The choice-wide behavioral association study: data-driven identification of interpretable behavioral components

David B. Kastner[1,4], Greer Williams[1], Cristofer Holobetz[1], Joseph P. Romano[2], Peter Dayan[3]

[1]*Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, CA 94143, USA*

[2]*Department of Statistics, Stanford University, Stanford, CA 94305, USA*

[3]*Max Planck Institute for Biological Cybernetics, Tübingen 72076, Germany*

[4]Lead Contact

Correspondence: David B. Kastner: david.kastner2@ucsf.edu

Abstract

Behavior contains rich structure across many timescales, but there is a dearth of methods to identify relevant components, especially over the longer periods required for learning and decision-making. Inspired by the goals and techniques of genome-wide association studies, we present a data-driven method—the choice-wide behavioral association study: CBAS—that systematically identifies such behavioral features. CBAS uses powerful, resampling-based, methods of multiple comparisons correction[1–3] to identify sequences of actions or choices that either differ significantly between groups or significantly correlate with a covariate of interest. We apply CBAS to different tasks and species (flies[4], rats[5], and humans[6]) and find, in all instances, that it provides interpretable information about each behavioral task.

Understanding how behavior differs between different groups of humans or other animals is critical for generating and testing hypotheses about the functional role of genes, regions of the brain, and neural circuits, and is central to characterizing neurological and psychiatric dysfunction[7]. However, behavior is highly complex, evolving over multiple timescales and exhibiting substantial path dependencies due to individual experience[8–12]. It is increasingly possible to automate behavioral paradigms, and for computational methods to revolutionize behavioral analyses[13–22]. The latter come in two main flavors[23,24]: model-based, or top-down approaches, and data-driven, or bottom-up approaches. The former are substantially more prevalent than the latter; we offer a partial corrective.

In model-based analyses, behavioral data, such as choices in a decision-making task, are processed under the specific assumptions of a hypothesis or model. If the model or hypothesis is correct, this is highly efficient, since large volumes of data can be reduced to a handful of parameters that index semantically meaningful phenotypes, such as learning rates or differential sensitivity to rewards or punishments. These parameters can then be compared between the groups. However, even when substantial effort is put into building multiple alternative models, and comparing them in a statistically rigorous manner, it remains possible that the best fitting model nevertheless fails to characterize the behavior properly, rendering nugatory any interpretation of group comparisons. Additionally, with such approaches, confirmation bias[25] presents a significant challenge for accurate interpretation. Furthermore, in the worst of cases, model and hypothesis-based analyses provides a post-hoc framework for explaining any difference that can be found in a behavioral dataset[26]. Modern machine learning methods[27,28] can provide useful lower bounds for how well hypothesis-driven models should fit, but these approaches lack interpretability and data efficiency.

Data-driven analyses start from the other end, taking, and then characterizing, behavior without relying on parametric assumptions, and making very few assumptions about how the data is generated. Some of these approaches are unsupervised, for instance finding clusters in behavioral space and defining them as 'syllables' which can subsequently be compared between groups[18,29,30]; others are more supervised, directly looking for discriminative differences between populations or individuals[13–16]. By not trying to force data into a limited set of parameters, these methods should be more sensitive; however, comparisons between groups pose severe statistical challenges, because of the complexity and dimensionality of behavioral datasets.

Here, we present the choice-wide behavioral association study (CBAS), a data-driven analysis method designed to identify relevant sequences of choices (or other discrete behavioral features) made by subjects. CBAS has two components: 1) breaking down behavior into a comprehensive language for comparison between two groups or correlation with a covariate of interest; and 2) using rigorous, resampling-based, statistical corrections to account for the resulting large number of comparisons and maintain statistical power despite correlations in the data.

Choice as a common discretization for behavior

In developing our analysis, we were motivated by the data-driven approach of genome-wide association studies (GWAS), whole exome sequencing (WES) and whole genome sequencing (WGS)[31]. In some ways, the state of behavioral analysis in systems neuroscience resembles the state of genetic analysis prior to GWAS/WES/WGS, where studies attempted to associate candidate genes with phenotypes. Candidate gene studies were often underpowered and failed to replicate[32–34], reminiscent of aspects of behavioral analyses[35].

