SEE locomotor behavior test discriminates C57BL/6J and DBA/2J mouse inbred strains across laboratories and protocol conditions

Neri Kafkafi [1, 3], Dina Lipkind [2], Yoav Benjamini [4], Cheryl L Mayo [3], Gregory I Elmer [3] & Ilan Golani [2]

1. Behavioral Neuroscience Branch, National Institute on Drug Abuse/IRP, Baltimore, Maryland

2. Department of Zoology, Tel Aviv University, Israel.

3. Maryland Psychiatric Research Center, Department of Psychiatry, School of Medicine, University of Maryland.

4. Department of Statistics, Tel Aviv University, Israel.

Abstract

Conventional tests of behavioral phenotyping frequently have difficulties differentiating certain genotypes and replicating these differences across laboratories and protocol conditions. In this study we explore the hypothesis that automated tests can be designed to quantify ethologically-relevant behavior patterns that would more readily characterize heritable and replicable phenotypes. We used SEE (Strategy for the Exploration of Exploration) to phenotype the locomotor behavior of the C57BL/6 and DBA/2 mouse inbred strains across three laboratories. The two genotypes differed in 15 different measures of behavior, none of which had a significant genotype ´ laboratory interaction. Within the same laboratory, most of these differences were replicated in additional experiments despite changing the test photoperiod phase and injecting saline. Our results suggest that well-designed tests may considerably enhance replicability across laboratories.

1. Introduction

Neurobehavioral genetics depends critically on the accuracy and consistency of behavioral measurements. The characterization of specific behaviors, sometime referred to as behavioral phenotyping, is essential for eventually associating them with particular gene loci. The need for behavioral phenotyping has resulted in the design of behavioral and physiological test batteries for mice (Crawley et al, 1997, 2000; Rogers et al, 1999). Considerable effort has been made to automate these tests, in order to increase the throughput needed for testing large numbers of animals and to avoid the effect of subjective human judgment. A locomotor behavior test in an unfamiliar arena is included in most test batteries, and can be conducted efficiently using standard, commercially available photobeam systems.

C57BL/6 (B6) and DBA/2 (D2) are two of the most commonly used inbred strains of laboratory mice. Consequently many behavioral differences between them were reported in the literature (for a review see Crawley et al., 1997). Since the locomotor behavior test is one of the most common tests, the locomotor behavior of these two genotypes has been reported in many studies. Common view holds that B6 is a high-activity strain while the D2 is an intermediate activity strain (Crawley et al., 1997; Cabib, 2002). This view, however, seems to represent only a broad average over studies, some reporting B6 to be significantly more active (e.g., Gorris and Abeelen, 1981; Elmer, Gorelick, Goldberg and Rothman, 1996; Logue, Owen, Rasmussen and Wehner, 1997; Bolivar, Caldarone, Reilly and Flaherty, 2000 (for females); Hatcher et al., 2001), many that did not detect a significant difference in activity (e.g., Jones, Radcliffe and Erwein, 1993; Womer, Jones, and Erwin, 1994; Tirelli and Witkin, 1994; Tolliver and Carney, 1995; Cabib and Boneventura, 1997; Bolivar et al., 2000 (for males); Rocha et al., 1998), and at least one (Rogers at al., 1999) that found D2 mice to have significantly higher activity. It should be noted that these studies used different arena sizes (longer dimension of 40 – 60 cm), arena shapes (square, rectangular or circular), light conditions (fully illuminated to complete darkness) and tracking techniques (photocells, photobeam and video). There

was no apparent relationship, however, between any of these factors and the ability to differentiate the two strains.

The problem of replicating behavioral results in other laboratories is neither particular to these two genotypes, nor to the locomotor behavior test. Consequently, Crabbe, Wahlsten and Dudek (1999) conducted a pioneering study in which they compared 8 genotypes in a battery of several standard behavioral tests across three laboratories. Despite the rigorous standardization of tests and housing protocols they reported many significant lab and lab × genotype interaction effects. One of their conclusions was that genotype differences found in a single laboratory might prove to be idiosyncratic to this laboratory. This conclusion might be interpreted as pointing to a major hindrance in current behavioral genetics research. As such, the importance of replicable phenotypes drives prominent database organizers to require submitted data to be validated in at least two different laboratories (Mouse Phenome Database; Paigen and Eppig, 2000).

Crabbe et al. (1999) included the B6 and D2 as part of the 8 genotypes in their study, but did not examine the differences between these two genotypes separately. In another pioneering step, however, they published all the raw results of their study in a web site, in a convenient format for downloading and analyzing. We have taken advantage of this feature to estimate the discriminative power and replicability of the new locomotor behavior method described in this study relative to a commonly used photobeam system, for these two genotypes. As we suggest in the discussion, both the multi-lab studies and the publication of the raw results on the web constitute a fruitful approach to behavioral phenotyping. When examining the results from this web site, most measures of the locomotor behavior test either did not detect significant differences between B6 and D2, or (in agreement with the general conclusion of their report) the differences were not consistent across the three laboratories.

The main remedy advocated so far for the lack of replicability is more careful standardization of test protocol, handling procedures and laboratory environment (Wahlsten, 2001; van der Staay and Steckler, 2002; Wahlsten, Crabbe and Dudek, 2001 but see Wurbel, 2000, 2002). This remedy, however, is expected to require a considerable effort (Wahlsten, 2001) since the level of standardization in the aforementioned study of Crabbe et al. was already much higher than is currently practiced in the field. We suggest

a complementary approach: design improved standard tests that can capture ethologically-relevant behavior patterns more precisely. Such tests may be more resistant to the laboratory environment and to small changes in protocol details. Locomotor behavior constitutes a good test case for this approach. Current locomotor behavior tests for mice, usually employing photobeam systems, are conducted in small cages of 40 - 60 cm width. They typically employ simple measures such as the distance traveled by the animal and the time spent in the center of the arena. These measures are cumulative and general, reflecting a common view that locomotor behavior is largely stochastic in nature, and can be quantified mainly by some measure of "general activity" (but see Paulus and Geyer, 1993 for a different viewpoint). In recent years, however, ethologically-oriented studies in rats (Eilam and Golani, 1989; Eilam, Golani and Szechtman, 1989; Golani, Benjamini and Eilam, 1993; Tchernichovski, Benjamini and Golani, 1996; Tchernichovski, Benjamini and Golani, 1998; Drai, Benjamini and Golani, 2000; Kafkafi, Mayo, Drai, Golani and Elmer, 2001) and more recently in mice (Drai, Kafkafi, Benjamini, Elmer and Golani, 2001; Benjamini, Drai, Elmer, Golani and Kafkafi, 2001; Kafkafi et al., submitted) found that locomotor behavior is highly structured and consists of typical behavior patterns. Once these patterns were isolated, they were found useful in psychopharmacological and psychobiological studies (Whishaw, Cassel, Majchrzak and Cassel., 1994; Cools, Ellenbroek, Gingras, Engbersen and Heeren, 1997; Gingras and Cools, 1997; Szechtman, Culver and Eilam, 1999; Whishaw, Hines and Wallace, 2001; Wallace, Hines and Whishaw, 2002). Based on these patterns, a Strategy for Exploring Exploration (SEE), complemented by respective software, was recently developed for the visualization and analysis of locomotor behavior data measured automatically by video tracking (Drai and Golani, 2001), and was proposed as a tool for behavioral phenotyping (Drai et al., 2001; Benjamini et al., 2001).

The SEE locomotor behavior test consists of both the SEE software and several testing procedures that were co-developed with it in order to enhance the accuracy, reliability and relevance of measurement. The SEE test has several properties that are suggested by the results of previous and present phenotyping projects to be important for genotype discrimination:

1. Large (2.50 m diameter) circular arena increases the area by 25 to 80 times that of common photobeam systems. Combined with a slightly better spatial resolution due to the use of video tracking, this means that the number of different locations that can be discriminated by the system is increased by a factor of approximately 100. The large arena also enables the animal to generate a much wider range of speeds, a key measure in the analysis. In addition, our results suggest that the open space of the large arena is more intimidating and consequently accentuates differences in wall hugging behavior.
2. A tracking rate of 25 or 30 frames per second is considerably higher than is currently practical with many photobeam systems. Such temporal resolution is important, since a mouse can accelerate, slow down and even stop and start again more than once during a single second.
3. Robust smoothing algorithms considerably reduce tracking noise and outliers, which are typical of the output of tracking systems of all types. Many endpoints, including the widely used distance traveled, are sensitive to such noise and artifacts.
4. The path of the animal is automatically segmented into discrete behavioral units with proven ethological relevance for rodents: progression segments separated by stops (lingering episodes) (Drai et al., 2000; Kafkafi et al., 2001). Most of SEE endpoints employ simple properties of such segments, such as their length, duration and maximal speed. Treating the path as a string of discrete, relevant units rather than a continuous series of coordinates allows a more straightforward analysis of complex structures.
5. The SEE language can be used to query, visualize and quantify complex properties of the behavior in a database including many sessions from many experiments, and easily design new endpoints for better genotype discrimination and replicability (Drai and Golani, 2001; Kafkafi et al, submitted; Kafkafi, submitted).
6. The issue of multiple comparisons, arising due to the use of many endpoints, is handled by the False Discovery Rate approach (Benjamini and Hochberg,

1995, Benjamini et al., 2001). This approach is preferable to either the too restrictive Bonferroni-like criterion or the too permissive approach of not controlling for multiple comparisons at all.

