Behavioural testing of standard inbred and $5HT_{1B}$ knockout mice: implications of absent corpus callosum

By: Douglas Wahlsten, John C. Crabbe , Bruce C. Dudek

Wahlsten, D., Crabbe, J. C., and Dudek, B. C. (2001). Behavioural testing of standard inbred and $5HT_{1B}$ knockout mice: implications of absent corpus callosum. *Behavioural Brain Research*, 125:23-32.

Abstract:

Rapid advances in biotechnology have created new demands for tests of mouse behaviour having both high reliability and high throughput for mass screening. This paper discusses several statistical and psychological factors pertinent to replication of results in different laboratories, and it considers the question of which inbred strains are best for test standardization. In this context, the problem of absent corpus callosum in the 129 strains is addressed with data from a recent study of six diverse tests of behaviour, and it is shown that effects of absent corpus callosum are usually nonsignificant and/or very small. Whether any 129 substrain is to be included in the list of standard strains depends on the goal of the standardization — collecting diverse phenotypic data on most available strains by a few expert investigators (the gold standard) or refining behavioural tests in order to establish a normal range of behaviour that can be used to judge a wider range of strains or even an individual mouse.

Keywords: Inbred strains; Anxiety; Locomotor activity; Gene–environment interaction; Hippocampal commissure; Test standardization; Serotonin receptor knockout

Article:

1. Introduction

Current interest in the question of standardizing tests of mouse behaviour is inspired by a number of recent developments. Extensive reviews of numerous tests of mouse behavior [9,10] have helped to make many of these techniques accessible to a wide variety of scientists, while conferences such as the meeting on ‘Behavioural Phenotyping of Mouse Mutants’ held at the University of Köln in February of 2000 have also drawn attention to these issues. Funding from the National Institutes of Health in the USA directed specifically to testing behaviours of inbred mouse strains and mutants has encouraged many investigators to undertake research in this area.

Rapid advances in biotechnology are exerting pressure on neurobehavioural genetics to provide better tools for assessing behaviour [36,38]. Lederhendler and Schulkin [20] recently noted: ‘As new genes are identified, scientific interest turns to their functions. To determine the function of genes, particularly how they contribute to behavior, molecular biologists need, and have been vigorously requesting, reliable phenotypes’. Large scale genetic screening programs to detect effects of induced mutations on behaviour are beginning to yield reams of data [4,27]. At the same time, there is discussion as to whether the screening programs might benefit from more extensive testing with standard in-bred strains [32] or a richer array of behavioural tests [20].

Two features of a useful behavioural test are paramount when our colleagues ask us for better measuring instruments. First, they want more efficient and sensitive tests suitable for high-throughput testing of hundreds or thousands of animals. Second, they want tests that are likely to yield similar results in the hands of different investigators. Both of these features may be enhanced by appropriate standardization of the tests. Each kind of test entails numerous features such as apparatus dimensions and materials, lighting conditions, trial duration, handling by the experimenter, and so forth. An informed choice of parameters ought to yield a more reliable and therefore more sensitive test. Use of a standard set of parameters in different labs should also enhance the repeatability of test results.

Whenever one seeks to standardize any test, a crucial factor is the choice of genetic strains on which to base the standard. It is not necessarily true that more strains in the standard sample result in better test standardization, because a more diverse sample is more likely to include strains with major neurological abnormalities. This issue, to be discussed in detail in this paper with regard to absence of the corpus callosum, brings into focus two goals of standardization. One might want to establish norms for behavioural variation, such that a strain or an individual animal later found outside the normal range would be considered aberrant, as is done in mutagenesis screening. Alternatively, one might want to establish benchmark data for a wide variety of strains in order to allow many other laboratories to evaluate their environments and procedures against genetically well defined standards. Both approaches would benefit from research explicitly designed to identify behavioural tests of high reliability, but they may differ in the attitude towards mouse strains that pre-form very poorly. If a strain afflicted by absent corpus callosum does poorly, we would like to know whether this indicates something fundamentally wrong with the mouse or instead reveals a flaw in the test itself that may be overcome by a better apparatus design or test procedure.

Some colleagues have expressed doubts about the wisdom of standardizing tests. Würbel [42], for example, disputes a ‘standardization fallacy’, the mistaken belief that standardization will eliminate individual variability in behaviour. There is no need to fret about this possibility, however. Strenuous attempts to render the laboratory environment perfectly uniform in order to expunge individual differences among genetically identical, inbred mice, have met with failure even for anatomical and physiological phenotypes [13]. The vast literature on mouse behaviour genetics accumulated over the past 40 years provides numerous examples of substantial within-strain variance in the confines of a single laboratory. If uniformity cannot be achieved within a single lab, there is no way that any kind of standardization will yield uniformity of behaviour across the entire field.

2. Replicability of test results — statistical aspects

Perhaps the strongest impetus towards standardization is the vexatious occurrence of divergent results when different laboratories work with the same or very similar genetic material. For example, three recent papers in *Nature Genetics* reported conflicting results concerning effects of corticotropin-releasing hormone (*Crh*) and one of its receptors (*Crhr2*) on mouse anxiety [1,7,18]. There are two potential sources of failures to replicate results across labs — statistical sampling error and systematic differences in the conditions of experimentation.

Sampling error arises from differences among individuals in a group. When these differences are substantial, as almost always happens, then the specific set of individuals chosen for testing may have a sample mean score that deviates considerably from the population mean, and two samples drawn from the same population will often have markedly different means. Usually we seek to minimize this problem by studying reason-ably large samples and conducting statistical tests to persuade ourselves and our colleagues that the observed effects are real. A typical procedure is to compare controls and mutants with a test of ‘significance’ at the $\alpha = 0.05$ level. If the apparent probability of obtaining a particular magnitude of difference between group means is less than 0.05, then the genetic effect is said to be statistically significant. What this jargon means is that the probability of such a large difference occurring because of sampling error is so low that we infer the difference must *not* have arisen merely from chance. This procedure is a null hypothesis test, where the null is that there is no genuine genetic effect. When the null is true, a Type I error occurs if we reject the null hypothesis. The experimenter chooses a criterion for Type I error, symbolized a, that makes such an error relatively rare.