GWAS/WES/WGS look for differences in base-pairs of the genome between groups of subjects. The large numbers of base pairs being compared in these studies necessitates statistical correction to enable reliable decisions about the significance of any differences found. The discrete nature of base-pairs and the ability to compare that set across subjects make GWAS/WES/WGS possible. To be able to develop a comparable method for behavioral analyses it is critical to identify an appropriate discretization of behavioral tasks. In many cases, choice provides just such a discretization (Fig 1). Indeed, there is evidence that, at a rather fundamental level, behavior occurs through discrete choices[18,36]. We use three tasks, in three diverse species—flies, humans, and rats—to show the breadth of applicability of CBAS.

Using choice as the basis of the comparison for the behavioral analysis requires an additional consideration beyond what is done for the genome with base-pairs. For GWAS/WES/WGS, an individual base pair can be a meaningful unit of information (although this is only a partial story[37,38]) that can be compared between subjects. This need not be the case for individual choices in behavior, whereby the choices that precede and follow a specific choice can change the meaning of that choice. In this case, the relevant behavioral feature is a whole sequence of choices. Therefore, CBAS does not just evaluate the occurrence of individual choices, but, rather, the occurrence of all sequences up to a certain, user-defined, length. In general, the longer the sequence length, the more data and computing time will be necessary.

Evaluating the occurrence of all sequences of choices up to a certain length, causes a complexity, when it comes to correcting for the many comparisons, since, the sequences can be highly correlated. Standard methods to correct statistically for multiple comparisons (e.g. Bonferroni, Holm, Benjamini-Hochberg, Benjamini-Yekutieli) are either incorrect or underpowered for correlated data. Therefore, we instead use an approach based on resampling to correct for the multiple comparisons, which retains power in the face of correlations[1].

CBAS identifies interpretable differences for fly y-maze

The first task we considered involves drosophila walking on a y-maze (Fig 1a). The movement of flies on the maze is tracked as they make the choice of going to either the left or

right arm after leaving the previous arm[4]. This left/right discretization of the task has enabled many conclusions about the genetic nature of individual variability[4,39–43]. Here, we compare the choices of two outbred strains of fly. Analyses in the original paper[4] identified some clear indications about the difference; however, the data provide a useful testing ground for CBAS since there are many subjects and the results are relatively low-dimensional.

When deciding to perform a data-driven genomics method, there are still decisions that need to be made. For instance, deciding to focus on only exome sequences in WES, or SNPs in GWAS. Similarly, to run CBAS, a few decisions need to be made about how to structure the data (Fig 2a). These decisions are important, and are not normally pre-determined directly by the data. Any conclusions drawn need to be interpreted in their light.

The first decision is the possible choices that will make up the sequences used in CBAS. We refer to this as the basic language for the application of CBAS. For the fly task, the language is left or right turns. Next, as described above, a decision needs to be made about the maximum number of choices in a row that will make up all the sequences used in CBAS. For the fly task, we chose a maximum sequence length of 10 choices in a row. That means that CBAS evaluates all sequences from length 1 – 10 that exist in the dataset.

CBAS then works through evaluating the average sequence count for each sequence, which we refer to as the rate of that sequence. The rate is calculated by counting the total number of times that sequence occurs in the population divided by the number of individuals in the population. To be precise about the occurrence of the sequence, a decision needs to be made about the number of trials over which this is counted, i.e. a criterion. The same criterion is applied to all subjects. For the fly task, we use the first 250 turns as that criterion. The last decision that needs to be made is what to do with subjects that do not reach the criterion. For the fly task, we exclude the subjects that do not reach criterion.

Given these decisions, we performed a CBAS comparing two outbred strains of flies, Cambridge-A (CA) and w1118, from a publicly available dataset[4]. The w1118 strain is the background strain for many transgenic flies[4]. For each sequence of left/right turns up to 10 turns long, the rate was calculated for the two strains (Fig 2b). Using the Romano-Wolf resampling-based multiple comparisons correction[1,2] (see methods), we determined which sequences the two strains utilize significantly differently (Fig 2c). As a significance threshold, we use median control of the false discovery proportion at 5%, which is comparable to 5% false discovery rate (FDR) for the resampling-based method[2] (see methods).

CBAS identifies many sequences that differ significantly between the CA and w1118 lines, some of which are shown in Fig. 2d. Upon inspection of the sequences, a clear, interpretable, difference is apparent between those sequences that occur more in the CA line and those that occur more in the w1118 line (Fig 2d&e). The CA line utilizes sequences with extended numbers of the same turn in a row, whereas the w1118 line utilizes sequences with more frequent changes in turn direction. Therefore, CBAS not only identifies that there is a difference between the two fly lines, but it also provides information to support an interpretation as to the nature of the difference.