In this study we provide an initial examination of the ability of the improved locomotor behavior test with SEE analysis to discriminate mouse genotypes in a replicable way, by phenotyping B6 and D2 mice across three laboratories.

2. Methods

The experiments were conducted in three laboratories: the National Institute of Drug Abuse/IRP laboratory in Baltimore (NIDA), the Maryland Psychiatric Research Center (MPRC) of the University of Maryland, and Tel-Aviv University (TAU). There were slight differences between the laboratories in arena size (due to room size limitation), tracking rate (due to the use of the European PAL video system in TAU instead of the American NTSC) and spatial resolution (due to camera parameters and height). These differences are summarized in Table 1 (left section). In addition, two other experiments were used to assess the replicability across experiments with different protocol conditions within the same lab (Table 1, right section): experiment MPRC/L was performed in the MPRC, but with the mice tested during the light phase of their photoperiod instead of the dark phase. Experiment NIDA/LS was performed in NIDA with the mice tested at their light phase and also injected with saline immediately before introducing into the arena. In addition, we included some results from a sixth experiment, NIDA/C, which compared C57BL/6J with CXBK/ByJ (an inbred recombinant strain originating from a cross between C57BL/6 and BALB/c) in NIDA. The time period between any two experiments within the same laboratory was at least three weeks. Other than the differences given in table 1, all conditions were equated as described below. The animals used in this study were maintained in facilities fully accredited by the American Association for the Accreditation of Laboratory Animal Care (AAALAC) (MPRC and NIDA) or by NIH Animal Welfare Assurance Number A5010-01 (TAU). The studies were conducted at all three locations in accordance with the Guide for Care and Use of Laboratory Animals provided by the NIH.

2.1 Animals

9-14 week old males from the inbred strains C57BL/6J (B6), DBA/2J (D2) and CXBK/ByJ (CXBK), shipped from Jackson Laboratories (B6, D2) or bred at IRP/NIDA (CXBK). Group sizes are given in parentheses in Table 1.

2.2 Housing

Animals were kept in a 12:12 light cycle, housed 2-4 per cage under standard conditions of 22°C room temperature and water and food ad libitum. The animals were housed in their room for at least 2 weeks before the experiment.

2.3 Tracking protocol

Each animal was brought from its housing room, introduced immediately into the arena and returned after the end of the 30 minutes session. The arena was a large (210-250 cm diameter), circular area with a non-porous gray floor and a 50 cm high, primer gray painted, continuous wall. The gray paint was especially chosen to provide a high-contrast background, enabling video tracking of black, white, brown and agouti-color mice without the need to dye or mark them. Several landmarks of various shapes and sizes were attached in different locations to the arena wall and to the walls of the room where the arena was located, in order to enable easy navigation for the mouse. The arena was illuminated with two 40 W neon bulbs on the ceiling, above the center of the arena. These light conditions were the same for experiments conducted in different phases of the photoperiod cycle.

Tracking was performed using a video camera installed on the ceiling, feeding directly into a PC computer running a Noldus EthoVision® video tracking system (Spink et al, 2001), using its subtraction mode. Tracking rates and spatial resolutions (the actual distance represented by a single pixel on the screen) are detailed in table 1. Coordinate files were exported from EthoVision® and analyzed using SEE.

	Across labs experiments			Within lab variations		
Experiment name	**NIDA**	**MPRC**	**TAU**	**MPRC/L**	**NIDA/LS**	**NIDA/C**
Strains (n)	B6 (8) D2 (8)	B6 (10) D2 (10)	B6 (9) D2 (9)	B6 (8) D2 (9)	B6 (5) D2 (6)	B6 (8) CXBK (8)
Laboratory	NIDA	MPRC	TAU	MPRC	NIDA	NIDA
Arena's diameter	250 cm	210 cm	250 cm	210 cm	250 cm	250 cm
Tracking rate	30/s	30/s	25/s	30/s	30/s	30/s
Spatial resolution	1.3 cm	1.0 cm	1.0 cm	1.0 cm	1.3 cm	1.3 cm
Cycle of testing	dark	dark	dark	light	light	light
Treatment	---	---	---	---	saline	---

Table 1: Differences in experimental groups and testing conditions in the six experiments. n for each group given in parentheses. The three experiments in the left section were used to test across lab replicability. The three experiments in the right section were used to test the replicability across experiments with different conditions within the same lab, by comparing with the left-section experiment of the same laboratory.

2.4 Path analysis

Robust smoothing (i.e., not affected by arbitrary outliers) of the animal path and speed is an important ingredient in the SEE test. We used the Lowess algorithm (Cleveland, 1977) as was described in Kafkafi et al. (2001) with some improvements. The main improvement consisted of adding an algorithm based on Repeated Running Median (RRM, see Tukey, 1977), which was used to smooth the path at the very low velocities without erasing the very short stops (arrests). In a single iteration the smoothed location for each data point is the median of the locations in a small time window around this point. This process is repeated for several iterations with different window sizes. We used 4 iterations with window sizes of 7, 5, 3 and 3 data points. This choice followed Tukey's (1977) guidelines and a comparison of the outcome to the actual videotaped behavior in several sampled sequences, in order to ensure that short arrests are not smoothed out. The results of the RRM smoothing were used only to isolate arrest intervals, which were defined as instances where the RRM smoothed location did no change for at least 0.2 seconds, and in these points the speed was defined as 0. Such short stops constitute an important part of the behavioral repertoire of rodents (Golani et al., 1993), especially of the small and fast-moving mouse (Drai et al., 2001), and most other

smoothing algorithms, while useful at higher speeds, would tend to smooth them out. For non-zero speeds, the smoothed locations and speed was estimated by the Lowess from the raw data as before (Kafkafi et al., 2001). The time window used for the Lowess was 0.4 s and the polynomial degree was 2. The above combined procedure was implemented in SEE Path Smoother, a stand-alone program available from the authors with or independently of the whole SEE package. For a general review of smoothing methods for tracking and discussion of their importance, see Hen, Sakov, Kafkafi, Golani and Benjamini, (submitted).

Segmentation of the smoothed path into lingering episodes and progression segments was done using the EM algorithm as in previous studies (Drai et al, 2000; Kafkafi et al, 2001) except for one important difference: the segmentation was always done into two components - lingering and progression - and we did not use the further division of progression into slow and fast movement ("2^{nd} an 3^{rd} gears"). The reason is that this sub-division is often not clear in mice, while the division into lingering and progression is very general. Most mice displayed a clearly bi-modal distribution of segment maximal speeds, with the threshold between the lingering and progression typically at values of 10 to 20 cm/s. As with the smoothing algorithms, the segmentation using the EM algorithm is currently implemented in a stand-alone program, which is available from the authors with or independently of the whole SEE package.

Visualization, analysis and calculation of behavioral measures were done with SEE (Drai and Golani, 2001), and with the assistance of two extension programs, the "SEE Experiment Explorer" and "SEE Endpoint Manager" (Kafkafi, submitted). The first was designed to assist with SEE querying any desired subsection of a database including many experiments, while the second standardizes SEE calculation of endpoints and the development of new endpoints. These programs are not necessary for the analysis but they make it much more efficient and user-friendly. Both are also available from the authors. 17 behavioral measures ("endpoints") were used in this study, and are listed in tables 2 and 3. In the approach suggested in this study the algorithms of these endpoints are not merely methods, but constitute an important result of the study. The algorithm for generating each endpoint is therefore reported in section 5, with its rationale and its success with reliably differentiating the locomotor behavior of B6 and D2.

2.5 Statistical methods

Standard transformations (log, square root, logit) were applied to the results in some of the endpoints (see table 2) so as to correct towards approximately normal distributions. Across labs results were analyzed by comparing experiments NIDA, MPRC and TAU using genotype × laboratory two-way ANOVA for each endpoint. We estimated the effect of each factor (genotype, laboratory, their interaction, and the "residuals" or individual animal) by the proportion of the variation attributed to that factor out of the total variation, i.e., the SS(Factor)/SS(Total), and assessed their statistical significance using F-tests. We also supported the above by calculating Ω^2 estimates for the effects of the factors. In spite of being generally less biased than the proportions of variation, they run into problems when the estimated effects are close to 0, as is the case here. Since elsewhere the values of Ω^2 were not smaller by more than 3% than the proportion of variation we report only the latter. Note that the proportion of genotypic variance is a relatively conservative estimate of the broad sense heritability, since some of the interaction variance may also be genetic, and part of the individual variance may be attributed to measurement error.

The testing of many endpoints in this study raises the problem of multiple comparisons. We approached the problem using the False Discovery Rate (FDR) as suggested by Benjamini and Hochberg, (1995) and Benjamini and Liu (1999). This approach calls for controlling the expected proportion of false discoveries among the discoveries (the number of erroneously rejected null hypotheses among the rejected ones). When all differences are not real this protects the experimenter against making even one false discovery, but otherwise FDR controlling procedures are more powerful than traditional multiple comparison procedures. See Benjamini et al. (2001) for detailed discussion of the approach with behavioral phenotyping and explanation of the procedures used to assess the significances obtained from the two-way ANOVA. The FDR was controlled separately for genotype, lab and interaction p-values, each at a level of 0.05.

