Received: 8 November 2019 | Revised: 13 May 2020 | Accepted: 20 May 2020

DOI: 10.1111/gbb.12676

ORIGINAL ARTICLE

Genes, Brain and Behavior

Mouse model systems of autism spectrum disorder: Replicability and informatics signature

Patricia Kabitzke[1,2] | Diana Morales[1,3] | Dansha He[1] | Kimberly Cox[1] | Jane Sutphen[1,4] | Lucinda Thiede[1,5] | Emily Sabath[1,6] | Taleen Hanania[1] | Barbara Biemans[7] | Daniela Brunner[1,8]

[1]PsychoGenics, Inc., Paramus, New Jersey

[2]The Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA

[3]Pfizer, Pearl River, NY

[4]Louisiana State University Health Sciences Center, New Orleans, LA

[5]Boehringer Ingelheim, Ridgefield, CT

[6]JRS Pharma, Patterson, NY

[7]Roche Innovation Center Basel, Basel, Switzerland

[8]Department of Psychiatry, Columbia University, New York, NY

Correspondence
Patricia Kabitzke, The Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, 75 Ames Street, Cambridge, MA.
Email: pkabitzke@gmail.com

Funding information
PsychoGenics; Roche; Simons Foundation

Abstract

Phenotyping mouse model systems of human disease has proven to be a difficult task, with frequent poor inter- and intra-laboratory replicability, particularly in behavioral domains such as social and cognitive function. However, establishing robust animal model systems with strong construct validity is of fundamental importance as they are central tools for understanding disease pathophysiology and developing therapeutics. To complete our studies of mouse model systems relevant to autism spectrum disorder (ASD), we present a replication of the main findings from our two published studies of five genetic mouse model systems of ASD. To assess the intralaboratory robustness of previous results, we chose the two model systems that showed the greatest phenotypic differences, the *Shank3/F* and *Cntnap2*, and repeated assessments of general health, activity and social behavior. We additionally explored all five model systems in the same framework, comparing all results obtained in this three-yearlong effort using informatics techniques to assess commonalities and differences. Our results showed high intra-laboratory replicability of results, even for those with effect sizes that were not particularly large, suggesting that discrepancies in the literature may be dependent on subtle but pivotal differences in testing conditions, housing enrichment, or background strains and less so on the variability of the behavioral phenotypes. The overall informatics analysis suggests that in our behavioral assays we can separate the set of tested mouse model system into two main classes that in some aspects lie on opposite ends of the behavioral spectrum, supporting the view that autism is not a unitary concept.

KEYWORDS

16p11.2, autism, behavior, Cacna1c, Cntnap2, informatics, mouse model systems, preclinical, replication, Shank3

Patricia Kabitzke and Daniela Brunner contributed equally to this study.

1 | INTRODUCTION

Autism spectrum disorder (ASD) has been linked to gene copy number and single nucleotide variation, findings that lead to a number of mouse model systems with etiological validity.[1-4] We previously completed and published two studies from a considerable effort using standard and informatics methods to phenotype five mouse model systems of ASD. The first study on the 16p11.2 and *Cntnap2* deletion mouse model systems,[5] and the second complementary study on two distinct *Shank3* knockout (KO) mouse model systems of Phelan-McDermid disorder, Feng's *Shank3*^{tm2Gfng} (hereafter *Shank3/F*) and Jiang's *Shank3*^{tm1Yhj} (hereafter *Shank3/J*) and a model of Timothy syndrome, the *Cacna1c* heterozygous (HET) mouse model system.[6]

Replication and robustness. Numerous papers, including our own,[5-12] have pointed at the difficulty of replicating results from preclinical studies, and offer various explanations including the likelihood of false positives because of small sample sizes. Whereas poor replicability could be a sign of the inherent variability of a model system, or simple absence of scientific rigor, it could also point to a more difficult to solve problem, namely lack of generalizability, which may in turn cast doubt on the translational value of these animal models. To address these potential issues, in this study we focus on two separate questions: Can we replicate the main results of our published studies? Do the five model systems lie on a behavioral continuum, or do they present idiosyncratic signatures?

In our previous publications, we chose to present all the results of our broad battery of behavioral endpoints and their independent analyses. Rather than using statistical corrections to reduce the experiment-wise type I error, we chose to do a real replication of the main findings. This route ensures fair assessment of effect sizes that may not be large yet may be of high scientific interest, which would be obliterated by simple statistical corrections. It should be noted too that *p*-value corrections should only be applied to results that are considered to be of the same "family," that is, tests that assess the likelihood of falsehood of the same hypothesis.[13] For example, within a phenotyping project, one could use two different measures of activity to assess if the model system has an abnormal motor pattern as compared with its control. These two measures are testing the same hypothesis, namely, "does this mouse model exhibit abnormal motor activity." Other aspects of the phenotype (e.g., social, cognitive, etc.) should be considered under other null hypotheses. It could be parsimonious, therefore, to assume that these other aspects of the phenotype have different underlying biological underpinnings better tested under independent hypotheses, which would not belong to the same "hypothesis family."

It is also important to differentiate between an exploratory analysis, such as a broad phenotyping effort, and a focused experimental exercise, such as a drug test for which a primary endpoint measure should be defined a priori.[14] Exploratory analyses should be taken as tentative poking of the phenotypical landscape to find where the signal resides and should be followed by replication. Thus, replication rather than statistical corrections should be the rule, contrary to the common practice to take single studies as proof of a phenomenon.[15] Taking this to heart, we bred independent groups of mice, from the *Shank3/F* and *Cntnap2* model systems and chose to repeat the exploratory analyses in the different batches to see whether phenotypic differences found in our published studies would replicate.[5,6] The SmartCube platform provided a comprehensive analysis. This test together with the reciprocal social interaction test and the urine-exposure open field test assessed behavior in a social setting. We added the standard open field test, not done before, to provide an assessment of activity levels in a nonsocial setting. Body weight measures helped to probe general health and our ability to replicate our own results. The objective of the studies here, for the first time, was to replicate the previous studies. Therefore, the methods of both study sets, from husbandry to the experimental details, are identical to the maximum degree possible.

ASD signatures: common elements or idiosyncratic features? A comprehensive analysis of behavior provides a panorama of complex function reflecting the downstream effects of the genetic insult or manipulation. Phenotypic analyses of model systems of ASD typically probe mouse functional domains putatively reflecting the core ASD triad, namely, social function, communication and repetitive behaviors.[16-19] In addition to those standard tests, we previously used additional proprietary platforms, SmartCube and NeuroCube, to discover previously unknown phenotypes. SmartCube captures 1/2 million points related to posture, position, trajectory, activity, behavioral sequences, response to stimulation and individual behaviors. This dataset is reduced to a several hundred informative features, constituting a rich content dataset amenable to machine learning mining techniques.[20-22] We took advantage of these features to run a comprehensive analysis of the results from the original SmartCube studies covering the five model systems and added the new results from the tests presented in this paper. We asked the following questions: Do the model systems lie in a complex plane or could their similarities and differences be well explained using a few dimensions? If the model systems are relatively similar, can we classify one mouse as being a mutant or a wild type (WT) by training the classifier on a different model system? Which model systems are confused with each other and which are distinct? In this comprehensive analysis, are replicate results similar to the original? These informatics exercises attempt to find core characteristics of the mutant mice that may be relevant to ASD-defining features in an agnostic and unbiased way.

2 | MATERIALS AND METHODS

2.1 | Ethics statement

PsychoGenics is an AAALAC accredited facility (Unit Number: 001213) and work is performed under PHS OLAW Assurance #A4471-01. This study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. The protocol was approved by the Committee on the Ethics of Animal Experiments of PsychoGenics. All efforts were made to minimize suffering and maximize animal welfare.

2.2 | Subjects

2.2.1 | *Shank3/F* line

Breeders: A cohort of 60 Het female and 30 Het male mice (JAX stock #017688, B6.129-*Shank3*<tm2Gfng>/J) was provided by The Jackson Laboratory at 11–13 weeks of age. *Line development prior to arrival at JAX*: Exons 13–16 of SH3/ankyrin domain 3 were replaced with a neomycin resistance (neo) cassette. The construct was electroporated into (129X1/SvJ × 129S1/Sv)F1-Kitl^{+}-derived R1 embryonic stem (ES) cells. Correctly targeted ES cells were injected into C57BL/6 blastocysts and the resulting chimeric males were bred to C57BL/6J females. The offspring were then intercrossed for five generations and maintained on a mixed C57BL/6J × 129 background prior to sending to The Jackson Laboratory. *Line maintenance at JAX*: Upon arrival, mice were additionally backcrossed to C57BL/6J inbred mice (Stock No. 000664) using a marker-assisted, speed congenic approach to establish this congenic line. Genome Scan results indicated that *Shank3* Feng breeders were fully congenic for C57BL/6J.

2.2.2 | *Cntnap2* line

Breeders: A cohort of 60 Het female and 30 Het male mice (JAX stock #017482, B6.129(Cg)-Cntnap2^{tm1Pele}/J), backcrossed to C57BL/6J for more than 10 generations, was provided by The Jackson Laboratory at 8–13 weeks of age.

2.2.3 | Breeding scheme

Mice were set in trios (2 females:1 male) and left together for 3 days. Breeders were 14–16 weeks of age when bred for the *Shank3/F* line and 11–16 weeks of age when bred for the *Cntnap2* line. See Methods S1, supporting information for genotyping results. Sex and genotype ratios were unbiased in all animal model systems. Animals were weaned at 4 weeks of age. Breeding success and pup survival were similar among all groups. All testing was done in male mice in order to allow for comparisons to findings obtained in the original studies. Age- and genotype-matched male (~P45) unfamiliar littermates were used as stimulus mice for the reciprocal social interaction test. These stimulus mice were used for the Grooming test at ~P50 to maintain the same sequence of testing used in other studies, where tests were performed longitudinally in the same animals (see Table 1).

TABLE 1 Time-course of tests

Test	~Age
Reciprocal social interaction	P45
Grooming test[a]	P50
Urine-exposure open field	P60
Standard open field	P75
SmartCube	P90

[a] Stimulus mice from the reciprocal social interaction test used—see Methods.

2.3 | General procedures

Mice were housed in OptiMice cages (Animal Care Systems, Inc.) on a 12/12 h light/dark cycle where 20–23°C room temperature and a relative humidity of 50% were maintained. Chow and water were provided ad libitum for the duration of the study and mice were checked twice daily for general health, food and water. Husbandry included enrichment, namely, shredded paper (Enviro-Dri; W.F. Fisher & Son Inc., NJ; Product 08ENV3000) and a nylabone (Bio-Serv, NJ; Product K3200). Breeders also received an amber-colored polycarbonate igloo for extra enrichment (Bio-Serv, NJ; Product K3328). On P0, pups were tattooed using nontoxic ink applied under the skin of their toe and a tail snip sample was taken for genotyping (see Methods S1). Once the genotype results were available (around P2), the litter size was culled down to n = 8 pups, removing mainly females via decapitation. Thus, litter size range was 3–8 after culling (average breeding yield was 6.9 pups/litter in the *Cntnap2* and 7.5 pups/litter in the *Shank3/F* mice). Animals were weaned in 2:2 mixed-genotype (KO and WT only), same-sex groups of four with shredded paper, one nylabone, and one polycarbonate amber-colored tunnel (3 7/8″ long × 2″ inside diameter; Bio-Serv, NJ: Product K3323) per cage. Testing occurred between 10:00 and 17:30 in separate experimental rooms, unless stated otherwise. Tests were conducted blind to genotype and the sequence of testing is indicated in Table 1. Euthanasia was required at first signs of illness, severe dehydration and/or emaciation defined as a loss of greater than 20% body weight with failure to regain weight while on a free feeding regimen, lack of righting reflex, catalepsy, morbidity, increased repetitive convulsions, respiratory distress or hemorrhage. Although no mice were euthanized for any of these reasons, mice were sacrificed at the end of the study using methods consistent with recommendations of the 2013 American Veterinary Medical Association Guidelines on Euthanasia. Carbon dioxide gas was used, and euthanasia was verified by observation of breathing and color of the animal, and by palpation of the heart in addition to the loss of reflexes. Required further verification of death was accomplished via cervical dislocation.

A goal of the project was to replicate the results already published, and therefore many of the experimental designs followed the original studies.

2.3.1 | Reciprocal social interaction test

Subject animals were isolated for 2 days before testing to increase social interest, determined after extensive literature review and pilots in WT mice and described before.[5,6] The day before testing, subject animals were individually habituated to the testing apparatus for 10 min and same-genotype male stimulus animals (that were unfamiliar and not of the same litter) were separately habituated to the apparatus in pairs. The day of testing, subject animals were placed in the testing apparatus for 5 min before an age- and weight-matched male stimulus animal was placed into the chamber. Behavior and ultrasonic vocalizations (USVs) were recorded for the pair for a total of 10 min. Ethovision XT (Noldus Information Technology, Wageningen, Netherlands) was used to measure distance, proximity and interaction between animals. We defined close proximity as the center of the bodies being between 1 and 5 cm apart and interactions when the distance from the nose to the other mouse body (nose, center and tail) was less than 1.5 cm. Interactions of the mice that were active (one mouse sniffing any part of the other mouse body), passive (recipient of the other subject's investigation) and reciprocal (both mice actively sniffing each other) were scored manually for the first 5 min of the test period.