Although great attention is often devoted to the appropriate level of Type I error in genetic research [19], this issue has little or nothing to do with test replicability. For one thing, the level of Type I error is insensitive to sample size; if there is truly no genetic effect, the probability of committing a Type I error is the same for paltry and huge samples. For another, replicability is a concern when there is a real effect to be detected, in which case the null hypothesis is a straw man. If there is a real effect, then the more plausible kind of error is Type II error, the failure to detect a real effect, which occurs with probability β. The level of Type II error is highly sensitive to sample size, being much lower for larger samples.

For a single study in one lab, the probability of correctly rejecting the null hypothesis of no genetic effect is $1 - \beta$, also known as the statistical power of the test. It is desirable that power be at least 80%, and a good argument can be made for 95% because this will make $\beta = \alpha = 0.05$. The probability that two independent labs replicate or obtain evidence of the same genetic effect is simply $(1 - \beta)^2$, and when each lab uses $1 - \beta = 0.8$, they should replicate a real effect on 64% of their tests. Only when power is 95% or more will probability of replication be 90% or more.

It is essential to address the question of statistical power during the design phase of the study, *before* the data are collected [37]. There are two crucial determinants of power — the true effect size and sample size. Effect size when comparing two groups can be ex-pressed as the coefficient δ, which indicates the number of standard deviations by which true group means differ [37]. Often it is possible to make a plausible estimate of the likely effect size by examining the sample means and standard deviations from a previous experiment. As shown in Fig. 1, the dispersion of sample effect sizes will tend to be quite wide unless sample sizes are large. For a small effect of $\delta = 0.5$, typical of quantitative trait locus (QTL) effects, there should be more than 30 animals per group to ensure that the sample effect size exceeds 0 and considerably more to ensure that the group difference is statistically significant. For a large effect of $\delta = 1.0$ standard deviation, on the other hand, samples from 20 to 30 mice per group should usually yield significant group differences. For even larger effects of $\delta > 2.0$ standard deviations, typical of those sought in mutagenesis screening studies [27], probability of replication should generally be very high for statistical reasons.

If more than one lab plans to examine the same genetic effect, it may be wise for them to use a more stringent criterion for Type I (α) error, because a false positive finding is more likely when many research groups test the same null hypothesis [37]. Using a smaller value of α will reduce the statistical power, however. The same difficulty confronts researchers when they conduct numerous null hypothesis tests within a single study, as we did in our comparison of the three labs [8]. The crucial consideration here is whether the null hypothesis is credible in the first place. Later in this paper, we compare brain sizes for mouse strains already known to differ substantially in brain size [25], and there is no good reason to deflate our α level when we are merely verifying established knowledge.

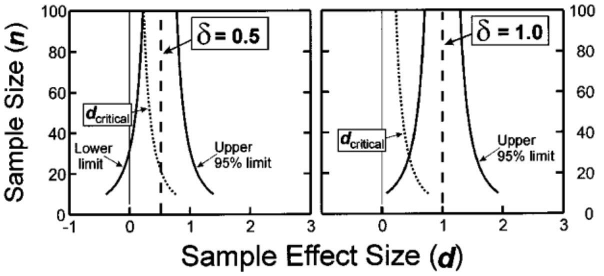

Fig. 1. Sample effect sizes (d) in relation to sample size (n) for two hypothetical values of true effect size (δ). The solid lines represent the lower and upper limits of d, such that about 95% of sample values will occur between those limits. The dotted line indicates the critical value of d required to reject the null hypothesis of no effect ($\delta = 0$) when a one-tailed t-test is used [37]. For a small genetic effect of $\delta = 0.5$ standard deviation, the chances that another lab will obtain a sample $d > 0$ is quite high when $n > 20$, but considerably larger samples are needed to ensure the effect will be judged statistically significant. For a substantial genetic effect with $\delta = 1.0$, chances that another lab will also find a significant effect are quite high, provided $n > 20$.

3. Replicability and the three components of standardization

If genetic and environmental effects on behaviour were perfectly additive, the same pattern of results of a genetic experiment should occur, even though environ-mental conditions in one lab lead to generally higher scores than in another lab. Thanks to the ubiquitous nature of gene –environment interaction, however, two labs may both employ large samples and reliable tests, yet fail to obtain the same results of a genetic experiment because the manifestation of genetic differences depends on the specific environment. Interaction is commonly observed in studies designed to detect it [6], and interaction is to be expected, given current knowledge about how genes function as parts of integrated systems [14,31]. In the presence of numerous factors leading to interaction, the only way to ensure similar outcomes of genetic experiments in different laboratories is to employ very similar, standardized conditions. If this goal is to be pursued seriously, three components of the process need to be standardized: the test situation itself, the laboratory environment prior to testing, and the genetic composition of the animal population.

The test situation includes the physical conformation of the apparatus and stimulus parameters and well as the test protocol that specifies in great detail how the animals are to be tested. The laboratory environment includes all features of the animal’s surroundings that impinge on it from conception until just prior to testing. It is known that many aspects of a simple test such as the elevated plus maze can influence results [17,26], and abundant data also demonstrate the important of diverse aspects of the lab environment [2,12,43]. In almost all instances when conditions are compared across laboratories, both the test situations and the lab environments are found to vary along many dimensions, and it is therefore impossible to know why two labs may fail to obtain the same results.

