

# Identity domains capture individual differences from across the behavioral repertoire

Oren Forkosh<sup>1,6</sup>, Stoyo Karamihalev<sup>1,2,6</sup>, Simone Roeh<sup>3</sup>, Uri Alon<sup>4</sup>, Sergey Anpilov<sup>1,2</sup>, Chadi Touma<sup>5</sup>, Markus Nussbaumer<sup>1</sup>, Cornelia Flachskamm<sup>1</sup>, Paul M. Kaplick<sup>1</sup>, Yair Shemesh<sup>1,2</sup> and Alon Chen<sup>1,2\*</sup>

**Personality traits can offer considerable insight into the biological basis of individual differences. However, existing approaches toward understanding personality across species rely on subjective criteria and limited sets of behavioral readouts, which result in noisy and often inconsistent outcomes. Here we introduce a mathematical framework for describing individual differences along dimensions with maximum consistency and discriminative power. We validate this framework in mice, using data from a system for high-throughput longitudinal monitoring of group-housed male mice that yields a variety of readouts from across the behavioral repertoire of individual animals. We demonstrate a set of stable traits that capture variability in behavior and gene expression in the brain, allowing for better-informed mechanistic investigations into the biology of individual differences.**

Individual differences are a hallmark of living organisms and central to our understanding of behavior and psychopathology. In humans, consistencies in emotional and behavioral expression have been extensively investigated and categorized by psychologists within the framework of personality traits<sup>1,2</sup>. In other species, however, the understanding of individual differences and the biological processes that underlie them has been hindered by an absence of strong conceptual foundations behind the trait establishment process and a lack of comprehensive behavioral screening paradigms.

Here, we propose to resolve these issues by using a computational framework for capturing and describing the space of individual behavioral expression and reducing diverse longitudinal behavioral data to trait-like dimensions. Personality traits can be thought of as having two essential characteristics: (1) they capture and represent a continuous gradient of differences between individuals of the same species and (2) they tend to be stable for individuals over time. Thus, a mathematical formulation of a trait directly informed by these properties would be a dimension that captures the maximum behavioral variability between individuals while maintaining minimum variability within individuals over time. We use the term ‘identity domains’ (IDs) to describe such traits obtained from decomposing a high-dimensional space of measured behaviors. Conceptually similar to principal component analysis (PCA), which identifies the directions of maximum variability, our linear discriminant analysis (LDA) decomposition-based approach seeks out dimensions that have maximal discriminative power and stability by maximizing the between-individual to within-individual variability ratio (Fig. 1a). We validate this framework in mice, one of the most commonly used model organisms in neuroscience and psychiatry research, and a species that readily allows for exploration of the biological underpinnings of individual differences.

## Results

**The social box paradigm.** To assess the broadest variety of ethologically relevant voluntary mouse behaviors, we used a long-term ‘social box’ living paradigm, wherein mice are housed in an

enriched, semi-naturalistic environment in groups of four and monitored over multiple days<sup>3,4</sup> (Fig. 1b–d; Supplementary Video 1). Automatic location tracking of individuals allowed for high-throughput behavioral data collection, with readouts consisting of both individual (for example, locomotion, exploration and foraging patterns) and social (for example, approaches, contacts and chases) behaviors. A total of 60 features per mouse per 12-h active phase were collected (Supplementary Fig. 1a). We initially monitored 42 groups of four young adult outbred male mice (a total of 168 animals) left undisturbed over a period of at least 4 days.

**LDA.** An initial analysis of the readouts from this dataset revealed a subset of behaviors that, in themselves, were discriminating between and/or stable within individuals (Supplementary Fig. 1b), which suggested that the social box paradigm could capture some of the information necessary for building IDs. We therefore proceeded to train our algorithm on this dataset.

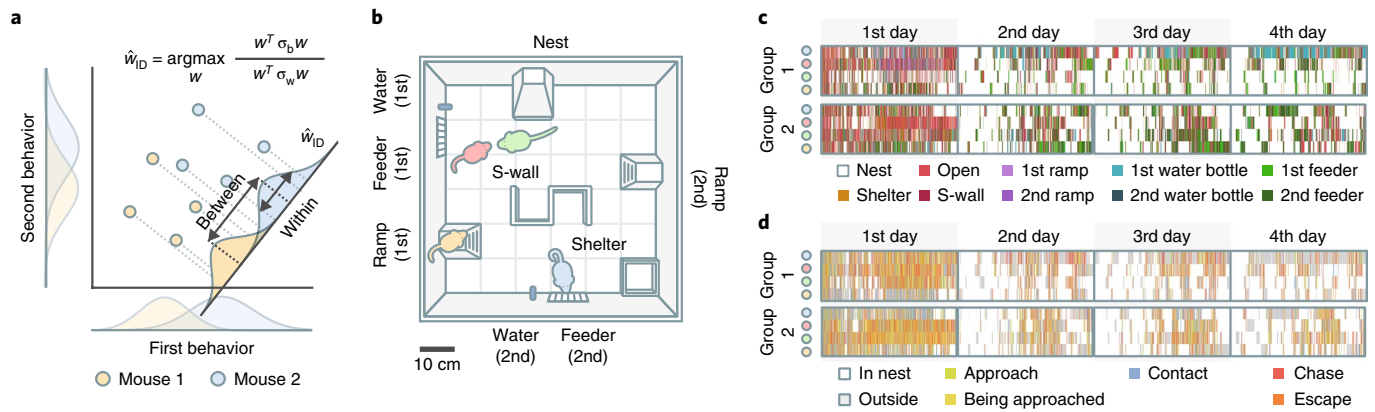
Since LDA is relatively resistant against redundancies in the input, we could afford to use all 60 initial behavioral readouts as input (Supplementary Fig. 1a), thus avoiding any bias that may have been introduced by a feature selection procedure. Likewise, while many readouts represent various normalizations of one underlying feature, these often carry additional information. For example, for ‘chase’ (the total number of chases) versus the same readout normalized to the time outside the nest, the latter holds additional meaning given the individual differences in time outside the nest.

Our analysis yielded four significant IDs that passed the threshold of less than 5% average overlap between individuals (Supplementary Fig. 2, ID5, the first dimension below this threshold, is shown for comparison). The dimensions produced this way were uncorrelated, although not necessarily orthogonal, thus resulting in four IDs, each spanning a different behavioral subspace (Fig. 2a).

To test the replicability of the four IDs, we used a separate dataset composed of control mouse measurements ( $n = 208$ ) in social boxes with a different layout (Supplementary Fig. 3a); this yielded only a subset of the current behavioral readouts (37 different readouts per

<sup>1</sup>Department of Stress Neurobiology and Neurogenetics, Max Planck Institute of Psychiatry, Munich, Germany. <sup>2</sup>Department of Neurobiology, Weizmann Institute of Science, Rehovot, Israel. <sup>3</sup>Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Munich, Germany.

<sup>4</sup>Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. <sup>5</sup>Department of Behavioral Biology, University of Osnabrück, Osnabrück, Germany. <sup>6</sup>These authors contributed equally: Oren Forkosh, Stoyo Karamihalev. \*e-mail: [alon\\_chen@psych.mpg.de](mailto:alon_chen@psych.mpg.de)



**Fig. 1 | From behavior to personality.** **a**, The task of finding stable and discriminative trait-like dimensions can be formulated as an optimization problem. We used LDA to reduce the multidimensional behavioral space by creating dimensions that maximize the ratio of inter-subject to intra-subject variability. **b**, Groups of four male mice, marked with dyes of four different colors for tracking purposes, were housed in an enriched environment where they could move and interact freely over multiple days. All of their movements were automatically tracked. Each arena contained a closed nest, two feeders, two water bottles, two ramps, an open shelter and an S-shaped separation wall in the center. **c,d**, Experimental movement (**c**) and social behavior (**d**) ethograms for two representative groups of mice show intra-individual consistencies and inter-individual variability. Each row represents a single mouse over time, and color codes indicate location (**c**) or either location or behavior (**d**).

mouse per active phase). The scores on the top four IDs obtained from this dataset correlated strongly with the respective original scores (Supplementary Fig. 3b). The strength of this relationship decreased steeply at ID5. We were therefore able to replicate the initial ID structure on an independent dataset despite the differences in setup and readouts.