2.3.2 | Grooming test

The grooming test is used to measure repetitive behavior during the first 2 h of the dark cycle. Mice were individually housed in standard mouse cages for 1 week before the test. One hour prior to the test, at 17:30, they were placed in the testing room to habituate. At 18:50, mice were transferred to a standard cage with clean bedding, red lights were turned on and white room lights were turned off and the testing cage was recorded for 2 h. After 2 h of videotaping, animals were group-housed in their original configurations, returned to the colony room and not used for any further testing.

2.3.3 | Urine-exposure open field test

Procedures were based on those described in the literature.[23] One week before the test, males were exposed to same-strain females for 5 min in a novel cage with fresh bedding. The day before the test a handful of soiled male bedding was placed in the female cages to induce estrus. Estrus was determined by visual inspection of the vaginal area. The open field was conducted in a dimly lit room. The adult male mice were placed in a clean open field, lined with paper (Strathmore Drawing Paper Premium, recycled, microperforated, 400 series; Strathmore Artist Papers, Neenah, Wisconsin) and containing some of its own home cage bedding in a corner of the arena. Open field activity was recorded for 60 min. At the end of the habituation period, the mouse was placed back in a clean polycarbonate cage with fresh bedding. The home cage bedding and any feces deposited by the mouse were removed from the open field. Urinary scent marks deposited on the paper during habituation were visualized under ultraviolet light and outlined with a pencil for subsequent quantification. Fifteen microliters of fresh female urine, pooled from 4 to 6 estrous females, was then pipetted onto the center of the Strathmore paper, and the mouse was placed back into the open field for 5 min. Open field activity and ultrasonic vocalizations were recorded. The marked sheets of Strathmore paper were treated with Ninhydrin spray (LC-NIN-16; Tritech Forensics, Inc., Southport, North Carolina) then left to dry for ∼12 h, which allowed the visualization of the urine traces as purple spots. Once dry, images were scanned and opened in ImageJ (U.S. National Institutes of Health, Bethesda, Maryland). Freehand selections of the circled areas (preexposure marking) were removed and copied into a new JPEG image. The preexposure and postexposure images were processed in 8-bit, with background subtracted, and converted to binary. Particles were analyzed at 1000-Infinity (pixels) and 0.00–1.00 (circularity), counted, and their area measured.

2.3.4 | Standard open field test

The open field test is used to assess motor activity. The open field chambers are Plexiglas square chambers (27.3 × 27.3 × 20.3 cm; Med Associates Inc., St Albans, Vermont) surrounded by infrared photobeam sources (16 × 16 × 16). The enclosure, divided into 16 squares by the infrared photobeams, is configured to split the open field into a center zone, constituted of the 4 central squares, and a periphery zone, constituted of the 12 surrounding adjoining sides, and the photocell beams were set to measure activity in the center and in the periphery of the open field chambers. Horizontal activity (distance traveled) and vertical activity (rearing) are measured from consecutive beam breaks. Animals are placed in the open field chambers for a 60 min session and returned back to the home cages after test completion.

2.3.5 | SmartCube test

SmartCube is a high-throughput automated behavioral platform that provides a comprehensive phenotypical assessment of mouse disease model systems and drug effects. Through computer vision, mechanical actuators, and machine learning techniques it identifies individual behaviors, postures, positions, trajectories, sequences of behaviors and responses to stimulation.[20,24,25] Stimulation comprises of a standard sequence of challenges presented in a 45 min session, including a ∼90 psi air puff startling stimulus, a motor challenge and an aversive mild-shock electric stimulus, and includes resting periods where mice are allowed to simply roam the cage. The sequence of stimuli and other environmental and experimental details never changes and therefore disparate study results can be compared with each other to a great extent. Behavior in general and specific responses to these stimuli are recorded through both force sensors and 3D video capture.

Digital videos of the subjects taken during the ∼45 min session are processed with computer vision algorithms to extract more than 1400 dependent measures including frequency and duration of behavioral states such as grooming, rearing, locomotion, immobility, and so forth, and many other features, postures, and body shape parameters obtained during the test session. Using machine learning techniques chosen to best separate pharmacological effects of reference compounds, the differential behavioral signatures of the mutant mice are then assessed quantitatively.[20,22,24,25] We plot data in the 2D space that best separates mutant from WT groups, representing the groups with their mean and standard error and using the overlap between groups as a discrimination index. Each mutant group was compared against its own WT group, with independent classifiers being trained for each mutant model against its corresponding WT group. We also ran two different analyses of model systems together, combining or not the replicas for the *Shank3/F* and *Cntnap2* model systems. To provide a robust common reference point, we combined all WT groups.

Mice were taken in their home cage to the SmartCube suite of experimental rooms where they remained until they were placed in the apparatus. A standard SmartCube protocol runs for a single session. After the session, mice were group-housed again and brought back to the colony room. Any abnormal behavior was noted.

2.4 | Data handling

For all tests, unless noted otherwise, statistical analyses consisted of one- or two-way ANOVAs (StatView for Windows Version 5.0.1, SAS Institute Inc., Cary, North Carolina) with genotype as a between-subjects factor and, when appropriate, age as a within-subject factor. No statistical tests were performed comparing the original and replication study results. Significant interactions between within- and between-subject factors were followed by simple main effects (SPSS, IBM). The level of significance was set at $p < 0.05$. No outliers were removed. For repeated measures ANOVAs, the data of a subject were removed when data were missing for such subject at a time point. Categorical data were analyzed with Mann Whitney and frequency data were analyzed with chi-square or Fisher exact as noted. All statistical tests and results are available in the supplementary material of this article (Methods S1).

2.4.1 | Data availability statement

In addition to the NC3Rs ARRIVE Guidelines Checklist, the data that support the findings of this study are openly available on the Open Science Framework platform at https://osf.io/4mujp/ (DOI: 10.17605/OSF.IO/4MUJP).[26]

2.4.2 | Bioinformatics for SmartCube

The features that define the phenotype (symptom descriptors) were identified and ranked using complex proprietary bioinformatics algorithms and an overall discrimination index was calculated. Graphical representations of the datasets corresponding to the groups compared were derived and a *p*-value was calculated to assess the statistical significance of the discrimination ratios. Top representative features are graphically presented to aid interpretation of differences (see Methods S1[22]).

Feature analysis: decorrelation and ranking. The outcome of a SmartCube run is a set of ~1400 features (behavioral parameters) that can be used for various analyses. Many of these features are correlated (e.g., rearing counts and supported rearing counts). Therefore, we form statistically independent combinations of the original features (further referred to as decorrelated features) that discriminate between the two groups more effectively. Each decorrelated feature extracts information from the whole cluster of the original features, so the new feature space has lower dimensionality. Next, we apply a proprietary feature ranking algorithm to score each feature discrimination power (ability to separate the two groups, e.g., control and disease). Ranking is an important part of our analyses because it weighs each feature change by its relevance: if there is a change in a feature that receives a low rank, this feature will automatically be discounted in our analyses, so we do not have to resort to the conventional "feature selection" approach and discard information buried in the less informative features.

Feature analysis: quantitative assessment of disease phenotype. In the new decorrelated feature space, the overlap between the "clouds" (Gaussian distributions representing the groups of mice in the ranked decorrelated features space) serves as a quantitative measure of separability between the two groups. For visualization purposes, we plot each cloud with its semi-axes equal to the one standard deviation (SD) along the corresponding dimensions. Note, however, that while the overlap between any two Gaussian distributions is always nonzero, it may not necessarily be seen at the "1-SD" level. As in overdetermined systems, the discrimination index sometimes can be artificially high, we calculate its statistical significance by estimating the probability that the result is because of chance.

Top features identification. Working back from the discrimination analysis we can identify the most important features that contribute the most to the separation between the two groups. Although statistical significance for differences between groups for the individual top features can be calculated, the alpha value for such statistical exercise cannot be set to the standard $p = 0.05$, as dozens of features are measured and combed for differences. Instead of over-interpreting such top features we present them in order to understand the mutant signatures but refrain from performing potentially misleading standard statistical tests.

3 | RESULTS

3.1 | Intra-laboratory replication

3.1.1 | SmartCube

In the replication study, the discrimination index between *Shank3/F* WT and KO mice reached 95% (Figure 1(A)). The probability that this discrimination value could be obtained by chance is negligible (Figure 1(B)). Top features contributing to this difference were:

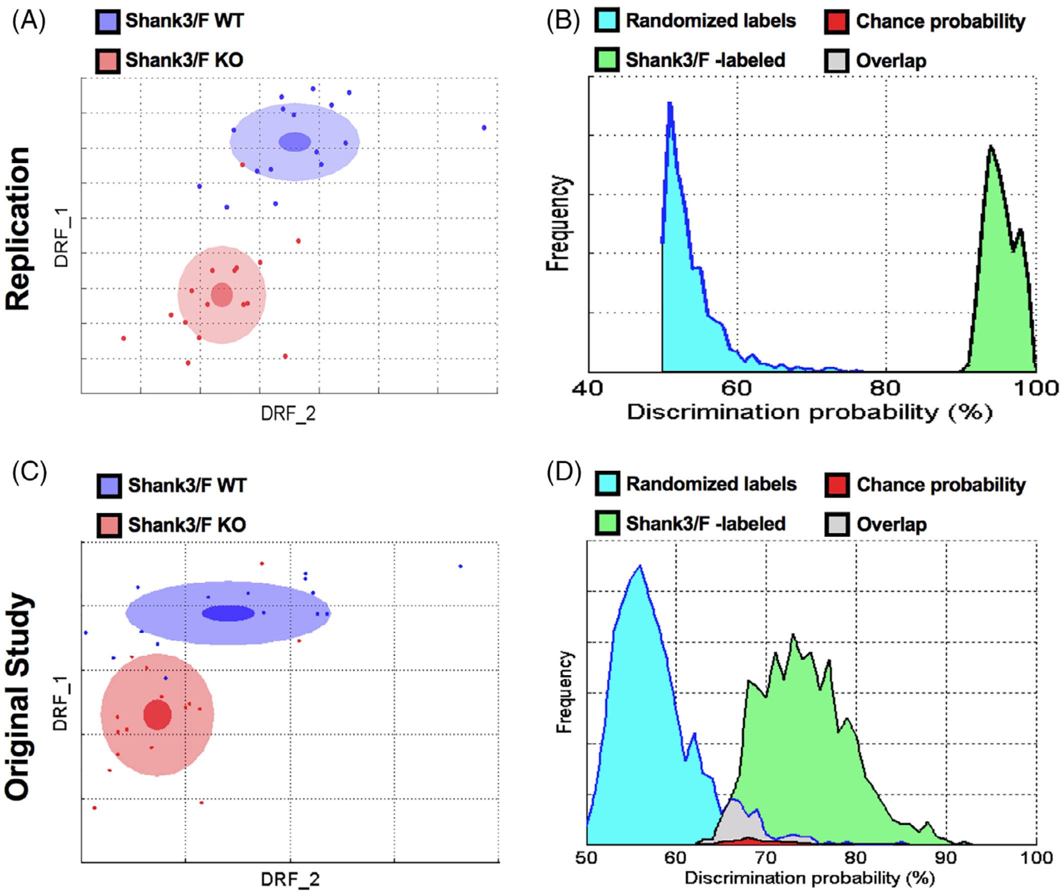


FIGURE 1 *Shank3/F* knockout (KO) and wild type (WT) littermates were very different in the SmartCube test across both studies. (A, C) 2D representation of the multidimensional space in which the two groups are best separated. The DRF1 and 2 axes represent statistically independent combinations of the raw features that best discriminate between the two groups whereas the overlap between the groups is a measure of discriminability (Methods S1). The dots represent individual mice (blue: WT; red: KO mouse). The center, small and large ellipses are the mean, standard error and standard deviation for each group, respectively. (B, D) Discrimination indexes are repeatedly calculated, and their distribution plotted, using either correct labels (green distribution) or randomized labels (blue distribution) such that the overlap between the two distributions (in red) represents the probability of obtaining the observed discrimination by chance. (A, B) In the replication study, the *Shank3/F* model separated well from the WT group with 95% discrimination and $p < 0.0001$. (C, D) In the original study, the *Shank3/F* groups separated from the WT group with 74% discrimination and $p = 0.02$. $n = 16$ mice per genotype/line (replication); $n = 15–16$ mice per genotype/line (original study)

reduced vertical movements, decreased startle, increased freezing, decreased sniffing and digging (Figures 3 and 4; Table S1). The discrimination between the *Cntnap2−/−* and WT mice reached 97% also with a very small *p*-value (Figure 2(A), (B)). The top features for this model were somehow opposite: hyperactivity, higher speed, increased movements, more missteps and increased startle (Figures 3 and 4; Table S1). Figures 1 and 2 show the previous original study results with 74 and 88% discriminations for the *Shank3/F* and *Cntnap2* model systems, respectively.