The study of Crabbe, Wahlsten and Dudek [8] addressed this problem by rigorously standardizing the details of the test apparatus and testing protocols for six commonly used behavioural tests and then conducting the tests simultaneously in three different laboratories using identical genetic groups. Although results were very similar for certain tests such as ethanol preference in a two-bottle test, large and significant strain by lab interactions were observed for several tests, including plus maze and cocaine activation. The interactions must have arisen from differences in the laboratory environments. Many aspects of the lab environment, such as the light–dark cycle and methods of handling, had been equated, but several apparently important factors that had not been equated evidently had strongly strain-specific effects.

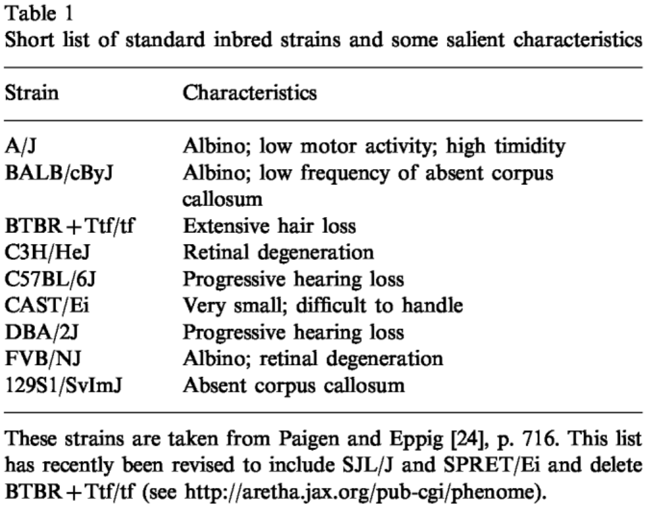

Table 1
Short list of standard inbred strains and some salient characteristics

Strain	Characteristics
A/J	Albino; low motor activity; high timidity
BALB/cByJ	Albino; low frequency of absent corpus callosum
BTBR + Ttf/tf	Extensive hair loss
C3H/HeJ	Retinal degeneration
C57BL/6J	Progressive hearing loss
CAST/Ei	Very small; difficult to handle
DBA/2J	Progressive hearing loss
FVB/NJ	Albino; retinal degeneration
129S1/SvImJ	Absent corpus callosum

These strains are taken from Paigen and Eppig [24], p. 716. This list has recently been revised to include SJL/J and SPRET/Ei and delete BTBR + Ttf/tf (see http://aretha.jax.org/pub-cgi/phenome).

The experience of the Crabbe et al. [8] study suggests that comprehensive standardization of both the test situation and the lab environment is not feasible. As argued elsewhere [38], a more attainable goal is to establish a set of standard tests that can serve as a benchmark or baseline for comparison across labs. Any lab that wishes to do so could then run the benchmark tests and compare results with their own local variants. This approach would not require each lab to adopt a single, world-wide standard, yet it would make available a standard that would aid interpretation of differences between labs. How such a benchmark or consensus version of a test might be achieved is a difficult challenge that has not yet been overcome.

4. Choosing standard strains

Any attempt to standardize a behavioural test or establish a benchmark must pay close attention to the population of animals on which the standard is to be based. The finest plastic and steel device tested in the cleanest and most humane animal facility will be of little advantage if an unwise choice of animal subjects is made. Inbred strains of mice are ideal material for standardization of behavioural tests because each strain is genetically uniform and relatively stable over a period of time. The genetic material was standardized decades ago, including breeding regimens and nomenclature, and dozens of strains are readily available from commercial breeders. Thus, the major question is how many and which strains should be used as the standard.

Paigen and Eppig [24] have recently proposed a short list of nine commonly used inbred strains that can serve as subjects for collecting standard phenotype data to be entered into the Mouse Phenome Database (http://aretha.jax.org/pub-cgi/phenome), along with an expanded list of more than 20 additional strains worthy of testing to detect diversity of behavioural phenotypes. These lists were the product of discussions among 36 mouse specialists held at the Jackson Laboratories on May 9–10, 1999. In Table 1, salient characteristics of the short-listed strains are noted.

It makes a great deal of sense to standardize a behavioural test with strains of mice that lack any major neurological abnormality. Then data on those strains can be used as a basis for judging when some other animal has a noteworthy deficit. Here we face an almost insurmountable challenge because so many of the common inbred strains possess one or more genetic defects that might influence behavioural test scores. Not one of the strains listed in Table 1 can be regarded as a standard of normality; each has some kind of noteworthy defect. Albinism itself has widespread physiological and behavioural consequences [11], and retinal degeneration eliminates all rods and then all cones in the retina, leaving greatly reduced visual function [16]. Despite these obvious shortcomings, the list of standard inbred strains has appeal because a minority of strains possesses any one defect, and comparisons across strains can yield important clues about the kinds of defects that may be most relevant for a particular behaviour. This pattern will be most readily interpreted for defects when at least two strains are afflicted, because one would expect both of them to show extreme scores in the same predicted direction. On the other hand, if only one strain has a particular defect and it shows an extreme score, a claim that the defect was the cause of the deviant behaviour is little more than a hunch, and further testing will certainly be needed to substantiate the relation.

Establishing a genetic correlation between an anatomical defect of brain development and variation in mouse behaviour is best done with a large number of inbred strains. A short list of strains as in Table 1 may yield data suggesting or hinting that retinal degeneration is a problem, for example, but far more evidence will be needed to make the case for a genuine, causal connection. There is one situation, however, where a possible neurological cause may be readily evaluated, and we would like to describe an instance in some detail. This method relies on non-genetic phenotypic variation within an inbred strain rather than differences among strains. It is especially important because it involves 129 strain mice that are most commonly employed for producing genetic knockouts.