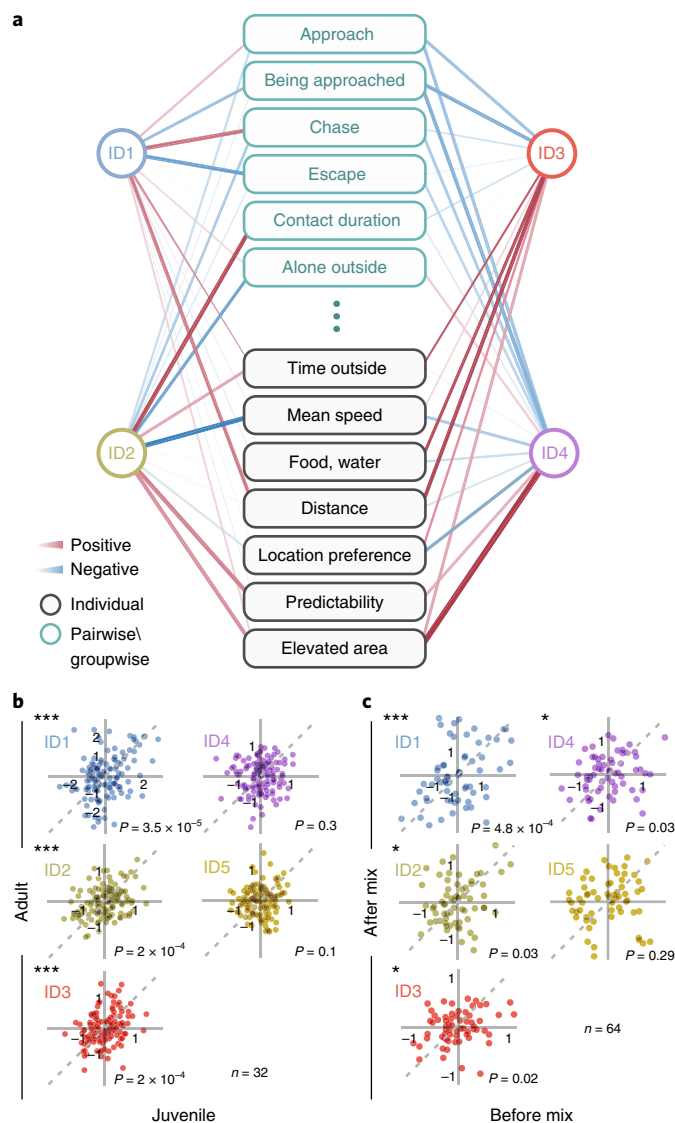
**IDs are stable over time, developmental stages and social contexts.** Having established and replicated the ID structure, we proceeded to experimentally validate the IDs. To assess the stability of ID scores over time, we first tested their self-similarity from an average of the first 4 days of the experimental period to the fifth day (Supplementary Fig. 4). All IDs fulfilled this criterion. We then tested their stability over developmental time by monitoring juvenile mice (8 groups, 4–5 weeks old) in the social box paradigm. The same mice were tested again as adults (15–16 weeks old). Individual scores on IDs 1–3 were stable over this prolonged period of time (Fig. 2b; Supplementary Fig. 5), which indicated that IDs assigned to juveniles captured individual differences that remained stable across developmental stages.

A major reason for the usefulness of investigating behavioral traits over specific behaviors is that traits more closely approximate the intrinsic properties of an individual and are therefore more robust against manipulations of the social environment. To test this property of IDs, mice from 16 groups (64 individuals) that had been assigned ID scores based on the 4-day baseline testing period were then shuffled into new groups, such that no mouse had ever been exposed to any of its new group members and were re-introduced to new arenas for another day of measurement (Supplementary Fig. 6). For adult male mice, this is a relatively dramatic and stressful manipulation and causes significant changes in many behavioral readouts, especially those related to general locomotion and aggression. Despite these changes, the scores for IDs 1–4, but not ID5, remained significantly self-similar to their baseline state (Fig. 2c). We additionally compared our model against a PCA run on the same dataset. In this analysis, only two out of the top four principal components (PCs) remained stable after this manipulation (Supplementary Fig. 7a–c). Thus, mice tended to maintain their distinguishing individual characteristics captured by IDs despite substantial changes to the social environment.

**IDs combine information from a variety of standard behavioral tests.** Having established that four IDs were stable over time and

across social contexts, we set out to assess their ability to predict a range of standard behaviors that are typically measured in classical mouse behavioral paradigms. To this end, we subjected mice with known ID scores to a battery of established behavioral assays (Fig. 3a,b). ID scores contained a significant portion of the information collected from classical tests (Fig. 3a). The pattern of correlations between the various tests and ID scores additionally suggested that IDs represent complex entities that could not be fully captured without comprehensive behavioral screening. Moreover, these relationships contribute to the notion that IDs carry information about the hidden factors that are co-modulated across the behavioral repertoire of an animal. For example, ID1 was correlated with a measure of dominance in the social hierarchy (David's score, Fig. 3c) and with features of locomotion in the open-field test (OFT) and of memory recall in the object recognition test. All of these behavioral readouts appear to be expressions of a common underlying trait.

**IDs capture transcriptomic variance in the brain.** A major advantage of animal models is the ability to mechanistically investigate the link between brain function and behavior. While the contributions of brain transcriptomic differences to human personality traits remain largely unexplored due to major technical difficulties in performing such studies, mouse ID scores may prove to be a useful proxy. To assess whether ID scores captured transcriptomic variance in the brain, we performed bulk RNA sequencing (RNA-seq) in mice that had been profiled in the social box ( $n=32$ ). For each individual, we sequenced the following three brain regions (Fig. 4a): the basolateral amygdala (BLA), the insular cortex (INS) and the medial prefrontal cortex (PFC). This yielded a total list of 13,073 genes that were jointly detectable in all three regions. For each region, we assessed the average variance explained across the gene set by all four IDs and compared it against a distribution derived from shuffling the ID scores across individuals. This test enabled us to not focus on specific associations between IDs and genes in a particular region, but rather to assess the ability of IDs to explain the variability in expression across the entire gene list. Strikingly, in all three regions, the IDs performed significantly better in their true configuration than would be expected by chance, which suggested that ID score assignment is remarkably fitting with regard to their association with gene expression (Fig. 4a). IDs contributed differently to the variance explained in each brain region (Fig. 4b), with occasional transcriptome-wide significant associations with specific genes (Fig. 4c,d). The same test



**Fig. 2 | Testing the IDs.** **a**, Running LDA in the 60-dimensional behavioral space resulted in four significant IDs. The IDs are uncorrelated between themselves and are represented by multiple overlapping behaviors (showing 13 representative behaviors out of 60; see Supplementary Fig. 9). The width of connecting lines reflects the strength of the correlation. **b**, ID scores of mice as juveniles (4–5 weeks old) remained significantly stable at adulthood (15–16 weeks old) for the first three IDs (each point represents the score of a single mouse on a single ID; one-sided permutation test,  $n = 32$  individuals; \*\*\* $P < 0.001$ ). **c**, ID scores of mice before and after being mixed into new social groups. Mice were significantly self-similar in their scores on IDs 1–4 across changing social environments. This relationship did not hold for ID5 (one-sided permutation test,  $n = 64$  individuals; \* $P < 0.05$ , \*\*\* $P < 0.001$ ).

performed using the top four PCs derived from the PCA did not yield significant findings in any of the three regions (Supplementary Fig. 7d). This was a surprising outcome, since the PCA-based analysis captures the maximum variance in behavior and may be expected to better predict a dynamic process such as gene expression. The finding that IDs outperform PCs in this instance implies that there may be genes with stable expression profiles that support stable individual differences in behavior.

**Personality space.** An important benefit that comes with having a known space of individual expression is the ability to search

that space for points of biological interest, which may represent behavioral specializations. We made a first attempt at this by using Pareto task inference on the two strongest IDs (ID1 and ID2). We found that ID1 and ID2 span a behavioral continuum on a triangle bounded by three personality archetypes (Fig. 5a). Such a configuration can be interpreted as a trade-off between three distinct evolutionary specializations, as previously shown for features of animal morphology<sup>5</sup> and *Caenorhabditis elegans* locomotion<sup>6</sup>. Analogous archetypes were found in the replication dataset.