The top features for the replication and historical studies are shown in Figures 3 and 4 (Table S1). It is clear that the two model systems show opposite differences or trends that are surprisingly consistent between the historical and replication studies. For example, whereas the *Shank3/F* KO mice showed increased freezing in both studies, the *Cntnap2−/−* mice exhibited less freezing. The WT groups, which are both C57BL/6J littermates, show similar freezing levels.

Figure 5 shows an analysis of all five model systems together. In the top panel (see Figure 5(A)) we combined the replicas for the *Shank3/F* and *Cntnap2* model systems, whereas in the bottom panel (see Figure 5(B)) the replicas are shown separated. To run these analyses, we trained the classifiers on all model systems at the same time, aiming at separating all of them from each other as much as possible. Surprisingly, models landed heavily on a line, with the combined WT group in the center. Thus, the *Shank3* model systems were on one side of the combined WT group, and the *Cntnap2* on the other side. The 16p11.2 and *Cacna1c* model systems did not show a strong signature. When replicas were run separately, they landed very much on top of each other and again on a line, suggesting good replication and opposite phenotypes for the *Shank3/F* and *Cntnap2* model systems.

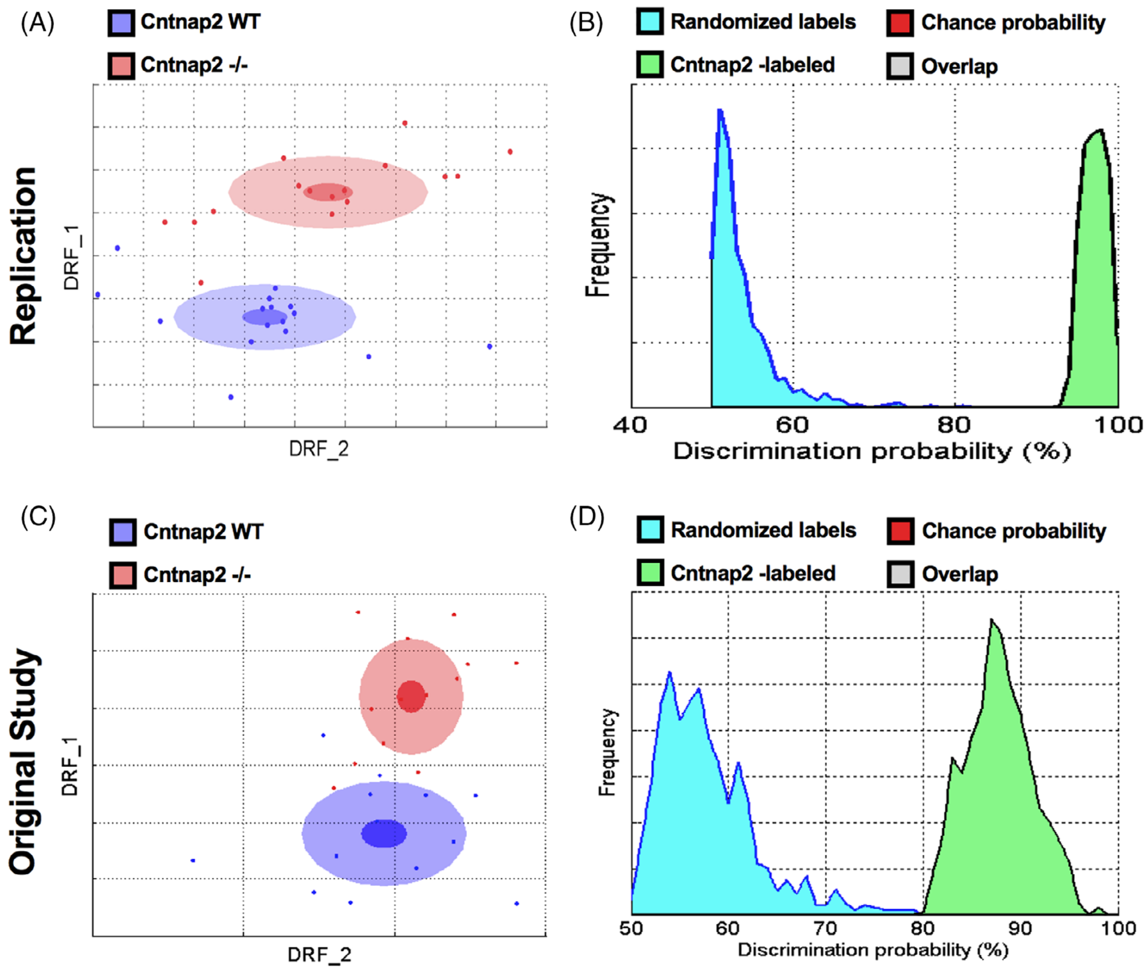


FIGURE 2 *Cntnap2*−/− and wild type (WT) littermates were very different in the SmartCube test across both studies. (A, B) In the replication study, the *Cntnap2* model separated well from the WT group with a 97% discrimination and $p < 0.0001$. (C, D) In the original study, the *Cntnap2* groups separated from the WT group with 88% discrimination and $p < 0.0002$. $n = 16$ mice per genotype/line (replication); $n = 13$ mice per genotype/line (original study)

3.1.2 | Body weight

Two-way ANOVAs showed that all groups significantly gained weight with age during the study [*Shank3/F*, age main effect: $F(2,60) = 1786$, $p < 0.0001$; *Cntnap2*, age main effect: $F(2,60) = 1316$, $p < 0.0001$; Figure 6; Table S2]. The *Shank3/F* KO mice tended to be slightly heavier than WT littermates, but this did not reach significance in the replication using Sidak's multiple comparisons test, unlike in the original study (at P90). A nonsignificant trend for decreased weight was also observed in the *Cntnap2*−/− mice, resembling the significant differences found in the first study (at P90). It should be noted, however, that in original study two cohorts of mice were needed to run all tests in the comprehensive phenotypic screen and therefore the sample size is double in size ($n = 28$–32 per group) compared with the replication ($n = 16$ per group).

3.1.3 | Reciprocal social interaction

To assess social behavior, we paired genotype- and age-matched male mice and allowed them to interact freely for 10 min. The replication results matched very closely the original study data (Table S3). *Shank3/F* mutant pairs were or tended to be closer to each other [$t(30) = 4.60$, $p < 0.0001$], and remained so for longer [$t(30) = 2.75$, $p < 0.01$], than the corresponding WT pairs (Figure 7(A), (B)). However, they also followed each other less [$t(30) = 9.50$, $p < 0.0001$] and showed less locomotion [$t(30) = 9.29$, $p < 0.0001$; Figure 7(C), (D)], suggesting an effect of hypoactivity on social behavior. Indeed, *Shank3/F* mutant pairs interacted less frequently with each other [back: $t(30) = 3.11$, $p < 0.01$] but when they did so, they interacted for longer than WT controls [front: $t(30) = 3.95$, $p < 0.001$; center: $t(30) = 3.84$, $p < 0.001$; back: $t(30) = 2.81$, $p < 0.01$; Figure 8(A), (B)], again, consistently with a hypoactive profile.

In addition to automated scoring, social interactions were hand-scored to isolate behaviors driven by the *subject* mouse (which was introduced to the testing chamber 5 min before the *stimulus* mouse). We found no differences in the rates of active (subject investigating the stimulus mouse), passive (stimulus investigating the subject mouse) or reciprocal (subject and stimulus mouse investigating each other) interactions (Figure 8(C)). However, increased time spent in

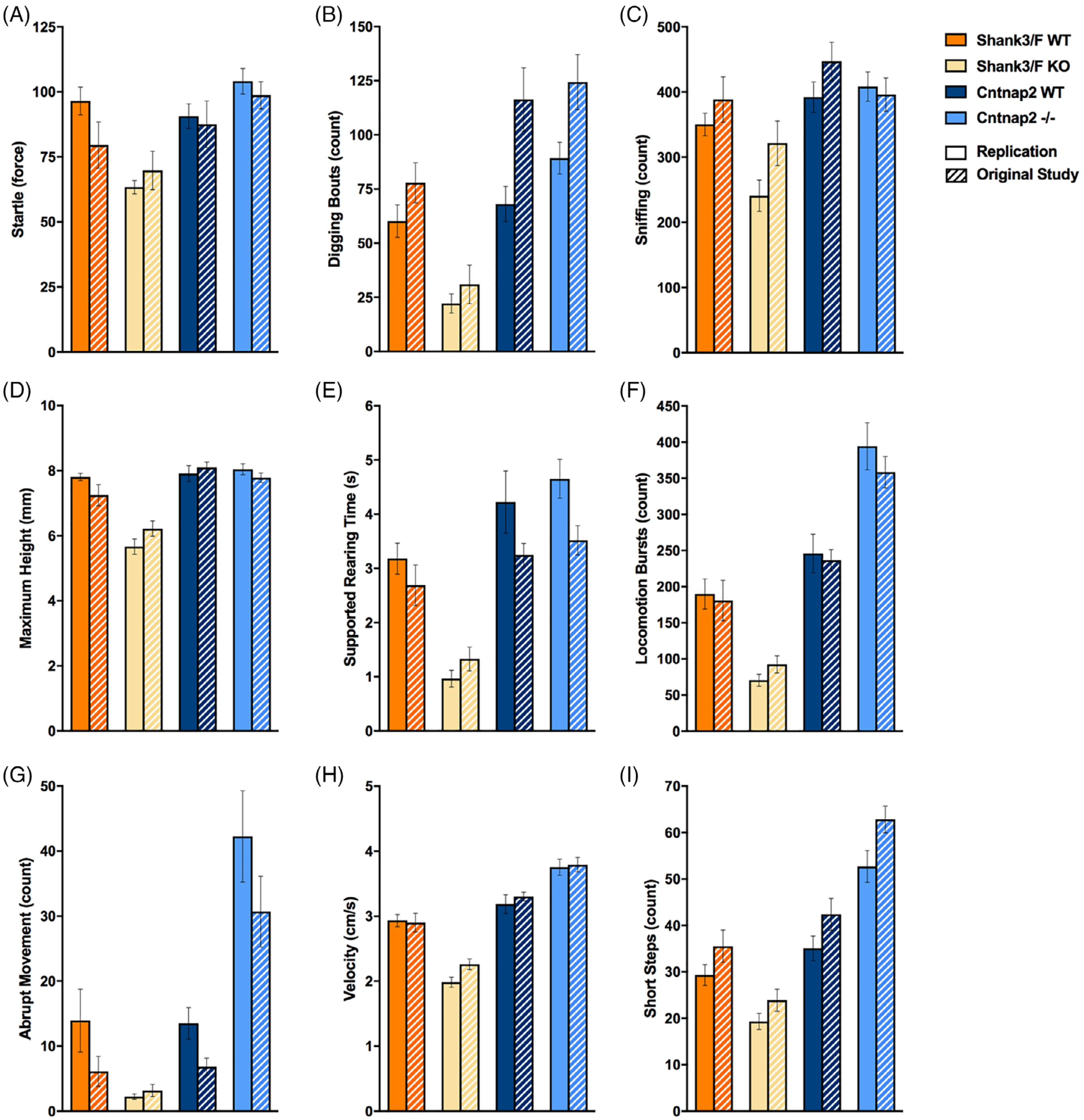


FIGURE 3 Top features in SmartCube showed changes in opposite directions for the models across both studies. The top features that separate the mutant and control groups (see Methods) are plotted for both models and for the replication (filled bars) and original (hashed bars) study. Remarkably, even with more than a year between the original and replication study, most of the results were extremely similar. Across models, many features showed decreases in the *Shank3/F* knockout (KO) mice compared with corresponding wild type (WT) mice, and increases or no effect in the *Cntnap2−/−* mice compared with WT littermates. These measures were either obtained from pressure sensors or extracted from the videos and scored using machine learning algorithms trained on human-labeled videos or hardcoded rules. *Shank3/F* KO mice showed (or tended to show) decreased startle to a tactile stimulus (A), number of digging bouts in the bedding provided (B), number of sniffing events (C), maximum height of the rump detected during the session (D), time spent rearing against the walls of the apparatus (E), bursts of locomotion (F), number of abrupt movements (G), average velocity (H) and number of short steps (I) compared with WT littermates, whereas the *Cntnap2−/−* model showed no change or increases in measures compared with WT littermates