The replicability across experiments with different conditions within the same lab was tested by comparing experiment NIDA with experiment NIDA/LS and comparing

experiment MPRC with MPRC/L (see Table 1), using the same procedures as for the three lab comparisons.

3. Results

Table 2 presents the results in each endpoint (group means and standard deviations) of the two genotypes in the three labs, NIDA, MPRC and TAU, and in the two additional experiments, MPRC/L and NIDA/LS (see table 1 for locations and conditions). Table 3 presents the genotype differences in each endpoint using genotype × lab two-way ANOVA as tested across NIDA, MPRC and TAU, and corrected for multiple comparison by FDR. Significant strain differences were found in 15 out of 17 endpoints. Out of these 15 endpoints, 7 had significant lab effects. The lab effects were much smaller than the genotype effects except for Home Base Relative Occupancy (see next section for the description of endpoints). None of the endpoints, however, had a significant genotype × lab interaction. This means that lab effects were all additive. That is, the genotype differences did not differ across labs even when the actual group means did. Regarding effect sizes, in 12 out of the 17 endpoints the genotypic variance was larger than the laboratory and interaction variances combined. In 7 out of these endpoints, the genotypic variance alone accounted for more than 50% of the total variance.

Experiment MPRC/L replicated the experiment in the MPRC, except that animals were tested during their light phase instead of their dark phase. The two experiments were compared using genotype × experiment two-way ANOVA. All the genotype differences that were significant in the across-lab comparison were also significant in this comparison except for one: the Diversity. The only endpoint that was significantly different between the two experiments was the Lingering Spatial Spread. None of the interaction effects were significant.

Experiment NIDA/LS replicated the experiment in NIDA, except that animals were tested during their light phase and also received a saline injection. Out of the 15 genotype differences that were significant in the across-lab comparison, only 4 were not significant in this comparison: the Maximal Segment Speed, the Diversity, the Number of Stops per Excursion and the Home Base Relative Occupancy. The only endpoint that was significantly different in experiment NIDA/LS was the Latency to Half Maximum Speed. As in the previous comparisons, none of the interaction terms were significant.

Endpoint	Units/ Transform	NIDA B6	NIDA D2	MPRC B6	MPRC D2	TAU B6	TAU D2	NIDA/LS B6	NIDA/LS D2	MPRC/L B6	MPRC/L D2
1. Distance Traveled	m	256	144	227	130	264	109	234	79	247	181
		(42)	(54)	(13)	(48)	(40)	(40)	(33)	(54)	(24)	(36)
2. Center Time	proportion	-0.89	-3.20	-0.99	-3.40	-0.79	-3.81	-0.54	-3.60	-0.60	-2.20
	Logit	(0.25)	(1.06)	(0.16)	(1.87)	(0.47)	(1.08)	(0.49)	(2.10)	(0.38)	(0.32)
3. Proportion of Lingering	proportion	0.49	0.80	0.50	0.80	0.30	0.70	0.53	0.87	0.47	0.67
		(0.10)	(0.10)	(0.03)	(0.10)	(0.03)	(0.10)	(0.07)	(0.09)	(0.06)	(0.06)
4. Number of Progression Segments	number	554	286	593	317	525	223	479	192	575	448
		(67)	(87)	(73)	(118)	(50)	(102)	(124)	(144)	(98)	(80)
5. Lingering Mean Speed	cm/s	3.60	2.30	2.90	2.00	2.00	1.50	3.49	1.48	3.10	1.90
		(0.40)	(0.70)	(0.40)	(0.40)	(0.20)	(0.50)	(0.32)	(0.57)	(0.43)	(0.18)
6. Lingering Spatial Spread	cm	1.34	1.51	1.28	1.60	1.00	1.36	1.46	1.53	1.19	1.24
	Square root	(0.18)	(0.29)	(0.07)	(0.21)	(0.06)	(0.35)	(0.21)	(0.15)	(0.13)	(0.06)
7. Length of Progression Segments	cm	28.9	29.9	23.4	24.6	28.0	31.7	31.8	27.4	26.8	26.2
		(7.7)	(9.9)	(4.3)	(5.0)	(5.1)	(12.8)	(8.8)	(7.6)	(4.4)	(5.4)
8. Diversity	number	132.0	120.0	116.0	111.0	140.0	120.0	123.6	91.4	115.9	114.8
		(6.9)	(29.3)	(2.5)	(14.9)	(5.9)	(36.9)	(15.8)	(54.7)	(3.3)	(7.1)
9. Maximal Segment Speed	cm/s	31.0	38.9	27.6	32.9	29.0	34.8	30.2	29.5	29.6	33.7
		(3.6)	(7.7)	(2.9)	(4.4)	(4.1)	(8.6)	(3.6)	(6.4)	(3.8)	(4.1)
10. Segment Acceleration	cm/s/s	3.2	3.6	3.2	3.5	3.0	3.2	3.1	3.3	3.2	3.4
	Log	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.2)	(0.4)	(0.1)	(0.1)
11. Rate of Turn	deg/s	25.6	20.8	31.6	24.6	28.2	19.0	25.8	19.1	30.6	24.2
		(1.4)	(3.1)	(2.4)	(3.1)	(2.7)	(1.9)	(0.3)	(1.9)	(1.7)	(2.1)
12. Radius of Turn	ratio	-0.62	-0.12	-0.74	-0.26	-0.80	-0.16	-0.68	-0.33	-0.71	-0.30
	Log	(0.07)	(0.11)	(0.11)	(0.17)	(0.14)	(0.12)	(0.09)	(0.24)	(0.10)	(0.08)
13. Home Base Relative Occupancy	ratio	2.0	2.3	1.1	1.6	1.6	2.0	2.0	2.5	1.2	1.5
	Log	(0.3)	(0.5)	(0.2)	(0.3)	(0.4)	(0.5)	(0.5)	(0.3)	(0.2)	(0.3)
14. Number of Excursions	number	6.0	4.5	5.5	4.4	5.3	3.6	5.9	3.5	5.3	4.9
	Square root	(0.5)	(0.8)	(0.5)	(0.9)	(0.4)	(0.9)	(0.7)	(1.4)	(0.2)	(0.5)
15. Number of Stops per Excursion	number	11.8	8.9	13.9	8.2	12.1	11.4	10.2	10.8	14.5	13.7
		(2.1)	(3.4)	(4.3)	(3.9)	(4.1)	(3.1)	(2.5)	(2.9)	(2.1)	(2.8)
16. Activity Decrease	m	-13.0	-26.0	-40.0	-20.0	-17.0	-21.0	9.6	21.1	-21.2	-32.8
		(13.0)	(20.0)	(16.0)	(16.0)	(22.0)	(23.0)	(6.5)	(13.1)	(17.6)	(19.0)
17. Latency to Half Maximum speed	s	2.9	4.5	3.2	4.2	3.6	4.9	4.3	5.7	3.6	4.3
	Log	(0.4)	(0.4)	(0.8)	(1.1)	(0.5)	(0.5)	(0.5)	(1.2)	(0.9)	(0.5)

Table 2: Strain means (SD) in the three laboratories, NIDA, MPRC and TAU, and for experiments NIDA/LS and MPRC/L. Experiment MPRC/L replicated the experiment in the MPRC, except that animals were tested during their light phase. Experiment NIDA/LS replicated the experiment in NIDA, except that animals were tested during their light phase and also received a saline injection. For additional differences in experiment conditions see Table 1. In endpoints in which transformations were used for the statistical analysis, the means and the SD are for the transformed variables.

Endpoint	Genotype			Laboratory			Interaction		
	F	p	%V	F	p	%V	F	p	%V
1. Distance Traveled	118.4	0.0001	68.3	1.1	0.335	1.3	2.4	0.1055	2.7
2. Center Time	86.5	0.0001	63.6	0.2	0.827	0.3	0.6	0.575	0.8
3. Proportion of Lingering	337.2	0.0001	76.6	26.8	0.0001	12.2	0.7	0.494	0.3
4. Number of Progression Segments	141.2	0.0001	71.6	3.8	0.0294	3.9	0.2	0.8086	0.2
5. Lingering Mean Speed	61.3	0.0001	33.3	33.3	0.0001	36.3	3.9	0.0227	4.3
6. Lingering Spatial Spread	34.1	0.0001	31.7	11.2	0.0001	20.9	1.4	0.25	2.7
7. Length of Progression Segments	0.6	0.444	1.1	3.3	0.0471	11.7	0.2	0.793	0.8
8. Diversity	5.2	0.0274	8.5	3.3	0.0448	10.8	0.7	0.5246	2.1
9. Maximal Segment Speed	15.8	0.0002	22.5	2.9	0.0656	8.2	0.3	0.715	1.0
10. Segment Acceleration	114.7	0.0001	51.2	29.0	0.0001	25.9	1.7	0.187	1.5
11. Rate of Turn	106.6	0.0001	52.1	21.6	0.0001	21.1	3.4	0.0403	3.4
12. Radius of Turn	247.1	0.0001	79.8	5.0	0.0105	3.2	2.3	0.1101	1.5
13. Home Base Relative Occupancy	12.2	0.001	11.9	21.0	0.0001	40.8	0.3	0.7095	0.7
14. Number of Excursions	54.0	0.0001	47.0	5.7	0.0062	9.9	0.7	0.483	1.3
15. Number of Stops per Excursion	9.9	0.003	15.6	0.7	0.4908	2.3	2.1	0.138	6.5
16. Activity Decrease	0.2	0.67	0.3	2.0	0.1486	6.6	3.8	0.0295	12.7
17. Latency to Half Maximum speed	46.7	0.0001	44.7	4.4	0.0183	8.3	0.5	0.5813	1.0

Table 3: Effect sizes as estimated by percent of variance (%V) attributable to genotype, laboratory and interaction in each endpoint, and their statistical significances expressed by the F-ratios and p-values, for the comparison across the three laboratories (experiments NIDA, MPRC and TAU, see Tables 1 and 2). Shaded effects were found to be significant while controlling the FDR at 0.05. Note that the percent of genotype variance is a conservative estimate of broad-sense heritability.