5. The problem of absent corpus callosum in strain 129 mice

One of the most dramatic neurological defects seen in any of the strains in Table 1 is absence of the corpus callosum (CC), a major axon pathway that normally interconnects the cerebral hemispheres and provides rapid transmission of information between the two sides of the cortex. The defect occurs at a relatively low frequency of less than 10% in the BALB/cByJ strain [41] but it afflicts more than 30% of the mice in many 129 strains. Somewhat surprisingly, the lack of a CC has no influence on mouse paw preference [5], but it does interfere with motor coordination on challenging tasks where speed of performance is crucial [28] Absent CC arises from recessive genetic defects at more than one locus [21,22,39], and the expression of the defect caused by certain targeted mutations depends on the genetic background [23].

Given the severe anatomical abnormality seen in the 129 strains, perhaps this strain should not be included in the short list of standard strains. Because of its common use as a source of embryonic stem cells for creating knockouts, however, there are compelling reasons to include it, at least for purposes of comparison with other strains. It is not clear which of the many substrains to include because of uncertainty about ancestries and evident genetic contamination in certain substrains [30,33], but the problem of absent CC is present in all substrains we have examined to date. The question that needs to be addressed is therefore whether absent CC in the 129 strains is likely to have significant impact on a wide range of behavioural tests.

If absent CC occurred in every mouse in a 129 strain, it would be necessary to study F_2 hybrid crosses or recombinant inbred strains in order to determine whether the inadequate performance of 129 on a behavioural test is causally related to absent CC or merely a spurious correlate of a curious anomaly. Fortunately, the defect shows ‘incomplete penetrance’, meaning that only a subset of genetically identical mice is afflicted. This interesting fact of forebrain development allows a convenient assessment of CC function within the strain. We would like to present new data from a recent study to illustrate this application. Details of the behavioural tests have been presented previously [8].

5. 1. Animals

Equal numbers of males and females of the inbred strains A/J, BALB/cByJ, C57BL/6J, DBA/2J, and the B6D2F1/J hybrid were obtained from the Jackson Lab-oratory, Bar Harbor, Maine, USA. The 129/SvEvTac strain was obtained from Taconic Farms. The 5HT1B knockout and its 129/Svter control strain [15] were generously provided by Dr Rene´ Hen at Columbia University, New York, USA. All mice were 11 weeks of age when testing began. Half of them had been bred in each laboratory from parents shipped about 3 months earlier, and half were shipped 5 weeks prior to testing.

The breeding and then testing were done simultaneously under very similar conditions at three sites— the State University of New York in Albany, Oregon Health Sciences University in Portland, and the University of Alberta in Edmonton. Many of the environmental conditions were rigorously equated, but some aspects of the environments in the three labs could not be equated.

5.2. Apparatus and testing protocols

Identical test apparatus and testing protocols were employed at the three sites. These were adopted after a long series of discussions and negotiations conducted mainly via email among the three labs. The specific parameters were not necessarily chosen because they would be optimal; rather, some were chosen because they could be conveniently used at all three sites. Full details of the tests are provided in the web site for this study (http://www.albany.edu/psy/obssr). Briefly, each mouse received one test each day in the same order, consisting of open field activity on Monday, elevated plus maze on Tuesday, accelerating rotorod on Wednesday, visible platform water escape learning on Thursday, cocaine activation of open field activity on Friday, rest on Saturday, two-bottle drinking acclimation on Sunday, and then two-bottle ethanol preference testings for 4 days, followed by euthansia and collection of brains for histology. The study was conducted with two replications, each requiring 2 weeks of testing, and starting 1 week apart. Replication effects were generally absent or very small and are not presented here. Within a day, the order of testing the different strains, sexes, and treatment conditions was randomized; different random orders were used in the three labs and for the replications.

Open field activity was assessed in a 15 min trial in the 40 × 40 × 30 cm high clear plastic Digiscan photo-cell beam apparatus made by AccuScan Inc. Testing was done in total darkness in order to equate lighting conditions in all three labs.

Elevated plus mazes (5 × 25 cm black plastic arms with a 5 × 5 cm central area, 2.5 mm clear plastic rim on open arms, 15 cm clear plastic walls on enclosed arms; arms 50 cm above the floor) were manufactured in the University of Alberta shop and shipped to the other two sites. Mice received one trial of 5 min duration under about 100 Lux illumination, and behaviour was scored from video tape.

Rotorod testing was administered on schedule to all mice according to a standard schedule, but a flaw in the surface of the rod at one site was so serious that many mice were able to grasp it tightly and remain on the rod without moving. This clever but inconsistent behaviour rendered the data uninterpretable, and further presentation of this task would not be edifying.

Water escape tanks 70 cm in diameter and 25 cm deep were cast in polyethylene in Edmonton and shipped to the other sites, along with a 15 cm black mesh platform placed in the centre of the tank that extended 5 mm above the surface. Water was kept within one degree of 25 ºC. The mouse was started at the edge of the tank at one of four randomly chosen compass locations and allowed 40 s to find the central platform. It received eight trials in 1 day with a 30 s intertrial interval.

Cocaine activation was assessed with a 15 min trial in the DigiScan apparatus immediately after an intraperitoneal injection of 20 mg/kg cocaine.

Ethanol preference was assessed with 6% (v/v) ethanol in local tap water versus local tap water. One reading was taken each day, and the position of the ethanol and tap water tubes was reversed after 2 days.

5.3. Histology

High quality histological assessment of mouse brain can be a very labour intensive enterprise that is not compatible with high throughput phenotyping. Accordingly, we devised a much faster method for studying CC defects that is nevertheless valid for identifying mice with very small or totally absent CC. We did not perfuse the mouse with fixative; instead, after euthanasia, the fresh brain was carefully removed from the skull and then immersed in either 4% buffered paraformaldehyde or 10% buffered formalin (from 37% stock formaldehyde solution). We are grateful to Dr Pierre Roubertoux for suggesting this modification of our earlier method. Brains collected in Albany and Portland were shipped to Edmonton for further processing. After at least 2 weeks in fixative, each brain was weighed to the nearest mg and then bisected along the midsagittal plane. The left half was stained en bloc with gold chloride using the method of Schmued [29] to reveal myelin. After fixation of the stain with 2% sodium thiosulfate, a video image of the surface of the brain was taken with the JAVA system from Jandel Scientific, and the cross-sectional areas of the corpus callosum, (CC), hippocampal commissure (HC), and anterior commissure (AC) were measured. The maxi-mum length of the CC was also recorded.

