Standardizing tests of mouse behavior: Reasons, recommendations, and reality

By: Douglas Wahlsten

Wahlsten, D. (2001). Standardizing tests of mouse behavior: reasons, recommendations and reality. *Physiology and Behavior*, 73: 695-704.

Made available courtesy of Elsevier: http://www.elsevier.com

*****Reprinted with permission. No further reproduction is authorized without written permission from Elsevier. This version of the document is not the version of record. Figures and/or pictures may be missing from this format of the document.*****

Abstract:

As more investigators with widely varying backgrounds enter the field of mouse behavioral genetics, there is a growing need to standardize some of the more popular tests because differences between laboratories in the details of behavioral testing and the pretesting environment can contribute to failures to replicate results of genetic experiments. It is argued here that we have sufficient knowledge to warrant a wise choice of a short list of standard strains and even details of apparatus and protocols for several kinds of behavioral tests. Equating the laboratory environment does not appear to be feasible. Instead, we need to learn what kinds of behavioral tests yield the most stable results in different labs and what kinds are most sensitive to the ubiquitous variations among test sites. Methods for making an informed choice of sample size for evaluating interactions between the laboratory environment and genotype are available and should be utilized in standardization trials. New resources for convenient sharing of data will greatly aid in collaborative and comparative studies involving several sites. Like the sequencing of an entire genome, test standardization is something that needs to be done only once if it is done properly, and the work will then benefit the field of behavioral and neural genetics for many years.

Keywords: Inbred strains; Knockouts; Behavioral genetics; Gene–environment interaction; Statistical power; Laboratory environment; Anxiety; Spatialmemory

Article:

1. Introduction

Desires to standardize tests of mouse behavior are inspired by burgeoning interest in mice as targets of molecular genetic studies and provoked by failures of several laboratories to replicate results of behavior–genetic experiments. Large numbers of scientists with little or no training in methods of behavioral testing hope to assess spatial memory, anxiety, and other constructs with apparatus and protocols that are often idiosyncratic to each laboratory. If genetic and environmental variation were truly additive in the algebraic sense, a freewheeling attitude towards details of testing would pose no problem because the patterns of results of genetic experiments would be essentially the same under most conditions, even though the overall mean scores might be higher in one lab than another because of environmental differences. Available evidence, however, indicates that genetic and environmental effects are often interactive [8,32,37,49,61], such that some genotypes respond more than others to specific features of the controlled environment in which they are assessed.

When interest is focused on a specific gene that is deliberately altered, it is crucial to know how general the results will be under conditions that differ from the original report. Four factors may substantially affect the results, two of these are themselves genetic [26]: (a) Flanking alleles carried by the strain that is the source of embryonic stem cells, usually a 129 strain, can complicate the interpretation of knockout effects, as discussed in detail elsewhere [16,25,64]. (b) Cells containing the targeted mutation are generally combined with nonmutant cells from another strain in a chimera and offspring are then backcrossed onto that strain background, and the apparent effects of the mutation may be altered by epistatic interaction with the genetic background [41]. The other two are environmental — the test situation itself and the pretesting laboratory environment. (c) The test situation consists of the physical test apparatus and stimuli impinging on the mouse during testing as well as the protocol of operations performed by the experimenter who administers the test. Many studies have found that seemingly minor changes in the test can have strain-dependent consequences [17,44,45,61]. (d) The laboratory environment comprises everything that impinges on the mouse before the start of testing, including conditions in the animal colony and methods of handling. Numerous features of the lab environment can have strain-specific effects, such as forceps handling [28], acidified water [42], postpartum pregnancy [58], or brief food deprivation [8]. Statistical interaction with genetic strain can be appreciable even when serious efforts are made to equate the lab environment [11].

Failures to replicate patterns of genetic results across laboratories can arise from any or all of these factors. In some instances the discrepancy may arise from one or two discrete factors [15], but it is sometimes very difficult to determine the reasons for disparate data. Consider recent attempts to map genes having relatively small influences (quantitative trait loci or QTLs) pertinent to locomotor activity. Flint et al. [21] examined an F_2 hybrid cross derived from C57BL/6J and BALB/cJ mice in a circular, white, open field measuring 60 cm in diameter, and they reported QTLs on chromosomes 1, 4, 12, and 15. Gershenfeld et al. [27] used the F_2 hybrid cross of C57BL/6J and A/J in a square, clear, open field measuring 42 × 42 cm, and they claimed QTLs on chromosomes 1, 10, and 19. The discordant findings might have arisen from different alleles of a progenitor strain, different apparatus configurations, or different lab environments. For unknown reasons, both studies failed to detect an effect of albinism on activity, contrary to many earlier studies [18,19,23,33], although some evidence suggests albinism effects are smaller on hybrid than inbred backgrounds [38]. Similar difficulties can arise in work with knockouts, as witnessed by three recent reports published back to back in Nature Genetics on conflicting results concerning effects of corticotropin-releasing hormone (Crh) and one of its receptors (Crhr2) on mouse anxiety [1,10,36].

Whereas the focus of much of the current research activity in this field is single gene effects, most of our knowledge about the role of the test situation and the laboratory environment in behavioral genetics comes from several decades of work with inbred strains. Inbred strains differ at many genetic loci, but their stable genetic compositions allow us to test hundreds or thousands of animals with the same genotype under various conditions and thereby ask whether genetic effects are dependent on the specific environment. The preferred design of this kind of experiment entails several inbred strains, with mice from each strain being subjected to several different environments. This kind of research could be done, at least in principle, by using several knockouts in the same experiment. Until genetically well-defined knockouts become more widely available in large numbers, however, the foundation of our field will continue to be built with inbred strains.

There is presently a complete absence of standard methods in mouse neurobehavioral genetics. Prospects and possibilities for adopting widely accepted standards are discussed in this paper, with emphasis on five aspects of the question — choosing standard strains, adopting common apparatus designs and test protocols, equating lab environments, employing adequate sample sizes, and sharing data in a common format. We need to have high quality standards available to serve as benchmarks against which a wide range of genetic and methodological variants can be compared in different laboratories. Having a good standard does not mean all researchers ought to employ these methods in every study. For those who do not wish to use the standard techniques routinely, they should be able to assess how similar the results with their own methods are to the benchmark. If seemingly minor alterations of a task in fact make little or no difference, this will be apparent by careful comparison with a standard method. Before such a comparison is possible, however, we must first have a good standard.

2. The population

First and foremost, standard methods entail collection of data on a standard set of strains, so that different research groups can be assured that they are working with the same genotypes. This stipulation does not imply that future research should be limited to a specific set of standard strains. On the contrary, genetic investigation benefits greatly from a rich diversity of genotypes. Nevertheless, use of standard strains in each laboratory is highly recommended at the outset of a series of studies in order to provide a reference point or baseline for comparison of other genotypes within and between laboratories. Suppose standard data are collected on strains A, B, C, and D at one site. Then a second lab finds that strain E yields what appears to be exceptionally high scores on similar tests. This could occur because of a unique set of alleles possessed by strain E or peculiarities of the test or the environment in the second lab. If the second lab also tested strains A, B, C, and D, however, interpretation of results would be greatly facilitated.