Note that the two within-lab comparisons above do not appropriately test the effect of the differences in their protocol, the testing phase and the saline injection, since these differences were not controlled factors within the same experiment, but differences of conditions between two experiments conducted at different occasions. Moreover, in principal the effect of repeating the experiment might have countered the effect of the different conditions in the second experiment, so as to create a false impression of replicability. Genotype differences that were significant in both within-lab comparisons, however, were also significant in the across-lab comparison. Together with the small size and lack of significance of the interaction terms, this strongly suggests that the genotype differences were replicated in the different experiments despite the differences in laboratory and conditions.

All the endpoints except for the last two, Activity Decrease and the Latency for Half Maximal Speed, can be measured in time bins of 5 minutes. Endpoints that discriminated the two strains over the whole session were mostly able to discriminate them in each of the 5 minute bins, as is demonstrated in figures 1, 2 and 3 for Distance Traveled, Center Time and Radius of Turn. Note that the development of the Radius of Turn seems to differentiate the CXBK mice in experiment NIDA/C from the other two strains. The CXBK started the session like the B6 and ended like the D2. In general, note that group variances were usually small even when measured in bins of 5 minutes, although group sizes were not large (5 to 10).

4. Endpoint list

Endpoint algorithms constitute an important result of this study, since replication of genotype differences across laboratories and conditions with a certain endpoint suggests that the algorithm of this endpoint captures a genotype-specific behavior pattern. In what follows we describe each endpoint separately with the rationale of using it, and present its success in reliably differentiating the two genotypes. The endpoints measure the behavior of individual mouse over a session. Many of the endpoints quantify properties of segments (either progression or lingering episodes) performed by the mouse throughout the session, and in these cases the median over all segments in the session were taken, unless mentioned otherwise.

***4.1 Distance Traveled*:** This is the total length of the smoothed path. This measure might seem trivial, as it is the primary endpoint used in most locomotor tests. In most of these tests, however, it is highly sensitive to the spatial and temporal resolution of the tracking because higher resolution can detect smaller meanderings of the path and thus increase its length considerably (Paulus and Geyer, 1993). Consequently, travel distances measured with different tracking parameters and arena sizes are usually not considered to be comparable. In addition, this measure is sensitive to recognition errors of the tracking system. For example, if the system erroneously identified the animal in the other side of the arena for only a single frame, the distance from the true location and back is added to the distance traveled, and such an error might happen many times during a session. The robust smoothing algorithms used in our method, however, take care of both these problems. In addition, the large arena may cause the B6 to increase their mileage and/or cause the D2 to decrease their mileage. This may explain why B6 mice traveled very significantly larger distances, about 250 m vs. 130 m in our setup. Furthermore, the strain means themselves, and not only the difference between them, were very consistent both across and within labs, despite the differences in arena size, protocol conditions and tracking parameters.

***4.2 Center Time*:** This endpoint measures the total time the animal spent at a distance of more than 15 cm from the wall of the arena. B6 mice spent significantly more time in the center than D2 mice. This difference as well as the strain means themselves were very

replicable both across labs and across experiments with different condition within labs, as shown by less than 1% of the total variation attributable to any lab, experiment or interaction effect, and none of which found statistically significant (table 3, and see also fig. 2). Such a difference was not reported for these two genotypes with smaller arenas, although most current photobeam systems measure center time regularly (and see fig. 4, middle raw). Since Center Time, in contrast to Distance Traveled, is not very sensitive to tracking resolution or tracking artifacts, it seems that the large arena used in our protocol accentuated the differences in wall hugging.

4.3 Proportion of Lingering Time equals the total duration of lingering episodes (i.e. stops, as computed by the segmentation EM algorithm) in the session as a proportion of the session duration. D2 mice had much longer lingering time, about 30 to 40 percent point more than B6, the difference being very significant. The lab effect was significant but within-lab differences were not.

4.4 Number of Progression Segments: The number of progression (movement) segments in the 30 min session, as computed by the segmentation process. Note that this number is, by definition, also the number of stops. B6 mice had significantly higher number of progression segments than D2 mice. Strain differences as well as strain means themselves were replicable across experiments with different conditions within labs (see also fig. 4, bottom raw).

4.5 Lingering Mean Speed equals the cumulative distance traveled in the lingering (stopping) mode, divided by the cumulative duration of lingering. It thus provides a rough measure of "local" mobility within stops, which probably consists mainly of scanning movements, few sideway and forward steps, rearing and "stretch-attend" behavior. B6 mice had higher lingering mean speed than D2 across and within laboratories. The large and very significant lab effect in this endpoint might reflect the sensitivity of small movement measurement to tracking conditions and parameters. This is also suggested by the lack of significant effect in the comparisons within laboratories. A similar pattern was also found with the Proportion of Lingering Time (see 4.3). As with all other comparisons, however, the genotype × lab interaction was not significant.

4.6 Spatial Spread of Lingering Episodes: the longest distance between any two points during a lingering episode (a stop). Since lingering episodes are frequently local

and circumscribed, their spatial spread is a more appropriate measure than their length (Drai et al., 2000). D2 had a significantly larger spatial spread. The significant lab effect may also be due to the lower reliability of tracking small lingering movements, but the interaction effect was not significant. Within the MPRC, the spatial spread was significantly smaller in experiment MPRC/L, in which the animals were tested during the light cycle of their photoperiod.

4.7 Length of Progression Segments: the path length of progression (movement) segments. No differences were found between the two genotypes in this measure. All group means were between 23 and 33 cm. This similarity seems to be a property of the B6 and D2 strains, since some of the additional strains we currently test do show significantly shorter or longer progression segments.

4.8 Diversity is the average distance between any two stops, weighted by the distribution of the duration of these stops (Tchernichovski et al., 1996). This measure captures the spatial and temporal scatter of stops. It is higher as a greater area is covered, as it is covered more homogenously, and as the duration of stopping is distributed more homogeneously over the arena. B6 had significantly higher diversity, but the proportion of variance attributed to the individual animal in this endpoint was very large, about 79% (Table 3), and the strain difference was not significant in both within-labs comparisons.

4.9 Maximal Speed of Progression Segments: The maximal speed attained during a segment was found to distinguish progression from lingering behavior (Drai et al., 2000, Kafkafi et al., 2001) and is therefore a reasonable choice for genotype discrimination. The median of segment maxima over all progression segments in the session was taken as the result of this animal. Despite the much higher distance traveled by B6 mice, D2 mice were significantly faster across labs with no significant lab effect. This difference was significant also in the comparison within the MPRC but not in the comparison within NIDA.

4.10 Segment Acceleration provides, for each progression segment, its maximal speed divided by its duration. It is thus a rough estimation of the acceleration in this segment (Kafkafi et al, submitted). D2 had significantly higher segment acceleration than B6 across labs. The laboratory effect was also significant. In both comparisons within labs genotype effects were significant but experiment and interaction effects were not.

4.11 Rate of Turn During Progression measures the amount of turning (change of progression direction) of progression in time. For each data point in the progression mode, this endpoint calculates the change of heading relative to the previous data point. Division by the tracking rate gives the rate of direction change in degree/s for each data point. The session value is the median of the absolute rate (i.e., no distinction between right and left turning) over progression during the whole session. This measure is not computed in the lingering mode, since the distance between consecutive data points in this mode might be very small, frequently much less than the spatial resolution of the tracking system, and thus the measured change in direction is likely to be meaningless. B6 Rate of Turn was very significantly higher than that of D2 across laboratories. Laboratory effect was also significant, but in both within labs comparisons genotype differences were significant while experiment and interaction effects were not.

4.12 Radius of Turn During Progression*:* This measure is computed by dividing, for each data point, the absolute turning rate (in degrees/s) by the speed (in cm/s), in order to get the curvature (in degrees/cm). The result is further multiplied by $180/\pi$ in order to get the radius of turning in cm. This is a measure of turning similar to the previous, but in relation to space and not to time. Note that it is possible to turn with the same radius but with different turning rates (by changing the speed) or with different turning radii but with the same turning rate. As with the turning rate, we use the median over all data points during the progression mode only. Furthermore, since one of our labs had an arena with a slightly smaller radius, and the animals progressed mainly along the wall, we calibrated the median radius of turn for each session by dividing it by the arena radius. The radius of turn (fig. 3) of the B6 was very significantly smaller in both across and within lab comparisons, about half of the arena radius, while the radius of turn of the D2 was almost as large as the arena radius. Note that the proportion of variance attributed to the genotype, which is a conservative estimation of broad sense heritability, was almost 80% in this endpoint, while merely 3.2% and 1.5% of the variance were attributed to lab and interaction effects respectively.