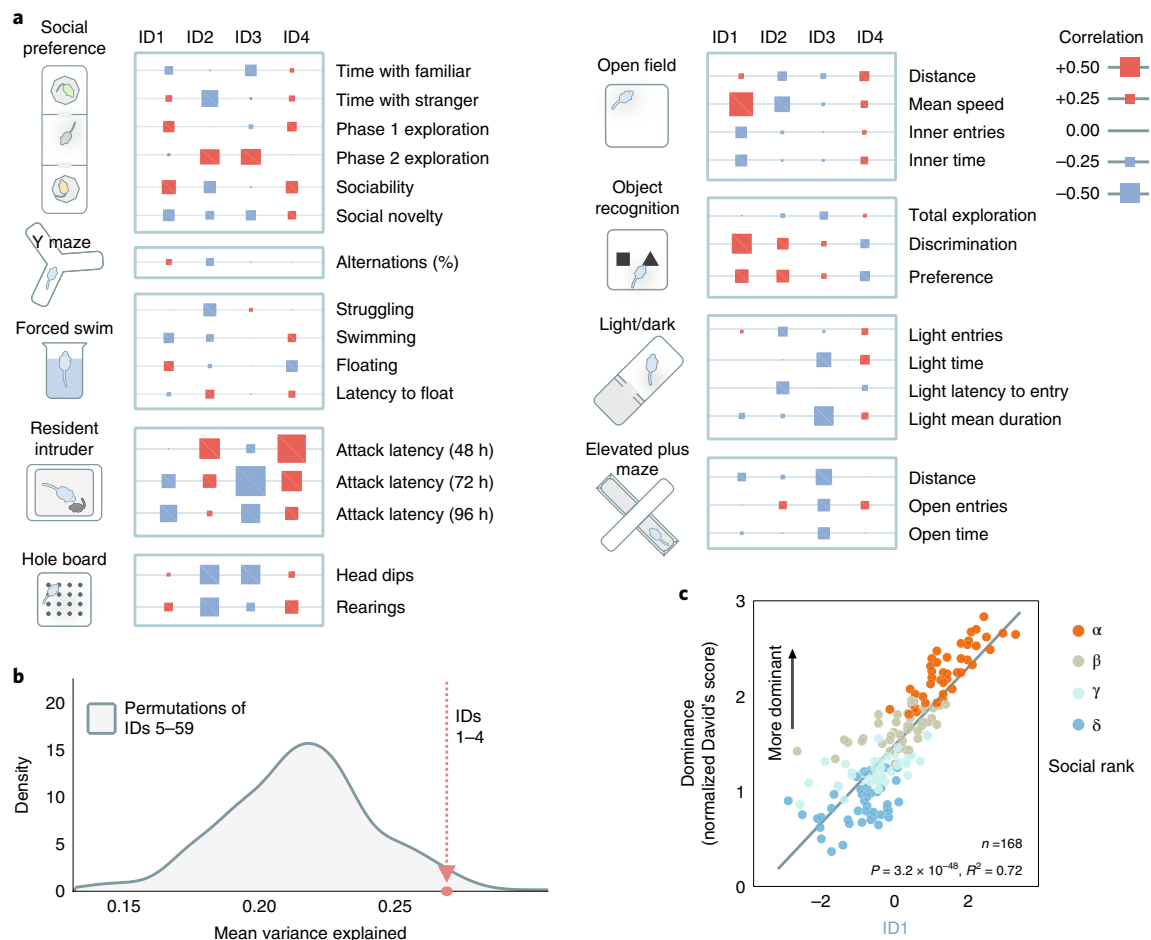
**IDs discriminate between genotypes.** Finally, we tested the ability of IDs to capture and discriminate between individuals with known genetically driven differences in behavioral tendencies. For this purpose, we used the high-anxiety versus normal-anxiety (HAB/NAB) model, wherein mice are selectively bred over multiple generations for different levels of anxiety-like behavior when subjected to the elevated plus-maze (EPM) test<sup>7</sup> (Supplementary Fig. 8a). We monitored heterogeneous groups composed of one HAB mouse and three NAB individuals each in the social box and assigned ID scores to them. We found that ID score assignment in its true configuration exhibited considerable power in discriminating between the genotypes (Fig. 5b; Supplementary Fig. 8b), with ID1 being the best predictor of genotype classification.

## Discussion

The systematic categorization of individual differences into constituent traits is an essential step toward more targeted explorations of their biological underpinnings. Human research in this direction, which typically uses questionnaires rather than behavioral input, has resulted in well-known and widely used taxonomies of human personality<sup>2,8</sup>. While the five-factor model, for example, replicates well across human cultures, some cross-cultural differences do exist<sup>9,10</sup>. Perhaps a more accurate analogy to what is usually referred to as personality in animals may be the concept of temperament. Individual differences in infants, for example, who are still relatively unaffected by the enculturation process, are typically described in psychology as temperamental traits<sup>11</sup>. We attempt to circumvent the terminological discussion by using the term ‘identity domain’ to refer to a trait-like dimension that is produced by maximizing the ratio of variability between individuals to the variability within. Regardless of terminology, each taxonomy is derived from a broad view of individual expression and assumes some degree of biological contribution to individual variability.

The structure of personality is often inferred by reducing the dimensions of a larger initial dataset. There are a variety of approaches available for dimensionality reduction, from the more commonly used PCA and factor analyses to various spectral techniques such as local linear embedding. Each method carries a set of underlying assumptions and may or may not require labeled data. Our choice of method was driven by theoretical considerations. We modeled IDs as dimensions with the following two main characteristics: the ability to capture individual differences (inter-individual) and stability over time (intra-individual). LDA capitalizes on the longitudinal aspect of behavioral data, reducing its dimensions so as to simultaneously maximize differences between individuals while minimizing variability within individuals over time. The importance of preserving stability during the decomposition process is demonstrated by the comparatively reduced performance of PCA on the same dataset (Supplementary Fig. 7).

LDA carries an underlying assumption of linearity, namely, that behaviors should map to IDs in a linear way. While nonlinear versions of the analysis are available, these require larger datasets and may produce noisier results. We attempted to overcome the limitations posed by this assumption by introducing nonlinearity to the model using different normalizations of the input behaviors and by



**Fig. 3 | IDs are reflected in multiple standard behavioral tests. a**, ID scores predict multiple readouts across standard tests (two-tailed Pearson's correlation statistic). Moreover, some standard test readouts are related to multiple IDs. The strength of each correlation is represented by the size of the squares (red, positive; blue, negative). **b**, The variance explained by ID scores in standard behavioral assays. IDs 1–4 explain significantly more variance in 'classical' behavioral test readouts than random sets of 4 from IDs 5–59. **c**, ID1 scores predict social dominance levels (David's scores), calculated based on the number and directionality of aggressive interactions ( $n=168$  individuals,  $R^2=0.72$ ,  $F(1, 166)=434.65$ ,  $P=3.19 \times 10^{-48}$ ).

warping the data to fit a normal distribution. Fitting the data to a normal distribution also reduced the negative impacts of outliers.

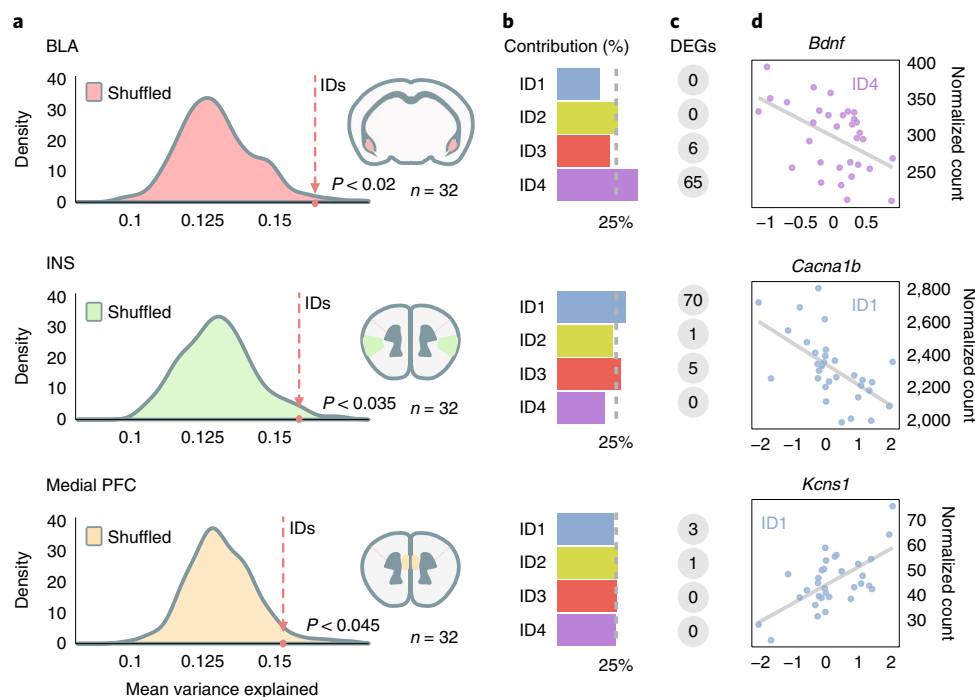
A general limitation of our approach is that the resulting ID space depends on the choice of behaviors used as input. We tried to contain this by expanding the behavioral space to everything we could observe and categorize. We additionally validated the ID structure in a separate dataset with different readouts used as input, and we showed that the specific list of readouts does not have a major impact on the outcome (Supplementary Fig. 3). Nevertheless, it is conceivable that the result of such an analysis may be swayed toward traits that are more easily detected using a given behavioral assessment device. Accordingly, in a setup in which groups of male mice cohabitate in an enclosed environment, the most prominent dimension (ID1) is related to social dominance, and individual differences such as those from the domains of learning and memory, which typically require stringent learning protocols, are likely to be underrepresented.