5.4. Results—measures of brain

Data were evaluated using a factorial analysis of variance (ANOVA) on eight genetic groups, two sexes, three sites, and two kinds of breeding (local or shipped), with four mice in each of the 96 groups and 281 degrees of freedom for the within-group variance. The sample size for this study was chosen to yield a high level of power (90%) for detecting a moderate strain by site interaction in which hypothetical rank orders of strains were invariant but the magnitude of strain differences varied across labs. Because the sample size required to detect an interaction is commonly greater than that needed to detect a main effect [35], power of tests of moderate main effects was consider-ably greater than 90%. While the strain by site interaction was assessed with 12 mice per group (two sexes × two shipping conditions × four mice), the sex main effect was assessed with about 190 mice per sex.

Given the small sample size within any one group, it was not possible to assess the normality of data and equality of variances on a group by group basis, but close inspection of the data indicated that most data were reasonably close to normally distributed and variances were not markedly heterogeneous for body size, brain size, or anterior commissure area. For the hippocampal commissure, there were a few extremely low scores in the three groups with 129 genes. The ANOVAs are summarized in Table 2 where only effects significant at $\alpha = 0.01$ or better are reported. Entries in the table are partial omega-squared values derived from the F ratio for each effect to express the percentage of variance attributable to the differences among group means if only that one factor is considered in the ANOVA. Strain differences were generally very large but a sex difference was seen only for body weight. There were no significant effects of breeding locally versus fresh shipping from the suppliers. Substantial differences among the three sites were seen for body and brain sizes as well as areas of the AC and HC. Because CC size was obviously not normally distributed (Fig. 2), effects were tested with ANOVA after removing mice with small CC (< 2 mm length), which tended to increase strain means for the three 129- derived groups and thereby reduce the apparent strain effect size. Brain size differences among sites were not a simple consequence of body size; average body sizes were largest in Portland (24.1 g) and smaller in Albany (22.9 g) and Edmonton (22.3 g), whereas average brain sizes were largest in Albany (0.490 g) but similar in Portland (0.466 g) and Edmonton (0.469 g). The site effect on anterior commissure size, on the other hand, was largely but not entirely accounted for by differences in overall brain size when multiple regression methods were used [3].

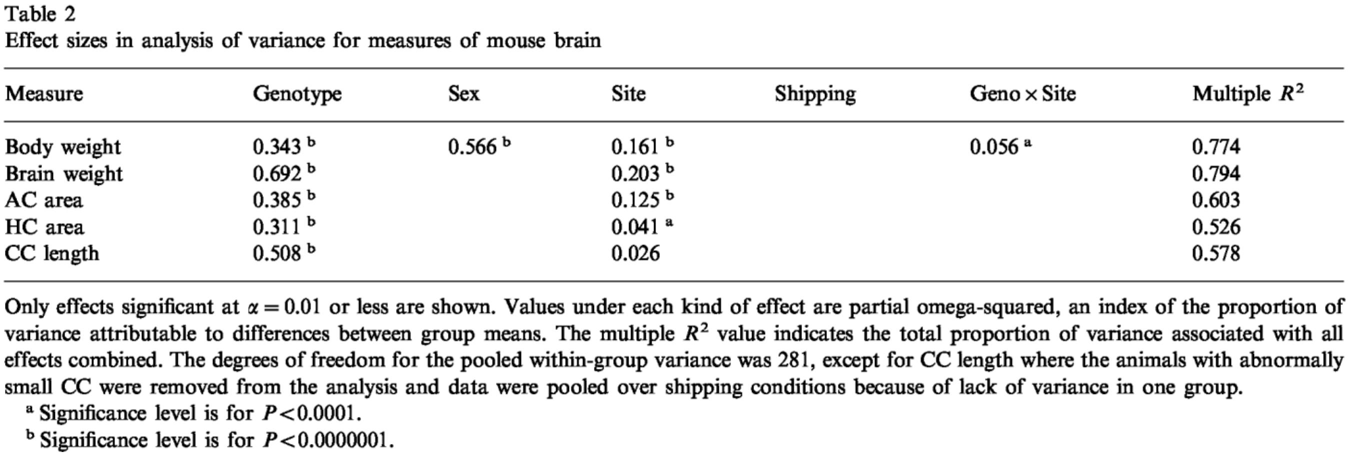
Table 2
Effect sizes in analysis of variance for measures of mouse brain

Measure	Genotype	Sex	Site	Shipping	Geno × Site	Multiple R^2
Body weight	0.343 [b]	0.566 [b]	0.161 [b]		0.056 [a]	0.774
Brain weight	0.692 [b]		0.203 [b]			0.794
AC area	0.385 [b]		0.125 [b]			0.603
HC area	0.311 [b]		0.041 [a]			0.526
CC length	0.508 [b]		0.026			0.578

Only effects significant at $\alpha = 0.01$ or less are shown. Values under each kind of effect are partial omega-squared, an index of the proportion of variance attributable to differences between group means. The multiple R^2 value indicates the total proportion of variance associated with all effects combined. The degrees of freedom for the pooled within-group variance was 281, except for CC length where the animals with abnormally small CC were removed from the analysis and data were pooled over shipping conditions because of lack of variance in one group.

[a] Significance level is for $P < 0.0001$.

[b] Significance level is for $P < 0.0000001$.