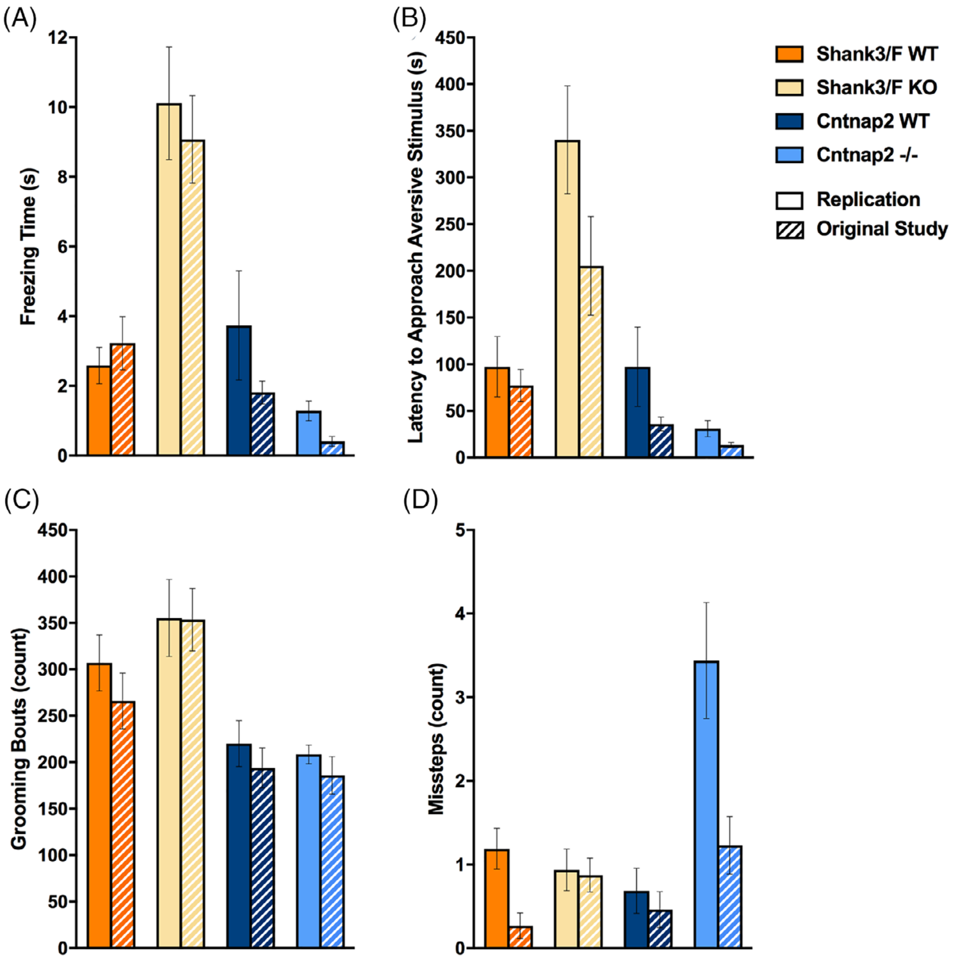


FIGURE 4 Additional SmartCube top features showed differences across the models that were consistent across both studies. The top features that separate the mutant and control groups are plotted for both models and for the replication and original (hashed bars) study. *Shank3/F* knockout (KO) mice showed (or tended to show) increased time freezing (A), latency to approach an aversive stimulus (B) and number of grooming bouts (C) compared with wild type (WT) littermates, whereas the *Cntnap2−/−* mice showed again no difference or decreases in measures compared with WT littermates. The number of missteps during a motor challenge (D) was the only top feature that showed inconsistent results across models and studies

reciprocal investigation was observed for *Shank3/F* KO mice in the replication [$t(30) = 2.38$, $p < 0.05$], similar to the original study (Figure 8(D)). The only measure seemingly not correlated and therefore possibly not confounded by activity in this test was ultrasonic vocalizations (Figure 7(E)). Although vocalizations were too infrequent to analyze in the original study, a trend of decreased vocalizations was observed in *Shank3/F* KO mice [$t(30) = 1.87$, $p < 0.08$]. No notable differences were observed between *Cntnap2−/−* and WT mice across either study with slight trends to show decreased vocalizations not reaching statistical differences (Figure 7(E)).

3.1.4 | Urine-exposure open field

To assess social behavior in a different way, not influenced by the behavior of companion mice, and potentially less confounded by levels of motor activity, we used an open field test in which male mice are exposed to urine of a female in estrous for 5 min after a 1 h baseline session (for all results see Table S4). Whereas sex-naïve mice readily scent-mark,[27,28] ultrasonic vocalizations require social experience. Therefore, we exposed all males to females 1 week prior to the open field test.

Shank3/F KO mice were hypoactive in this test as compared with the WT, showing less locomotion around the chamber and in the center during both the baseline [chamber: $t(30) = 7.10$, $p < 0.0001$; center: $t(30) = 8.30$, $p < 0.0001$] and urine exposure [chamber: $t(30) = 7.00$, $p < 0.0001$; center: $t(30) = 5.30$, $p < 0.0001$] sessions in the replication study (Figure 9). This pattern did not reach statistical significance during the baseline session in the original study (Figure 9(A), (C)) but showed the same difference for the urine test phase (Figure 9(B), (D)). *Shank3/F* KO mice additionally showed less time in the center of the open field as compared with the WT in both the baseline [$t(30) = 4.12$, $p < 0.001$] and the urine-exposure [$t(30) = 2.85$, $p < 0.01$] session and fewer ultrasonic vocalizations [$t(30) = 3.19$, $p < 0.01$; Figure 10(A), (B)]. These results were identical in the original and replication studies. Although there were no differences in overall chamber scent marking, it was significantly decreased in the center after urine exposure in the replication study [$t(30) = 2.87$, $p < 0.01$] and showed a very similar trend in the original study, suggesting that this behavior may be indicative of a social deficit (Figure 10(C), (D)).

Cntnap2−/− mice did not exhibit a clear hyperactive phenotype in this test. As in the original study, distance traveled during the baseline session was lower in the KO mice but not statistically so [*t*

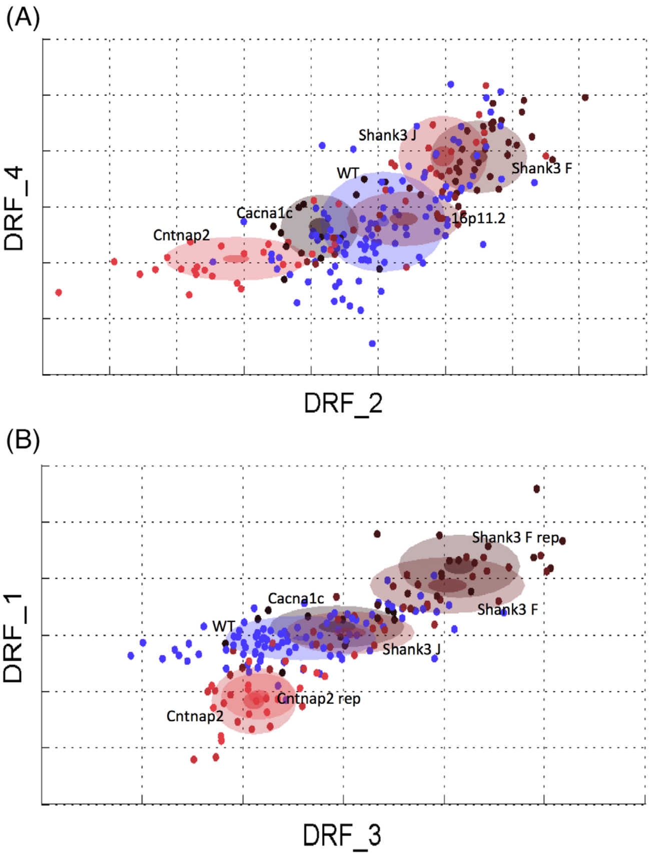


FIGURE 5 “Cloud” analysis across autism spectrum disorder (ASD) mouse models. (A) All five ASD mouse models plotted in the same multidimensional space with mutant replicas (for the *Shank3/F* knockout and *Cntnap2−/−* model systems) and all wild type (WT) littermates pooled. (B) The four mouse models on C57BL/6J genetic background plotted in the same multidimensional space with mutant replica results separated but WT littermates pooled. *n* = 16 mice per genotype/line (replication); *n* = 13–16 mice per genotype/line (original studies); dots are individual animals

(30) = 1.99, *p* < 0.06; Figure 9(A), (C)]. Behavior during the exposure session was inconsistent, with *Cntnap2−/−* mice showing more locomotion around the chamber as compared with the WT in the original study and slightly less locomotion in the center in the replication study [*t*(30) = 1.92, *p* < 0.07; Figure 9(B), (D)]. These mutant mice also showed less time in the center, but only during the urine exposure session and only in the replication study [*t*(30) = 2.86, *p* < 0.01; Figure 10(A)]. In the historical data, there were no differences or trends. Ultrasonic vocalizations were reduced in the original study but a trend in the same direction did not reach significance in the present study (Figure 10(B)). Overall scent marking was slightly reduced in the original study but otherwise did not show any differences (Figure 10(C), (D)).

Thus, the *Shank3/F* mice showed an anxiety-like phenotype reflected in reduced time and activity in the center of the chamber, reduced social ultrasonic response as compared with the WT, and reduced scent marking near to a female stimulus, whereas results in the *Cntnap2* model system were not consistent.

3.2 | Inter-laboratory replication and assessment of convergent calidity

3.2.1 | Open field—standard test of general activity

To provide a standard measure and characterization of activity we ran a 60 min open field test (for all results, see Table S5). Whereas in our previous studies we assessed the development of motor function in juvenile mice, we focused here on assessment of the young adults (∼P75), to provide a separated assessment of activity at the age when we assessed social behavior, as activity may be a confounding factor for some of these tests. We cannot, therefore, assess direct replicability of activity test per se, rather we extend measures of activity to an

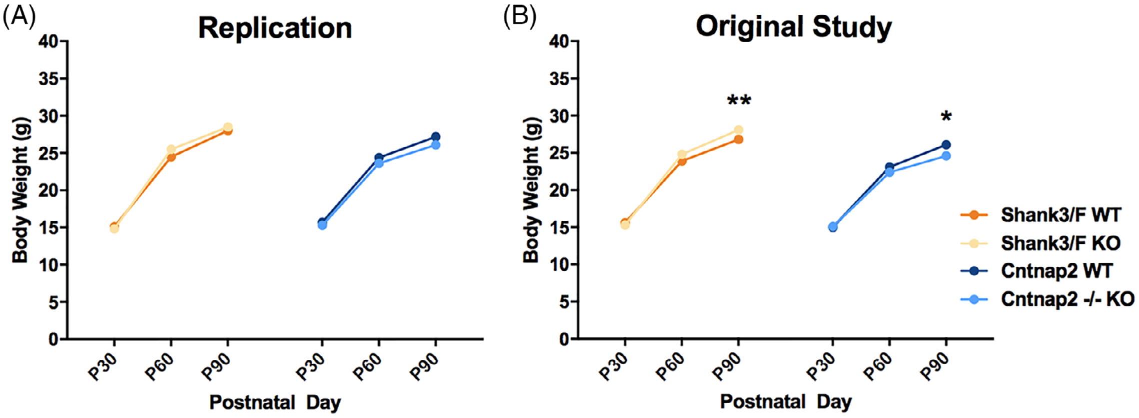


FIGURE 6 Body weight. No differences in body weight were found between mutant and wild type mice for either model in the replication study. Note errors bars are hidden by the graphing symbols. Data shown are means ± SEM; *n* = 16 mice per genotype/line (replication); *n* = 28–32 mice per genotype/line (original study); compared with wild type (WT): *$p < 0.05$, **$p < 0.01$. KO, knockout