Data-driven methods invariably need more data than hypothesis-driven methods. To estimate the sample size needed for appropriately powered experiments to detect any difference between these two fly lines, we took advantage of the large dataset to resample groups of smaller sizes from the data and recalculate the CBAS for each set of resampled groups. We generated an estimate of the power for different number of flies per group (Fig 2f) (see methods), by comparing the number of significant sequences identified by CBAS when comparing CA to w1118 to the number of significant sequences identified by CBAS when comparing the lines to themselves (Fig S1b). With 40 flies per group, CBAS has an estimated power >80% to distinguish the CA and w1118 lines.

Graceful decay of CBAS output with decreasing group size

The power calculation concerns the ability of CBAS to distinguish between the two strains. A more refined question is what the nature and comprehensiveness of the collection of significantly different sequences would be as smaller numbers of flies per group are analyzed. This would determine our ability to derive interpretations from fewer subjects.

We first evaluated the number of significant sequences identified from smaller group sizes as a fraction of the total number of significant sequences identified with the full dataset (Fig 3a). As expected, with smaller group sizes, we find fewer overall sequences; however, across the range of group sizes evaluated, the median fraction of sequences was larger than the fraction of the population being used in the smaller group CBAS (Fig 3a). This indicates that, for this dataset at smaller group sizes, the proportion of sequences identified by CBAS grows faster than the proportion of subjects in the CBAS. This provides a rapid increase in the amount of information provided by CBAS as sample size increases.

We next evaluated the fraction of the sequences identified by CBAS for the smaller sized groups that are not identified in the CBAS on the full dataset. Since we control the false discovery proportion and not the family-wise error rate, we do not expect this value to be zero; however, it was consistently small across all the group sizes evaluated (Fig 3b). The medians across the different repeats of the same sample sizes across the different sample sizes are all less than 2%.

Then, we evaluated the similarity of the sequences identified by CBAS with nonoverlapping sets of subjects with the same group size (Fig 3c). At smaller groups sizes, this overlap could be quite low (median 24.4%), even though there was sufficient evidence to discriminate the strains (Fig. 2f). This means that the specific sequences identified by CBAS can be quite variable from one experiment to the next, especially at smaller group sizes. Given that, we sought to understand if the structure of the sequences comported with the conclusion from the full dataset—that the CA line uses more of the same turns in a row and the w1118 line more frequent changes in direction of turn. Even though the overlap between CBAS on different groups might not be that high (Fig. 3c), the sequences identified still follow the same structure as the full dataset in regard to number of the same turns in a row, (Fig 3d). As the group size increases the output of CBAS consistently becomes more similar to the results from the full dataset (Fig 3e), licensing more generalized conclusions.

CBAS provides evidence for testing model-based analysis

The second task we considered, is the two-step task developed to test the interplay between model-based and model-free (reinforcement) learning (Fig 1b)[44]. In this task, subjects choose between pairs of images at two different stages. The image choice at the first stage governs the pair of images from which the subject can chose at the second stage. Following image choice at stage two, reward is delivered based on dynamic probabilities associated with each of the images. Model-based analyses of variants of this task have led to conclusions about the algorithms used in different brains regions[44,45] and differences that underlie psychiatric symptoms[6]; however, the interpretations are not without controversy[46], and the complexity of behavior makes it difficult to evaluate ways in which the model might be missing features in the data.

To test the ability for CBAS to provide useful information in such a situation, we evaluated the open-source dataset from Gillan et al[6]. In that work, a large population of human subjects performed the two-step task, and also answered a series of psychiatric symptom questionnaires. The authors performed factor analysis on the different questionnaires and found that factor 2, which they associated with intrusive thoughts and compulsive behavior, had significant association with the way subjects performed that task. They found that subjects' factor 2 score was negatively associated with model-based decision making, leading to their conclusion that the greater the loading on intrusive thoughts and compulsive behavior for a person, the less likely they were to utilize model-based decision making on the two-step task.