Origins of the body and brain size effects are not known for certain because the variation among lab environments was multifactorial. Mice drank local tap water and breathed local, fresh air. Although they all were fed Purina 5001 diet, this kind of diet can involve local differences in the specific sources of protein and other constituents of the diet [34]. The objective of the comparative study was not to equate lab environments. Rather, it was designed to determine whether the ubiquitous, largely unavoidable differences between laboratories would alter the pattern of genetic differences evident in test scores. For several behaviours, there was significant strain by site interaction [8], but for the features of brain presented here, no interaction was seen.

Size of the CC can be expressed conveniently in terms of its total length at the midsagittal plane (Fig. 2). Apart from one BALB/cByJ mouse with a short CC, absent or very short CC was seen only in the 129 strains and the knockout. Frequency of absent or very small CC ranged from 30% in Edmonton to 42% in Albany, but the site difference was not significant ($\chi^2 = 1.\ 3$, df = 2, $P > 0.\ 10$). Most cases of defective CC involved total absence, but a wide range of phenotypic expression was apparent for all three 129 strain groups, and the 2 mm cutoff for abnormality was somewhat arbitrary [39].

5.5. Results—measures of behaviour

Given the variability in CC size within a 129 strain, possible effects on behaviour can be addressed either by dichotomizing at the 2 mm criterion and conducting *t*-tests or computing non-parametric correlations between CC length and test scores. These analyses were done by pooling data for the three 129 groups, which did not differ significantly among themselves with respect to either the CC or most behavioural measures.
This yielded one large 129 strain group with 139 mice, 52 of which clearly had abnormal CC (see Table 3). That sample size was adequate to detect a substantial effect size of $\delta = 0.7$ with 90% power when $\alpha = 0.01$. Using more liberal but conventional values of $\alpha = 0.05$ and $\beta = 0.2$, $n = 52$ would be adequate to detect a small effect size of $\delta = 0.5$. The *t*-tests on dichotomized data suggested only one possible effect of the CC defect-a reduction in percent time in the open arms of the plus maze. The Spearman correlations pointed to several other small correlations that were just beyond the critical value for significance at $\alpha = 0.01$, but these were more difficult to interpret because the values were also influenced by mice whose CC sizes were definitely within the normal range of variation. Scatterplots of scores involved in three of the correlations with $r > 0.2$ are shown in

Fig. 3. For open field activity (Fig. 3A), the result was influenced by one very active mouse with a long CC. For wall hugging (Fig. 3B), the spread of scores was similar throughout the range of CC values but there was a moderate concentration of values near 100% for mice with no CC. Percent time in the open arms of the plus maze (Fig. 3C) was at 100% for several mice with normal CC, and this extreme score invariably occurred for animals that entered an open arm at the start of the trial and remained immobile. Wall hugging scores near 100% in the open field and open arm time near 100% in the plus maze almost always appeared in mice that were inactive. Hence, it would be hazardous to take these small correlations as a sign that absent CC increased anxiety.

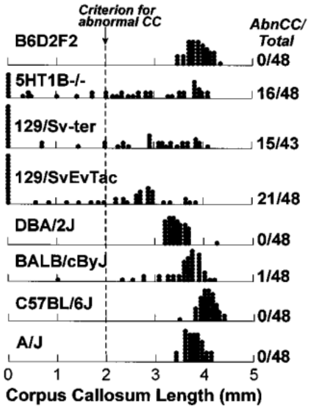

Fig. 2. Length of the corpus callosum (mm) measured at the midsagittal plane for mice of eight genetic groups tested in three different laboratories [8]. Only mice derived from the inbred strain 129 showed a high frequency of absent CC. An animal with CC length < 2.0 mm was considered abnormal for the purpose of dichotomizing and conducting *t* tests, but this criterion was somewhat arbitrary. Other criteria yielded similar conclusions, but using 0 mm as the criterion would intermix undoubtedly abnormal CC having only a few axons traversing the interhemispheric fissure with clearly normal CC.

Although these data do not prove that the lack of a CC in 129 strain mice has no effect on behaviour, they certainly indicate that any effect must be quite small for these particular tests. When using these sample sizes, an observed effect size of about $d = 0.2$ yields a 95% confidence interval for the true effect size δ that ranges from − 0.16 to 0.56. While it may not be entirely safe to ignore the problem of absent CC when studying behaviours that are not believed to be closely related to interhemispheric transfer, the magnitude of the effect is not likely to be very large. Thus, if the research focuses on genetic effects that themselves are clearly substantial, it may be implausible to believe that the genetic effect is a mere artifact of absent CC in 129 strain mice.

Table 3
Tests of effects of abnormal CC on behavior of 129-derived mice

Measure	Normal CC ($n = 87$)	Abnormal CC (<2mm) ($n = 52$)	*d*	r_s
Open field cm in 15 min	2635 ± 1549	2320 ± 1382	0.22	0.24
Open field — vertical movements	34.8 ± 32.0	29.8 ± 33.1	0.16	0.26
Open field — % time near walls	78.4 ± 14.3	80.9 ± 15.6	−0.3	−0.22
Cocaine activation — increase from Day 1, cm in 15 min	2736 ± 3231	2569 ± 3025	0	0.02
Plus maze — total entries in 5 min	13.7 ± 8.4	12.3 ± 8.0	0.17	0.16
Plus maze — % time in open arms	42.2 ± 29.8	26.9 ± 26.1	0.55*	0.23
Water maze — mean latency (sec) on 8 trials	8.3 ± 5.8	10.0 ± 7.3	−0.3	−0.24
Water maze — improvement over first 4 trials (sec)	3.9 ± 7.2	4.9 ± 6.8	−0.1	−0.13
Ethanol preference — g/Kg body weight consumed over 4 days	3.7 ± 2.3	3.6 ± 2.4	0	0
Ethanol preference — ratio on days 2 and 4	0.34 ± 0.28	0.38 ± 0.31	−0.1	−0.1

Table gives the group mean ± standard deviation. Effect size *d* is the difference between group means divided by pooled standard deviation. Only the value for plus maze % time in open arms reached a respectable level of significance ($t = 2.99$, $df = 137$, $P = 0.008$, two-tailed). Spearman r_S is calculated between CC length in mm and test score; a value greater than 0.2 or less than −0.2 would be needed to suggest statistical significance at $\alpha = 0.01$ for a two-tailed test.