In addition, while the social box represents a relatively enriched and interactive environment, it is nevertheless modest and simplistic compared to the natural habitats of mice. We therefore suspect that the four IDs described here do not capture the entire extent of individual differences that mice might reveal in more complex and diverse situations. Likewise, the finding of four unique IDs only sets a lower bound on the total number of IDs that may best describe

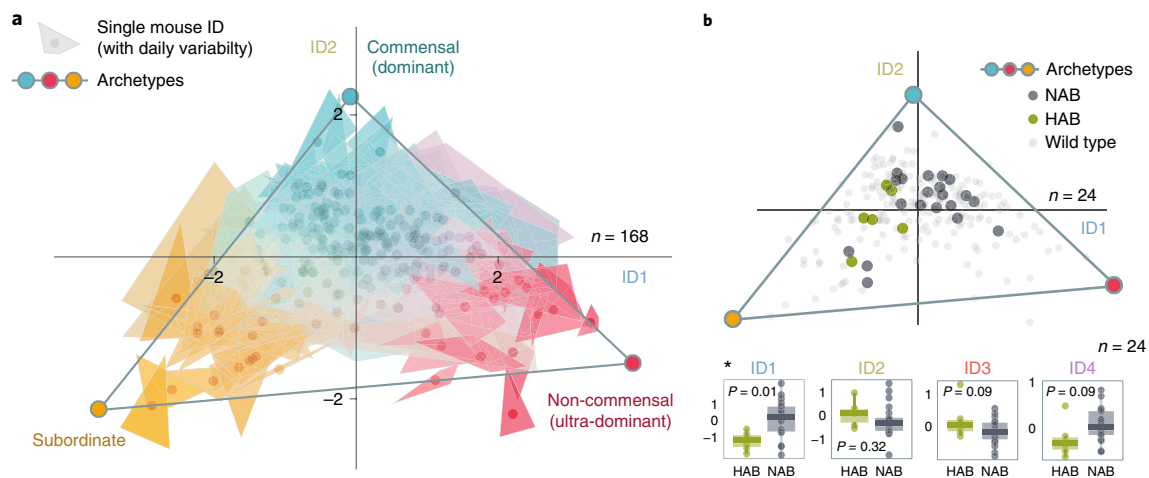
the entire space of behavioral expression of individual differences in mice. More IDs may emerge given a broader range of observations.

An important advantage of this approach, however, is that there is no principal limit on the number and breadth of additional readouts that can be used to supplement our 60-dimensional behavioral space. In addition, there is no principal objection against expanding the initial dataset beyond behavioral readouts to also include physiological (for example, heart rate and blood pressure) or neural (for example, electro-encephalography) input. We anticipate that as the inputs become more varied, covering additional sections of the phenotypic repertoire, the resulting ID space will come ever closer to the true underlying structure of individual differences in mice.

It is important to note that the personality space we describe was not arbitrarily structured. Instead, we found a well-defined triangle at the intersection of ID1 and ID2. One plausible explanation for this geometry comes from the concept of Pareto optimality. For a biological system that faces multiple tasks, this is a state wherein improvements on one task are only possible at the expense of another. In nature, evolutionary trade-offs are a driving force for Pareto optimality. It was recently shown that optimal states translate to polygons in trait spaces, with the corners representing archetypal optimality in a single evolutionary task<sup>5</sup>. For example, the famous Darwin's finches were shown to have three morphologically distinct beak-shape archetypes that are optimized for the consump-



**Fig. 4 | IDs carry information on gene expression in the brain.** **a**, RNA-seq results from the BLA, the INS and the PFCs of mice collected after baseline behavioral assessment. Plotted are the distributions of mean  $R^2$  values for 200 models with 4 shuffled IDs as predictors. The ID scores explain significant amounts of variance in the transcriptome (permutation test, true  $R^2$  value indicated by red circle and broken arrow line;  $n = 32$  mice). **b**, The relative contributions of each ID to the transcriptomic variance explained in each region (the broken line reflects the 25% expectation given equal contributions). **c**, Numbers of differentially expressed genes (DEGs) per ID after multiple testing correction. **d**, Expression of a representative gene from each region plotted against individual scores on the associated ID.



**Fig. 5 | Personality space.** **a**, The personality space captured by the first two IDs forms a triangle with three archetypes at its corners ( $t$ -ratio test,  $P = 0.006$ ; other ID combinations and different archetype numbers did not yield significant results). These archetypes may correspond to three behavioral strategies that mice exhibit in nature: commensal, non-commensal (ultra-dominant) and subordinate. The ID scores of each individual are represented by a trapezoid, with scores on each day marking the four corners (a triangle is depicted if the fourth point falls inside the shape). Thus, the size of each trapezoid reflects the stability of ID scores for each individual over time (smaller means more stable). **b**, Mice bred for high anxiety-related behavior (HAB) and normal anxiety-related behavior (NAB) were assessed using the IDs. A significant genotype effect was detected by ID1 (two-tailed  $t$ -test,  $t(25) = -2.86$ ,  $P = 0.01$ ,  $n = 24$  individuals;  $*P < 0.05$ ) but not by other IDs (ID2:  $t(25) = -1.01$ ,  $P = 0.32$ ; ID3:  $t(25) = 1.79$ ,  $P = 0.09$ ; ID4:  $t(25) = 1.76$ ,  $P = 0.09$ ). Boxes represent the 25%, 50% (median) and 75% quantiles, and whiskers span from minima to maxima. The Pareto space also reveals that HAB mice tend to be more subordinate (top).

tion of different foods<sup>5</sup>. Archetype analysis has been used to infer the functional identities of cell types based on their gene-expression profiles<sup>12</sup>, and to detect activity patterns related to motor tasks in human functional MRI experiments<sup>13</sup>. The existence of a three-

archetype structure at the intersection of ID1 and ID2 indicates that IDs may be capturing evolutionarily meaningful variation in individual differences. While we were able to describe the archetypes in light of the behavioral associations of ID1 and ID2, inferring the

selection pressures that may have led to such specializations is much less straightforward. Further research is needed to better assess the origins and meaning of the presented archetype structure.

Personality is a complex entity that reflects stable individual differences and, in so doing, maps the space of phenotypic variability. Here, we showed that IDs provide a bias-free surrogate measure of personality obtained directly from behavioral data. IDs show considerable stability over time, developmental stages and across social contexts. They allow quantitative exploration of personality differences in organisms in which such analyses were previously inaccessible. By drawing on consistent inter-individual differences, IDs captured the essential features of personality, thus offering access to a biologically meaningful and evolutionarily relevant meta-behavioral phenotype.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information, details of author contributions and competing interests, and statements of code and data availability are available at <https://doi.org/10.1038/s41593-019-0516-y>.

Received: 21 September 2018; Accepted: 16 September 2019;  
Published online: 4 November 2019

### References

1. Eysenck, J. H. *The Structure of Human Personality* (Methuen & Co., 1953).
2. McCrae, R. R. & Costa, P. J. *Personality in Adulthood: A Five-Factor Theory Perspective* (Guilford Press, 2002).
3. Shemesh, Y. et al. High-order social interactions in groups of mice. *eLife* **2**, e00759 (2013).
4. Shemesh, Y. et al. Ucn3 and CRF-R2 in the medial amygdala regulate complex social dynamics. *Nat. Neurosci.* **19**, 1489–1496 (2016).
5. Shoval, O. et al. Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* **336**, 1157–1160 (2012).
6. Gallagher, T., Bjorness, T., Greene, R., You, Y. J. & Avery, L. The geometry of locomotive behavioral states in *C. elegans*. *PLoS One* **8**, e59865 (2013).
7. Krömer, S. A. et al. Identification of glyoxalase-I as a protein marker in a mouse model of extremes in trait anxiety. *J. Neurosci.* **25**, 4375–4384 (2005).
8. Butcher, J. N. Minnesota Multiphasic Personality Inventory. in *The Corsini Encyclopedia of Psychology* (eds Weiner, I. B. & Craighead, W. E.) <https://doi.org/10.1002/9780470479216.corpsy0573> (2010).
9. McCrae, R. R., Costa, P. T., Del Pilar, G. H., Rolland, J.-P. & Parker, W. D. Cross-cultural assessment of the five-factor model. *J. Cross Cult. Psychol.* **29**, 171–188 (1998).
10. Triandis, H. C. & Suh, E. M. Cultural influences on personality. *Annu. Rev. Psychol.* **53**, 133–160 (2002).
11. Rothbart, M. K. Measurement of temperament in infancy. *Child Dev.* **52**, 569–578 (1981).
12. Hart, Y. et al. Inferring biological tasks using Pareto analysis of high-dimensional data. *Nat. Methods* **12**, 233–235 (2015).
13. Hinrich, J. L. et al. Archetypal analysis for modeling multisubject fMRI data. *IEEE J. Sel. Top. Signal Process.* **10**, 1160–1171 (2016).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

## Methods

**Animals.** All animal experiments were approved by the Animal Care and Use Committees of either the Government of Upper Bavaria (Munich, Germany) or the Weizmann Institute of Science (Rehovot, Israel).

Male CD-1 (ICR) mice aged 8–12 weeks during the assessment were used for all experiments, with the exception of the HAB/NAB animals (see below). The animals were housed in an specific-pathogen-free facility in temperature-controlled rooms under standard conditions with a 12h light–dark cycle (lights on at 8:00). After weaning, non-sibling individuals were randomly assigned to be housed in groups of four per cage. At around 7–8 weeks of age, the mice were transferred to the behavioral testing rooms and painted. All animals were housed in a temperature-controlled environment with food and water available *ad libitum*.