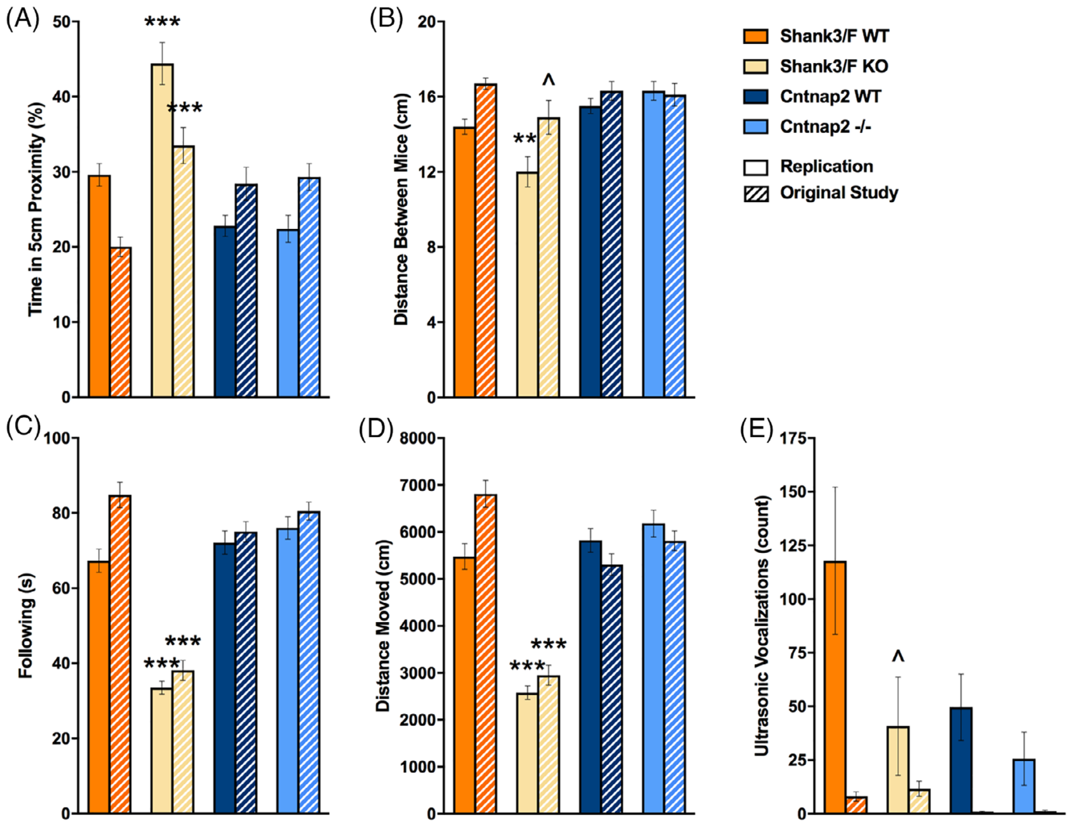


FIGURE 7 Proximity, activity and vocalizations emitted during reciprocal social interaction. Measures are shown for the replication and original (hashed bars) study. (A) *Shank3/F* knockout (KO) pairs spent more time in close proximity to each other than the corresponding wild type (WT) pairs in both the original and replication study, whereas no phenotypic differences were seen in the *Cntnap2* model across both studies. (B) *Shank3/F* KO pairs were in closer proximity to each other than the corresponding WT pairs in the replication study, whereas the difference was only marginal in the original study. There were no differences in the *Cntnap2* model across both studies. (C) The time following each other was decreased in the *Shank3/F* KO as compared with the WT mice across both studies. The *Cntnap2* model did not show a phenotypic difference in either study for this measure. (D) Distance moved showed a very similar pattern to following. (E) *Shank3/F* KO mice tended to emit fewer vocalizations than WT littermates in the replication but not in the original study, whereas no significant differences were observed in the *Cntnap2* model in either study. Data shown are means ± SEM; n = 16 mice per genotype/line (replication), n = 14–16 mice per genotype/line (original study); compared with WT: ^$p < 0.08$, **$p < 0.01$, ***$p < 0.001$

older age. Distance covered was reduced overall in both model systems [*Shank3/F*: $t(30) = 8.83$, $p < 0.0001$; *Cntnap2*: $t(30) = 2.25$, $p < 0.05$; Figure 11(A), (C)]. Rearing [$t(30) = 3.31$, $p < 0.01$; Figure 11(B)], and all measures in the arena center [time: $t(30) = 2.81$, $p < 0.01$; distance moved: $t(30) = 8.46$, $p < 0.0001$; rearing: $t(30) = 3.74$, $p < 0.001$; Figure 12] were only decreased in the *Shank3/F* KO mice, relative to the WT control mice.

3.2.2 | Grooming—test of repetitive behavior

Following the Peça et al.[29] report of increased time grooming in the *Shank3/F* KO mice, we assessed grooming in long sessions of 120 min (for all results see Table S6). Whereas both model systems showed increased grooming frequency [*Shank3/F*: $t(29) = 2.79$, $p < 0.01$; *Cntnap2*: $t(29) = 3.46$, $p < 0.01$; Figure 13(B)] compared with WT littermates, the increase in time only reached significance for the *Cntnap2*−/− model system [$t(29) = 2.50$, $p < 0.05$; Figure 13(A)].

4 | DISCUSSION

4.1 | Findings across the three studies

Animal model systems are needed in neuroscience and drug discovery to better understand the fundamental pathology and pathogenesis of disease, identify and validate drug targets, and screen potential therapeutics. The use of animal models, however, presents many challenges. In ASD, for instance, there are many models based on different genetic findings, from gene mutations to copy number variation, for which phenotyping efforts typically concentrate on three main domains of ASD, namely, repetitive behavior, communication and social deficits. More often than not, researchers assume implicit homologies of the substrates underlying the symptoms in humans and those in rodents, using a simple face value similitude principle. For instance, social behavior in mice is used to model social deficits in children with ASD, and ultrasonic vocalizations to model communication deficits. Apart from the implications of such homology assumptions,

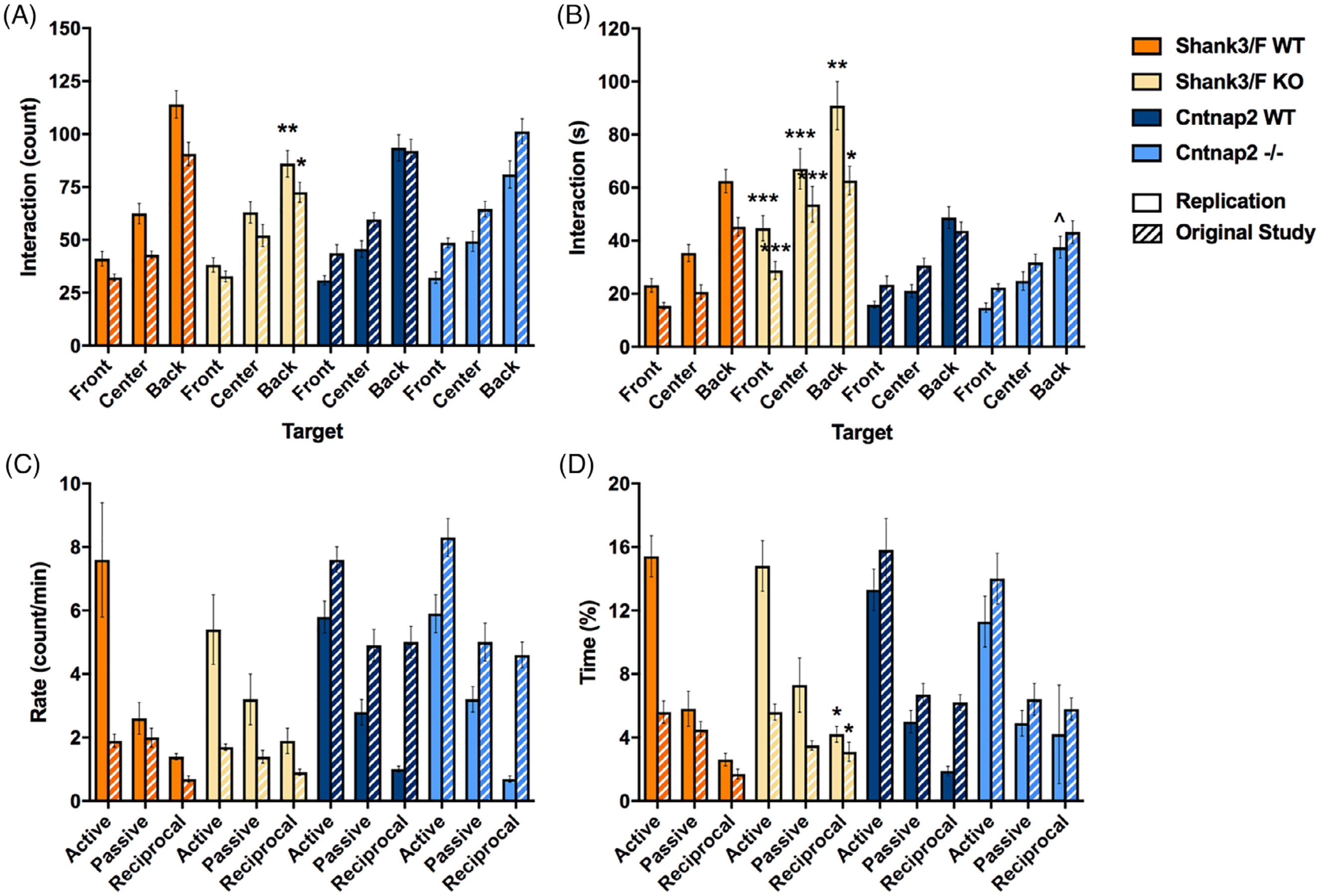


FIGURE 8 Social interactions. Measures are shown for the replication and original (hashed bars) study. (A) *Shank3/F* knockout (KO) pairs showed fewer interactions (approaches towards each other's back) than the corresponding wild type (WT) pairs, and no differences were observed for *Cntnap2* pairs. These results were identical in both the original and the replication studies. (B) The time interacting, however, showed opposite patterns in the *Shank3/F* mice, with the KO pairs interacting for a shorter duration than the WT pairs. Again, no differences were observed in the *Cntnap2* pairs. Once again, the results were identical in the original and replication studies. (C) Hand scoring did not capture differences in the frequency of interaction in either study ("Rate"), for either model. (D) The percent time interacting, however, showed increased reciprocal interactions in the *Shank3/F* KO mice compared with WT littermates, and, again, no differences in the *Cntnap2* model in both studies. Data shown are means ± SEM. These results were identical across the two studies (compared with WT: ^*p* < 0.08, **p* < 0.05, ***p* < 0.01, ****p* < 0.001)

which can and should be challenged, this simple approach implies that all models of ASD will present with similar behavioral patterns. Our project aimed at comparing five different ASD models comprising a large phenotypic effort[5,6] and examining these implicit assumptions. In addition, we studied the robustness and reproducibility of the results. Therefore, we present here a brief summary of all our previous results, replications of several new results using the Shank3/F and Cntnap2 ASD models, and conclusions from machine learning based tests that identified putative common behavioral features across the five model systems investigated.

Development. In terms of development, all model systems showed a rather normal and robust progression (see Table 2). The only exception was the 16p11.2 HET mice, which were smaller than the corresponding WT mice. This is consistent with the finding that this model system is lethal in the C57BL/6J genetic background,[30] suggesting some serious early prenatal or postnatal issues. *Shank3/F* KO mice were slightly heavier, consistent with previous publications,[31] although the replication study showed just a non-significant trend, whereas the *Cntnap2* mutant mice were slightly underweight.

Neonatal ultrasonic vocalizations. The measure of USV was variable suggesting that much larger sample sizes may be required to find robust differences (see Table 2). Alternatively, longer isolation sessions can be used, although the risk of causing unknown long term changes needs to be examined carefully if the subjects are to be reused later in development.[32]

Startle response. *Shank3/F* KO mice showed lower startle levels than the corresponding WT mice in two different tests (ASR and SmartCube tests), whereas the *Cancnac1* mice showed a similar pattern of reduced startle response compared with WT mice in the SmartCube test only (see Table 2). These findings are of interest given that *Shank3* has been suggested to directly contribute to Rett

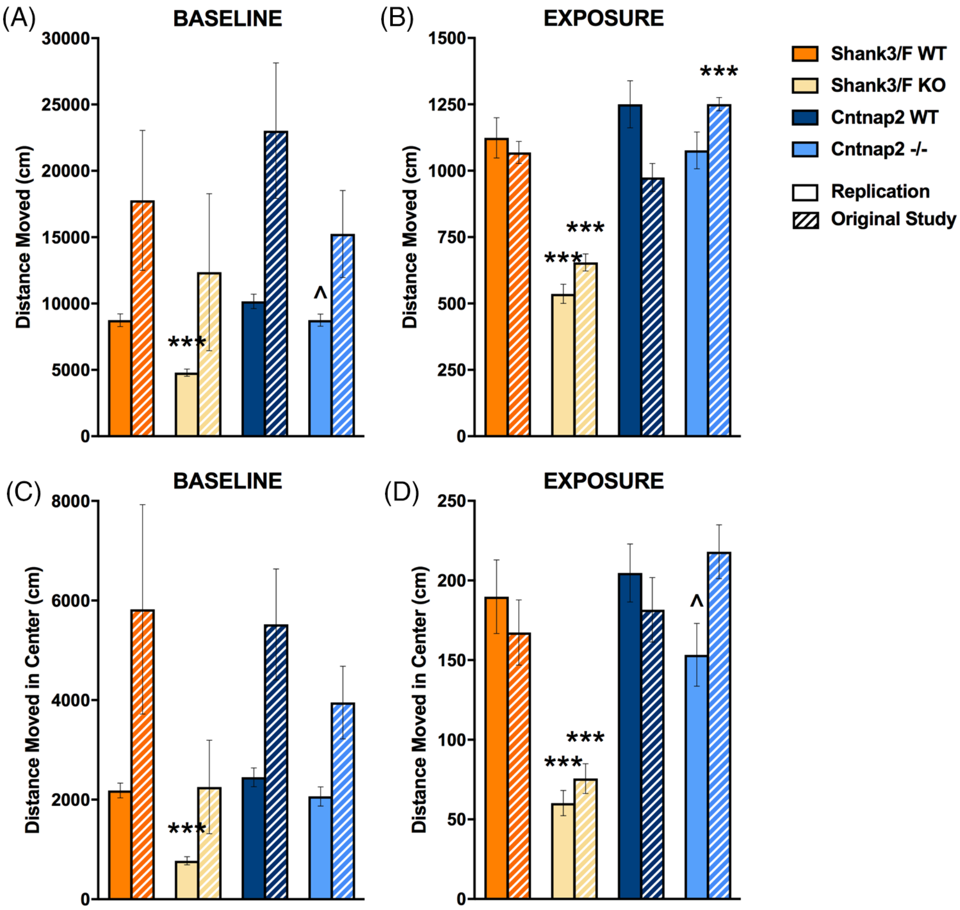


FIGURE 9 Open field activity during the baseline and urine-exposure sessions. Measures are shown for the replication and original (hashed bars) studies. (A) During the baseline session, *Shank3/F* knockout (KO) mice moved less and *Cntnap2*−/− tended to move less around the chamber than wild type (WT) littermates during the replication but not in the original study. (B) *Shank3/F* KO mice moved around the chamber less than WT mice during the exposure session in both studies while *Cntnap2*−/− mice moved around the chamber more than WT mice in the original study only. (C) *Shank3/F* KO mice moved less in the center than WT mice during the baseline session in the replication but not the original study, whereas there were no differences between KO and WT in the *Cntnap2* model. (D) *Shank3/F* KO mice moved less in the center than WT mice during the urine-exposure session in both studies while *Cntnap2*−/− mice only tended to move less in the center during urine exposure in the replication study. Data shown are means ± SEM; *n* = 16 mice per genotype/line (replication); *n* = 15–16 mice per genotype/line (original study); compared with WT: ^*p* < 0.07, ****p* < 0.001