5.6. Results — effects of the serotonin 1B receptor knockout

As a case in point, the same methods used to judge CC effects were then used to judge the knockout effect by comparing the $5HT_{1B}$ − / − homozygotes with 129/Sv-ter controls. This analysis revealed two interesting effects. Brain weight of the knockouts averaged 491 mg versus 466 mg for the controls, a difference of *d*= 1.14 standard deviations that was obviously significant (*P* = 0.000001). Among the many behavioural measures, only in the open field was a knockout effect apparent; the − / − homozygotes travelled further in 15 min (2965 versus 2249 cm, *d* = 0.53, *P* = 0.02), and the − / − homozygotes exhibited a higher level of rearing (52.7 versus 28.1 movements, *d* = 0.77, *P*= 0.001). For neither of these behavioural measures was there a difference within either genetic group between mice with normal and abnormal CC.

5. 7. Conclusion about absent CC in strain 129 mice

In many ways, 129 mice are not a good choice for standardization of any behavioural test because they often show extreme scores in comparison with other inbred strains and because of continuing uncertainty about the best substrain to use. These problems arise independently of the challenge of absent CC, which by itself makes them less than ideal as a standard for judging what is normal and abnormal behaviour. The 129 strain will nevertheless be found on most re-searchers' short list of inbred strains for standard testing solely because they are now so widely used in neurobehavioural genetics to produce knockouts. In this context, it would be wise to check for effects of absent CC in most studies. Only when a fairly large base of information from several labs is available, one large enough to allow meta-analysis, can we with greater confidence assert that there is no problem because of absent CC. The anatomical defect looks severe to the human observer, whereas the *t*-tests of behavioural effects are often less worrisome. Rather than dismiss the problem of absent CC as unworthy of concern in a general sense, it would be more helpful to expand the investigations until we can state that certain behaviours are clearly affected by CC absence, whereas others show little or no effects.

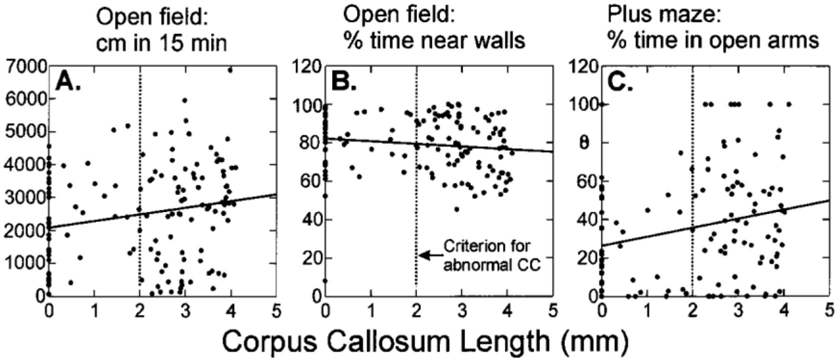

Fig. 3. Scatterplots of values for all 139 mice in the three groups (129/SvEvTac, 129/Sv-ter, $5HT_{1B}$ − / −) derived from strain 129, with the criterion for abnormally small CC (2 mm) shown as a vertical line. As indicated in Table 3, the three measures showed marginally significant Spearman correlations with CC length, but none of these correlations accounted for a substantial portion of variance in test score. (A) Distance travelled in the open field was slightly longer for mice with normal CC mainly because of one extremely high value. (B) Wall hugging was slightly higher on average for mice with no CC, but data were difficult to interpret because extreme wall hugging scores near 100% were invariably seen in animals with very low activity scores. (C) Percent time in the open arms of the elevated plus maze showed a preponderance of values at 100% for mice with normal CC, but the 100% score invariably occurred for animals that entered one open arm and froze there for the entire trial. Similarly, scores of 0% were mainly seen in mice that froze.

6. Different meanings of test standardization

The problem of absent CC in strain 129 mice brings into focus two approaches to standardization. The first, exemplified by Paigen and Eppig [24], seeks to collect a wide range of phenotypic data on numerous inbred strains in order to create a database of strain characteristics, much as has already been done for allelic variants in different strains. According to those authors, physiological and behavioural testing would best be done 'by the most expert and willing laboratories', thereby establishing a kind of gold standard for phenotypes. This approach is best served by testing as many inbred strains as possible, including those with obvious neurological defects. The emphasis is placed on genetically standardized strains, and it is presumed that good measures of a wide range of behaviours are already available.

The other approach seeks to standardize the tests on a small subset of strains or even hybrid mice lacking gross abnormalities in order to define the limits of normal and abnormal behaviour. This approach is most likely to progress rapidly if tests of known reliability and validity are employed [36,38]. For many kinds of tests, this may require further parametric studies with a goals of optimizing the test conditions and making it easier for other labs to adopt the same methods. Refinement of the tests would best be done prior to undertaking large scale mutagenesis screening experiments and surveys of almost all available inbred strains.

While we recognize that these approaches to standardization are complementary, we are following primarily the second option in our present research. We believe that careful study of the psychometric and parametric aspects of the many available tests of mouse behaviour will enable us to identify a relatively small array of demonstrably good tests that are sensitive to a large portion of the total genetic variance among the more common inbred strains in specific behavioural domains. Such a test array would then facilitate further research with a wider variety of genotypes as well as investigation of important environmental effects on test performance. Hopefully this kind of approach to the problem of standardization will encourage other laboratories to adopt the same methods, but we do not believe it would be feasible or desirable to attempt to impose a single standard on a diverse field of research.