There were two different groups to compare in the fly dataset. In this dataset the relevant comparison is correlation with a covariate (i.e. factor 2 score from the psychiatric symptom questionnaires)[6]. Therefore, we extended CBAS to identify the sequences that significantly correlate with this particular factor score (see methods). As with the CBAS applied to the fly data, there are decisions that need to be made to apply CBAS to this human dataset (Fig 4a). For this CBAS, the language is comprised of 8 different units: choosing image 1 or 2; making a choice within set A or set B, distinguishing A and B depending on whether a reward was delivered (shown in the figure as bold and underlined); and making 'no choice' (either at the first or second image), which occurs in the data, albeit rarely. Following Akam et al.[45] we collapsed image 3 and 4 into set A and image 5 and 6 into set B because the specific images within the set are not relevant for the critical decision at the first stage. We ran CBAS on all sequences up to 4 choices long and evaluated the rate of the sequences over the 400 choices of the dataset (200 trials of the two-step task). In the open-source dataset, there was no subject who did not reach criterion.

CBAS evaluated the correlation between the usage of each sequence and the factor 2 score (Fig 4b&S2a). Multiple sequences were significantly correlated (Fig 4c&d). To understand the output of CBAS, we review the expectation associated with the hypothesis-driven analysis of this experiment. This task was designed to evaluate the interplay between model-based and model-free decision making[44]. Model-based decision making develops an understanding of the structure of the world (i.e. the model) and makes choices based on that understanding. Model-free decision making makes choices based on reinforced past successful actions, without developing an understanding of the structure of the world. A way to see the difference between these two decision-making schemes is when a reward is delivered at set A ('**A**' in our language) after having come from image 2. A regular model-free learner will reinforce the visits

to both 2 and A and will therefore be more likely to choose 2 on the next trial. The model-based learner will have an understanding that choosing image 1 on the next opportunity makes it more likely to return to set A (because of the common/rare transition structure in the environment) and will therefore chose object 1[44].

Instead of identifying correlations with sequences that relate to either the model-free or model-based learning, CBAS instead identifies a positive correlation with many sequences involving being rewarded at A and then selecting image 2 or being rewarded at B and then selecting image 1 (Fig 4d). Just as it was helpful with the fly data to find complete sets of sequences that were significant (e.g. all of the sequences that occurred in the data with 10 left or right turns in a row were identified as happening more in CA than w1118, etc.) we were also able to find practically complete sets of sequences that were identified as being positively correlated with the factor 2 score. Practically all of the sequences in the dataset that contained 1**A**2 or 2**B**1, were significantly correlated with this score; whereas none of the sequences with 1**B**1 was identified as significant (and there were no instances of 2**A**2 in the dataset) (Fig4e&S2b).

Sequences 1**A**2 or 2**B**1 can be classified as anti-model-based decisions. The subjects get rewarded after choosing from the common side (A from 1 or B from 2), but then selects the image that will rarely bring them back to the previously rewarded side (2 from A or 1 from B). CBAS therefore identifies that anti-model-based learning is a prominent feature that correlates with intrusive thought and compulsive behavior loading. Further experimentation will be needed to understand the interplay between this mode of decision making, the task, and these symptoms scales.

As with the fly data, we could evaluate the sample size needed to reach a given statistical power for identifying any significant correlation between the factor 2 score and this task. We resampled smaller group sizes and compared the CBAS for the true relationship between the sequence counts and factor 2 score to a randomly generated set of factor 2 score values that was drawn from the same distribution as the data (Fig S2c). A sample size of 900 individual provides a power >80% to detect significant correlations between the factor 2 score and the subject (Fig 4f), which compares favorably to the sample size of ~1,200 – 1,600 subjects that Gillan et al. calculated to generate their dataset[6].

CBAS identifies a phenotype consistent with ASD in *Scn2a* haploinsufficient rats

The third task we considered, involves spatial alternation behavior in rats. For this task, to get reward, the rats must alternate between pairs of arms of a track whilst visiting a different arm of the track in between (Fig 1c). Spatial alternation is a common behavioral paradigm for phenotyping and neurophysiology, and the discretization of the behavior into the arms chosen by the animals forms the basis of many of the conclusions from these studies[47–51]. However, our recent work calls into question the assumptions and hypothesis that motivate the standard analysis for spatial alternation behavior[5,52]. Even though we developed reinforcement learning agents to fit individual behavior, those agents showed clear differences from the way the animals learned[5], limiting their use for phenotyping. Therefore, we applied CBAS.

We sought to discriminate wild-type (WT) rats from those haploinsufficient for *Scn2a* (*Scn2a*$^{+/-}$), a high confidence, large effect, autism spectrum disorder (ASD) risk gene[53,54]. We collected a dataset of over 200 rats performing six different spatial alternation contingencies using our previously described automated behavioral system[5] (see methods). Each spatial alternation contingency was defined by the three arms of the track where alternation needed to occur to get reward. For example, if the contingency was at arms 2, 3, and 4, reward would be provided for every arm visit in the sequence **3**-**4**-**3**-**2**-**3**-**4**.