4.13 Home Base Relative Occupancy measures how much the animal stays in its most preferred location. This endpoint is computed by multiplying the total lingering duration by the number of lingering episodes for each place in the arena. This calculation

is performed in bins of 10° along the perimeter of the arena. The home base was defined as the bin having the maximum value of this multiplication. It was usually, but not always, the place where the mouse was first introduced. The occupancy of the home base is given as the ratio between this maximum and the mean over all 10° bins. Note that "occupancy" in this algorithm use both the time of staying and the number of stops, as found by Eilam and Golani (1989). D2 mice had significantly higher home base occupancy across laboratories. This endpoint was, however, the only one in which the lab effect was much larger than genotype effect. The genotype differences were also significant in the within MPRC comparison but not in the within NIDA comparison.

4.14 Number of Excursions*:* A round trip starting and ending at the home base is an excursion, which is a natural unit of rodent exploration (Eilam and Golani, 1989; Golani et al., 1993; Tchernichovski and Golani, 1995; Tchernichovski et al., 1998; also recently employed in Whishaw et al., 2002 and Wallace et al., 2002). B6 mice had significantly more excursions, about 30 vs. 20 of the D6 mice, both across and within laboratories. The lab effect was also significant, but the size of its effect was about a fifth of the strain effect.

4.15 Number of Stops per Excursion: The mean number of stops in each excursion. This endpoint was found to be useful with Long-Evans rats (Golani et al., 1993). It was also used to distinguish between rats showing high and low response to novelty under dexamphetamine (Cools et al., 1997; Gingras and Cools, 1997). Session median was typically between 8 and 15, which is perhaps slightly larger than that of rats, and B6 had significantly more stops per excursion across laboratories. This difference was also significant in the comparison within the MPRC but not in the comparison within NIDA.

4.16 Activity Decrease measures the difference in activity between the first and second 15 minutes of the session. No significant differences were found between B6 and D2 in this regard. Interestingly, this measure yielded the most significant difference between B6 and BALB/c strain (a single-lab study, unpublished data).

4.17 Latency to Half of Maximal Speed quantifies how fast the activity of the animal builds up during the beginning of the session. It measures the time, from the start of the session, it took the animal to exceed for the first time a speed equaling half of the maximal speed it attained during the whole session. B6 mice were significantly faster to

take off, both across and within labs, typically around 10 seconds, while D2 typically took around 100 seconds. In experiment NIDA/LS this endpoint was significantly higher, comparing with the other experiment in this lab. This was probably a result of the saline injection in NIDA/LS, since no such change was detected in the comparison within the MPRC.

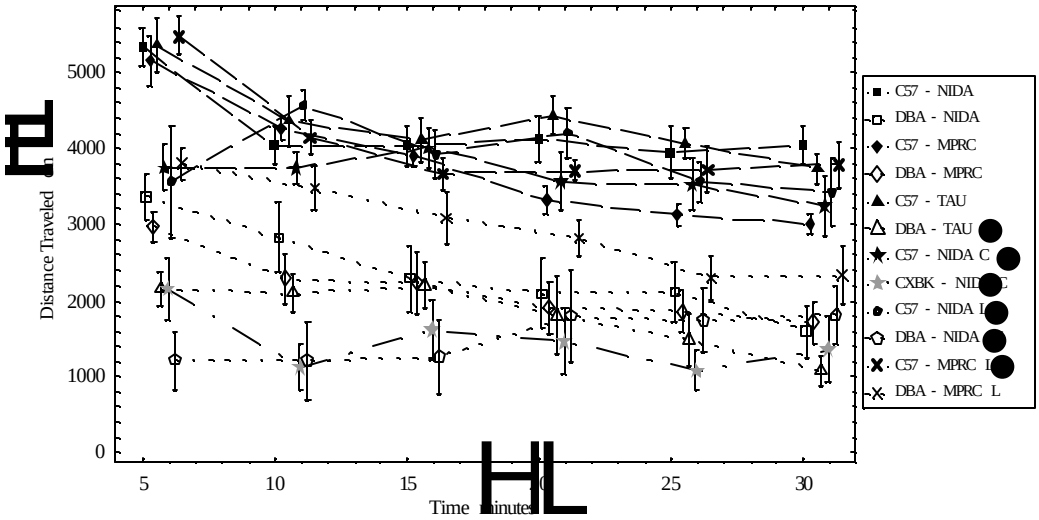

Fig. 1: The distance traveled in the six experiments (group means ± SE) in time bins of 5 minutes. Dark symbols with dashed lines: B6 groups. Open symbols with dotted lines: D2 groups. Grey star with dash-dot line: a single CXBK group. Experiment MPRC/L replicated the experiment in the MPRC, except that animals were tested during their light phase instead of their dark phase. Experiment NIDA/LS replicated the experiment in NIDA, except that animals were tested during their light phase and also received a saline injection. Experiment NIDA/LS compared B6 with CXBK mice in NIDA with animals tested during the light phase. For additional differences in experiment conditions see Table 1.

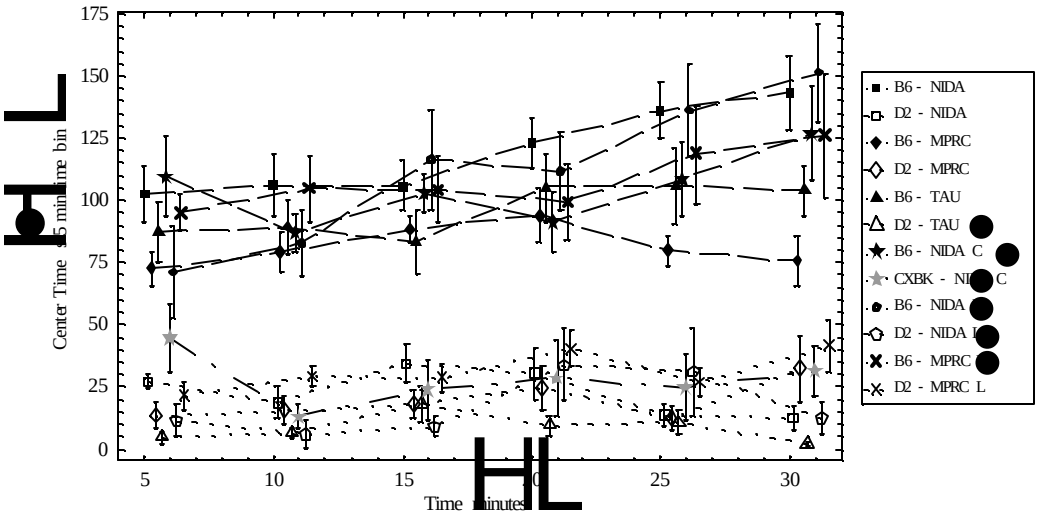

Fig. 2: Center time in the six experiments (group means ± SE) in time bins of 5 minutes. Dark symbols with dashed lines: B6 groups. Open symbols with dotted lines: D2 groups. Grey star with dash-dot line: a single CXBK group. Experiment captions (NIDA, MPRC, etc.) as in Fig. 1.

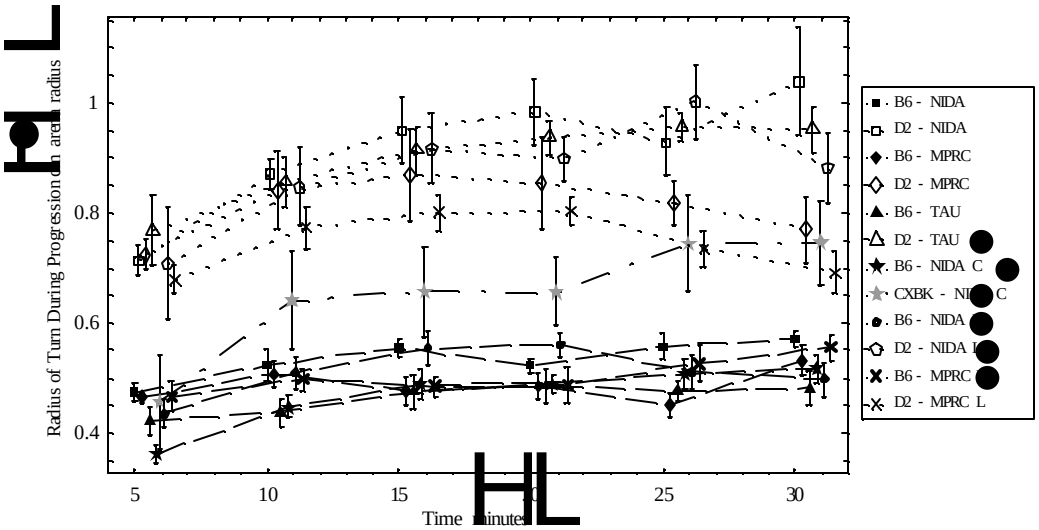

Fig. 3: Radius of turn during progression in the six experiments (group means ± SE) in time bins of 5 minutes. Dark symbols with dashed lines: B6 groups. Open symbols with dotted lines: D2 groups. Grey star with dash-dot line: a single CXBK group. Experiment captions (NIDA, MPRC, etc.) as in Fig. 1.