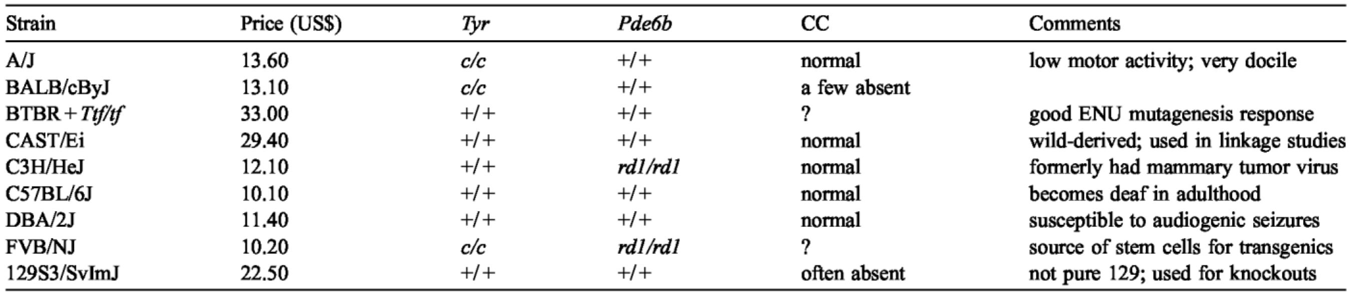
Mouse geneticists long ago recognized the importance of adopting standard strains and breeding schemes [50]. By far the best method for insuring replicability of genotypes across labs is brother–sister inbreeding for at least 60 generations to achieve genetic purity. The major question is how many and which of the numerous inbred strains should be adopted for collection of standard data. Because of the potential importance of strain-specific results, more than one strain should certainly be employed. At the same time, limitations of time and funds preclude the study of a dozen or more strains in many labs. Three criteria aid in making a short list of good candidates: The strains should represent a variety of different ancestries [2], they should be among those commonly used in this area of research, and they should be relatively free from severe neurological defects or bizarre behavioral characteristics. Common use is easily tabulated from the research literature and is also apparent in the price of the mice from commercial suppliers.

Three neurological defects warrant special concern. Retinal degeneration is a consequence of a recessive allele of the phosphodiesterase 6B, cGMP, rod receptor, beta poly-peptide gene (Pde6b; see Mouse Genome Database at www.informatics.jax.org). Although they retain sensitivity to light mediated by cones for some time after rods have succumbed [30,34] and may exhibit normal circadian rhythms in melatonin production [31], retinal degenerate strains usually perform poorly on tests involving visual cues. Such strains are therefore valuable for demonstrating the validity of tests of pattern vision, but they are not ideal for establishing standards for the normal range of inbred mouse behavior. Albinism (c allele of the tyr, tyrosinase gene) has widespread physiological effects [14], including reduced ipsilateral pathways in the visual system. Unfortunately, so many common strains are albino that we cannot avoid them. The strains BALB/c and 129 suffer absence of the corpus callosum (CC) with various degrees of penetrance. Whereas this is a dramatic defect anatomically, effects on many behaviors are often very subtle or non-existent [3,6,63] and motor coordination deficits appear primarily on the most difficult tasks [52]. Thus, it is reasonable to include strains having a moderate frequency of absent CC, especially because those with incomplete penetrance provide an internal check that can be assessed easily in future studies.

A group of mouse researchers met May 9–11, 1999, at the Mouse Strain Database Summit at the Jackson Laboratory and drafted a recommended short list of nine inbred strains, supplemented by more comprehensive lists of less commonly used strains. This work grew to become the Mouse Phenome Project [43] with a website for the Mouse Phenome Database (MPD) that lists recommended strains (www.jax.org/phenome). Pertinent characteristics of the 1999 short list of strains are given in Table 1. There is almost unanimous support for using the C57BL/6 and DBA/ 2 strains because they have been widely employed for many years, are inexpensive, readily available from many suppliers, and breed well. While several of the other strains may be suitable for investigation of cancer, cardiac physiology, or the immune system, they sometimes pose problems for work on behavior. For example, CAST/Ei mice are notoriously difficult to handle. They can be tested in many situations [40] but may become intractable in tests involving mild stress. They are therefore useful not as a standard but as an extreme genotype to be rated relative to the standard. BTBR + T*tf*/*tf* is currently difficult to obtain in large quantities and very few data are available on its brain or behavior. The best substrain is also subject to debate. BALB/cJ males tend to fight with little provocation, the strain is a poor breeder, and it has a high frequency of absent CC; whereas BALB/cByJ is less aggressive, breeds better, and has a low frequency of absent CC. Nevertheless, BALB/cJ is the most commonly used substrain. C3H/HeJ, besides being retinal degenerate, until recently carried mouse mammary tumor virus, whereas the C3HeB/FeJ strain is free from the virus. Within the past year, however, the Jackson Lab began shipping C3H/HeJ mice that lack the virus [7]. As for the best 129 substrain, these animals exist in a bewildering variety (see discussion at jaxmice.jax.org/html/nomenclature/nomen _ 129.shtml), certain of which (129/SvJ) have been genetically contaminated [53,57], and confusion persists about many 129 substrains. The nomenclature site lists Jackson Labs stock 002448 as 129S3/SvImJ, but the MPD site lists the same stock as 129S1/SvImJ.

All things considered, a defensible, very short list of strains for standardization of behavioral tests is A/J, BALB/cByJ, C57BL/6J, DBA/2J. Some labs may also want to include a 129 strain such as 129S3/SvImJ. Even this array will be inadequate for certain tests because of strain-specific peculiarities. For example, on the submerged platform water-maze, the A/J strain is an implacable wall hugger [11] and the BALB/c and 129 strains often float for extended periods [22,65]. When a given test yields meaningful data for only a minority of a very short list of standard strains, perhaps this casts aspersions more on the test than the mouse genes.

Table 1
Short list of recommended inbred strains



Strain	Price (US$)	*Tyr*	*Pde6b*	CC	Comments
A/J	13.60	*c/c*	+/+	normal	low motor activity; very docile
BALB/cByJ	13.10	*c/c*	+/+	a few absent	
BTBR + *Ttf/tf*	33.00	+/+	+/+	?	good ENU mutagenesis response
CAST/Ei	29.40	+/+	+/+	normal	wild-derived; used in linkage studies
C3H/HeJ	12.10	+/+	*rd1/rd1*	normal	formerly had mammary tumor virus
C57BL/6J	10.10	+/+	+/+	normal	becomes deaf in adulthood
DBA/2J	11.40	+/+	+/+	normal	susceptible to audiogenic seizures
FVB/NJ	10.20	*c/c*	*rd1/rd1*	?	source of stem cells for transgenics
129S3/SvImJ	22.50	+/+	+/+	often absent	not pure 129; used for knockouts

This list was proposed by the Strain Database Summit at the Jackson Labs in 1999 and later revised slightly [43]. A more recent revision (see www.jax.org/phenome) dropped BTBR from the A list and added SJL/J and SPRET/Ei. Prices are in US dollars at 6 weeks of age from The Jackson Laboratory (see jaxmice.jax.org). *Tyr* is the tyrosinase gene; *c/c* mice are albino. *Pde6b* is the gene for the retinal enzyme phosphodiesterase 6B, cGMP, rod receptor, beta polypeptide; the *rd1/rd1* genotype leads to degeneration of all rods in the retina and deficits in vision.