The language for the spatial alternation CBAS was visiting each arm of the track and not getting reward and visiting the goal arms and getting reward. That means that within each contingency there were a total of 9 possibilities—6 unrewarded arms and 3 rewarded arms. For example, if the contingency was at arms 2, 3, and 4, the language is composed of visiting arms 1, 2, 3, 4, 5, or 6 and not getting rewarded or visiting arms **2**, **3**, or **4**, and getting rewarded. Each contingency was considered separately, as visiting a sequence of arms during one contingency likely means something different than visiting those same arms in a different contingency. We chose to evaluate all sequences up to six choices long. For the criterion, we used the trial at which each animal reached 100 perfect performance sequences four choices long for each contingency, and we included animals that did not reach the criterion (Fig 5a) (see methods).

In running the CBAS on this dataset, we evaluated the rate difference of >86,000 sequences across the six different contingencies, with a total of 1,476 sequences being identified as significantly different between the *Scn2a*$^{+/-}$ and WT littermates (Fig 5b&c). To interpret common features of those sequences, we sought categories of sequences for which a substantial fraction was identified as being significantly different between the groups: the strategy that was informative for the fly and human datasets (Fig 2e&4e). One common feature of spatial alternation behavior is the 3-arm structure of each contingency (e.g. arms 2, 3, and 4 are the only arms with the potential to be rewarded during contingency A and E). Therefore, in all contingencies, we evaluated all sequences that exclusively contained all sets of three arms. For example, within a contingency we identified all sequences exclusively containing arms 2, 3, and 4, which means that every sequence within that category only contains visits to arms 2, 3, and 4 (but does not have to visit all of the arms). We did this with sets of three arms, independent of whether the choice was rewarded or not, as well as the set of three arms in a contingency that were all rewarded. There are 21 sets of three arms for each contingency, and, consistent with the relevance of these sets for the behavior, across all contingencies they contained only ~25% of all the sequences that the rats performed during the experiment, but ~75% of all of the significant sequences (Fig S3a).

Some of these sets of 3 arms are clearly interpretable. One such is the set of three arms from the current contingency, which contains all sequences where the animal exclusively visits the arms of the contingency but could make errors in the order of those arm visits (Fig 5d; left). A second interpretable set consists of three arms from the current contingency that are rewarded, which contains all sequences where the animal consistently performs the task correctly. A third and fourth are the sets of three arms exclusively containing the previous (Fig 5d; right), or the one before previous, contingency arms, which contain sequences where the animals repeat prior actions that are no longer optimally rewarded.

For all sets of three arms, in all the contingencies we tabulated the fraction of the total sequences in the set that were significant either for WT > *Scn2a*$^{+/-}$ or WT < *Scn2a*$^{+/-}$. We then asked if any of these fractions were larger than would be expected from randomly distributing the significant sequences across all possible sequence types within the entire category of sets of three arms (see methods) and found consistent patterns in how the WT and *Scn2a*$^{+/-}$ rats differed (Fig 5e). In 4 out of the 6 contingencies, *Scn2a*$^{+/-}$ rats showed increased usage of sequences containing the current contingency, and in 3 out of the 6 contingencies, *Scn2a*$^{+/-}$ rats also showed increased usage of sequences related to prior contingencies (either 1 or 2 contingencies back). By contrast, in 4 out of the 6 contingencies the WT rats showed increased usage of sequences containing the rewarded arms of the current contingency. This indicates that WT rats are ultimately better at performing spatial alternation behavior than *Scn2a*$^{+/-}$ rats, and that *Scn2a*$^{+/-}$ rats show difficultly transitioning from prior actions, repeating sequences from prior contingencies, possibly consistent with restrictive and repetitive actions, which forms one of the diagnostic criteria for ASD[55].

As with the fly and human data, we could evaluate the sample size needed to reach a given statistical power to detect any difference between these genotypes. We resampled smaller group sizes and compared the CBAS for the comparison between the WT and *Scn2a*$^{+/-}$ rats to the comparison between WT and itself and *Scn2a*$^{+/-}$ and itself (Fig S3b). A sample size of 30 rats per group provides a power >80% to detect a difference between the two genotypes (Fig 5f).

