_{1B}^{+/+} and ^{-/-} genotypes in our study is the genetic polymorphisms among the three contributing 129 substrains. One control procedure to minimize such genetic drift is to maintain the mutation with heterozygote matings, which has the added advantage of controlling for maternal effects (Hen, 1999). The ideal comparison would then involve littermates that should differ consistently only in genotype at the locus in question (Phillips et al., 1999).

There are other solutions to this problem as well (Gerlai, 2001, and references therein). Although our study clearly does not provide a sensitive and thorough test of the sensitivity of a null mutation to different environmental conditions, we believe it provides enough evidence of environmental sensitivity to suggest caution in interpreting the effects of such a manipulation based on results from a single laboratory. It should not, therefore, be surprising that three different laboratories obtained markedly different elevated plus maze results when testing CRHR2 knock-out mice, produced independently and tested under nonmatched conditions (Bale et al., 2000; Coste et al., 2000; Kishimoto et al., 2000; see Crabbe, 2001 for discussion).

GENERAL CONCLUSIONS

At the outset of this study, we asked whether three different labs that tested the same behaviors of the same strains in the same way would obtain the same results. We offer no simple answers to this complex issue, but several conclusions are warranted.

1. Analysis of our data in greater depth revealed essentially the same pattern of results as our initial report (Crabbe et al., 1999). Strain \times lab environment interactions were substantial for several measures of open field activity, cocaine activation of motor behavior, and elevated plus maze behavior, whereas noteworthy interactions were not observed for visible platform water escape and ethanol preference.
2. Although the interaction term in the ANOVAs was clearly significant and moderate to large for several measures, the main effect of strain was very large for almost every measure, and robust differences between the most extreme-scoring strains were generally observed in all three labs. Substantial effects of lab environments did not suppress or obscure effects of genetic variation, although they did alter the pattern of moderate genetic effects.
3. Different strains were responsible for interaction effects involving different behaviors. We did not find that certain strains were generally susceptible or resistant to lab environment effects across most tasks.
4. Strain \times lab environment interaction effects were not confined to behavior. Body weight also showed interaction effects, although brain weight and forebrain commissure sizes did not (Wahlsten et al., 2001).
5. Sources of these lab environment effects are unknown, but one viable hypothesis can be proposed. Different experimenters at the three labs probably presented idiosyncratic arrays of odor cues and handled the mice somewhat differently. This factor needs to be studied systematically within a lab. Control of many studies can probably be improved by using the same person for testing mice in all groups in all phases of an experiment. It has been suggested that robots could eliminate lab differences in odors and handling methods, but we are not sanguine about an approach that would reduce the scope and complexity of behaviors that could be assessed.
6. Another experimenter effect can occur when people make different judgments or ratings of behaviors. For example, there were indications that the experimenter in Edmonton stopped the watch slightly sooner for a water escape. This kind of observer effect can give rise to a main effect of lab environment but is not likely to generate an interaction with strain, however. Training observers in different labs to identical criteria might reduce such an effect, but automated scoring with photocells or video tracking seems more promising for equating ratings across labs.
7. Several features of apparatus, protocols, and lab environment that were equated in our study could contribute to even larger interaction effects if they are allowed to vary among testing sites. Traffic and noise in the colony and test rooms, time of day for testing, and proximity to the time of cage changing could influence results. It seems likely that different features of the lab environment will prove to be crucial for different measures. Season of testing and density of mice in the cage influence levels of thermal nociception in mice (Chesler et al., in press), while nibbling on folate-rich corn cob bedding can suppress skeletal defects in transgenic mice (Pennisi, 2002).
8. Whereas the fine details of apparatus and procedures are often described well in neuroscience journals, details of the lab environment are usually scant or cursory. The lab environment ought to be presented in greater detail and viewed as an integral part of any study, one that can have a substantial influence on results.

FUTURE INVESTIGATIONS

The generalizability and replicability of our study are presently unknown. We are currently conducting a partial replication in two labs (Edmonton, Portland) as part of the Mouse Phenome Project. It is unknown how similar results would be for many other traits not tested here. One experiment explicitly comparing multiple inbred strains for electroconvulsive seizure thresholds found good agreement of strain sensitivities in two laboratories (Frankel et al., 2001), but this involved a reflex rather than a voluntary behavior. A sophisticated mathematical analysis of rodent exploratory behavior in an open arena (Drai and Golani, 2001) found strong evidence of stability for some, but not all, measures in three laboratories (Golani and Benjamini, personal communication).