CC = corpus callosum.

Some of the difficulties involved in work with inbred strains may be overcome by working with F_1 hybrid mice, partly because many of the genes known to cause neurological defects are recessive. The B6D2F1 cross between C57BL/6 and DBA/2 is readily available from several suppliers, but it is not spared the hearing defects that afflict its parent strains.

All on the short list suggested here can be purchased from The Jackson Laboratory, which is a more convenient supplier for researchers in Canada and the United States than else-where in the world. Many colleagues in Europe prefer to work with mice such as the Orl substrains from the CNRS facility in Orléans, France. The SHIRPA protocol is being standardized on six strains (BALB/cOlaHsd, C3H/HeNHsd, C57BL/ 6JOlaHsd, CBA/CaOlaHsd, DBA/2OlaHsd, 129/SvHsd), four being Ola substrains maintained by Harlan-Olac UK [48], and SHIRPA is now being used to screen for defects in a variety of mutations [4,46]. Substrain differentiation [54,55] proceeds slowly but inexorably and will eventually lead to divergence of phenotypic means for genetic reasons. Although it is not advocated here that the global community of mouse researchers should always work with Jax mice, there is much to recommend a systematic comparison of the Jax substrains with the local substrain as part of a standardization study. By air, a trip to London, Frankfurt, or Paris differs little from a trip to Edmonton, Alberta, and those European climates are closer to Bar Harbor, ME, than is northern Canada. Airplanes carry freight in both directions, and researchers in North America could equally well import substrains from Europe. It would be worthwhile to conduct systematic comparisons among substrains maintained by suppliers such as Charles River Laboratories, Harlan Sprague Dawley, and Taconic. Presuming there are no marked differences between substrains in a preliminary study, our confidence in the genetic similarity of the animals for purposes of behavioral research will be enhanced. Extensive data on microsatellite alleles in various substrains would also aid interpretation of research in different countries.

Thus, while no standard set of inbred strains has yet been adopted by any consensus of researchers in this field, enough progress has been made to warrant a well-informed debate that could soon lead to, if not outright consensus, then at least a decisive majority vote.

3. The test situation: apparatus and protocols

Numerous tests of behaviors in many domains are readily available for work with mice [12], but a wide variety of lab-specific implementations prevails for most tests in common use. No recognized standard for any apparatus can be cited. Even for the simplest conceivable device, the featureless open field, apparatus polymorphism prevails. Some are square and others circular; some are clear and others opaque; some are brightly lit and others totally dark; some have tops and others are open; some are enclosed in sound-attenuating boxes and other sit on a table top in a room with technicians hustling back and forth and a radio blaring. At least eight commercial manufacturers of test apparatus sell open field activity devices (Table 2), yet no two are the same, despite the lack of patent protection for so simple an apparatus. Diversity also obtains for the Morris water-maze where swimming pools range from 35 to 200 cm in diameter, are white or black, contain clear water or water made opaque with one of many substances (several of which cause the fur to become wet quickly), use room temperature or heated water, and are placed in settings with cues of various sizes and distances from the tank in different labs. For the elevated plus maze, many parameters are known to influence the data [35,47]. Available information collected in several labs suggests that specific details of most kinds of apparatus are important [13], but there are virtually no thorough and systematic studies wherein several inbred strains were observed under a wide variety of conditions on a single kind of test.

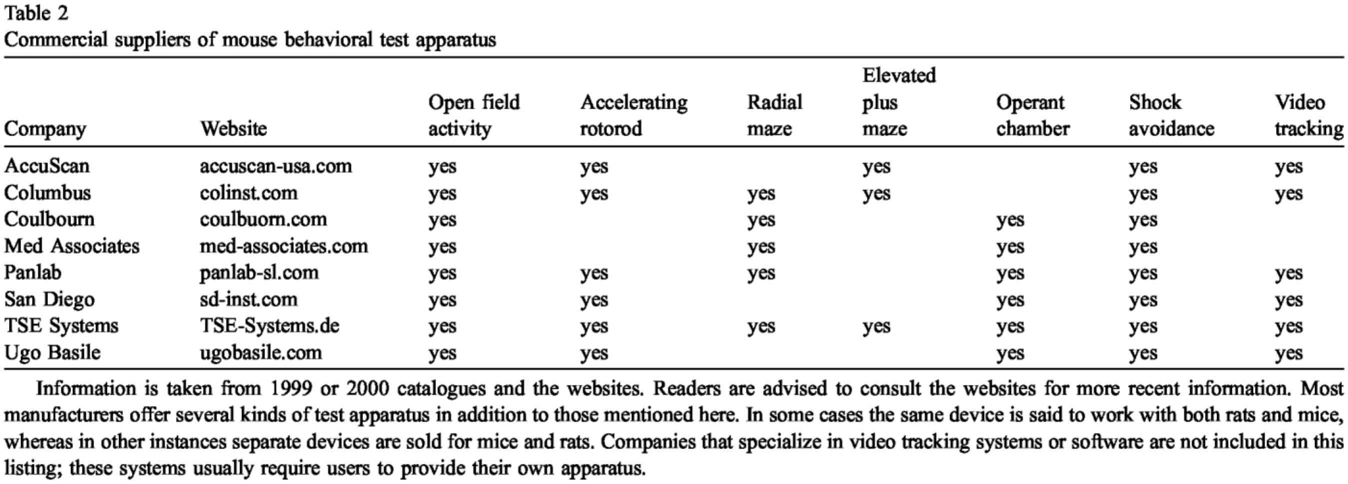


Table 2
Commercial suppliers of mouse behavioral test apparatus

Company	Website	Open field activity	Accelerating rotorod	Radial maze	Elevated plus maze	Operant chamber	Shock avoidance	Video tracking
AccuScan	accuscan-usa.com	yes	yes		yes		yes	yes
Columbus	colinst.com	yes	yes	yes	yes		yes	yes
Coulbourn	coulbuorn.com	yes		yes		yes	yes	
Med Associates	med-associates.com	yes		yes		yes	yes	
Panlab	panlab-sl.com	yes	yes	yes		yes	yes	yes
San Diego	sd-inst.com	yes	yes			yes	yes	yes
TSE Systems	TSE-Systems.de	yes	yes	yes	yes	yes	yes	yes
Ugo Basile	ugobasile.com	yes	yes			yes	yes	yes

Information is taken from 1999 or 2000 catalogues and the websites. Readers are advised to consult the websites for more recent information. Most manufacturers offer several kinds of test apparatus in addition to those mentioned here. In some cases the same device is said to work with both rats and mice, whereas in other instances separate devices are sold for mice and rats. Companies that specialize in video tracking systems or software are not included in this listing; these systems usually require users to provide their own apparatus.