5. Comparison with a photobeam test

The multi-laboratory approach first taken by Crabbe et al. (1999) prompted our efforts to explore improved tests as a mean for achieving higher replicability. To this end, we have furthered their discussion by directly comparing the replicability of our results with the replicability of the locomotor behavior results measured by AccuScan™ Digiscan photobeam systems in complete darkness, as published in their website at http://www.albany.edu/psy/obssr3/ . This is especially needed since the SEE test is not yet standardized and commercialized as the photobeam tests, requiring both more space to be conducted and motivation for replacing the familiar with the less familiar. It is thus reasonable to ask if the gain in performance is large enough to justify such an investment. Moreover, the SEE test cannot be said to be more replicable than the photobeam system merely because this study did not find significant genotype × lab interactions while the Crabbe at al (1999) did. This difference could be attributed a higher power of the second study to detect interactions, since it included more genotypes and larger (when collapsed over sex) group size.

In order to equalize strains, sexes, session duration and group size we used only the photobeam test B6 and D2 males in all three laboratories, only the first 15 minutes of each SEE test session, and only 8 randomly chosen animals from each SEE test group. We compared the genotype differences found in three corresponding analogous endpoints: Horizontal Distance, Center Time and Number of Movements in the photobeam system with the respective Distance Traveled, Center Time, and Number of Progression Segments in the SEE test. No transformations were used in this comparison with any of the endpoints.

Fig. 4 displays the results of this comparison with the columns representing the proportion of variance attributed to genotype, laboratory, genotype × laboratory interaction and individual animals. The variances attributed to the genotype are seen to be large for all three measures in the SEE test, and are highly statistically significant ($p<0.0001$). The variance attributed to the interaction in the SEE test is too small to estimate graphically in fig. 4 (all $p > 0.29$). It was 1.2% in the Distance Traveled (vs.

11.5% in the photobeam test's Horizontal Distance, where $p < 0.05$), 0.54% in the Center Time (vs. 2.57 % in the photobeam test's Center Time) and 0.59% in the Number of Progression Segments (vs. 4.6% in the photobeam test's Number of Movements. The lab effects were also smaller in the SEE test than in the Photobeam test, and generally less significant (in the SEE test: $p=0.05$ for Center Time and Number of Progression Segments, $p>0.49$ for Distance traveled; in the photobeam test: $p<0.01$ for Horizontal Distance, $p=0.0001$ for Center Time, and $p>0.16$ for Number of Movements). The locomotor behavior test with SEE analysis thus increased genotypic effect and decreased laboratory, interaction and individual effects several folds. Note that, in contrast with the photobeam test, this advantage was achieved without any special effort to equalize housing and testing conditions (table 1).

The third comparison of the photobeam system's number of movements with SEE's number of progression segments (Fig. 4, bottom two graphs) also illustrates some of the problem with using significance in the two-way ANOVA for assessing across-lab replicability. The genotype difference in the photobeam system is actually significant while the laboratory and interaction effects are not. With SEE the genotype effect is much more significant but the lab effect is also marginally significant. This is mainly due to the higher discrimination power of SEE, as is evident from smaller group-variances in the graph and smaller proportion of individual variance in the column. Since F-ratios of all factors in the standard ANOVA are computed relative to this pooled within-group variance, SEE did better with detecting the genotype difference, but also with detecting the lab difference. Assessing replicability by lab and interaction significance without considering also the genotype significance thus penalizes methods with higher discrimination power, and in this case might have led one to believe that SEE analysis was less replicable across laboratories, although the graphs shows clearly that the opposite is true.

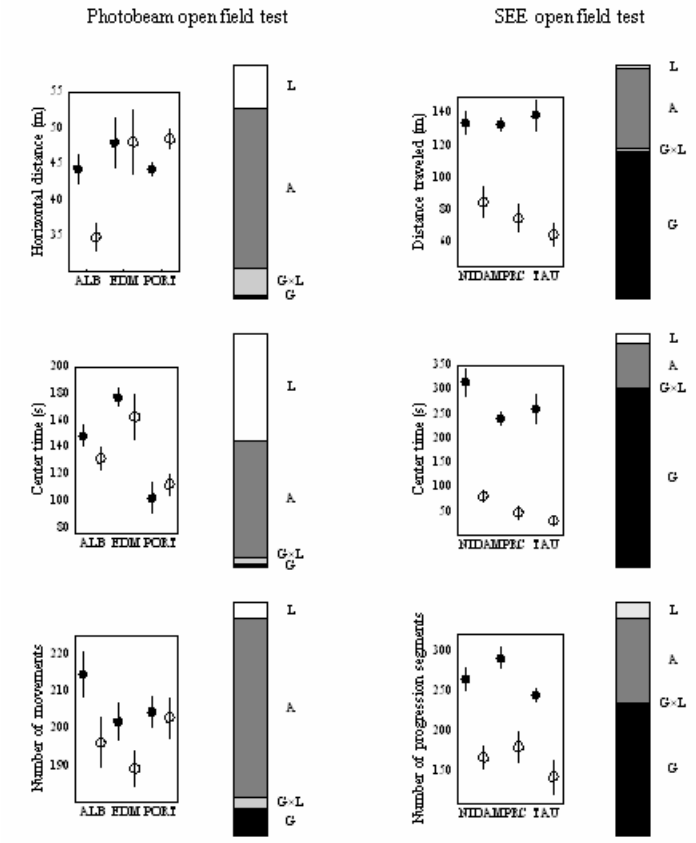

Fig. 4: Comparison between a photobeam test (left) and the SEE test (right) of the differences detected between B6 and D2 across three laboratories in three corresponding analogous endpoints: Horizontal distance vs. Distance traveled (top), Center time vs. Center time (middle), and Number of movements vs. Number of progression segments (bottom). Closed circles: B6 groups. Open circles: D2 groups. Error bars represent Standard Errors. Photobeam results are from a study described in Crabbe et al., 1999. Laboratories: Albany (ALB), Edmonton (EDM), Portland (PORT). Laboratories in SEE test: NIDA, MPRC, TAU. Columns at the right represent the proportion of variance attributed to each factor. L: laboratories, G: genotypes, G×L: genotype × laboratory interaction, A: individual animals. Note that the interaction parts G×L in the Center Time and Number of Progression Segments in the SEE test (middle and bottom right) are too small to be noticed in their columns. Data for both tests are based on 15 min session duration and n = 8 males, and no transformations are used.

It is important to note that by comparing the performances in endpoints such as Horizontal Distance in the photobeam system and Distance Traveled in the SEE test we do not imply that these endpoints necessarily measured the same behavior in the two tests, in light of the considerable differences in test conditions. In fact, the photobeam results included many additional endpoints, such as vertical activity and number of movement, number of clockwise and counter clockwise revolutions and stereotypy count. None of these, however, detected much larger genotype differences than the three endpoints shown in fig 4. The only differences that were highly replicable were the advantage of the B6 in several endpoints, such as the horizontal distance and number of movements, during the first 5 minutes of the session only. The same pattern can be seen with the distance traveled measured in a single laboratory experiment by Jones at al. (1993). This agrees with our results in the measure of Latency to Half Maximum Speed, suggesting that B6 are faster starters than D2. It appears that in the photobeam systems, this initial advantage of the B6 disappears after the first 5 minutes, when the two strains both covered the entire small arena. In the large arena used in this study, however, genotype differences were generally maintained, as is illustrated in figures 1, 2 and 3, even after 30 minutes.

6. Discussion

The B6 and D2 are two of the most commonly used inbred strains of laboratory mice. Consequently, many behavioral differences between them have been reported in the literature. The results of most reports, however, have not been methodically corroborated in more than one laboratory. The findings of Crabbe et al (1999) suggest that such differences could well prove to be idiosyncratic to the specific laboratory. The purpose of the experiments described in this report was to explore the heritability and replicability of endpoints derived from ethologically based studies of exploratory behavior (Eilam and Golani, 1989; Eilam et al, 1989; Golani et al., 1993; Tchernichovski et al., 1998; Drai et al., 2000; Kafkafi et al, 2001; Drai et al., 2001; Benjamini et al., 2001; Kafkafi et al., submitted).

Replicable differences between the two genotypes were found in several different aspects of locomotor behavior. B6 mice traveled longer distances, performed more stops (and, by definition, more progression segments) and spent less time in lingering. The increase in distance traveled by the B6 mice was achieved by increasing the number of progression segments, not by increasing the typical length of progression segments. D2 mice, however, had higher maximal speeds and accelerations. This combination of differences is counter-intuitive to a "general activity" view of locomotor behavior. According to such a view, optional measures of activity, such as the distance traveled or the number of beam breaks in different photobeam arrangements, are most likely to be correlated, and thus the exact definition of an activity measure is not that crucial. "Activity", as measured by a variety of hardware and algorithms, is the endpoint reported in most locomotor behavior studies, and is used to assess the effect of many drugs and treatments. Our results suggest, however, that it artificially combines several different aspects of behavior that are not necessarily correlated in a trivial way.