One problem demanding further attention is the relative reliability or repeatability of different tests within a single laboratory. A test with low reliability yields data that are strongly influenced by the fluctuations of behavior from minute to minute or day to day, and such data will tend to make the genetic influence as well as interactions appear relatively small. In our data, strain differences in percent time in the open arms of the elevated plus maze were notably small, but we lacked evidence on the reliability of that measure. Perhaps the 5-min trial was not long enough to confer adequate reliability. It is highly desirable when comparing different tasks and measures of behavior that the tests have closely comparable reliabilities based on the same standard population of mouse strains. Reliability can be increased by using a longer trial or more trials, although caution is needed because validity of measures may also be altered when longer test sessions are used.

Animal models retain an advantage over human populations in the study of genotype by environment interaction because genotype can be replicated and held constant, while manipulating environment. Research on humans is severely limited in its ability to accomplish this. Heath et al. (2002) review some of the factors that make the detection of genotype \times environment interaction effects in psychiatric disorders so difficult. Fortunately, as the identity of specific genes affecting specific behavioral characteristics or disease traits in humans become known, this situation should improve. It is clearly desirable to know what environmental factors may serve a protective role, or as triggers, in individuals known to carry genetic variants that function as risk factors in disease development. For example, certain gene variants involved in ethanol metabolism protect against the development of alcoholism (Harada et al., 1983; Thomasson

et al., 1991; Borrás et al., 2000) by making alcohol intake aversive (e.g., nausea), and environmental factors such as peer pressure, divorce, or being raised in an environment with an alcoholic, may have little impact on this kind of genetic effect.

In humans and laboratory animals alike, the environmental factors that can most readily alter brain development and behavior will depend on the specific gene. Deliberate manipulation of developmental outcomes will be feasible only when the genes involved in multifactorial genotype by environment interactions are understood. Studies of inbred mouse strains can provide strong evidence of the importance of interactions, but studies of single gene effects will be needed to unpack the complex pathways involved in statistical interactions (see Johnston and Edwards, 2002). In mice, further progress in this endeavor will require identification of the genes responsible for behavioral differences between inbred strains (Phillips and Belknap, 2002; Phillips et al., 2002).

We thank R. H. Kant of AccuScan for the generous loan of equipment.

REFERENCES

- Anagnostopoulos AV, Mobraaten LE, Sharp JJ, Davisson MT. 2001. Transgenic and knockout databases: Behavioral profiles of mouse mutants. *Physiol Behav* 73:675–689.
- Bale TL, Contarino A, Smith GW, Chan R, Gold LH, Sawchenko PE, Koob GF, Vale WW, Lee KF. 2000. Mice deficient for corticotropin-releasing hormone receptor-2 display anxiety-like behaviour and are hypersensitive to stress. *Nat Genet* 24:410–414.
- Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. 2001. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* 125:279–284.
- Biddle FG, Eales BA. 1999. Mouse genetic model for left-right hand usage: Context, direction, norms of reaction, and memory. *Genome* 42:1150–1166.
- Boehm SL, Schafer GL, Phillips TJ, Browman KE, Crabbe JC. 2000. Sensitivity to ethanol-induced motor incoordination in 5-HT(1B) receptor null mutant mice is task-dependent: implications for behavioral assessment of genetically altered mice. *Behav Neurosci* 114:401–409.
- Bolivar V, Cook M, Flaherty L. 2000. List of transgenic and knockout mice: Behavioral profiles. *Mamm Genome* 11: 260–274.
- Borrás E, Coutelle C, Rosell A, Fernandez-Muixi F, Broch M, Crosas B, Hjelmqvist L, Lorenzo A, Gutierrez C, Santos M, Szczepanek M, Heilig M, Quattrocchi P, Farres J, Vidal F, Richart C, Mach T, Bogdal J, Jornvall H, Seitz HK, Couzigou P, Pares X. 2000. Genetic polymorphism of alcohol dehydrogenase in europeans: the