Purchase of commercial apparatus is one means of achieving consistency, and this has appeal because there are some excellent implementations of the open field test, for example, that are already used in many labs. Beyond the open field, however, the selection of devices becomes problematic. Two difficulties are evident in this realm. First and most fundamental, apparatus manufacturers are typically brilliant in working with plastic, stainless steel, and transistors, while world-class expertise in testing mouse behavior resides mainly in public research institutions. Researchers should not look to manufacturers to set standards for apparatus or software for data acquisition and analysis. On the contrary, the research community should advise the manufacturers about the critical variables and optimal parameters for a particular test. Manufacturers could then compete to create the most efficient, convenient, reliable, and aesthetic implementation. Second, high quality commercial apparatus is often too expensive to be widely adopted as a standard in many labs. The cost factor itself inspires a variety of local shortcuts.

The process of apparatus development typically proceeds in a manner that spawns diversity. After deciding on the behavioral domain to be studied, researchers read the literature, attend conferences, and note the features of any apparatus that has given good results in the labs of respected colleagues. They may even visit another lab to learn crucial but unpublished details of a test. Then, a rough sketch is presented to the local shop, where technicians introduce their own innovations. The prototype is delivered to a graduate student who runs pilot tests with a few spare mice and asks for modifications to compensate for mice that prove to be too clever or stubborn to give valid data. Eventually, through an iterative process, a device is adopted, often under intense pressures of thesis or grant proposal deadlines, that yields results the lab director considers good enough for the present purpose. Once a large body of data is accumulated with the specific apparatus, the lab adheres to that configuration religiously until the retirement of the lab director in order to make later results comparable with their own basic findings. Close comparability with other labs is not guaranteed by this process, although similarity may increase over the years through a form of convergent evolution.

Specific details of the test protocol may also be of great importance, even when absolutely identical apparatus is adopted in two or more labs. Within each lab, we carefully implement step-by-step procedures that guarantee all animals in the study will be handled in the same manner by the experimenters, given the same time between trials, exposed to the same odor cues, and so forth. Detailed instructions are typed and distributed to those who do the testing, and the same instructions are used to train new people in the rituals of a specific lab year after year. Young scientists who establish their own labs often replicate the procedures they learned under the tutelage of eminent masters during graduate or postdoctoral study. Yet, when we compare protocols across labs, we often discover that different masters use different procedures. It is possible to adopt a common protocol in several labs, as was done in the study by Crabbe et al. [11], and this kind of exercise can really open our eyes to things we commonly do by habit without any compelling reason. Publication of procedures can provide helpful guide-lines, but we often find that descriptions of tests as presented in Current Protocols in Neuroscience, for example, are not sufficiently detailed to insure all labs do precisely the same things when performing a test.

If the research community acknowledges the need for some kind of standardization, how might this be achieved? In my opinion, the crucial starting point is agreement on the meaning of test standardization. This does not require that all labs be pressured to use identical apparatus and protocols in their research. A single standard would be resisted because of investment in the current apparatus as well as egotistical inertia, and it would also deprive the field of variety that can lead to important discoveries. What we need is a specific apparatus configuration with a rigorously implemented protocol that can serve as a convenient standard for comparison with the many extant variants and commercial devices. By assessing a small number of standard inbred strains with the standard test, a benchmark can be established. Then any lab can run a trial to compare data from their local device with results from the standard test. Likewise, commercial manufacturers can calibrate their apparatus against the standard and use the data in their advertisements.

It is not essential that a standard apparatus be truly optimal, although it should be free from major flaws. By comparing a fixed standard with many variants, we will gain valuable information about the size of deviations in numerical data that arise from variations in the test itself. Hopefully, many labs will find comfort in evidence that their own apparatus yields values reasonably close to the benchmark. Some labs may even find cause to proclaim their own test superior to the standard.

Optimization is an elusive quarry. There are simply too many variables in even a simple behavioral test to make feasible a comprehensive factorial study of, say, five inbred strains tested under many conditions. The work of Peeler [44] provides a case in point because it was one of the most comprehensive studies in this literature. Males of the BALB/cByJ and C57BL/6ByJ progenitors plus their seven recombinant inbred strains were tested at three times of day on 2 days using shuttle boxes with three kinds of barriers between the compartments; the study thus involved 9 × 3 × 3 = 81 independent groups of mice. Crabbe et al. [1 1] examined males and females of eight strains tested in three labs after shipping from a supplier or local breeding, for a total of 2 × 8 × 3 × 2 = 96 groups (each lab studied only 32 groups). Bulman-Fleming et al. [5] studied males and females of two inbred strains and two reciprocal F1 hybrids, each of which was derived from ovaries grafted into either an inbred or hybrid mother followed by surrogate fostering at birth to either an inbred or hybrid mother, for a total of 2 × 4 × 2 × 2 = 32 groups. Given my own experience with two of these studies and conversations with Peeler (personal communication), I conclude that studies with more than 100 independent groups exceed not only our budgets but also our abilities to counterbalance effectively and organize complex experiments.

For practical reasons, perhaps a better alternative is to establish a standard test and then examine the importance of parametric variations one or two variables at a time in a single study. This denies us data about optimal combinations of or complex interactions among various features of a test. While such information might be valuable, there is no reason to believe that conclusions about optimal combinations would be the same in different labs, as discussed in the next section.

Deciding on a reasonable standard may prove easier than we fear. When leading experts in a particular behavioral domain are assembled in one room, the accumulated experience of these few individuals can quickly lead to agreement on many items, leaving time for vigorous debate on contentious points, some of which may need to be resolved by experimentation. For example, a working group convened by the Office of Behavioral and Social Science Research at NIH in Bethesda, MD, on July 13–14, 1998, made considerable progress in the domains of activity, anxiety, and reflex development in a span of less than 2 days but then disbanded without reaching a workable solution.

Perhaps a standard test could be achieved through a six-step process. (a) Assemble a small group of respected researchers, each having long experience with a specific domain and real enthusiasm for the common goal of establishing a standard. International representation would make broad acceptance of conclusions more likely. (b) Prior to the first meeting, group members exchange reprints and discussion papers, and they draft a list of important variables to be standardized. (c) Convene the group for a week. Ideally, they should be sequestered until they deliver a consensus standard. Meeting near a lab where ideas can be checked with live mice might be productive. (d) Each group member then takes the draft standard back to his or her lab and tests a small number of mice of the standard strains to verify the quality of the data. It would not be necessary or even helpful at this point to expect that labs obtain the same results. The crucial thing is that each expert should be satisfied that the putative standard apparatus and protocol are free from any major flaws. Negotiations on improvements can then be conducted via e-mail and telephone. (e) Next, the draft standard is posted on several websites and comments are invited. After a few labs not involved in the drafting of the standard have tried the test, the group will either return to step (d) or move on to the full standardization trial. (f) Finally, sufficiently large samples of the standard strains need to be evaluated on the test in several labs and the data submitted for peer review and publication. The labs participating in this step need not be the ones that helped to design the standard. Hopefully the trial could be done in several countries, but this would not be absolutely essential for success of the enterprise.