pathology, and mouse model systems of Rett show a robust startle deficit.[33] Although out of this paper's scope, it is tempting to speculate that the link between the startle deficit in the *Cacna1c* and *Mecp2* mouse model systems originates from BDNF modulation of hippocampal neurogenesis, which is affected in both model systems.[34,35] Different SHANK3 KO constructs could lead to different phenotypes, and this may be the reason that *Shank3* from Jiang's lab did not show a startle deficit.[36] The *Cntnap2* mutant mice showed an increase in startle in SmartCube as compared with the WT controls but not in the standard test, a finding that would also require replication. Interestingly, PPI was increased in both the *Shank3/F* and *Cntnap2* mutant mice, a finding opposite to what one would expect for schizophrenia, but consistent with some findings in ASD.[37]

Repetitive behavior. We used several tests to assess repetitive behavior across the three studies (see Table 2). An interesting pattern emerged in the *Shank3/F* mice, which showed lower marble burying, digging, and sniffing compared with WT littermates but higher grooming frequency, consistent with previous reports.[31] The *Cntnap2* mice also showed higher grooming compared with WT littermates, consistent with previous findings, but no changes in marbles buried.[38] Few tests in this battery showed similar patterns between the *Shank3/F* and *Cntnap2* model systems, the two models out of the five studied that presented the more robust phenotypes, and thus it is of particular interest to confirm perseverative grooming. This is another simple endpoint measure with good translation potential, as the circuitry involved is very conserved across species.[39] Opposite patterns in these tests suggest differing neuroanatomical and neurotransmitter support, arguing against the interchangeable use of these tests for assessment of perseveration. Furthermore, the fact that grooming is apparently not correlated with locomotion (for which the two models go in opposite directions), makes it less likely that the grooming test is confounded by this basic behavioral trait.

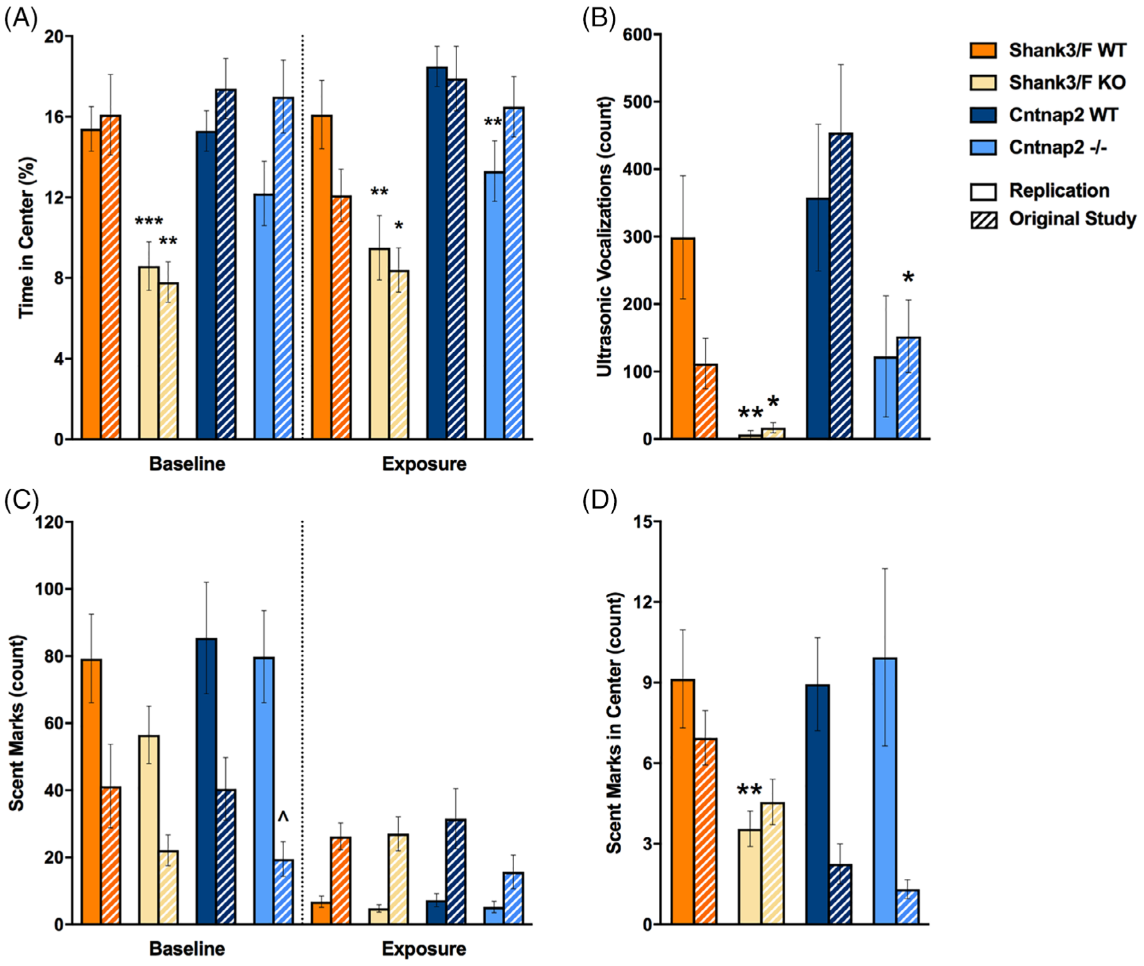


FIGURE 10 Time in the center, ultrasonic vocalizations and scent marking in the urine open field test. Measures are shown for the replication and original (hashed bars) study. (A) *Shank3/F* knockout (KO) mice spent less time in the center than wild type (WT) mice during both testing sessions across both studies, while *Cntnap2−/−* mice only spent less time in the center during urine exposure compared with WT mice in the replication study. (B) *Shank3/F* KO male mice vocalized less than WT mice in both studies, whereas *Cntnap2−/−* mice vocalized less than WT mice only in the original study. (C) KO mice from both models in both studies scent marked similarly to WT mice during both baseline and urine exposure sessions. (D) *Shank3/F* KO male mice scent marked less than WT mice in the center during the urine-exposure session in the replication study only while there were no differences between KO and WT in the *Cntnap2* model. Data shown are means ± SEM (compared with WT: ^*p* < 0.07, **p* < 0.05, ***p* < 0.01, ****p* < 0.001)

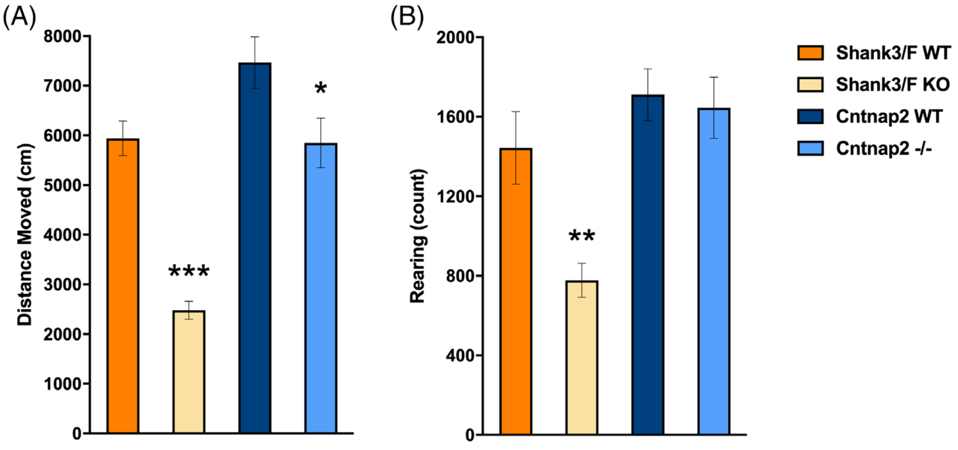


FIGURE 11 Overall behavior in the standard open field test. *Shank3/F* knockout (KO) mice moved (A) and reared (B) less than wild type (WT) littermates. *Cntnap2−/−* mice moved less (A) than WT littermates but did not differ in rearing frequency (B). Data shown are means ± SEM; *n* = 16 mice per genotype/line; compared with WT: **p* < 0.05, ***p* < 0.01, ****p* < 0.001

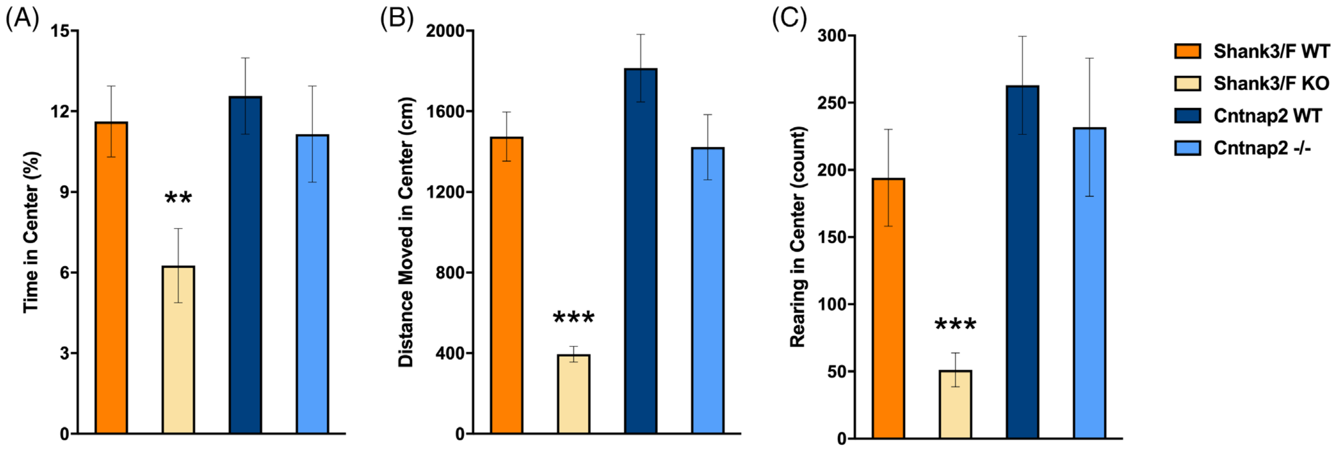
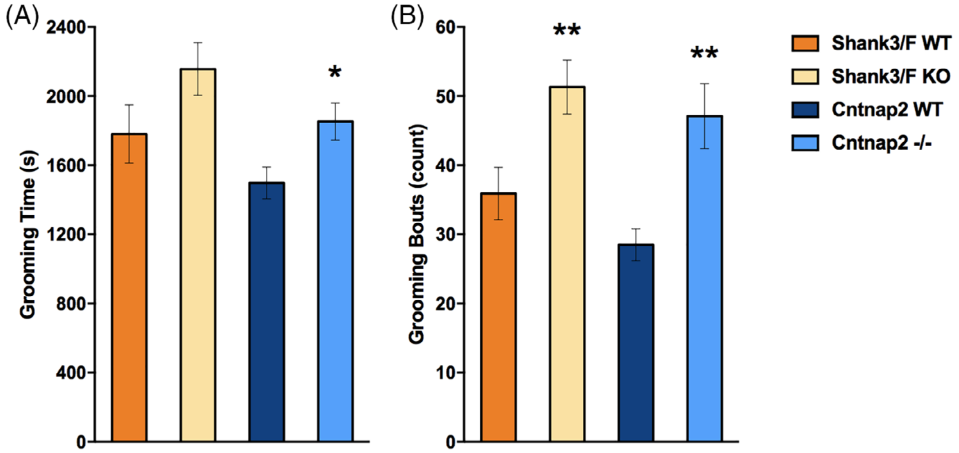