We have presented a general behavioral analysis method, CBAS, for identifying interpretable behavioral components that is grounded in sequences of choices made by subjects. There has been significant progress, in recent years, in tracking and analyzing, short timescale actions of subjects[13–22]; however, we lack methods for generally analyzing and interpreting sequences of these actions or long-run choices and decisions of subjects during behavior. CBAS provides just such a method and is applicable across a wide array of species and different behavioral paradigms (Fig 1). It can be used to test models and hypotheses (Fig 2&4), and, for instance, using generalizable simple principles about the relationship between sequences and task contingencies, it can generate hypotheses in complex behaviors where reliable computational understanding has yet to emerge (Fig 5). Through taking advantage of large-scale data collection and rigorous statistical methods, CBAS has the potential to transform our use of behavior in a comparable way to the ways that GWAS/WES/WGS changed the paradigm for genomic studies.

Methods

Animals: All experiments on rats were conducted in accordance with University of California San Francisco Institutional Animal Care and Use Committee and US National Institutes of Health guidelines. Rats were fed standard rat chow (LabDiet 5001). To motivate the rats to perform the task, reward was sweetened evaporated milk: 25 g of sugar per can (354 ml) of evaporated milk (Carnation). The rats were food restricted to ~85% of their basal body weight.

The *Scn2a* mutant rats were generated due to funding from the Simons Foundation Autism Research Initiative. Long Evans *Scn2a* mutant animals (LE-*Scn2a*^{em1Mcwi};

342 RRID:RGD_25394530) were generated at the Medical College of Wisconsin and shipped to the
343 University of California San Francisco for this study. Briefly, a single guide RNA targeting the
344 sequence GTGAAATCCAACCAATTCCA sequence within exon 5 of *Scn2a* was mixed with Cas9 (*S.*
345 *pyogenes*) protein (QB3 MacroLab, UC Berkeley) and injected into the pronucleus of fertilized
346 Long Evans (Crl:LE, Charles River Laboratories) embryos. Among the resulting offspring, a
347 mutant founder was identified harboring a net 4-bp deletion allele consisting of a 10-bp
348 deletion (rn7: chr3:50,364,411-50,364,420) along with a 6-bp insertion of TTCACT, inducing a
349 frameshift in the coding sequence predicted to truncate the normal protein after 193 amino
350 acids. The founder was backcrossed to the parental Crl:LE strain to establish a breeding colony.

351 **Spatial alternation behavior:** The automated behavior system for spatial alternation behavior
352 was previously described[5]. There are different symbols on each arm of the track serving as
353 proximal cues, and there are distal cues distinguishing the different walls of the room.
354 Pneumatic pistons (Clippard) open and close the doors. Python scripts, run through Trodes
355 (Spike Gadgets), control the logic of the automated system. The reward wells contain an
356 infrared beam adjacent to the reward spigot. The automated system uses the breakage of that
357 infrared beam to progress through the logic of the behavior. In addition to the infrared beam
358 and the spigot to deliver the reward, each reward well has an associated white light LED.

359 Each cohort of rats is divided into groups of four (or three) animals. The same groups
360 were maintained throughout the duration of the experiment. Within a group, a given rat is
361 always placed in the same rest box, and the four rats of a group serially perform the behavior.
362 The rats have multiple sessions on the track each day. Prior to beginning the first spatial
363 alternation contingency, the rats experience multiple days and sessions where they get
364 rewarded at any arm that they visit (provided it is not an immediate repeat). During this period
365 of the behavior, the duration of a session is defined by a fixed number of rewards, or a fixed
366 amount of time on the track (15 minutes), whichever came first. During the alternation task the
367 duration of a session was defined either by a fixed number of center arm visits and at least one
368 subsequent visit to any other arm, or a fixed amount of time on the track (15 minutes),
369 whichever came first.

370 The algorithm underlying the spatial alternation task is such that three arms on the track
371 have the potential for reward within a given contingency, for example during a contingency at
372 arms 2-3-4, arms 2, 3, and 4 have the potential to be rewarded, and arms 1, 5, and 6 do not. Of
373 those three arms we refer to the middle of the three arms as the center arm (arm 3 in the
374 above example) and the other two arms as the outer arms (arms 2 and 4 in the above example).
375 Reward is delivered at the center arms if and only if: 1) the immediately preceding arm whose
376 reward well infrared beam was broken was not the center arm. Reward was delivered at the
377 outer two arms if and only if: 1) the immediately preceding arm whose reward well infrared
378 beam was broken was the center arm, and 2) prior to breaking the infrared beam at the center
379 arm, the most recently broken outer arm infrared beam was not the currently broken outer
380 arm infrared beam. The one exception to the outer arm rules was at the beginning of a session,
381 if no outer arm infrared beam was broken prior to the first infrared beam break at the center
382 arm, then only the first condition had to be met.