A viable standard apparatus needs to be fairly simple to build and instructions must be easy to follow, so that a wide range of labs will be able to build and run the benchmark test. Alternatively, the apparatus might be built and sold at a modest price. Any one lab would need only one or two copies of the standard apparatus for comparison with the benchmark. This would be true for small-scale labs as well as large facilities that already possess dozens of a particular version of the test. One can even imagine apparatus in the form of a kit that could be taken apart, sterilized, and then shipped from lab to lab.

Probably the most difficult task would be standard scoring of the behavior stream. Automation is essential for large-scale research but is a barrier to standardization because of complexity and expense. Whatever method is deemed best, certainly the behaviors should be videotaped where feasible for repeated assessment and possible scoring at a single site. If the path of the animal can be reduced to a series of *x*- and *y*-coordinates with a time stamp, the data file can later be subjected to sophisticated track analysis using, for example, David Wolfer's WinTrack program (www.dpwolfer.ch/wintrack) or ethological analysis as advocated by Golani et al. [29] and Drai et al. [20].

4. The lab environment

Use of a standard apparatus and protocol can effectively equate the test situation across a wide range of sites, yet this will not guarantee the same phenotype means for the standard strains. As the recent study by Crabbe et al. [11] revealed, significant and substantial interactions between eight mouse strains and three labs occurred, especially for tests of anxiety, despite a rigorously equated test situation. Certain tests did give very similar results in all three labs, alcohol preference drinking for instance. Because the study was apparently the first of its kind and rather few behavioral domains were assessed, it is not possible to know why certain behaviors were so sensitive to the lab environment. Alcohol preference was the only behavior measured over a period of 24 h with a bare minimum of human handling, whereas the other five tests involved brief trials with handling before and after each trial. Thus, different handling or people unique to each lab could have been an important source of Strain × Lab interaction. Because only one test minimized handling, this point remains largely speculative.

The handling hypothesis could be assessed by monitoring home cage activity around the clock and, for half the mice, removing them periodically for an activity test trial in a novel environment — repeating the entire study in several labs. If researchers are serious about identifying specific features of the lab environment that are responsible for discrepant results, they must be prepared to run experiments simultaneously in several labs so that data are not contaminated by fluctuations arising from season of testing or cohort differences from commercial suppliers.

In the Crabbe et al. [11] study run simultaneously in Albany, Portland, and Edmonton, shipping, mating, and in most cases weaning were done on the same day in the three labs, and testing itself was started within the same hour. Cages were changed at the same time on the same days. The experiment demonstrated that simultaneity is possible, but the regimentation was also exceedingly difficult, and each lab subsequently vowed never to join such a tightly choreographed dance again. Furthermore, despite strenuous efforts to equate many other features of the animal husbandry, it simply was not possible to equate everything.

Whereas standardization of the genetic composition of the mice and the test situation is both possible and desirable, standardization of the lab environment is effectively impossible. The different physical layouts of most labs are quite literally set in stone. Especially in large research operations where many experiments are in progress at the same time, researchers will not be willing to change the housing and handling conditions for all mice in order to achieve the goal of one comparative study that will probably not be repeated. In many settings, the researcher also confronts the higher authority of animal care facility directors and institutional ethics committees that often insist things be done their way. Experts in rearing and testing mice sometimes must yield to the judgment of veterinarians trained in the methods of removing fur balls and kidney stones from spoiled house cats.

Although equating lab environments is not on the agenda, certain things can be done to help us better understand this source of variation. The usual description of the mouse's environment in a journal article is paltry at best. A superior accounting could be achieved by filling in details on a longer list of features of husbandry, then making this available via a website maintained either by the lab itself, the journal that publishes the article, or an organization of researchers. A detailed data form and database on lab environments has recently been established by Arndt and Surjo for the Mutant Mice Behaviour network (www.medi-zin.uni-koeln.de/mmb-network). Knowing some of the major environmental differences between labs may inspire plausible and testable hypotheses about origins of discrepant results. Controlled studies of the lab environment and comparisons across labs will be greatly facilitated if we can employ standard strains and standard test situations.

5. Statistical power and sample size

Possessing a list of salient features of the lab environment, it will then be worthwhile to assess formally the impact of variations in several of these. Special attention should be devoted to the magnitude of effects. We need to know more than mere *P* values for significance. Effect size, expressed as a fraction or number of standard deviations by which two groups differ, is a more informative index of the potency of an alteration in the environment. Small, moderate, and large effects correspond to effect sizes (*d* in samples) of about 0.5, 0.75, and 1.0 S.D., respectively [62]. The literature already provides vast information on effect sizes from a great number of targeted mutations and inbred strain differences. It would be very helpful to know the comparable effect sizes for alterations in the lab environment. A short list of variables is offered in Table 3.

The critical question for mouse behavior testing is not merely whether the lab environment has a statistically significant effect on behavior. The main effect in an analysis of variance (ANOVA) may be quite fascinating; for example, mice were generally more active and less anxious in Edmonton than in Albany and Portland [11]. Nevertheless, for genetic analysis, dramatic environmental effects pose no threat to the validity of our conclusions unless there is differential impact depending on genotype. We need to know whether the pattern of results of genetic experiments depends on the specific lab that does the experiment. This question is addressed by the interaction term in the ANOVA, not the main effect of environment. How to test for interaction is addressed in depth in most graduate-level statistics courses and texts. Procedures for doing an ANOVA are well implemented in many computer programs and have become routine in our field.

The greatest shortcoming is that the typical application of ANOVA to assess interaction tends to be quite insensitive to real interactions [59]. When the null hypothesis that effects are additive is evaluated, the probability of a Type II error (failure to reject a false null hypothesis) often is very high for the interaction effect. For many kinds of interaction, substantially larger sample sizes are required to confer adequate power on the test of interaction than are needed to detect the main effects in the same factorial experiment. Methods for meeting this challenge are presented elsewhere [9,59,60,62].