**Painting.** The fur of each mouse was painted to enable identification by automatic video tracking. Painting was carried out under mild isoflurane anesthesia using commercially available semi-permanent hair dyes of the following three colors: Pillarbox Red, Voodoo Blue and Sunshine (Tish & Snooky's NYC). A fourth color, a green hue, was achieved by mixing the latter two dyes. The dyes were applied using a paint brush. Excess color was removed from the animal's fur with tissues. The period under anesthesia was typically no longer than 10 min.

Mice were single-housed for 1–2h after painting and subsequently reunited with their cage mates. A minimum of 3 days of recovery and habituation was allowed following this procedure before the mice were introduced into the social arenas.

**Social box setup.** Mice were studied in a specialized arena designed for automated tracking of individual and group behavior. Each arena housed a group of four male mice. The arena consisted of an open 60 × 60 cm box and included the following objects: a covered nest, an open shelter, an S-shaped wall, two water bottles, two feeders and two elevated ramps. Food and water were available *ad libitum*. The arenas were illuminated at 2 lux during the dark phase (12h) and at 200 lux (using light-emitting diode lights) during the light phase (12h). A color-sensitive camera (Manta G-235C, Allied-Vision) was placed 1 m above the arena and recorded the mice during the dark phase. Mouse trajectories were automatically tracked off-line using specially written software in Matlab (Mathworks).

To validate the IDs and ensure repeatability, we also computed the ID scores for mice that were recorded in arenas of a different design<sup>3</sup>. These alternative arenas were 75 × 50 cm and included a covered nest, a closed shelter (which was smaller than the nest and had only one entry), two elevated ramps, two feeders, a single water bottle, an elevated block that is away from the walls and a Z-wall.

**Identification and classification of interactions between mice.** We automatically identified and classified interactions between mice as events in which the distance between two mice ( $d$ ) was less than 10 cm. We then used the movement direction of one mouse relative to another mouse ( $\theta$ ) to identify the nature of the contact for either of the mice. If for mouse A, the projection of the direction of its movement relative to mouse B was small enough ( $|\tan(\theta) \cdot 2d| < \theta_1$ , for  $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$ ), then it was considered as moving toward B. If  $|\tan(\theta) \cdot 2d| < \theta_2$ , for  $\frac{\pi}{2} < \theta < \frac{3\pi}{2}$ , then it was moving away from it. Otherwise, it was assumed that the mouse was idle with respect to the other mouse ( $\theta_1$  and  $\theta_2$  were found by optimization).

To classify aggressive and non-aggressive contacts, we first used a hidden Markov model<sup>14</sup> to identify post-contact behaviors in which mouse A was moving toward B, and B was moving away from A (A was following B). We then used 500 manually labeled events to learn statistical classifiers of aggressive and non-aggressive post-contact behavior. For each event, we estimated a range of parameters (including individual and relative speed, and distance) and optimized a quadratic discriminant classifier<sup>15</sup> (a  $k$ -nearest neighbor algorithm based on these parameters) and a decision-tree classifier (which used these parameters at each tree intersection)<sup>16</sup>. We found that for a test set of 1,000 events, none of these classifiers was individually accurate enough. However, a combined approach, in which we labeled an event as 'aggressive' if any of the classifiers labeled it as such, gave ~80% detection with 0.5% false alarms.

**David's score for dominance.** We used the normalized David's score to assign each individual with a continuous measure of its social rank<sup>17</sup>. David's score assumes a linear hierarchy whereby each pair from the group includes a more-dominant and a less-dominant individual. The score is based on the measure of the fraction of interactions in which mouse  $i$  chased mouse  $j$  relative to the total number of agonistic interactions, which we denote as  $P_{ij}$ . David's score of each individual is the sum

$$DS_i = w^i + w_2^i - l^i - l_2^i$$

where  $w^i = \sum_{j \neq i} P_{ij}$  is the sum of the fraction of times that mouse  $i$  has 'won' (that is, was the chaser) and  $w_2^i = \sum_{j \neq i} w^j P_{ji}$  is a similar sum weighted by the  $w^j$  of the other mice. Meanwhile,  $l^i = \sum_{j \neq i} P_{ji}$  is the sum of the fractions of 'losses' (escapes) and, accordingly,  $l_2^i = \sum_{j \neq i} w^j P_{ji}$  is its weighted sum. The score is then normalized to be between 0 and  $N-1$  (where  $N$  is the number of subjects, which is four in our case) by using the following formula:

$$\text{NormDS} = \frac{1}{N} \left( DS + \frac{N(N-1)}{2} \right)$$

**LDA.** LDA is a method for finding an optimal linear separator between two or more clusters, or, equivalently, finding a subspace in which the differences between the clusters are the greatest<sup>18</sup>. Such a subspace can serve as a lower dimensional representation of the data much like the commonly known PCA is used for dimensionality reduction. There are two notable differences between LDA and PCA. First, unlike PCA, LDA is a supervised method that requires that each data point would be labeled (in our case, the labels are the identities of the mice). Second, while PCA only seeks to maximize the variance of the entire data, LDA also minimizes the variability within each class (here, minimizing the variability in the behaviors of each mouse across the 4 days of the experiment). These properties of LDA make it suitable for capturing stable behavioral traits.

In mathematical terms, linear discriminant analysis works by finding a projection matrix  $W$  that both maximizes the variability between clusters  $\Sigma_b$  and minimizes the variability within a cluster  $\Sigma_w$  (see below for a formal definition of  $\Sigma_b$  and  $\Sigma_w$ ). The ratio between the projection of these two is known as the Fisher–Rao discriminant, which is defined as follows:

$$w = \operatorname{argmax}_w \operatorname{tr} \left( \frac{w^T \Sigma_b w}{w^T \Sigma_w w} \right)$$

which can be simplified by reformulating it using a Lagrange multiplier  $\lambda$  such that

$$\mathcal{L}(w) = w^T \Sigma_b w - \lambda (w^T \Sigma_w w - 1)$$

The solution to this Lagrangian is obtained at its stationary points with respect to  $w$ , or

$$\frac{d}{dw} \mathcal{L}(w) = 2\Sigma_b w - 2\lambda \Sigma_w w = 0$$

and the result is the matrix of eigenvectors of  $\Sigma_w^{-1} \Sigma_b$

For comparison, PCA can be defined in a similar way as follows:

$$w = \operatorname{argmax}_w \frac{w^T \Sigma w}{w^T w}$$

where we only take into account the total covariance of the data ( $\Sigma$  is actually equal the sum of  $\Sigma_b$  and  $\Sigma_w$  up to a factor; see below). Similar to LDA, this quotient is solved by rewriting the equation as a Lagrangian and finding the stationary points of

$$\mathcal{L}(w) = w^T \Sigma w - \lambda (w^T w - 1)$$

The solution to the PCA problem is obtained by finding the top eigenvectors of  $\Sigma$ .

Using LDA to infer the IDs in our study, we started with an ensemble of  $1 \times 60$  dimensional behavioral vectors, one vector for each mouse  $m$  and for each day  $d$ ,

$$b_m(d) = (b_m^1(d), b_m^2(d), \dots, b_m^{60}(d))$$

where each dimension  $b_m^i(d)$  represents one normalized measured behavior (for example, the time that mouse  $m$  spent outside on the  $i$ th day; see the software available through the link in the Code availability statement). The mouse index  $m$  is in the range  $m = 1, \dots, M$ , where  $M$  is the total number of mice ( $M = 168$ ) and the days  $d$  vary between  $d = 1, \dots, D$ , where  $D$  is the total number of days ( $D = 4$ ). We performed LDA on this ensemble and obtained a set of basis vectors. Choosing only eigenvalues that indicate a greater contribution of between-individual over within-individual variability (Supplementary Fig. 2), we arrived at four eigenvectors for the projection matrix that spans the ID space  $E^{\text{ID}} = (e_1^{\text{ID}}, \dots, e_4^{\text{ID}})$ . Thus, we could now represent the IDs of any mouse  $m$  as a  $1 \times 4$  dimensional vector (since there are four IDs) such that

$$\text{id}_m = \langle E^{\text{ID}} \cdot b_m(d) \rangle_d$$

where the average  $\langle \cdot \rangle_d$  is carried out along all 4 days of the experiment (see the software available through the link in the Code availability statement).

**Fisher–Rao discriminant.** The Fisher–Rao discriminant is a measure of how separable two or more clusters are. It assumes that clusters are more distinct when the distance between them  $\Sigma_b$  is large relative to the size of each cluster by itself  $\Sigma_w$ . In mathematical terms, the Fisher–Rao discriminant is defined as the ratio of these two quantities

$$D_{\text{FR}} = \operatorname{tr} \left( \frac{\Sigma_b}{\Sigma_w} \right)$$

where  $\operatorname{tr}()$  is the trace function (which is needed in the case of a multivariate space in which  $\Sigma_w$  and  $\Sigma_b$  are matrices). The variability (or distance) between clusters is defined as

$$\Sigma_b = \sum_{c=1}^C n_c (\mu_c - \mu) (\mu_c - \mu)^T$$

and the variability within (or size of) the clusters is defined as

$$\Sigma_w = \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu}_{c_i})(\mathbf{x}_i - \boldsymbol{\mu}_{c_i})^T$$

where  $\boldsymbol{\mu}$  is the global mean,  $\boldsymbol{\mu}_{c_i}$  is the mean of the cluster associated with the  $i$ th sample  $\mathbf{x}_i$ ,  $n_i$  is the number of samples in cluster  $c_i$ ,  $n = \sum n_i$  is the total number of samples, and  $C$  is the number of clusters.