FIGURE 12 Behavior in the center during the standard open field test. *Shank3/F* knockout (KO) mice spent less time (A), moved (B) and reared (C) less in the center of the open field as compared with wild type (WT) control mice. *Cntnap2−/−* mice did not differ from WT control mice on any measure in the center of the open field. Data shown are means ± SEM (compared with WT: ***p* < 0.01, ****p* < 0.001)

FIGURE 13 Grooming. *Cntnap2−/−* mice spent more time (A) and showed more bouts (B) of grooming than wild type (WT) littermates, whereas *Shank3/F* knockout (KO) mice just showed more bouts (B) of grooming. Data shown are means ± SEM; *n* = 15–16 mice per genotype/line; compared with WT: **p* < 0.05, ***p* < 0.01



Locomotor activity. Consistent with previous reports,[31,38] several model systems showed a decrease in locomotor activity compared with their WT controls (see Table 3). Whereas the two *Shank3* and the *Cacna1c* model systems were consistently hypoactive across different tests as compared with their respective WT mice, in contrast with literature findings the *Cntnap2* mice were hypoactive in the open field, did not show strong differences in the urine open field or marble burying, but was nevertheless hyperactive in the SmartCube test. This SmartCube test was designed to provide strong stimulation and cause strong reactions (such as a defensive burying response to an electrical probe, and a startle response to tactile stimulation), and thus it is possible that the pattern of activity in the *Cntnap2* mice reflect environmental reactivity, whereas the *Shank3/F* mice show a low endogenous level of activity, regardless of environmental stimulation. Our informatics analysis suggested that, at least in the SmartCube test, there is a signature continuum, with the model systems lining up in the following order: *Shank3/F* -> *Shank3/J* -> *Cacna1c* -> all WT & 16p11.2 -> *Cntnap2*. Interestingly, independent of the distance covered during locomotion, most model systems moved faster in the NeuroCube test as shown in our original publications[5,6] as compared with their WT controls. Although we expect that motor incoordination and general activity measures will be consistent for the same model system in the two systems, it is possible that SmartCube and NeuroCube induce different activities through opposite emotional reactivity. The NeuroCube system includes a rather dim blue shade light, and it has no challenges or stimulation, whereas SmartCube hardware uses bright white light in addition to the challenging stimuli. Thus, it is possible that, even if two different model systems are motorically similar in the nonanxiogenic environment, one reacts with hypoactivity and the other with hyperactivity in an anxiogenic arena. SmartCube did reproduce for a second time the published findings of lower rearing in *Shank3/F* KO mice compared with WT mice,[29] consistent with its overall hypoactive pattern. An initial goal of the informatics analysis was to attempt to capture those features that were common to all the model systems under study. That is, despite all the differences reported in the literature and in our own studies, what could we find that defined them all as mouse models of ASD? Our informatics analysis, however, suggested that the mutant systems were better separated in two classes, one with a hyperactive and one with a hypoactive profile. Hence, a classifier trained on one class would fail to classify the other class. We could not, therefore, identify features over and above the model systems phenotypic differences.

TABLE 2 Summary of developmental, sensory-motor, cognitive and repetitive behavioral results across the ASD animal models studied

	Development						Sensory-motor			Cognition		Repetitive					
	Isolation test						Startle			T-maze			SmartCube[a]			Grooming	
Gene/copy number variation	Coordination	Ultrasonic vocalizations	Thermoregulation	Geotaxis[a]	Neonatal body weight	Postnatal body weight	Standard startle	SmartCube startle[a]	Prepulse inhibition	Acquisition	Reversal	Marbles buried	Digging bouts	Sniff counts	Grooming bouts	Time	Bouts
SHANK3/F KO	≈	≈	↑	≈	≈	↑≈	↓	≈↓	↑	≈	≈	↓	↓↓	≈↓	↑≈	≈	↑
SHANK3/J KO	↑	≈	≈	≈	≈	≈	≈	≈	≈	≈	≈	≈	↓	≈	≈	-	-
Cacna1c HET	≈	≈	≈	↕	≈	≈	≈	↓	≈	≈	≈	≈	≈	≈	≈	-	-
16p11.2 HET	≈	≈	≈	≈	↓	↓	≈	≈	≈	≈	≈	≈	↑	≈	≈	-	-
Cntnap2 KO	↓	≈	≈	↑	≈	↓≈	≈	≈↑	↑	↑	≈	≈	≈↑	≈≈	≈≈	↑	↑

Note: Directions of effects are shown comparing each model system to their corresponding WT controls across the three studies (upward arrow: increased, downward arrow: decreased, ≈: no difference). Two symbols occupying a single cell reflect both the original study and replication results; "-" indicates no data collected for that model.

Abbreviations: ASD, autism spectrum disorder; HET, heterozygous; KO, knockout; WT, wild type.

[a] Quantitative measure (see the description of test assessment in Methods).

TABLE 3 Summary of activity/motor function results across the ASD animal models studied

	Activity/motor function													
	Standard open field				Distance moved			SmartCube[a]				Neonatal		NeuroCube	
Gene/copy number variation	Distance moved	Rearing	Time in center	Distance in center	Social interaction	Urine open field baseline	Marble-burying	Time mobile	Supported rearing time	Abrupt movement frequency	Short step frequency	Pup righting	Pup rolling	Gait	Average speed
SHANK3/F KO	↓	↓	↓	↓	↓↓ˆ	≈↓	↓	↓↓	↓↓	≈ ↓	↓↓	≈	≈	≈	↑
SHANK3/J KO	-	-	-	-	↓	≈	≈	≈	≈	≈	≈	≈	≈	≈	↑
Cacna1c HET	-	-	-	-	↓	↓	≈	↓	≈	↓	≈	≈	≈	≈	≈
16p11.2 HET	-	-	-	-	≈	≈	↑	≈	≈	≈	↑	↓	≈	↓	↑ˆ
Cntnap2 KO	↓	≈	≈	≈	≈≈	≈↓ˆ	≈	↑↑	≈≈	↑↑	↑↑	↓	↑	↓	↑

Note: Directions of effects are shown comparing each model system to their corresponding WT controls across the three studies (upward arrow: increased, downward arrow: decreased, ≈: no difference, ˆ: non-significant trend). Two symbols occupying a single cell reflect both the original study and replication results. "-" indicates no data collected for that model.

Abbreviations: ASD, autism spectrum disorder; HET, heterozygous; KO, knockout; WT, wild type.

[a] Quantitative measure (see the description of test assessment in Methods).

TABLE 4 Summary of social behavioral results across the ASD animal models studied

	Social																		
	Three chamber		Reciprocal social interaction							Urine-exposure open field									
Gene/ copy number variation	Sociality	Recognition	Proximity time	Proximity distance	Following time	Interaction count	Interaction time	Reciprocal interaction time	Ultrasonic vocalizations	Marking count baseline	Marking count exposure	Marking count center exposure	Distance moved baseline	Distance moved exposure	Distance moved center baseline	Distance moved center exposure	Center time baseline	Center time exposure	Ultrasonic vocalizations
SHANK3/ F KO	↓	↓	↑↑	↓↓	↓↓	↓↓	↑↑	↑↑	≈↓	≈≈	≈≈	≈↓	≈↓	↓↓	≈↓	↓↓	↓↓	↓↓	↓↓
SHANK3/ J KO	≈	≈	≈	≈	↓	≈	↑	≈	≈	≈	≈	≈	≈	≈	≈	≈	↓	↓	↓
Cacna1c HET	≈	≈	↑	≈	↓	≈	↑	≈	≈	≈	≈	≈	↓	↓	↓	↓	≈	↓	≈
16p11.2 HET	≈	≈	≈	↓	≈	≈	↑	≈	≈	≈	≈	≈	≈	↑	≈	≈	↑	≈	≈
Cntnap2 KO	≈	≈	≈≈	≈≈	≈≈	≈≈	≈↓	≈≈	≈≈	↓≈	≈≈	≈≈	≈↓	↑≈	≈≈	≈↓	≈≈	≈↓	↓↓

Note: Directions of effects are shown comparing each model system to their corresponding WT controls (upward arrow: increased, downward arrow: decreased, ≈: no difference, dashed shorter arrow: non-significant trend). Two symbols occupying a single cell reflect both the original study and replication results.

Abbreviations: ASD, autism spectrum disorder; HET, heterozygous; KO, knockout; WT, wild type.

Social behavior. We used three different tests across this series of studies to analyze behavior in a social context: the three chamber, reciprocal interaction and urine-open field tests (see Table 4). We previously reported that *Shank3/F* KO mice showed no preference for the social stimuli in the three-chamber test but that perhaps genotypic differences did not reach significance because of the high variability in the WT group.[6] The lack of significance in the other model systems, however, did not seem confounded by any other factor and probably reflects no deficit (or undetectable deficits) in our experimental setup. In the reciprocal social interaction test, we found that mice that were hypoactive (*Shank3/F*, *Shank/J* and *Cntnap2*) follow each other less, but interact for longer periods of time. This suggests a confounding effect of activity, in particular for the *Shank3/F* and *Cacna1c* models, as they were consistently hypoactive in different tests of motor activity. The 16p11.2 mice also showed closer proximity and more time spent interacting, clearly not a social "deficit." Finally, the *Cntnap2* mice were not significantly different from the corresponding WT group across any of the reciprocal social interaction measures. It should be noted that in the first study of this series,[5] the effect of stimulus genotype was interrogated in this test and it was found that the results with the WT stimulus were identical to those with the same genotype stimulus. In addition, one concern might be that social interaction levels would be artificially reduced in the mutant pairs, however those results are quite comparable to the WT pairs. Social behavior in rodents, and associated learning and recall processes, primarily depends on olfactory function, whereas in primates, the most salient cues are received from visual and auditory inputs reviewed in Ref. 40 suggesting distinct fundamental neurobiological substrates. Patterns of activation of oxytocin and vasopressin-like receptors in social contexts, however, suggest conserved neural networks among different species.[41] Indeed, vasopressin and oxytocin modulate social recognition in both rodents and primates,[42,43] suggesting common circuitry despite different sensory input modalities. Common neurotransmitter systems, and potentially common downstream circuitry, make the translation from rodent preclinical to the human case possible, despite qualitative differences in the perceptual apparati used by the different species.

Our results point to several difficulties with the reciprocal social interaction test, starting with the fact that time following and interaction frequency differences between mutant and WT mice were always in the opposite direction of interaction time and proximity differences. That is, mice that follow and interacted more also spent less time interacting and remain further away. This pattern is better explained by motor activity rather than by differences in social drive. In human, importantly, social behavior is not a unitary domain and presents differently in different disorders. For example, in Williams and Down syndrome, autism characteristics coexist with hyper-, not hypo-, sociality.[44,45] Moreover, social behavior patterns are varied and change according to the context.[46] Social behavior seats atop a number of other processes, such as basic and higher-order perceptual processes, emotional states and reactivity, behavioral control, anxiety, and so forth. Hyper sociality in Williams syndrome, for example, is better explained by decreased behavioral control, and not by either lack of emotion recognition or increased social approach, a finding that required an extensive neuropsychological assessment.[44] The current approach of modeling all social deficits expected in autism with a single test in mice is clearly an oversimplification. A different approach, based on efforts to identify underlying constructs and their biological underpinnings, such as Frith's affiliation, dominance, and other "startup kits"[47] is needed.

Juvenile and adult ultrasonic vocalizations. Measurements of USVs in the reciprocal social interaction test showed no significant phenotypic differences across the five model systems, whereas, in the urine-exposure open field test, interesting decreases in the number of USVs emitted were observed in the *Shank3/F*, *Shank3/J* and *Cntnap2* KO mice compared with their WT counterparts (see Table 4). These results suggest that USV emission may be a measure sensitive to social context in a test probably not so obviously confounded by activity.

5 | CONCLUSIONS

Despite many of our findings not replicating with results published in the literature, we found overall excellent replication of most of the results from our previous publications, using the same protocols and animal models and often after consultation with the originating laboratories. In the present collection of studies, WT controls' data were robust and consistent, although there were exceptions. The direction of the genotype effects, and often the effect size, was also very comparable. Thus, independent from the absolute values found in the studies, the conclusions were rather consistent. For the husbandry of genetically manipulated animals it is standard in our lab to house them in mixed genotypes, to provide a normal stimulus for allogrooming and huddling. Although there is a risk that phenocopying may attenuate some of the phenotype, this is preferable to a phenotype that is secondary to other issues. Thus, we argue that the problem in the lack of replicability of results resides more in the differences between labs, protocols, husbandry, data handling, and statistical analysis, that lead to different outcomes, than in the variability of the animal models themselves.