For the running of the behavior, the infrared beam break determined an arm visit; however, the rats sometimes go down an arm, get very close to the reward wells, but do not break the infrared beam. Therefore, for all the analyses described for the rats, an arm choice is defined as when a rat gets close to a reward well. These times were extracted from a video recording of the behavior. These, effective, missed pokes were more frequent at the beginning of a contingency. This proximity-based definition of an arm visit added additional arm visits to those defined by the infrared beam breaks, and none of them could ever be rewarded, nor alter the logic of the underlying algorithm. However, because of the non-Markovian nature of the reward contingency, they could affect the rewards provided for subsequent choices.

A total of 121 WT (66 males, 55 females) and 120 *Scn2a*$^{+/-}$ (66 males, 54 females) rats were run on the spatial alternation task. WT and *Scn2a*$^{+/-}$ rats were littermates and were housed together prior to their being food restricted before the behavior. During the behavior and food restriction the rats were single housed. 1 WT rat died after finishing the first spatial alternation contingency, 1 WT rat died after finishing the fourth spatial alternation contingency, and 1 *Scn2a*$^{+/-}$ rat died after finishing the first spatial alternation contingency. The data from the animals that died was included up until their expiration. For the first contingency: 5/121 WT rats and 15/120 *Scn2a*$^{+/-}$ rats did not reach the CBAS criterion. For the second contingency: 4/120 WT rats and 19/119 *Scn2a*$^{+/-}$ rats did not reach the CBAS criterion; for the third contingency: 4/120 WT rats and 12/119 *Scn2a*$^{+/-}$ rats did not reach the CBAS criterion; for the fourth contingency: 2/120 WT rats and 12/119 *Scn2a*$^{+/-}$ rats did not reach the CBAS criterion; for the fifth contingency: 4/119 WT rats and 14/119 *Scn2a*$^{+/-}$ rats did not reach the CBAS criterion; and for the sixth contingency: 3/119 WT rats and 12/119 *Scn2a*$^{+/-}$ rats did not reach the CBAS criterion.

Romano-Wolf resampling based multiple comparisons correction: We follow the terminology and description laid out in Clarke et. al[56] to describe the Romano-Wolf multiple comparison correction. First, we describe the way the method corrects for the family-wise error rate (FWER) and then explain how the procedure is extended to provide median control of the false discovery proportion. FWER control at a level of α means that across all comparisons there is a α percent chance of having at least one false positive rejection of a null hypothesis. The Romano-Wolf procedure provides FWER control through resampling the data[1]. It tests a total of S hypotheses.

It is not generally known if the Romano-Wolf procedure controls for type III, or directional, errors. Type III errors are errors in the sign, or direction, of the conclusion. For example, if a statistical test provided information to reject the null hypothesis $\theta_1 = \theta_2$, and you then concluded that $\theta_1 > \theta_2$, when in fact $\theta_1 < \theta_2$. Therefore, instead of running a single two-tailed test for each sequence, we run two one-tailed tests for each sequence. Therefore, the total number of hypotheses tested, S, is twice the total number of sequences being compared. For CBAS those hypotheses take one of two forms: 1) the rate of each sequence (r_s) is the same between two groups $\Delta r_s = 0$, or 2) that there is no correlation (ρ_s) between each sequence and a covariate of interest, $\rho_s = 0$. For the one-tailed versions of each hypothesis, we ask if $r_{s_1} - r_{s_2} > 0$ and $r_{s_2} - r_{s_1} > 0$ for the comparison CBAS, where r_{s_N} is the rate of sequence s for group N, or if $\rho_s > 0$ and $\rho_s < 0$ for the correlational CBAS.

The first step in the procedure is to create a studentized test statistic for each hypothesis. The studentization is different based on whether the CBAS is comparing two groups or calculating a correlation. In the case where two groups are being compared:

$$t_s = \frac{\Delta r_s}{\sigma_s} \tag{7}$$

where σ_s is the standard error of Δr_s, which we calculate by combining the standard error of the mean of the rate for each group using error propagation, i.e. $\sigma_s = \sqrt{{\sigma_{s_1}}^2 + {\sigma_{s_2}}^2}$, where σ_{s_N} is the standard error of the occurrence rate of sequence s for group N.