B6 mice spent much longer time away from the wall than D2 mice. They typically took only several seconds from the start of the session to reach half of the maximal speed attained in the session, while D2 typically took a minute or so. B6 mice progression was less straight, both in relation to space (turning radius) and time (turning rate). Most

differences found over the whole 30min session were also prominent in most of the 5min time bins of the session.

A conservative estimate of the broad-sense heritability was higher than 50% with 7 endpoints out of the 17 tested in this study, despite being conducted over three laboratories with slightly different conditions. Two complementary approaches for mapping genes affecting behavioral traits are Quantitative Trait Loci (QTL, e.g., Flint et al., 1995; Turri, Talbot, Radcliffe, Wehner and Flint, 1999; Belknap at al., 2001) and mutagenesis (e.g., Takahashi, Pinto and Vitaterna, 1994; Nolan at al., 2000). Mutants and recombinant inbred strains (the BXD strains) derived from B6 and D2 play important role in both QTL and mutagenesis studies conducted in mice. With both approaches, the discrimination power and replicability of phenotyping crucially affect the quality of the results. Both approaches may thus gain from employing the SEE test.

Interestingly, mice of both strains were not more active when tested during the dark phase of the photoperiodic cycle than during the light phase. This, however, may be a result of the light in the testing arena. The contrast between the dark housing room and the lighted arena might have had some inhibiting effect that balanced the higher activity expected during the dark phase. Note that it is not possible to conclude about the effect of the photoperiod phase or saline injection since they were not manipulated within a single experiment in a single laboratory. The overall results strongly suggest, however, that a considerable proportion of our endpoints were much more sensitive to genotype differences than to conditions such as photoperiod phase and saline injections.

Since all experiments in this study included only two strains each, it was not possible to measure correlations between endpoints across strains. We currently study this issue in a multi-lab experiment involving 10 strains. While we expect certain endpoints to be usually correlated, the particular behavioral phenotype that characterizes certain genotypes may well be the absence of such a correlation. The CXBK results from experiment NIDA/C supports this possibility. For example, it can be argued that the Radius of Turn during progression (fig. 3) is probably inversely correlated with the Center Time (fig. 2), since animals walking mostly along the wall are likely to have a progression radius similar to the arena radius, and indeed the D2 mice, who spent most of their time near the wall, also had a progression radius consistently close to the arena

radius, while the B6 mice, who spent more time in the center, consistently had a progression radius of about half the arena radius. Hypothetically, however, an animal can walk near the wall in tight arcs, or travel through the center in straight progression segments. Note that the Center Time of the CXBK group in experiment NIDA/C seems not to differ from that of the D2 groups in the other experiments (fig. 2), but their Radius of Turn was much smaller than that of the D2 during the first half of the session (fig. 3), closer to that of the B6.

Multi-lab experiments are currently assessed using the standard method of genotype × laboratory two-way ANOVA. In this model, large size and high significance of lab effects, and even more so of the genotype × laboratory interaction effects, are considered as indicating a replicability problem (Crabbe et el., 1999; Wahlsten, 2001). Such significance, however, should not be considered separately from the size and significance of the genotype effect. Methods with higher discrimination power are likely to increase significance of all effects because of the small within groups variance. Good replicability should thus be indicated by a proportion of variance due to genotype that is several folds larger than that due to the laboratory and especially the interaction (e.g., fig. 4). This problem may be solved by using a mixed model instead of a fixed model ANOVA, with the laboratories (and therefore also the interaction) regarded as a random factor. We are currently engaged in adapting this strategy to the problem at hand.

In general, the SEE locomotor behavior test was much more sensitive to the differences between B6 and D2 than the standard locomotor photobeam test, as demonstrated here using a direct comparison with data published from a multi-lab study by Crabbe at al. (1999), while the sensitivity to the laboratory effect and genotype × laboratory interaction was generally lower. It is difficult to determine how much of the increased performances should be attributed to the different setup, to the different tracking system, to the robust smoothing or to the more sophisticated analysis. With most endpoints it was probably an effect of more than one of the above factors and their interaction. In addition to the much larger arena in this study (210 – 250 cm vs. 45 cm) it is also important to note that the animals in the photobeam test were run in complete darkness, as opposed to usual room lights in the SEE setup. In the case of the Center Time especially it is probable that both the large arena and the lights were the main

factors that uncovered the big difference between the strains. With rats we noticed that darkness typically lowers the range of speeds used by the animal considerably (unpublished data), thus limiting the its behavioral repertoire and lowering the quality of SEE analysis, which is highly dependent on the speed. Arena size and light level, as most properties of the setup in this study, were deliberately chosen to increase the resolution and the performance of the analysis.

In order to encourage the use of this new test we developed a simple stand-alone program that automatically performs the analysis described in this paper, including the smoothing, segmentation and computation of all the endpoints described above. This program, as the SEE notebook itself, are available from the authors. It of course remains to be shown that the high performances of the SEE test can be generalized to other common and important genotypes, an issue that we are currently studying. Note also that SEE analysis can be easily upgraded. New endpoints can be readily developed in SEE and current endpoints can be updated, based on ongoing results (Kafkafi et al, submitted). In our experience the study of additional strains tends to promote this process because each new strain displays more clearly some patterns that highlight additional properties of locomotor behavior. The performance of the SEE test is thus likely to be further improved in the future by the diverse efforts of the research community.

Standardization is the usually advocated remedy for the problem of phenotyping replicability across laboratories. It is estimated, however, to require a substantial and highly-coordinated effort by many laboratories (Wahlsten, 2001), and the level of standardization needed for satisfactory replicability is yet to be demonstrated. Moreover, behavioral tests that are sensitive to any small change in protocol, even within the same laboratory, are likely to make research much more difficult. In contrast, the high replicability of the SEE test in this study was achieved despite several differences in tracking parameters and using only mild standardization of housing conditions. Within labs, experiments conducted during the light instead of the dark cycle and with saline injections did not blur most genotype differences, and did not significantly interact with any of them. This suggests that tests and measures that are specifically designed to capture ethologically-relevant behavior patterns are likely to be more resistant to

environment manipulations, and thus constitute a fruitful approach for behavioral phenotyping.

Acknowledgments: The SEE software is available from the authors upon request. It requires, however, the *Mathematica*™ programming environment by Wolfram Research, Inc. A simpler, stand-alone program for automatic smoothing, segmentation and calculation of endpoints is also available. We thank Noldus Information Technology for the use of their EthoVision® system in Tel-Aviv University. This research is supported by NIH grant #1 R01 NS40234-01.

References

Belknap, J. K., Hitzemann R, Crabbe J. C., Phillips, T. J., Buck, K. J. & Williams R. (2001). QTL analysis and genomewide mutagenesis in mice: complementary genetic approaches to the dissection of complex traits. *Behavior Genetics* 31(1), 5-15.

Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B*, **57,** 289-300.

Benjamini, Y., Liu, W. (1999). A step-down multiple hypothesis testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* 82(1-2), 163-170.

Benjamini, Y., Drai, D., Elmer, G. I., Golani, I. & Kafkafi, N. (2001). Controlling the false discovery rate in behavior genetics research. *Behavioral Brain Research* 125, 279-284.

Bolivar, V. J., Caldarone, B. J., Reilly, A. A & Flaherty, L. (2000). Habituation in the open field: a survey of inbred strains and F1 hybrids. *Behavior Genetics* 30 (4), 258-293.

Cabib, S. & Bonaventura, N. (1997). Parallel strain-dependent susceptibility to environmentally-induced stereotypies and stress-induced behavioral sensitization in mice. *Physiology Behavior* 61(4), 499-506.

Cabib, S. (2002) The contribution of studies in inbred strains of mice to understanding of a hyperactive phenotype. *Behavioral Brain Research* 130, 103-109.

Cleveland, W. S. (1977) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistics Association* 74, 829-836.

Cools, A.R., Ellenbroek, B.A., Gingras, M. A., Engbersen, A. & Heeren D (1997) Differences in vulnerability and susceptibility to dexamphetamine in Nijmegen high and low responders to novelty: a dose-effect analysis of spatio-temporal programming of behavior. *Psychopharmacology*, 132, 181-187

Crabbe, J.C., Wahlsten, D., & Dudek, B.C. (1999). Genetics of mouse behavior: Interactions with lab environment. *Science*, 284, 1670-1672.

Crawley, J. N. (2000). What's Wrong with My Mouse? Behavioral Phenotyping of Transgenic and Knockout Mice. NY: Wiley-Liss.

Crawley, J.N., Belknap, J.K., Collins, A., Crabbe, J.C., Frankel, W., Henderson, N., Hitzemann, R.J., Maxson, S.C., Miner, L.L., Silva, A.J., Wehner, J.M., Wynshaw-Boris, A., and Paylor, R. (1997). Behavioral phenotypes of inbred mouse strains. *Psychopharmacology*, 132, 107-124.

Drai, D., Benjamini, Y. & Golani, I. (2000) Statistical discrimination of natural modes of motion in rat exploratory behavior. *Journal of Neuroscience Methods*, 96, 119-131.