Note that the sum of the within-variability and between-variability is proportional to the total covariance of the data ( $\Sigma$ )

$$\Sigma = \frac{\Sigma_b + \Sigma_w}{n - 1}$$

In our case, and using the nomenclature defined in the previous section, as each cluster represents the measured behaviors of one mouse  $m$  along 4 days, the Fisher–Rao discriminant equals

$$D_{FR} = \text{tr} \left( \frac{\Sigma_b}{\Sigma_w} \right) = \text{tr} \left( \frac{\sum_{m=1}^M D(\boldsymbol{\mu}_m - \boldsymbol{\mu})(\boldsymbol{\mu}_m - \boldsymbol{\mu})^T}{\sum_{d=1}^D \sum_{m=1}^M (\mathbf{b}_m(d) - \boldsymbol{\mu}_m)(\mathbf{b}_m(d) - \boldsymbol{\mu}_m)^T} \right)$$

where, again,  $m = 1, \dots, M$  is the mouse index ( $M = 168$ ),  $d = 1, \dots, D$  is the day index ( $D = 4$ ),  $\mathbf{b}_m(d)$  is the behavioral vector of mouse  $m$  at day  $d$  (vector of size  $1 \times 60$ ), and

$$\boldsymbol{\mu}_m = \frac{1}{D} \sum_{d=1}^D \mathbf{b}_m(d) \text{ and } \boldsymbol{\mu} = \frac{1}{M} \sum_{m=1}^M \boldsymbol{\mu}_m$$

**IDs.** Personality cannot be directly measured, but it can be inferred from the behavior. We used the measured behavior of the mice for each of the 4 days they spent in the social box. The normalized measured behaviors of mouse  $m$  on day  $d$  are denoted by a vector  $\mathbf{x}_{m,d}$  of dimension 60 (the number of behaviors; an explanation of the normalization procedure follows). The ID  $I_m$  defines a distribution on the behavior space in the following way:

$$\mathbf{x}_{m,d} = A I_m + \boldsymbol{\varepsilon}$$

where  $A$  is a matrix linking the IDs to the behaviors, and  $\boldsymbol{\varepsilon}$  is a distribution term (or noise due to variability or external factors). To estimate the IDs, this equation was reversed to find a  $W$  that would give us

$$I_m = W(\mathbf{x}_{m,d} + \boldsymbol{\varepsilon})$$

Note that since  $A$  is not usually square,  $W$  is not simply a reversal of  $A$ .

To find  $W$ , which in turn would give us the IDs, we used LDA. Here the variability-within is defined as the variability of the same individual mouse on different days, or

$$\Sigma_w = \sum_{m=1}^M \sum_{d=1}^D \left( \mathbf{x}_{m,d} - \sum_{d'=1}^D \mathbf{x}_{m,d'} \right) \left( \mathbf{x}_{m,d} - \sum_{d'=1}^D \mathbf{x}_{m,d'} \right)^T$$

where  $M$  is the total number of mice ( $M = 168$ ) and  $D$  is the total number of days ( $D = 4$ ). Accordingly, the variability-between is the variability between mice or

$$\Sigma_b = D \cdot \sum_{m=1}^M \left( \sum_{d=1}^D \mathbf{x}_{m,d} - \boldsymbol{\mu} \right) \left( \sum_{d=1}^D \mathbf{x}_{m,d} - \boldsymbol{\mu} \right)^T$$

The link between the behavior and the IDs is obtained in the same way as in the classical LDA by solving

$$W = \text{argmax}_{W'} \text{tr} \left( \frac{W'^T \Sigma_b W'}{W'^T \Sigma_w W'} \right)$$

Once we have  $W$ , we found  $A$  by solving it as a linear regression problem of the form

$$\mathbf{x}_{m,d} = A I_m + \boldsymbol{\varepsilon}$$

We tested 42 groups of mice in 6 separate batches (8 groups of mice were tested simultaneously in 8 separate social arenas each time). To avoid batch effects and data drifts, we used quantile normalization on each behavior for each batch on each day separately. We quantile-normalized the data to have a normal distribution by computing the quantile of each sample and then computing the inverse of the normal cumulative distribution function (also known as the ‘probit’ function). If two or more samples were identical before the normalization, they were all assigned the same value after normalization.

**Pareto optimality.** To survive and reproduce, animals are constantly confronted with tasks such as finding food or evading predators. There is often a trade-off between tasks so that the success of an animal in one task has to come at the expense of its performance on another. Recent work has shown that the best phenotypes are the weighted average of archetypes, which are phenotypes that specialize in one task<sup>3,12</sup>. These phenotypes can either be morphological, as for the beak sizes of ground finches, or behavioral phenotypes.

The shape of the phenotype space is determined by the number of archetypes or by the number of tasks the animal faces. In case of two archetypes, optimal phenotypes would fall on the line connecting the two archetypes, while if there are three archetypes, the phenotypes would be contained inside a triangle, and so on.

One direct outcome of this theory is that looking at the entire phenotype space makes it possible to deduce the location of the archetypes and thereby the different biological challenges that an animal can face. The positions of the archetypes are found using a hyperspectral unmixing algorithm. The data are first centered to have zero mean, and then projected using PCA into a subspace with dimension  $n - 1$  (where  $n$  is the number of archetypes). Then, an unmixing algorithm is used to fit a  $n$  vertices polytope that best fits the data. Here, we used minimal volume simplex analysis, which is suitable for relatively small datasets since it does not allow for outliers. The analysis was performed in Matlab using the package Pareto task inference package.

**Social box behavioral readouts.** We collected a total of 60 different readouts for each mouse on each day. Due to the linearity of LDA, some of the behaviors we measured were computed with several different normalizations. The most common normalizations we used were the total time in the arena (12 h; abbreviated total), the time outside the nest (outside), and for interactions, the total number of contacts (contacts). Definitions of the readouts are provided below.

*Pairwise.*

- Time outside (1): the fraction of time that the mouse spends outside the nest. Normalizations: total time (%).
- Frequency of visits outside (2): the rate at which the mouse exits the nest. Normalizations: total time (per 1 h).
- Foraging correlation (3): the correlation between the times that the mouse is outside the nest and the times that another mouse is outside the nest, averaged over all mice. For example, the foraging correlation between two mice would equal one if the mouse is always outside the nest when the other mouse is outside, and also enters the nest whenever the other mouse enters the nest. The correlation would be  $-1$  whenever the mouse is outside and the other mouse is inside the nest. Normalizations: (3) none (arbitrary units (a.u.)).
- Contact rate (4, 5): the number of contacts the mouse had. A contact is defined as two mice being less than 10-cm apart while both are outside the nest. Normalizations: (4) total time (per 1 h), (5) time outside (per 1 h).
- Time in contact (6): the fraction of time that a mouse is in contact with other mice while outside the nest. Normalizations: (6) time outside (per 1 h).
- Median/mean contact duration (7, 8): median or mean duration of contacts. The contact duration does not include the times when the mouse approached, moved away from or chased the other mouse. Normalizations: (7, 8) none (s).
- Follow (12, 18, 24): a follow is a contact that ended with one mouse going after another mouse until disengagement. Follows can be either aggressive (chases) or non-aggressive. Normalizations: (12) number of contacts (a.u.), (18) time outside (per 1 h), (24) total time (per 1 h).
- Being followed (13, 19, 25): the number of times a mouse is followed at the end of a contact. It can be either in an aggressive (chases) or non-aggressive manner. Normalizations: (13) number of contacts (a.u.), (19) time outside (per 1 h), (25) total time (per 1 h).
- Chase (10, 16, 22): chases are interactions that ended with the mouse pursuing another mouse in an aggressive manner. Aggressiveness was determined using a classifier that was trained on labeled samples (see Methods). Normalizations: (10) number of contacts (a.u.), (16) time outside (per 1 h), (22) total time (per 1 h).
- Escape (11, 17, 23): the number of time that the mouse was aggressively chased by another mouse. Normalizations: (11) number of contacts (a.u.), (17) time outside (per 1 h), (23) total time (per 1 h).
- Non-aggressive follow (14, 20, 26): the number of times the mouse followed another mouse at the end of a contact in a non-aggressive way. Normalizations: (14) number of contacts (a.u.), (20) time outside (per 1 h), (26) total time (per 1 h).
- Non-aggressively being followed (15, 21, 27): the number of times the mouse was followed by another mouse at the end of a contact in a non-aggressive way. Normalizations: (15) number of contacts (a.u.), (21) time outside (per 1 h), (27) total time (per 1 h).
- Approach (28–32): an approach is a directed movement of the mouse toward another mouse that ends in contact. Not all interactions necessarily start with an approach, while others might start mutually, with both mice approaching each other. Normalizations: (28) none (a.u.), (29) time outside (per 1 h), (30) time outside with one or more mice (per 1 h), (31) number of contacts (a.u.), (32) total (per 1 h).
- Being approached (33–35): the number of times the mouse was approached by another mouse. Normalizations: (33) number of contacts (a.u.), (34) time outside (per 1 h), (35) none (a.u.).
- Approach escape (36): the fraction of contacts in which the mouse initiated the contact and ended up being chased. Normalizations: (36) number of aggressive contacts (a.u.).
- Difference between approaches and chases (9): the total number of chases subtracted from the total number of approaches. Normalizations: (9) none (a.u.).