- ADH2*2 allele decreases the risk for alcoholism and is associated with ADH3*1. *Hepatology* 31:984–989.
- Bouwknicht JA, Hijzen TH, van der GJ, Maes RA, Hen R, Olivier B. 2000. Ethanol intake is not elevated in male 5-HT(1B) receptor knockout mice. *Eur J Pharmacol* 403: 95–98.
- Bowers BJ, Collins AC, Wehner JM. 2000. Background genotype modulates the effects of g-PKC on the development of rapid tolerance to ethanol-induced hypothermia. *Addict Biol* 5:47–58.
- Brown RE, Stanford L, Schellinck HM. 2000. Developing standardized behavioral tests for knockout and mutant mice. *ILAR. J* 41:163–174.
- Castanon N, Scearce-Levie K, Lucas JJ, Rocha B, Hen R. 2000. Modulation of the effects of cocaine by 5-HT1B receptors: a comparison of knockouts and antagonists. *Pharmacol Biochem Behav* 67:559–566.
- Chesler EJ, Wilson SG, Lariviere WR, Rodriguez-Zas SL, Mogil JS. 2002. The relative influence of organismic and laboratory environmental factors influencing a behavioral trait. *Nat Neurosci* (in press).
- Clément Y, Calatayud F, Belzung C. 2002. Genetic basis of anxiety-like behaviour: a critical review. *Brain Res Bull* 57:57–71.
- Cohen J. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Collins RL. 1975. When left-handed mice live in right-handed worlds. *Science* 187:181–184.
- Coste SC, Kesterson RA, Heldwein KA, Stevens SL, Heard AD, Hollis JH, Murray SE, Hill JK, Pantely GA, Hohimer AR, Hatton DC, Phillips TJ, Finn DA, Low MJ, Rittenberg MB, Stenzel P, Stenzel-Poore MP. 2000. Abnormal adaptations to stress and impaired cardiovascular function in mice lacking corticotropin-releasing hormone receptor-2. *Nat Genet* 24:403–409.
- Crabbe JC. 2001. Use of genetic analyses to refine phenotypes related to alcohol tolerance and dependence. *Alcohol Clin Exp Res* 25:288–292.
- Crabbe JC. 2002. Genetic contributions to addiction. *Annu Rev Psychol* 53:435–462.
- Crabbe JC, Phillips TJ, Feller DJ, Hen R, Wenger CD, Lessov CN, Schafer GL. 1996. Elevated alcohol consumption in null mutant mice lacking 5-HT1B serotonin receptors. *Nat Genet* 14:98–101.
- Crabbe JC, Wahlsten D, Dudek BC. 1999. Genetics of mouse behavior: interactions with laboratory environment. *Science* 284:1670–1672.
- Crawley JN. 2000. What's wrong with my mouse? Behavioral phenotyping of transgenic and knockout mice. New York: Wiley-Liss.
- Crawley JN, Paylor R. 1997. A proposed test battery and constellations of specific behavioral paradigms to investigate the behavioral phenotypes of transgenic and knockout mice. *Horm Behav* 31:197–211.
- Crawley JN, Belknap JK, Collins A, Crabbe JC, Frankel W, Henderson N, Hitzemann RJ, Maxson SC, Miner LL, Silva AJ, Wehner JM, Wynshawboris A, Paylor R. 1997. Behavioral phenotypes of inbred mouse strains: implications and recommendations for molecular studies. *Psychopharmacology* (Berlin) 132:107–124.