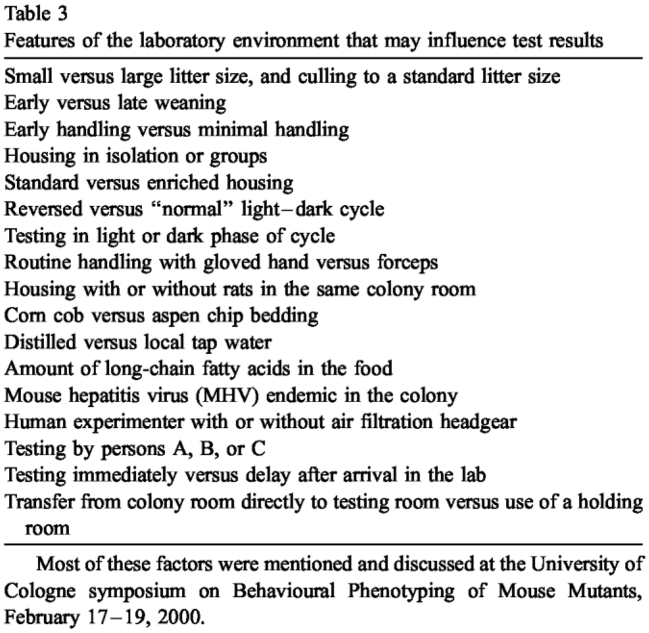


Table 3

Features of the laboratory environment that may influence test results

Small versus large litter size, and culling to a standard litter size
Early versus late weaning
Early handling versus minimal handling
Housing in isolation or groups
Standard versus enriched housing
Reversed versus "normal" light–dark cycle
Testing in light or dark phase of cycle
Routine handling with gloved hand versus forceps
Housing with or without rats in the same colony room
Corn cob versus aspen chip bedding
Distilled versus local tap water
Amount of long-chain fatty acids in the food
Mouse hepatitis virus (MHV) endemic in the colony
Human experimenter with or without air filtration headgear
Testing by persons A, B, or C
Testing immediately versus delay after arrival in the lab
Transfer from colony room directly to testing room versus use of a holding room

Most of these factors were mentioned and discussed at the University of Cologne symposium on Behavioural Phenotyping of Mouse Mutants, February 17–19, 2000.

Whereas criteria for large and small effects (d values) of a single treatment are widely accepted, criteria for a substantial, noteworthy interaction are generally lacking in our field. The following proposal provides a guideline for sample sizes needed to detect interaction in a 2 × 2 factorial design. Suppose a knockout (— / —) and its control (+/+) are tested in two labs (A, B). Table 4 shows hypothetical group means in the left panel that express purely additive effects of genotype and lab environment. The right panel shows a kind of interaction that I believe our ANOVA should be able to detect. Specifically, the knockout effect is *twice as large* in Lab B than in Lab A. Of course, there might be no knockout effect in one lab and a large effect in the other, but in such a case the interaction would be obvious. We need to devote careful attention to the more common situation where the interaction portrayed in a graph looks substantial to the educated eye of the scientist but less so in the ANOVA table. The example of Gene × Environ-Environment interaction in Table 4 involves a moderate knockout effect size of 0.75 S.D. in Lab A but a large effect size of 1.5 S.D. in Lab B. Control (+/+) values in Labs A and B also differ by 0.75 S.D. In this specific situation, the method of Wahlsten [60] reveals that about 7 mice per group will be needed to detect the genetic main effect or the lab main effect with 95% power (5% chance of Type II error), whereas 46 mice in each of the four groups will be needed to detect the gene–environment interaction with the same level of power. Generally speaking, *when the mutation has twice the magnitude of effect in one lab versus another, at least six times as many mice are needed to insure a significant interaction term in the ANOVA than are needed simply to detect the main effects*.

For large factorial designs involving many mouse strains and several environments, computations are somewhat more elaborate [9], but the need for larger samples to detect the interaction effect persists [59]. In the case of interaction terms with more than one degree of freedom, so many patterns of results are possible that a universal criterion for a noteworthy interaction is difficult to imagine. What kind of interaction is worthy of the expense to find it probably depends very much on the topic being studied and the importance of interaction for theories in the field.

The sample size issue is decisive in the context of test standardization. We want to know if results are closely comparable across several labs or instead are lab-dependent. It would be foolish to invest considerable money and effort in perfecting standard tests but then compare labs using sample sizes that are insufficient to detect the very phenomenon we set out to explore.

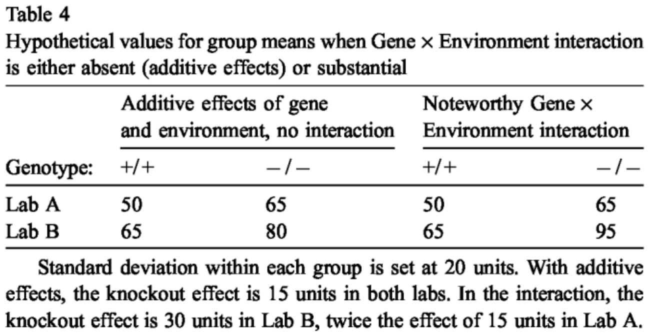


Table 4
Hypothetical values for group means when Gene × Environment interaction is either absent (additive effects) or substantial

	Additive effects of gene and environment, no interaction		Noteworthy Gene × Environment interaction	
Genotype:	+/+	−/−	+/+	−/−
Lab A	50	65	50	65
Lab B	65	80	65	95

Standard deviation within each group is set at 20 units. With additive effects, the knockout effect is 15 units in both labs. In the interaction, the knockout effect is 30 units in Lab B, twice the effect of 15 units in Lab A.

6. Sharing raw data

Before the advent of the Internet and the World Wide Web, the endpoint of most research was publication of a concise journal article, followed by discrete burial of the raw data in an office file cabinet or a landfill. The curiosity of consumers of the results was strictly limited by the kind of data analysis chosen by the authors.

Now it is possible to post the raw data on a website for downloading and reanalysis by colleagues far from the home lab, as was done by Crabbe et al. [11] at www.alba-ny.edu/psy/obssr. We provided files that could be read directly into the SPSS or SYSTAT statistical packages. Posting raw data in ASCII (*.txt or *.dat) format that can be imported into almost any spreadsheet or statistics pro-gram would probably be a better option. Files in the Excel (*.xls) format are used by many researchers, but there are reservations about proclaiming the product of a specific corporation to be the world wide standard.

Standard testing of standard strains in several labs calls for publication of data in a standard format that can quickly be assembled into a large file for analysis. One can envision other valuable applications of a standard format. For example, a journal could require that raw data be provided to reviewers prior to publication and then be posted on the journal's website after the study has passed peer review. Granting agencies could achieve greater return for their investment of public funds if they required labs to archive their data in a central facility that is accessible by qualified researchers. In this respect, mouse geneticists are not restricted by policies of privacy or informed consent of their subjects. Organizations of mouse researchers have already established voluntary public archives for data collected in many countries, including the MPD (www.jax.org/ phenome), the Mutant Mice Behaviour network (www.me-dizin.uni-koeln.de/mmb-network), and the MyMouse consortium (www.mymouse.org). The formats and contents of these sites differ considerably, and it remains to be seen whether the entire field will eventually adopt a common standard for reporting data.

7. Conclusions

The need for some kinds of standardization is apparent from the broad interest in the topic at the present time and attendance at meetings organized to discuss the matter. As data from large-scale mutagenesis screening studies of behavioral phenotypes begin to appear [5 1], more questions are being asked about the most appropriate kinds of tests [39] and the most appropriate baseline criteria for judging abnormal behavior [56].

Despite lengthy deliberations, firm steps towards standardization have been few and poorly coordinated, and some investigators continue to have reservations about the wisdom of standardization. For instance, Würbel [66] claims that "If standardization were fully effective, inter-individual variation within study populations would decrease to zero ... " and each experiment would amount to a single-case study. In my opinion, extreme homogeneity will not result from standardization, because we already know that individual anatomical, physiological, and behavioral variation usually exists within a highly inbred strain in the confines of a single lab, even when strenuous efforts are made to achieve a uniform environment [24]. Standardization is likely to lead to more reproducible or interpretable results of complex experiments done in different laboratories, despite the presence of ineradicable within-strain variance.