There has been much recent discussion surrounding the issues of the replicability and reproducibility of preclinical results.[10,15,48] We do strongly agree that exploratory results should be confirmed in the same lab through replication but to do this, incentives around funding and publication must be strengthened. Between labs, reproducing confirmed findings with slight environmental and experimental deviations should lead to greater understanding of the robustness of models and associated phenomena. None of this is possible, of course, without complete and transparent reporting of results and, optimally, access to datasets underlying both positive and negative (often unpublished) results along with associated data such as animal health records.[49,50] The rapid evolution of increasingly sophisticated computational modeling techniques and inexpensive data storage are beginning to converge with the efforts made by several groups to develop a global ontology and common data elements to harmonize massive amounts of preclinical data.[51-55] By analyzing data *en masse*, the hope is that researchers will be able to develop hypotheses on more solid empirical ground[56] and drug developers can more accurately determine what model systems and tests to pursue; both spending less time and financial resources following independent underpowered studies (see e.g., Ref. 57).

In summary, we reiterate our strong belief that mouse model systems of human disease that present with etiological validity (i.e., where there is homology between cause of pathology in both human and animal model system) and construct validity (i.e., there is homology of the pathological process) are fundamental tools for the understanding of gene function, pathophysiology, and are necessary for drug and treatment development. We further argue, based on a complete review of data obtained from three comprehensive studies, that simple measures of known biology such as startle reactivity and self-grooming may provide a venue to link rodent and human pathology. Other measurements, namely social behavior, require refinement and a proven homology to specific constructs underlying the idiosyncratic social profiles described for each syndrome. The face validity approach that supports the use of simple social tests brings little promise for ASD, an area of research in critical need of robust and replicable preclinical science.

ACKNOWLEDGMENTS

We are grateful to Judy Watson-Johnson, Vanessa Rivera and Melinda Ruiz for breeding efforts. Thanks to Ian Russell for assistance with all video-based systems and to Mitch Silverstein who offered invaluable support in apparatus development. This project was funded by the Simons Foundation and Roche as a contract to PsychoGenics. Simons Foundation, Roche, and PsychoGenics contributed to the conception and design of the experiments and to the writing of the manuscript. All aspects of experimental design, data collection and analysis and decision to publish the manuscript were made jointly by Roche, PsychoGenics and the Simons Foundation. D. B. and T. H. hold PsychoGenics shares and contributed to the invention of the PsychoGenics high-throughput technologies. There are no patents, products in development or marketed products related to this study to declare.

ORCID

Patricia Kabitzke https://orcid.org/0000-0003-3305-0795

REFERENCES

1. Moy SS, Nadler JJ. Advances in behavioral genetics: mouse models of autism. *Mol Psychiatry*. 2008;13(1):4-26.
2. Ey E, Leblond CS, Bourgeron T. Behavioral profiles of mouse models for autism spectrum disorders. *Autism Res*. 2011;4(1):5-16.
3. Jiang YH, Ehlers MD. Modeling autism by SHANK gene mutations in mice. *Neuron*. 2013;78(1):8-27.
4. Kazdoba TM, Leach PT, Crawley JN. Behavioral phenotypes of genetic mouse models of autism. *Genes Brain Behav*. 2016;15(1):7-26.
5. Brunner D, Kabitzke P, He D, et al. Comprehensive analysis of the 16p11.2 deletion and null Cntnap2 mouse models of autism spectrum disorder. *PLoS One*. 2015;10(8):e0134572.
6. Kabitzke PA, Brunner D, He D, et al. Comprehensive analysis of two Shank3 and the Cacna1c mouse models of autism spectrum disorder. *Genes Brain Behav*. 2018;17(1):4-22.
7. Peng R. The reproducibility crisis in science: a statistical counterattack. *Significance*. 2015;12(3):30-32.
8. Golani I, Wexler Y, Benjamini Y. The demand for replicability of behavioral result: from burden to asset. paper presented at: Measuring Behavior; 2014.
9. Jarvis MF, Williams M. Irreproducibility in preclinical biomedical research: perceptions, uncertainties, and knowledge gaps. *Trends Pharmacol Sci*. 2016;37:290-302.
10. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res*. 2015;116(1):116-126.
11. Munafo M, Noble S, Browne WJ, et al. Scientific rigor and the art of motorcycle maintenance. *Nat Biotechnol*. 2014;32(9):871-873.
12. Brunner D, Balci F, Ludvig EA. Comparative psychology and the grand challenge of drug discovery in psychiatry and neurodegeneration. *Behav Processes*. 2012;89(2):187-195.
13. Cabin RJ, Mitchell RJ. To Bonferroni or not to Bonferroni: when and how are the questions. *Bull Ecol Soc Am*. 2000;81(3):246-248.
14. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biol*. 2014;12(5):e1001863.
15. Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365-376.
16. Happé F, Ronald A. The 'fractionable autism triad': a review of evidence from behavioural, genetic, cognitive and neural research. *Neuropsychol Rev*. 2008;18(4):287-304.
17. Silverman JL, Yang M, Lord C, Crawley JN. Behavioural phenotyping assays for mouse models of autism. *Nat Rev Neurosci*. 2010;11(7):490-502.
18. Moy SS, Nadler JJ, Magnuson TR, Crawley JN. Mouse models of autism spectrum disorders: the challenge for behavioral genetics. *Am J Med Genet C Semin Med Genet*. 2006;142C(1):40-51.
19. Crawley JN. Mouse behavioral assays relevant to the symptoms of autism. *Brain Pathol*. 2007;17(4):448-459.
20. Brunner D, Nestler E, Leahy E. In need of high-throughput behavioral systems. *Drug Discov Today*. 2002;7(18 suppl):S107-S112.
21. Tecott LH, Nestler EJ. Neurobehavioral assessment in the information age. *Nat Neurosci*. 2004;7(5):462-466.
22. Alexandrov V, Brunner D, Hanania T, Leahy E. High-throughput analysis of behavior for drug discovery. *Eur J Pharmacol*. 2015;750:82-89.
23. Wohr M, Roullet FI, Crawley JN. Reduced scent marking and ultrasonic vocalizations in the BTBR T+tf/J mouse model of autism. *Genes Brain Behav*. 2011;10(1):35-43.
24. Roberds SL, Filippov I, Alexandrov V, Hanania T, Brunner D. Rapid, computer vision-enabled murine screening system identifies neuropharmacological potential of two new mechanisms. *Front Neurosci*. 2011;5:103.
25. Houghten RA, Pinilla C, Giulianotti MA, et al. Strategies for the use of mixture-based synthetic combinatorial libraries: scaffold ranking, direct testing in vivo, and enhanced deconvolution by computational methods. *J Comb Chem*. 2008;10(1):3-19.
26. Kabitzke P, Morales D, He D, Cox K, Sutphen J, Thiede L, Sabath E, Hanania T, Biemans B, Brunner D. Mouse Model Systems of Autism Spectrum Disorder: Replicability and Informatics Signature [dataset]; 2019. https://doi.org/10.17605/OSF.IO/4MUJP.
27. Lehmann ML, Geddes CE, Lee JL, Herkenham M. Urine scent marking (USM): a novel test for depressive-like behavior and a predictor of stress resiliency in mice. *PLoS One*. 2013;8(7):e69822.
28. Novotny M, Harvey S, Jemiolo B. Chemistry of male dominance in the house mouse, *Mus domesticus*. *Experientia*. 1990;46(1):109-113.

29. Peça J, Feliciano C, Ting JT, et al. Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. *Nature*. 2011;472(7344):437-442.
30. Horev G, Ellegood J, Lerch JP, et al. Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proc Natl Acad Sci U S A*. 2011;108(41):17076-17081.
31. Wang X, McCoy PA, Rodriguiz RM, et al. Synaptic dysfunction and abnormal behaviors in mice lacking major isoforms of Shank3. *Hum Mol Genet*. 2011;20(15):3093-3108.
32. Scattoni ML, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in adult BTBR T+tf/J mice during three types of social encounters. *Genes Brain Behav*. 2011;10(1):44-56.
33. Waga C, Asano H, Sanagi T, et al. Identification of two novel Shank3 transcripts in the developing mouse neocortex. *J Neurochem*. 2014;128(2):280-293.
34. Larimore JL, Chapleau CA, Kudo S, Theibert A, Percy AK, Pozzo-Miller L. Bdnf overexpression in hippocampal neurons prevents dendritic atrophy caused by Rett-associated MECP2 mutations. *Neurobiol Dis*. 2009;34(2):199-211.
35. Lee TT, Hill MN, Lee FS. Developmental regulation of fear learning and anxiety behavior by endocannabinoids. *Genes Brain Behav*. 2016;15(1):108-124.
36. Zhou Y, Kaiser T, Monteiro P, et al. Mice with Shank3 mutations associated with ASD and schizophrenia display both shared and distinct defects. *Neuron*. 2016;89(1):147-162.
37. Madsen GF, Bilenberg N, Cantio C, Oranje B. Increased prepulse inhibition and sensitization of the startle reflex in autistic children. *Autism Res*. 2014;7(1):94-103.
38. Bader PL, Faizi M, Kim LH, et al. Mouse model of Timothy syndrome recapitulates triad of autistic traits. *Proc Natl Acad Sci U S A*. 2011;108(37):15432-15437.
39. Kalueff AV, Stewart AM, Song C, Berridge KC, Graybiel AM, Fentress JC. Neurobiology of rodent self-grooming and its value for translational neuroscience. *Nat Rev Neurosci*. 2016;17(1):45-59.
40. Behrendt R-P. *Neuroanatomy of Social Behaviour: An Evolutionary and Psychoanalytic Perspective*. London, UK: Karnac Books; 2011.
41. Johnson ZV, Young LJ. Oxytocin and vasopressin neural networks: implications for social behavioral diversity and translational neuroscience. *Neurosci Biobehav Rev*. 2017;76(Pt A):87-98.
42. Insel TR. Oxytocin – a neuropeptide for affiliation: evidence from behavioral, receptor autoradiographic, and comparative studies. *Psychoneuroendocrinology*. 1992;17(1):3-35.
43. Carter CS. Oxytocin pathways and the evolution of human behavior. *Annu Rev Psychol*. 2014;65:17-39.
44. Porter MA, Coltheart M, Langdon R. The neuropsychological basis of hypersociability in Williams and Down syndrome. *Neuropsychologia*. 2007;45(12):2839-2849.
45. Klein-Tasman BP, Phillips KD, Lord C, Mervis CB, Gallo FJ. Overlap with the autism spectrum in young children with Williams syndrome. *J Dev Behav Pediatr*. 2009;30(4):289-299.
46. Moss J, Howlin P, Hastings RP, et al. Social behavior and characteristics of autism spectrum disorder in Angelman, Cornelia de Lange, and Cri du Chat syndromes. *Am J Intellect Dev Disabil*. 2013;118(4):262-283.
47. Frith U. Time we identified cognitive phenotypes for the social deficits in autism; 2016. http://frithmind.org/blog/2016/05/26/time-we-identified-cognitive-phenotypes-for-the-social-deficits-in-autism/.
48. Kafkafi N, Agassi J, Chesler EJ, et al. Reproducibility and replicability of rodent phenotyping in preclinical studies. *Neurosci Biobehav Rev*. 2018;87:218-232.
49. Brunner D, Balci F, Kabitzke P, Hill H. Consensus Preclinical Checklist (PRECHECK): Experimental Conditions – Rodent Disclosure Checklist. *Int J Comp Psychol*. 2016;29(1):1-5.
50. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol*. 2010;8(6):e1000412.
51. Wang Q, Liao J, Hair K, et al. Estimating the statistical performance of different approaches to meta-analysis of data from animal studies in identifying the impact of aspects of study design. *bioRxiv*. 2018;256776.
52. Hume S, Chow A, Evans J, et al. CDISC SHARE, a Global, Cloud-based Resource of Machine-Readable CDISC Standards for Clinical and Translational Research. *AMIA Jt Summits Transl Sci Proc*. 2017;2018:94-103.
53. Lapinlampi N, Melin E, Aronica E, et al. Common data elements and data management: remedy to cure underpowered preclinical studies. *Epilepsy Res*. 2017;129:87-90.
54. Smith DH, Hicks RR, Johnson VE, et al. Pre-clinical traumatic brain injury common data elements: toward a common language across laboratories. *J Neurotrauma*. 2015;32(22):1725-1735.
55. Ferguson AR, Nielson JL, Cragin MH, Bandrowski AE, Martone ME. Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nat Neurosci*. 2014;17(11):1442-1447.
56. Haefeli J, Ferguson AR, Bingham D, et al. A data-driven approach for evaluating multi-modal therapy in traumatic brain injury. *Sci Rep*. 2017;7:42474.
57. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. *PLoS Biol*. 2015;13(6):e1002165.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Kabitzke P, Morales D, He D, et al. Mouse model systems of autism spectrum disorder: Replicability and informatics signature. *Genes, Brain and Behavior*. 2020;19:e12676. https://doi.org/10.1111/gbb.12676