Drai, D,, Kafkafi, N., Benjamini, Y., Elmer, G. I. & Golani, I. (2001). Rats and mice share common ethologically relevant parameters of exploratory behavior. *Behavioral Brain Research*, 125 (1-2), 133-140.

Drai, D. & Golani I (2001). SEE: a tool for the visualization and analysis of rodent exploratory behavior. *Neuroscience and Biobehavioral Reviews* 25 (5), 409-426.

Eilam, D., Golani, I. & Szechtman, H. (1989) D2 Agonist quinripole induces perseveration of routes and hyperactivity but no perseveration of movements. *Brain Research*, 460, 255-267.

Eilam, D. & Golani I (1989) Home base behavior of rats (Rattus norvegicus) exploring a novel environment. *Behavioral Brain Research* 34, 199-211.

Elmer, G. I., Gorelick, D. A., Goldberg, S. R. & Rothman R. B. (1996) Acute sensitivity vs. context-specific sensitization to cocaine as a function of genotype. *Pharmacology Biochemistry and Behavior* 53, 623-628.

Flint, J., Corley, R., DeFries, J. C., Fulker, D.W., Gray, J.A., Miller, S. & Collins, A.C. (1995). A simple genetic basis for a complex psychological trait in laboratory mice. *Science*, 269(5229):1432-5.

Gerlai, R. (1996) Molecular genetic analysis of mammalian behavior and brain processes: caveats and perspectives. *The Neurosciences*, 8: 153-161.

Gingras, M.A. & Cools, A.R. (1997) Different behavioral effects of daily or intermittent dexamphetamine administration in Nijmegen high and low responders *Psychopharmacology* 132: 188-194.

Golani, I. (1992). A mobility gradient in the organization of vertebrate movement: The perception of movement through symbolic language. *Behavior and Brain Sciences*, 15,249-308.

Golani, I., Benjamini, Y., & Eilam, D. (1993). Stopping behavior: constraints on exploration in rats (Rattus norvegicus). *Behavioral Brain Research*, 53,21-33.

Gorris, L. G. M. & Abeelen, J. H. F. van (1981) Behavioral effect of (-)naloxon in mice from four inbred strains. *Psychopharmacology* 74, 355-359.

Hatcher J. P., Jones D. N. C., Rogers, D. C., Hetcher, P. D., Reavill C., Hagan, J. J. & Hunter, A. J. (2001) Development of SHIRPA to characterize the phenotype of gene-targeted mice. *Behavioral & Brain Research* 125, 43-47.

Hen, I., Sakov, A., Kafkafi, N., Golani, I. & Benjamini, Y. (submitted), The dynamics of spatial behavior: How can robust smoothing techniques help?

Jones, B.C., Reed, C. L., Radcliffe, R. A. & Erwein, G. (1993) Pharmacogenetics of cocaine: I. Locomotor activity and self-selection. *Pharmacogenetics*, 3, 182-188.

Kafkafi, N. (in oress). Extending SEE for large-scale phenotyping of mouse open-field behavior. *Behavioral Research Methods, Instruments & Computers.*

Kafkafi, N., Mayo, C. L., Drai, D., Golani, D. & Elmer, G. I. (2001). Natural segmentation of the locomotor behavior of drug-induced rats in a photobeam cage. *Journal of Neuroscience Methods*, 109, 111-121.

Kafkafi, N., Pagis, M., Lipkind, D., Mayo, C. L., Benjamini, Y., Elmer, G. I. & Golani, D. (in press). Darting behavior: a quantitative movement pattern for discrimination and replicability in mouse locomotor behavior. *Behavioral Brain Research.*

Logue, S. F., Owen, E.H., Rasmussen, D. L. & Wehner, J. M. (1997) Assessment of locomotor activity, acoustic and tactile startle, and prepulse inhibition of startle in inbred mouse strains and F1 hybrids: implications of genetic background for single gene and quantitative trait loci analyses. *Neuroscience*, 80(4), 1075-86.

Nolan, P. M., Peters, J., Strivens, M., Rogers, D., Hagan, J., Spurr, N., Gray, I. C., Vizor, L., Brooker, D., Whitehill, E., Washbourne, R., Hough, T., Greenaway, S., Hewitt, M., Liu, X., McCormack, S., Pickford, K., Selley, R., Wells, C., Tymowska-Lalanne, Z., Roby, P., Glenister, P., Thornton, C., Thaung, C., Stevenson, J. A., Arkell, R., Mburu, P., Hardisty, R., Kiernan, A., Erven, A., Steel, K.P., Voegeling, S., Guenet, J. L., Nickols, C., Sadri, R., Nasse, M., Isaacs, A., Davies, K., Browne, M., Fisher, E. M., Martin, J., Rastan, S., Brown, S. D., Hunter, J. (2000) A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nature Genetics* 2000 25(4): 440-3.

Paigen, K., Eppig, J. T. (2000). A mouse phenome project. Mamm Genome. 11(9), 715-7.

Paulus, M. P. & Geyer, M. A. (1993) Three independent factors characterize spontaneous rat motor activity. *Behavioral Brain Research* 53(1-2), 11-20.

Rocha, B. A., Odom, L. A., Barron, B. A., Ator, R., Wild, S. A., Forster, M. J. (1998) Differential responsiveness to cocaine in C57BL/6J and DBA/2J mice. *Psychopharmacology (Berl)*, 1998, 138, 82-8.

Rogers, D. C., Jones, D. N., Nelson, P. R., Jones, C. M., Quilter, C. A., Robinson, T. L., Hagan, J. J. (1999) Use of SHIRPA and discriminant analysis to characterise marked differences in the behavioural phenotype of six inbred mouse strains. Behav Brain Res 105(2): 207-1.

Spink, A. J., Tegelenbosch, R. A. J. & Buma, M. O. S, Noldus, L. P. J. J. (2001). The EthoVision video tracking system – a tool for behavioral phenotyping of transgenic mice. *Physiology and Behavior*, 73(5), 731-734.

Szechtman H, Culver K, Eilam D (1999) Role of dopamine systems in obsessive-compulsive disorder (OCD): Implications from a novel psychostimulant-induced animal model. *Pol J Pharmacol* 51 (1): 55-61.

Takahashi, J. S., Pinto, L. H. & Vitaterna, M. H. (1994). Forward and reverse genetic approaches to behavior in the mouse. *Science*, 264(5166):1724-33.

Tchernichovski, O., Benjamini, Y. & Golani, I. (1996) Constraints and the emergence of freedom in the ontogeny of rat exploratory behavior. *Behaviour* 133, 519–539.

Tchernichovski, O. & Golani, I. (1995) A phase plane representation of rat exploratory behavior. *Journal of Neuroscience Methods* 62(1-2), 21-7.

Tchernichovski, O., Benjamini, Y. & Golani, I. (1998), The dynamics of long term exploratory behavior in the rat, part I. *Biological Cybernetics* 78 (6), 423-432.

Tirelli, E., Witkin, J. M. (1994). Verticalization of behavior elicited by dopaminergic mobilization is qualitatively different between C57BL/6J and DBA/2J mice. *Psychopharmacology* (Berl). 116(2), 191-200.

Tolliver, B. K., Carney, J. M. (1995). Locomotor stimulant effects of cocaine and novel cocaine analogs in DBA/2J and C57BL/6J inbred mice. *Pharmacology Biochemistry Behavior*, 50(2), 163-9.

Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading, Mass.

Turri, M. G., Talbot, C.J., Radcliffe, R. A., Wehner, J. M. & Flint, J. (1999) High-resolution mapping of quantitative trait loci for emotionality in selected strains of mice. *Mammlian Genome* (1999) 11, 1098-101.

van der Staay, F. J. & Steckler, T. (2002). The fallacy of behavioral phenotyping without standardization. *Genes, Brain and Behavior* 1, 9-13.

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

Wahlsten, D., Crabbe, J. C. & Dudek, B. C. (2001). Behavioral testing of standard inbred and 5HT1B knockout mice: implications of absent corpus callosum. Behavioral Brain Research 125, 23-32.

Wallace, D. G., Hines, D. J. & Whishaw, I. Q. (2002). Quantification of a single exploratory trip reveals hippocampal formation mediated dead reckoning. *Journal of Neuroscience Methods*, 30, 113.

Whishaw, I.Q., Cassel, J. C., Majchrzak, M., Cassel, S., Will, B. (1994). "Short-stops" in rats with fimbria-fornix lesions: evidence for change in the mobility gradient. *Hippocampus* 4 (5), 577-582.

Whishaw, I. Q., Hines D.J. & Wallace D.G. (2001) Dead Reckoning (path integration) requires the hippocampal formation: evidence from spontaneous exploration and spatial learning tasks in light (allothetic) and dark (idiothetic) tests. *Behavioral Brain Research* **127**, 49-69.

Womer, D. E., Jones, B. C., Erwin, V. G. (1994). Characterization of dopamine transporter and locomotor effects of cocaine, GBR 12909, epidepride, and SCH 23390 in C57BL and DBA mice. *Pharmacology, Biochemistry and Behavior,* 48(2), 327-335.

Wurbel, H. (2000) Behaviour and the standardization fallacy. *Nature Genetics* 26(3), 263.

Wurbel, H. (2002) Behavioral phenotyping enhanced – beyond (environmental) standardization. *Genes, Brain and Behavior* 1, 38.