It is argued here that we have sufficient knowledge to warrant a wise choice of a short list of standard strains and even details of apparatus and protocols for several kinds of behavioral tests. Equating the laboratory environment does not appear to be feasible. Instead, we need to learn what kinds of behavioral tests yield the most stable results in different labs and what kinds are most sensitive to the ubiquitous variations among test sites. Methods for making an informed choice of sample size for evaluating laboratory-specific genotype effects are available and should be utilized in standardization trials. Resources for convenient sharing of data that are now being constructed will greatly aid in collaborative and comparative studies involving several sites. Given that research funding agencies, pharmaceutical and biotechnology companies, commercial breeders, and apparatus manufacturers would all benefit from successful standardization and large-scale phenotyping exercises, it should be possible to solicit support for a major initiative in this direction. Like the sequencing of an entire genome, test standardization is something that needs to be done only once if it is done properly, and the benefits of the work will be felt widely in the field of behavioral and neural genetics for many years.

References

[1] Bale TL, Contarino A, Smith GW, Chan R, Gold LH, Sawchenko PE, Koob GF, Vale WW, Lee K-F. Mice deficient for corticotropin-releasing
hormone receptor-2 display anxiety-like behaviour and are hypersensitive to stress. Nat Genet 2000;24:410–4.

[2] Beck JA, Lloyd S, Hafezparast J, Lennon-Pierce M, Eppig JT, Festing MFW, Fisher EMC. Genealogies of mouse inbred strains. Nat Genet 2000;24:23–5.

[3] Bishop K, Kruyer A, Wahlsten D. Agenesis of the corpus callosum and voluntary wheel running in mice. Psychobiology 1996;24: 187–94.

[4] Brown SD, Nolan PM. Mouse mutagenesis — systematic studies of mammalian gene function. Hum Mol Genet 1998;7:1627–33.

[5] Bulman-Fleming B, Wahlsten D, Lassalle JM. Hybrid vigour and maternal environment in mice: I. Body and brain growth. Behav Proc 1991;23:21–33.

[6] Bulman-Fleming B, Wainwright PE, Collins RL. The effects of early experience on callosal development and functional lateralization in pigmented BALB/c mice. Behav Brain Res 1992;50:31–42.

[7] C3H strain free of exogenous MMTV. Jax Notes No. 480:8, 2000.

[8] Cabib S, Orsini C, Le Moal M, Piazza PV. Abolition and reversal of strain differences in behavioral responses to drugs of abuse after a brief experience. Science 2000;289:463–5.

[9] Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale (NJ): Erlbaum, 1988.

[10] 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 JJ, Rittenberg MB, Stenzel P, Stenzel-Poore MP. Abnormal adaptations to stress and impaired cardiovascular function in mice lacking corticotropin-releasing hormone receptor-2. Nat Genet 2000;24:403 – 9.

[11] Crabbe JC, Wahlsten D, Dudek BC. Genetics of mouse behavior: interactions with laboratory environment. Science 1999;284:1670–2.

[12] Crawley JN. What's wrong with my mouse? Behavioral phenotyping of transgenic and knockout mice. New York: Wiley-Liss, 2000.

[13] Crawley JN, Belknap JK, Collins A, Crabbe JC, Frankel W, Henderson N, Hitzemann RJ, Maxson SC, Miner LL, Silva AJ, Wehner JM, Wynshaw-Boris A, Paylor, R. Behavioral phenotypes of inbred mouse strains: implications and recommendations for molecular studies. Psychopharmacology 1997;132:107–24.

[14] Creel D. Inappropriate use of albino animals as models in research. Pharmacol, Biochem Behav 1980;12:969–77.

[15] Crestiani F, Martin JR, Möhler H, Rudolph U. Resolving differences in GABAA receptor mutant mouse studies. Nat Neurosci 2000;3:1059.

[16] Crusio WE. Gene-targeting studies: new methods, old problems. Trends Neurosci 1996;19:186–7.

[17] Crusio WE, Schwegler H, Brust I. Covariations between hippocampal mossy fibres and working and reference memory in spatial and non-spatial radial maze tasks in mice. Eur J Neurosci 1993;5:1413–20.

[18] DeFries JC, Hegmann JP, Weir MW. Open-field behavior in mice: evidence for a major gene effect mediated by the visual system. Science 1966;154:1577–9.
[19] DeFries JC, Hegmann JP. Genetic analysis of open-field behavior. In: Lindzey G, Thiessen DD, editors. Contributions to behavior– genetic analysis. The mouse as a prototype. New York: Appleton-Century-Crofts, 1970. pp. 23–56.
[20] Drai D, Elmer G, Benjamini Y, Kafkafi N, Golani I. Phenotyping of mouse exploratory behavior. Paper presented at the symposium on Behavioural Phenotyping of Mouse Mutants, University of Cologne, February 19, 2000.
[21] Flint J, Corley R, DeFries JC, Fulker DW, Gray JA, Miller S, Collins AC. A simple genetic basis for a complex psychological trait in laboratory mice. Science 1995;269:1432–5.
[22] Francis DD, Zaharia MD, Shanks N, Anisman H. Stress-induced disturbances in Morris water-maze performance: interstrain variability. Physiol Behav 1995;58:57–65.
[23] Fuller JL. Effects of the albino gene upon behaviour of mice. Anim Behav 1967;15:467–70.
[24] Gärtner K. A third component causing random variability beside environment and genotype. A reason for the limited success of a
30 year long effort to standardize laboratory animals? Lab Anim 1990;24:71–7.
[25] Gerlai R. Gene-targeting studies of mammalian behavior: is it the mutation or the background genotype? Trends Neurosci 1996;19: 177–81.
[26] Gerlai R. Targeting genes associated with mammalian behavior: past mistakes and future solutions. In: Crusio WE, Gerlai R, editors. Hand-book of molecular– genetic techniques for brain and behavior re-search. Amsterdam: Elsevier, 1999. pp. 364–75.
[27] Gershenfeld HK, Neumann PE, Mathis C, Crawley JN, Li X, Paul SM. Mapping quantitative trait loci for open-field behavior in mice. Behav Genet 1997;27:201 – 10.
[28] Ginsburg BE. Genetic parameters in behavioral research. In: Hirsch J, editor. Behavior– genetic analysis. New York: McGraw-Hill, 1967. pp. 135–53.
[29] Golani I, Benjamini Y, Eilam D. Stopping behavior: constraints on exploration in rats (Rattus norvegicus). Behav Brain Res 1993;53: 21–33.
[30] Goto M, Eibhara S. The influence of different light intensities on pineal melatonin content in the retinal degenerate C3H mouse and the normal CBA mouse. Neurosci Lett 1990;108:267–72.
[31] Goto M, Oshima I, Tomita T, Ebihara S. Melatonin content of the pineal gland in different mouse strains. J Pineal Res 1989;7:195–204.
[32] Gottlieb G. Normally occurring environmental and behavioral influences on gene activity: from central dogma to probabilistic epigenesis. Psychol Rev 1998;105:792–802.
[33] Henry KR, Schlesinger K. Effects of the albino and dilute loci on mouse behavior. J Comp Physiol Psychol 1967;63:320–3.
[34] Hicks D, Sahel J. The implications of rod-dependent cone survival for basic and clinical research. Invest Ophthalmol Vis Sci 1999;40: 3071–4.
[35] Hogg S. A review of the validity and variability of the elevated plus-maze as an animal model of anxiety. Pharmacol, Biochem Behav 1996;54:21–30.
[36] Kishimoto T, Radulovic J, Radulovic M, Lin CR, Schrick C, Hoosh-mand F, Hermanson O, Rosenfeld MG, Spiess J. Deletion of Crhr2 reveals an anxiolytic role for corticotropin-releasing hormone recep-tor-2. Nat Genet 2000;24:415 – 9.
[37] Lassalle J-M. Les interactions entre génotype et environement. Psychol Fr 1986;31:205–11.
[38] Lassalle J-M, Le Pape G. Differential effects of the albino gene on behavior according to task, level of inbreeding, and genetic back-ground. J Comp Physiol Psychol 1981;95:655–62.
[39] Lederhendler I, Schulkin J. Behavioral neuroscience: challenges for the era of molecular biology. Trends Neurosci 2000;23:451 – 4.
[40] LeRoy I, Roubertoux PL, Jamot L, Maarouf F, Tordjman S, Mortaud S, Blanchard C, Martin B, Guillot PV, Duquenne V. Neuronal and behavioral differences between Mus musculus domesticus (C57BL/ 6JBy) and Mus musculus castaneus (CAST/Ei). Behav Brain Res 1998;95:135–42.