*Individual.*

- Region of interest exploration (37, 38): quantifies the amount of exploration the mouse is doing. Measured as the entropy of the probability of being in each of the ten regions-of-interest (ROIs). Mice that spend the same amount of time in all regions will be given the highest score, while mice that spend all their time in a single ROI will be scored zero. When normalized to the time outside, the computation of the entropy also differed by ignoring the probability of being inside the nest. Normalizations: (37) none (bits), (38) time outside (bits $h^{-1}$ ).
- Grid exploration (59): quantifies the amount of exploration the mouse is doing. Analogous to ROI exploration, grid exploration was determined using entropy; however, instead of looking at the ROIs, we divided the arena into a 6×6 grid (10×10 cm; a total of 36 possible locations). Normalizations: (59) none (bits).
- Predictability (60): measures how predictable the paths that the mouse takes as the mutual information between its current and previous location in the arena. For that, the arena was divided into a 6×6 grid (10×10 cm; a total of 36 possible locations), and for each cell, we computed the probabilities of it moving to any of the adjacent cells. Normalizations: (60) none (bits).
- Distance (58): the total distance traveled by the mouse while outside the nest. To smooth the tracking, the mice locations were sampled once every second. Normalizations: (58) none (m).
- Median/mean speed (54, 55): the median or mean speed while outside the nest. To smooth the computation of the speed, the locations of mice were sampled once every second. Normalizations: (54, 55) none (m s $^{-1}$ ).
- Tangential velocity (56): the tangential component of the speed or the part of speed perpendicular to the previous direction of movement. Normalizations: (56) none (m s $^{-1}$ ).
- Angular velocity (57): the rate of change in the direction of the mouse. Normalizations: (57) none (rad s $^{-1}$ ).
- Food or water (39, 40): the time spent in the feeder or water zones. Normalizations: (39) total time (a.u.), (40) time outside (a.u.).
- Food (41): the time spent next to the feeders. Normalizations: (41) time outside (a.u.).
- Water (42): the time spent next to the water bottles. Normalizations: (42) time outside (a.u.).
- Feeder preference (43): the time spent at (or near) the feeder adjacent to the nest (feeder 1) relative to the farther-away feeder (feeder 2). Normalizations: (43) none (a.u.).
- Water preference (44): time spent near the water bottle adjacent to the nest (water 1) relative to the farther-away water bottle (water 2). Normalizations: (44) none (a.u.).
- Elevated area (45, 46): the time spent on an elevated object in the arena: ramps or block. Normalizations: (45) total time (a.u.), (46) time outside (a.u.).
- Open area (47): the time spent in the open area (outside the nest and in any of the ROIs). Normalizations: (47) time outside (a.u.).
- Shelter (48): the time spent in the shelter, which is a box without a roof and is closed only on its sides. Normalizations: (48) time outside (a.u.).
- Ramps (49): the time spent on the elevated ramps. Normalizations: (49) outside (a.u.).
- S-wall (50): the time spent in the S-wall. Normalizations: (50) time outside (a.u.).
- Distance from walls (51): the average distance from the walls while in the open area. Normalizations: (51) none (cm).
- Distance from nest (52): the average distance from the nest (while outside the nest). Normalizations: (52) none (cm).
- Alone outside (53): the fraction of time the mouse is outside while all other mice are in the nest. Normalizations: (53) total time (a.u.).

**Standard behavioral assays.** All behavioral tests were performed within 2 weeks following the social arena assessment in the same test room in two batches of eight groups each using two different test sequences. During this time, the mice were housed in their original groups in standard polycarbonate cages. For all tests that were assessed by human observers, the observers were blinded to the specific ID scores of the mice.

Timeline 1 consisted of the OFT and novel object recognition (NOR) test, the social preference test (SPT) and the dark-light transfer (DaLi) test. Timeline 2 consisted of the OFT and NOR test, followed by the EPM, the DaLi test, the SPT and the forced-swim test (FST). For both timelines, each test was followed by a minimum of 48 h of rest.

**OFT and NOR test.** The OFT and NOR tests were performed in 60×60 cm boxes under minimal illumination (2–3 lux) in 3 sessions. Each animal was introduced to the arena for 5 min then briefly removed and reintroduced for an additional 5 min to the same arena, now with two identical objects placed at predetermined locations (acquisition phase). Finally, each animal was reintroduced to the arena after a 4-h delay for 5 min (retrieval phase), with one of the two identical objects replaced by a novel one<sup>18</sup>. Object preference was calculated as the novel/familiar object ratio, while the discrimination index was calculated as the ratio of preference to the retrieval phase total exploration time.

**DaLi test and EPM tests.** Anxiety-like behaviors were assessed using the DaLi or EPM tests performed in standard behavioral apparatuses. The illumination of the light sections of each apparatus was set at 200 lux, and the duration of both tests was set to 5 min.

**FST.** Behavioral despair was measured using the FST. Each animal was placed in a 2 liter transparent beaker filled halfway with 23 °C water for 6 min. Floating, swimming and struggling times were manually scored by experienced observers.

**SPT.** The SPT was performed under low illumination (2 lux) over three sessions in a three-chamber apparatus. This consisted of a middle chamber connected to two chambers on each side by a door. An empty metal grid cone was placed in the center of each of the two side-chambers. During the first session, the doors to the side chambers were closed and each test mouse was introduced into the middle chamber and allowed to habituate for 5 min. The doors were subsequently opened and a stimulus CD-1 mouse was placed under one of the metal grid cones, the other remaining empty, for 10 min. Sociability was calculated based on this session as the ratio of time spent in the chamber with the stimulus mouse to the time in the chamber with the empty cone, weighted by the total time spent in both of these chambers. Finally, in a subsequent 10-min session, a different stimulus mouse was placed under the other cone. Preference for social novelty was assessed using the ratio of time with novel versus familiar mouse weighted by total time in either chamber.

**Y-maze alternation task.** Working memory was assessed using the Y-maze alternation task. Each mouse was introduced for 5 min into a Y-shaped three-arm apparatus with distinguishing visual cues on the walls at the end of each arm. The proportion of spontaneous non-repeated subsequent entries into each of the three arms (alternations) from the total number of three-arm entries (including repeat entries) was used as the final readout.

**Resident-intruder test.** Aggression toward an unfamiliar intruder mouse was assessed using a resident-intruder paradigm. For this test, each mouse was single-housed in a fresh type 2 cage. At each time point, 48, 72 and 96 h after single-housing, an intruder C57/BL6 mouse was introduced into the cage. Latency to the first aggressive interaction was assessed by an observer. Each trial was interrupted after the first overtly aggressive confrontation or after 15 min.

**Holeboard exploration test.** Exploratory behavior was measured using the holeboard exploration test. Each mouse was introduced for 5 min into a 40×40 cm arena surrounded by transparent Plexiglas walls with 4×4 equally spaced holes on the floor. The number of head dips into any hole and the number of rearings during the test interval were assessed by an observer.

**RNA-seq.** Animals were killed via an overdose of isoflurane, and the brains were then dissected and flash-frozen. Tissue samples were cryopunched using a 1-mm diameter punching tool at bregma 1.98 mm (600- $\mu$ m depth for the PFC and the INS) or an 0.8-mm diameter punch at bregma –0.7 mm (1,000- $\mu$ m depth for the BLA) according to Paxinos and Franklin<sup>19</sup>. Total RNA was isolated using a miRNeasy micro kit (Qiagen) after homogenization using a Bullet Blender (Next Advance). Residual genomic DNA was removed using a Turbo DNA-free kit (Ambion, Invitrogen). RNA integrity (RIN) and absence of DNA were confirmed using an Agilent RNA ScreenTape (4200 TapeStation, Agilent; eRIN values all >7.8) and a Qubit DNA High sensitivity kit, respectively. Sequencing libraries were prepared using an Illumina TruSeq Stranded Total RNA Library Preparation HT kit using mammalian RiboZero Gold following the standard protocol starting from 300 ng (PFC), 500 ng (INS) and 375 ng (BLA) of total RNA using 12 cycles of PCR amplification. Libraries were quality checked using Bioanalyzer DNA High Sensitivity chips (Agilent Technologies) and quantified using a KAPA Library Quantification kit (KAPA Biosystems). Sequencing was performed on six lanes of an Illumina HiSeq4000 PE 2×100 (Illumina) multiplexing all samples.