- Dartmouth professor sees link between fluoride and behavior. 2002. *Manchester Union Leader*, July 2, p. 1.
- Drai D, Golani I. 2001. SEE: a tool for the visualization and analysis of rodent exploratory behavior. *Neurosci Biobehav Rev* 25:409–426.
- Erlenmeyer-Kimling L. 1972. Genotype-environment interactions and the variability of behavior. In Ehrman L, GS Omenn GS, Caspari E, editors. *Genetics, environment and behavior*. New York: Academic Press.
- Enserink M. 1999. Fickle mice highlight test problems. *Science* 284:1599–1600.
- Fox WM. 1965. Reflex-ontogeny and behavioural development of the mouse. *Anim Behav* XIII:234–241.
- Frankel WN, Taylor L, Beyer B, Tempel BL, White HS. 2001. Electroconvulsive thresholds of inbred mouse strains. *Genomics* 74:306–312.
- Fuller JL. 1964. Measurement of alcohol preference in genetic experiments. *J Comp Physiol Psychol* 57:85–88.
- Gerlai R. 2001. Gene targeting: technical confounds and potential solutions in behavioral brain research. *Behav Brain Res* 125:13–21.
- Gold LH. 1999. Hierarchical strategy for phenotypic analysis in mice. *Psychopharmacology* (Berlin) 147:2–4.
- Gottlieb G. 1998. Normally occurring environmental and behavioral influences on gene activity: from central dogma to probabilistic epigenesis. *Psychol Rev* 105:792–892.
- Harada S, Agarwal DP, Goedde HW, Ishikawa B. 1983. Aldehyde dehydrogenase isozyme variation and alcoholism in Japan. *Pharmacol Biochem Behav* 18(Suppl 1): 151–153.
- Hen R. 1999. Letter. *Science* 285:2068–2069.
- Henderson ND. 1970. Genetic influences on the behavior of mice can be obscured by laboratory rearing. *J Comp Physiol Psychol* 72:505–511.
- Henderson ND. 1976. Short exposures to enriched environments can increase genetic variability of behavior in mice. *Dev Psychobiol* 9:549–553.
- Hogg S. 1996. A review of the validity and variability of the elevated plus-maze as an animal model of anxiety. *Pharmacol Biochem Behav* 54:21–30.
- Hyde LA, Crnic LS, Pollock A, Bickford PC. 2001. Motor learning in Ts65Dn mice, a model for Down syndrome. *Dev Psychobiol* 38:33–45.
- Immen W. 1999. Laid-back Edmonton mice have scientists puzzled. *Globe Mail* June 4:A1–A6.
- Irwin S. 1968. Comprehensive observational assessment: Ia. A systematic, quantitative procedure for assessing the behavioral and physiological state of the mouse. *Psychopharmacologia* 13:222–257.
- Johnston TD, Edwards L. 2002. Genes, interactions, and the development of behavior. *Psychol Rev* 109:26–34.
- Kishimoto T, Radulovic J, Radulovic M, Lin CR, Schrick C, Hooshmand F, Hermanson O, Rosenfeld MG, Spiess J. 2000. Deletion of *crhr2* reveals an anxiolytic role for