[41] Magara F, Mu¨ller U, Lipp H-P, Weissmann C, Staliar M, Wolfer DP. Genetic background changes the pattern of forebrain commissure defects in transgenic mice underexpressing the P-amyloid-precursor pro-tein. Proc Natl Acad Sci USA 1999;96:4656–61.
[42] Maxson SC. Methodological issues in genetic analysis of an agonistic behavior (offense) in male mice. In: Goldowitz D, Wahlsten D, Wimer RE, editors. Techniques for the genetic analysis of brain and behavior: focus on the mouse. Amsterdam: Elsevier, 1992. pp. 349–73.
[43] Paigen K, Eppig JT. A Mouse Phenome Project. Mamm Genome 2000;11:715–7.
[44] Peeler DF. Shuttlebox performance in BALB/cByJ, C57BL/6ByJ, and CXB recombinant inbred mice: environmental and genetic determinants and constraints. Psychobiology 1995;23:161–70.
[45] Poderycki MJ, Simoes JM, Todorova MT, Neumann PE, SeyfriedTN. Environmental influences on epilepsy gene mapping in EL mice. J Neurogenet 1998;12:67–86.
[46] Rafael JA, Nitta Y, Peters J, Davies KE. Testing of SHIRPA, a mouse phenotypic assessment protocol, on Dmdmdx and Dmdmdx3cv dystro-phin-deficient mice. Mamm Genome 2000; 11:725 – 8.
[47] Rodgers RJ, Dalvi A. Anxiety, defence and the elevated plus-maze. Neurosci Biobehav Rev 1997;21:801 – 10.
[48] Rogers DC, Jones DNC, Nelson PR, Jones CM, Quilter CA, Robinson TL, Hagan JJ. Use of SHIRPA and discriminant analysis to characterise marked differences in the behavioural phenotype of six inbred mouse strains. Behav Brain Res 1999;105:207–17.
[49] Roubertoux PL, Nosten-Brtrand M, Carlier M. Additive and interactive effects between genotype and maternal environments: concepts and facts. Adv Study Behav 1990;19:205–47.
[50] Russell ES. A history of mouse genetics. Annu Rev Genet 1985;19: 1–28.
[51] Sayah DM, Khan AH, Gasperoni TL, Smith DJ. A genetic screen for novel behavioral mutations in mice. Mol Psychiatry 2000;5:369–77.
[52] Schalomon PM, Wahlsten D. Wheel running behavior is impaired by both surgical section and genetic absence of the mouse corpus callosum. Brain Res Bull [in press].
[53] Simpson EM, Linder CC, Sargent EE, Davisson MT, Mobraaten LE, Sharp JJ. Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. Nat Genet 1997;16:19–27.
[54] Stiedl O, Radulovic J, Lohmann R, Birkenfeld K, Palve M, Kammermeier J, Sananbenesi F, Spiess J. Strain and substrain differences in context- and tone-dependent fear conditioning of inbred mice. Behav Brain Res 1999;104:1–12.
[55] Swerdlow NR, Martinez ZA, Hanlon FM, Platten A, Farld M, Auerbach P, Braff DL, Geyer MA. Toward understanding the biology of a complex phenotype: rat strain and substrain differences in the sensorimotor gating-disruptive effects of dopamine agonists. J Neurosci 2000;20:4325–36.
[56] Tarantino LM, Gould TJ, Druhan JP, Bucan M. Behavior and muta-genesis screens: the importance of baseline analysis of inbred strains. Mamm Genome 2000;11:555–64.
[57] Threadgill DW, Yee D, Matin A, Nadeau JH, Magnuson T. Genealogy of the 129 inbred strains: 129/SvJ is a contaminated inbred strain. Mamm Genome 1997;8:441–2.
[58] Wahlsten D. Mice in utero while their mother is lactating suffer higher frequency of deficient corpus callosum. Dev Brain Res 1982;5: 354–7.
[59] Wahlsten D. Insensitivity of the analysis of variance to heredity– environment interaction. Behav Brain Sci 1990;13:109–61.
[60] Wahlsten D. Sample size to detect a planned contrast and a one degree-of-freedom interaction effect. Psychol Bull 1991;110:587–95.
[61] Wahlsten D. Single-gene influences on brain and behavior. Annu Rev Psychol 1999;50:599–624.
[62] Wahlsten D. Experimental design and statistical inference. In: Crusio WE, Gerlai RT, editors. Handbook of molecular–genetic techniques for brain and behavior research. Amsterdam: Elsevier, 1999. pp. 41–57.
[63] Wahlsten D, Crabbe JC, Dudek BC. Behavioural testing of standard inbred and 5HT1B knockout mice: implications of absent corpus callosum. Behav Brain Res [in press].
[64] Wolfer DP, Lipp H-P. Simple experimental solutions to the genetic background and flanking gene problems. Paper presented at the symposium on Behavioural Phenotyping of Mouse Mutants, University of Cologne, February 19, 2000.

[65] Wolfer DP, Stagljar-Bozicevic M, Errington M, Lipp H-P. Spatial memory and learning in transgenic mice: fact or artifact. News Physiol Sci 1998;13:118–22.
[66] Würbel H. Genetics of behaviour: the standardization fallacy. Nat Genet 2000;26:263.