Sequencing was performed on a HiSeq4000 to generate 100 base pair paired-end reads. Read quality was checked using FastQC<sup>20</sup>, and adapters were then trimmed using cutadapt<sup>21</sup>. For quantification of transcript expression levels, Kallisto<sup>22</sup> was executed using Gencode M11 annotation and collapsed to gene level.

Count data were prefiltered for low counts at a threshold of  $\geq 20$  counts per sample in a minimum of 31 out of 32 samples per region. In addition, the top three most highly expressed genes were excluded, which resulted in a total list of 13,073 genes. Differential gene expression analyses were performed in DESeq2 (ref. <sup>23</sup>). Heteroscedasticity in the data was reduced using the DESeq2 regularized logarithm transformation. The plate row of each sample was identified as a potential batch effect and corrected for using the package limma<sup>24</sup>.

The normalized and log-transformed count data were then used as the outcome of a linear model with either the real or the shuffled ID scores (day 1 in the social box) as predictors ( $\text{rlog}(\text{count}) = \text{ID1} + \text{ID2} + \text{ID3} + \text{ID4} + \epsilon$ ). Total variance explained by the model and the fraction of variance explained by each individual predictor were estimated using the variancePartition package<sup>25</sup>. Two hundred models with shuffled ID scores were run per region to generate a distribution of mean variance explained across all genes, and the model with the real ID scores was tested against this distribution using a one-sample *t*-test.

**Statistics.** Statistical analyses were performed using Matlab or R. Most test statistics were derived from permutation tests (unless otherwise specified). For these tests, the relevant statistics were derived by permuting individual group-belonging to create a null distribution. Normality of the underlying data distributions was typically not assumed, except in cases when parametric tests were used (for example, analysis of variance (ANOVA)) as indicated where applicable. Differential RNA expression analyses were conducted with the assumption of a negative binomial distribution, adjusting significance levels for multiple comparisons using Benjamini–Hochberg correction. The tests used and the relevant test statistics are specified in the figure legends. The sample size of the main cohort used to infer IDs was chosen to maximize the specificity and applicability of the resulting model, going far beyond the sample sizes typically used for behavioral experiments. Sample sizes for the ID validation experiments were similar to or larger than those reported in previous publications<sup>4,5</sup>.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### Data availability

The RNA-seq data for this project have been deposited to the NCBI's Sequence Read Archive (SRA) under the following accession number: [PRJNA542512](https://www.ncbi.nlm.nih.gov/sra/PRJNA542512). The datasets generated during and/or analyzed during the current study are available from the corresponding author upon request.

### Code availability

All the code used in the Matlab LDA implementation, including a demonstration of its use on the results from the original cohort of mice ( $n = 168$ ), is publicly available at the following link: <https://orenforkosh.github.io/IdentityDomains/>. The color-based video tracking system will be made available upon request. Likewise, the self-similarity tests implemented in Matlab and the R code used in the RNA-seq data analysis will be made available upon reasonable request.

### References

14. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
15. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classification* (Wiley, 2012).
16. Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. *Classification and Regression Trees* (CRC Press, 1984).
17. De Vries, H., Stevens, J. M. G. & Vervaecke, H. Measuring and testing the steepness of dominance hierarchies. *Anim. Behav.* **71**, 585–592 (2006).
18. Leger, M. et al. Object recognition test in mice. *Nat. Protoc.* **8**, 2531–2537 (2013).
19. Franklin, K. B. J. & Paxinos, G. *The Mouse Brain in Stereotaxic Coordinates* (Academic Press, 1997).
20. Andrews, S. et al. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (2010).
21. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
22. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
23. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
24. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
25. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).

### Acknowledgements

The authors thank N. Eren, I. Couzin and C. Wotjak for their assistance, advice and constructive criticism. They thank M. Engel for her technical assistance with the RNA-seq experiment. Thanks are also given to J. Keverne for professional English editing, formatting and scientific input. Their thanks also go to O. Maoz for his unique insights into the mathematics and their interpretation. Finally, the authors would like to extend special thanks to the recently passed Chaya Tannor for fascinating discussions on human personality. A.C. receives financial support from serving as the Vera and John Schwartz Family Professorial Chair at the Weizmann Institute and as the head of the Max Planck Society—Weizmann Institute of Science Laboratory for Experimental Neuropsychiatry and Behavioral Neurogenetics. This work is supported by the following grants and agencies (to A.C.): a FP7 Grant from the European Research Council (260463); the Israel Science Foundation (1565/15); the ERANET Program; the Chief Scientist Office of the Israeli Ministry of Health; the Federal Ministry of Education and Research (01KU1501A); Roberto and Renata Ruhman; Bruno and Simone Licht; the I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation (grant no. 1916/12); the Nella and Leon Benoziyo Center for Neurological Diseases; the Henry Chanoch Kreuter Institute for Biomedical Imaging and Genomics; the Perlman Family Foundation, founded by Louis L. and Anita M. Perlman; the Adelis Foundation; the Marc Besen and the Pratt Foundation; and the Irving I. Moskowitz Foundation. S.K. is supported by the International Max Planck Research School for Translational Psychiatry (IMPRS-TP).

### Author contributions

O.F. and S.K. designed the experiments, analyzed the results and wrote the manuscript. C.T. contributed to the design of the behavioral experiments. M.N., C.F. and P.M.K. assisted in experiments. S.R. performed the preprocessing of the RNA-seq data and contributed to the final analyses. U.A., S.A. and Y.S. contributed to the manuscript. A.C. supervised and supported the project.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41593-019-0516-y>.

**Correspondence and requests for materials** should be addressed to A.C.

**Peer review information** *Nature Neuroscience* thanks Ann Kennedy and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Videos from the social arena were recorded and encrypted using the Gecko GigE Video Recorder (version 1.9.9.4).

Data analysis

LDA was implemented in Matlab (version 2017a, the Mathworks, Inc.). Likewise, Pareto Task Inference and the self-similarity permutation-based tests were performed in Matlab using the Pareto Task Inference (ParTI) package. Bioinformatic analyses for the RNA-sequencing dataset as well as multiple standard statistical tests, were performed in R (version 3.4). The open-source packages used in the analyses were as follows: FastQC, cutadapt, Kallisto, DESeq2, limma, variancePartition

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon request.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The initial sample size for creating the ID structure was informed by the need to balance the number of behavioral features which go into the algorithm against the number of individuals available. The initial experiments were thus on 16 groups (n = 64) animals, such that more individuals would be available than behavioral features. As this initial experiment generated hypotheses regarding stability and relationships to classical tests, we continued adding the "baseline" days of new cohorts to the training pool for the ID algorithm, such that the final pool is derived from 168 individuals.
Data exclusions	No data exclusion was performed in the major (n = 168) cohort used to create the ID structure. Likewise, no data were excluded in the ID score versus standard behavioral tests set of experiments. A single group was excluded from the high- normal-anxiety selectively bred mouse cohort, since the high-anxiety mouse in that group was considered a strong outlier on ID1 based on pre-established criteria (> 2 SD away from the group mean, data available upon request). This exclusion ultimately increased the statistical significance of the group-test for ID1, yet it rendered the findings of the corresponding test for ID3 insignificant.
Replication	The central validation of our method in the current paper involves training an LDA algorithm on a set of behavioral data. Therefore, in addition to a number of cross-validations internal to this dataset, we performed an external replication (described in the manuscript) using new data with different readouts from a similar system. We then confirmed that the top-ranking Identity Domains produced in this data were strongly related to the original IDs when mapped onto the same space.
Randomization	For most experiments, no individual allocation was necessary, as there were no experimental groups. In the case of the group-mix experiment, a scheme of how the mice were shuffled between groups is provided in the supplement. This allocation was blind to any information about individual behavior, only taking into account fur color, which was necessary for automatic behavioral tracking. In the case of the high-anxiety & normal-anxiety selectively bred line of animals, group allocation was such that a single high-anxiety mouse was grouped with 3 randomly selected non-littermate normal-anxiety mice.
Blinding	All videos generated from the the social arenas were tracked automatically after the end of each experiment. Thus prior knowledge about the behavioral features of any individual was available a priori and no blinding was necessary. This also applies in the case of the standard behavioral set of tests - the ID characterization of each animal was not available prior to the execution and scoring of the behavioral test battery.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

# Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Male CD-1 mice between the ages of 8 and 12 weeks old (at the start of each experiment) were used in the central ID cohort (n = 168). Similarly, CD-1 males starting at a younger age (4-5 weeks old, n=32) were used in the juvenile/adult ID score comparisons. The adult (ca. 12 weeks old, n = 24) male mice that were selectively bred for differential levels of anxiety-like behavior (HAB/NAB model) also originated from a CD-1 background.
Wild animals	No wild animals were used in the reported set of experiments.
Field-collected samples	No field-collected samples were used in the reported set of experiments.
Ethics oversight	All animal experiments were approved by the Animal Care and Use Committees of either the Government of Upper Bavaria (Munich, Germany) or the Weizmann Institute of Science (Rehovot, Israel).

Note that full information on the approval of the study protocol must also be provided in the manuscript.