- corticotropin-releasing hormone receptor-2. *Nat Genet* 24:415–419.
- McClearn GE, Rodgers DA. 1959. Differences in alcohol preference among inbred strains of mice. *Q J Stud Alcohol* 20:691–695.
- McIlwain KL, Merriweather MY, Yuva-Paylor LA, Paylor R. 2001. The use of behavioral test batteries: Effects of training history. *Physiol Behav* 73:705–717.
- Moldin SO, Farmer ME, Chin HR, Battey JF Jr. 2001. Trans-NIH neuroscience initiatives on mouse phenotyping and mutagenesis. *Mamm Genome* 12:575–581.
- Moldin SO, Gottesman II. 1997. At issue: genes, experience, and chance in schizophrenia—positioning for the 21st century. *Schiz Bull* 23:547–561.
- Paigen K, Eppig JT. 2000. A mouse phenome project. *Mamm Genome* 11:715–717.
- Pennisi E. 2002. Good diet hides genetic mutations. *Science* 296:1011.
- Phillips TJ, Belknap JK. 2002. Complex-trait genetics: emergence of multivariate strategies. *Nat Rev Neurosci* 3:478–485.
- Phillips TJ, Belknap JK, Hitzemann R, Buck K, Cunningham CL, Crabbe JC. 2002. Harnessing the mouse to unravel the genetics of human disease. *Genes Brain Behav* 1:14–26.
- Phillips TJ, Crabbe JC, Metten P, Belknap JK. 1994. Localization of genes affecting alcohol drinking in mice. *Alcohol Clin Exp Res* 18:931–941.
- Phillips TJ, Hen R, Crabbe JC. 1999. Complications associated with genetic background effects in research using knockout mice. *Psychopharmacology (Berlin)* 147:5–7.
- Plomin R, Crabbe J. 2000. DNA. *Psychol Bull* 126:806–828.
- Rogers DC, Fisher EMC, Brown SDM, Peters J, Hunter AJ, Martin JE. 1997. Behavioral and functional analysis of mouse phenotype: SHIRPA, a proposed protocol for comprehensive phenotype assessment. *Mamm Genome* 8:712–713.
- Rogers DC, Jones DN, Nelson PR, Jones CM, Quilter CA, Robinson TL, Hagan JJ. 1999. Use of SHIRPA and discriminant analysis to characterise marked differences in the behavioural phenotype of six inbred mouse strains. *Behav Brain Res* 105:207–217.
- Rogers RJ, Dalvi A. 1997. Anxiety, defence and the elevated plus-maze. *Neurosci Biobehav Rev* 21:801–810.
- Sapolsky RM. 2000. Genetic hyping. *The Sciences* 40:12–15.
- Simpson EM, Linder CC, Sargent EE, Davisson MT, Mombaaten LE, Sharp JJ. 1997. Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. *Nat Genet* 16:19–27.
- Sokolowski MB, Wahlsten D. 2001. Gene-environment interaction and complex behavior. In: Chin HR, Moldin SO, editors. *Methods in genomic neuroscience*. Boca Raton, FL: CRC Press, p 3–27.
- Southwick CH, Clark LH. 1968. Interstrain differences in aggressive behavior and exploratory activity of inbred mice. *Commun Behav Biol A* 1:49–59.
- Surjo D, Arndt SS. 2001. The Mutant Mouse Behaviour network. A medium to present and discuss methods for the behavioural phenotyping. *Physiol Behav* 73:691–694.
- Takahashi JS. 1996. What's wrong with my mouse? New interplays between mouse genetics and behavior. Washington, DC: Society for Neuroscience.
- Tarantino LM, Gould TJ, Druhan JP, Bucan M. 2000. Behavior and mutagenesis screens: the importance of baseline analysis of inbred strains. *Mamm Genome* 11:555–564.
- Thomasson HR, Edenberg HJ, Crabb DW, Mai XL, Jerome RE, Li TK, Wang SP, Lin YT, LuRB, Yin SJ. 1991. Alcohol and aldehyde dehydrogenase genotypes and alcoholism in Chinese men. *Am J Hum Genet* 48:677–681.
- Thompson WR. 1953. The inheritance of behavior: behavioral differences in fifteen mouse strains. *Can J Psychol* 7:145–155.
- Tordoff MG, Bachmanov AA, Friedman MI, Beauchamp GK. 1999. Testing the genetics of behavior in mice. *Science* 285:2069.
- Toye AA, Cox R. 2001. Behavioral genetics: anxiety under interrogation. *Curr Biol* 11:R473–R476.
- Turri MG, Datta SR, DeFries J, Henderson ND, Flint J. 2001. QTL analysis identifies multiple behavioral dimensions in ethological tests of anxiety in laboratory mice. *Curr Biol* 11:725–734.
- van der Staay FJ, Steckler T. 2002. The fallacy of behavioral phenotyping without standardisation. *Genes Brain Behav* 1:9–13.
- Wahlsten D. 1990. Insensitivity of the analysis of variance to heredity-environment interaction. *Behav Brain Sci* 13:109–120.
- Wahlsten D. 1999. Single-gene influences on brain and behavior. *Annu Rev Psychol* 50:599–624.
- Wahlsten D. 2001. Standardizing tests of mouse behavior: Reasons, recommendations, and reality. *Physiol Behav* 73:695–704.
- Wahlsten D, Gottlieb G. 1997. The invalid separation of effects of nature and nurture: lessons from animal experimentation. In: Sternberg RJ, Grigorenko EL, editors. *Intelligence, heredity and environment*. Cambridge: Cambridge University Press, p 163–192.
- Wahlsten D, Crabbe JC, Dudek BC. 2001. Behavioral testing of standard inbred and 5HT1B knockout mice: implications of absent corpus callosum. *Behav Brain Res* 125:23–32.
- Wehner JM, Radcliffe RA, Bowers BJ. 2001. Quantitative genetics and mouse behavior. *Annu Rev Neurosci* 24:845–867.
- Würbel H. 2000. Behaviour and the standardization fallacy. *Nat Genet* 26:263.
- Würbel H. 2002. Behavioral phenotyping enhanced—beyond (environmental) standardization. *Genes Brain Behav* 1:3–8.
- Xu X, Domino EF. 1994. Genetic differences in the locomotor response to single and daily doses of phencyclidine in inbred mouse strains. *Behav Pharmacol* 5:623–629.
- Yerkes RM. 1907. *The dancing mouse*. New York: Macmillan.