References

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

[2] Benefiel AC, Greenough WT. Effects of experience and environment on the developing and mature brain: implications for laboratory animal housing. ILAR J 1998;39:5–11.

[3] Bishop KM, Wahlsten D. Sex and species differences in mouse and rat forebrain commissures depend on the method of adjusting for brain size. Brain Res 1999;815:358–66.

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

[5] 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.

[6] 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.

[7] Coste SC, Kesterson RA, Heldwein KA, Stevens SL, Heard AD, Hollis JH, Murray SE, Hill J K, 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. Nature Gen 2000;24:403–9.

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

[9] Crawley JN. Whats Wrong with My Mouse? Behavioral Pheno-typing of Transgenic and Kockout Mice. New York: Wiley-Liss, 2000.

[10] Crawley JN, Belknap JK, Collins A, Crabbe JC, Frankel W, et al. Behavioral phenotypes of inbred mouse strains: implications and recommmendations for molecular studies. Psychopharmacology 1997;132:107–24.

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

[12] Francis DD, Meaney MJ. Maternal care and the development of stress responses. Curr Opin Neurobiol 1999;9:128–34.

[13] Gairtner 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 Animals 1990;24:71–7.

[14] Gottlieb G. Normally occurring environmental and behavioral influences on gene activity: From central dogma to probabilistic epigenesis. Psychol Rev 1998;105:792–802.

[15] Hen R. Testing the genetics of behavior in mice. Science 1999;285:2069–70.

[16] Hicks D, Sahel J. The implications of rod-dependent cone survival for basic and clinical research. Invest Ophthalmol Vis Sci 1999;40:3071–4.

[17] 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.

[18] Kishimoto T, Radulovic J, Radulovic M, Lin CR, Schrick C, Hooshmand F, Hermanson O, Rosenfeld MG, Spiess J. Deletion of Chrhr2 reveals and anxiolytic role for corticotropin-releasing hormone receptor-2. Nature Gen 2000;24:415–9.

[19] Lander E, Kruglyak L. Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. Nature Gen 1995;11:241–7.

[20] Lederhendler I, Schulkin J. Behavioral neuroscience: challenges for the era of molecular biology. Trends Neurosci 2000;23:451–4.

[21] Lipp HP, Wahlsten D. Absence of the corpus callosum. In: Driscoll P, editor. Genetically-Defined Animal Models of Neurobehavioral Dysfunction. Birkhaiuser-Boston, 1992:217–52.

[22] Livy DJ, Wahlsten D. Tests of genetic allelism between four inbred mouse strains with absent corpus callosum. J Hered 1991;82:459–64.

[23] Magara F, Muiller U, Lipp H-P, Weissmann C, Staliar M, Wolfer DP. Genetic background changes the pattern of forebrain com-missure defects in transgenic mice underexpressing the B-amyloid-precursor protein. Proc Natl Acad Sci USA 1999;96:4656–61.

[24] Paigen K, Eppig JT. A mouse phenome project. Mamm Genome 2000;11:715–7.

[25] Roderick TH, Wimer RE, Wimer CC, Schwartzkroin PA. Ge-netic and phenotypic variation in weight of brain and spinal cord between inbred strains of mice. Brain Res 1973;64:345–53.

[26] Rodgers RJ, Dalvi A. Anxiety, defence and the elevated plus-maze. Neurosci Biobehav Rev 1997;21:801–10.

[27] Sayah DM, Khan AH, Gasperoni TL, Smith DJ. A genetic screen for novel behavioral mutations in mice. Mol Psychiatry 2000;5:369–77.

[28] Schalomon PM, Wahlsten D. Wheel running behavior is impaired by both surgical section and genetic absence of the mouse corpus callosum. Brain Res Bull, 2001, in press.

[29] Schmued LC. A rapid, sensitive histochemical stain for myelin in frozen brain sections. J Histochem Cytochem 1990;38:717–20.

[30] 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. Nature Gen 1997;16:19–27.

[31] Sokolowski MB, Wahlsten D. Gene-environment interaction and complex behavior. In: Chin H. and Moldin S.O., editors. Methods in Neurogenetics. CRC Press, 2001, in press.

[32] Tarantino LM, Gould TJ, Druhan JP, Bucan M. Behavior and mutagenesis screens: the importance of baseline analysis of inbred strains. Mamm Genome 2000;11:555–64.

[33] 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.

[34] Tordoff MG, Bachmanov AA, Friedman MI, Beauchamp GK. Testing the genetics of behavior in mice. Science 1999;285:2069.

[35] Wahlsten D. Insensitivity of the analysis of variance to heredity-environment interaction. Behav Brain Sci 1990;13:109–20.

[36] Wahlsten D. Single-gene influences on brain and behavior. Ann Rev Psychol 1999;50:599–624.

[37] Wahlsten D. Experimental design and statistical inference. In: Crusio WE, Gerlai R T, editors. Handbook of Molecular-genetic Techniques for Brain and Behavior Research. Amsterdam: Elsevier, 1999:41–57.

[38] Wahlsten D. Standardizing tests of mouse behaviour: reasons, recommendations, and reality. Physiol Behav 2001; in press.

[39] Wahlsten D, Schalomon PM. A new hybrid mouse model for agenesis of the corpus callosum. Behavl Brain Res 1994;64:111– 7.

[40] Wahlsten D, Schalomon PM, Crabbe J, Dudek B. Behavioral effects of absent corpus callosum and the $5HT_{1B}$-/- knockout in strain 129 mice. Soc Neurosci Abstr 1999;25:69.

[41] Wahlsten D, Smith G. Inheritance of retarded forebrain commissure development in fetal mice: results from classical crosses and recombinant inbred strains. J Hered 1989;80:11–6.

[42] Wuirbel H. Genetics of behaviour: the standardization fallacy. Nature Gen 2000;26:263.

[43] Wuirbel H. Standard housing interferes with normal brain development in laboratory rodents: implications for welfare and research. Submitted.