In the case where the correlation is being calculated the studentized test statistic[3] is:

$$t_s = \frac{\sqrt{n}\hat{\rho}_n}{\hat{\tau}_n} \tag{8}$$

where:

$$\hat{\rho}_s = \frac{\sum X_i Y_i - n\bar{X}_s\bar{Y}}{\sqrt{\sum(X_i - \bar{X}_n)^2 \sum(Y_i - \bar{Y})^2}} \tag{9}$$

$$\hat{\tau}_s = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_s)^2(Y_i - \bar{Y})^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_s)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{10}$$

For eq. 8, 9, and 10, n is the number of subjects for which the correlation is being calculated, X_i and Y_i and the values of the metrics being correlated (in our case, the sequence count and factor 2 score for each individual respectively), $\bar{X}_s$ is the mean sequence count for the specific sequence being considered, and $\bar{Y}$ is the mean of the covariate of interest (factor 2 score, which is the same for any sequence).

When two groups are being compared, we resample from the entire population with replacement (separately for each group) and build up a null distribution by bootstrapping M times. The test statistic from the mth bootstrap sample for $m = 1, \ldots M$ is:

$$t_s^{*,m} = \frac{\Delta r_s^{*,m}}{\sigma_s^{*,m}} \tag{11}$$

where, $\Delta r_s^{*,m}$ is the difference in the rate of each sequence whilst resampling, with replacement, from the entire population, ignoring the group labels. $\sigma_s^{*,m}$ is the accompanying standard error of the resampled groups. The resampled group sizes are the same as the two groups of interest.

In the case where the correlation is being calculated, the test statistic based on the mth bootstrap sample is:

$$t_s^{*,m} = \frac{\sqrt{n}\hat{\rho}_s^{*,m}}{\hat{\tau}_s^{*,m}} \tag{12}$$

where, $\hat{\rho}_s^{*,m}$ is the correlations of each sequence whilst resampling, with replacement, from the entire population, ignoring the group labels. $\hat{\tau}_s^{*,m}$ is the accompanying normalization factor of the resampled groups. The resampled group size is the same as the original.

We used a value of $M = 10{,}000$. Importantly, for each individual resampling, m, the same resampled set is used for all sequences.

The test statistics, and their accompanying estimators are ordered from largest to smallest values. This creates a $M \times S$ matrix where each column contains all the estimators of the test statistics. The first column contains the estimators from the largest test statistic, the second column contains the estimators from the second largest, etc.

To define the distribution for which each test statistic is compared, which then determines the adjusted p-value, the following algorithm is used. The first sequence considered is the one with the maximum test statistic, t_s. Its comparison distribution is defined as the maximum value within each row of the matrix of estimators of the test statistic:

$$\max(t^{*,m}) = \max\{t_1^{*,m}, \dots, t_S^{*,m}\} \tag{13}$$

which provides a total of M values, $t^{*,m}$ (there is no longer an association with s, because these values can come from a resampling of any of the sequences). Using those M values, the adjusted p-value is calculated as follows:

$$p_s^{adj} = \frac{\#\{\max(t^{*,m}) \geq t_s\} + 1}{M + 1} \tag{14}$$

After calculating p_s^{adj} for the first sequence, the column with the test statistic estimators generated from the first sequence is removed from the matrix. This now leaves a matrix that is $M \times (S - 1)$. The above procedure is then used to calculate p_s^{adj} for the sequence with the second largest test statistic, and then its column of test statistic estimators is removed, etc until all p_s^{adj} have been calculated. Following the algorithm described in Clarke et al.[56], we enforce monotonicity of the p-values by resetting the p-value for each sequence:

$$p_s^{adj} = \max\{p_s^{adj}, p_{s-1}^{adj}\} \tag{15}$$

This is done prior to calculating the p-value for the next sequence.

To control the false discovery proportion (i.e., control the number of false positives divided by the total number of hypotheses rejected), the idea of k-FWER is introduced[2]. For

control of the FWER, k is equal to 1, and that leads to an α percent chance that there is at least 1 false positive among all hypotheses rejected (most commonly $\alpha = 0.05$). If k equals 2, then there is an α percent chance that there are at least 2 false positives among all hypotheses rejected. Therefore, to get control of the false discovery proportion we need to find the k that provides the proportion of interest given the number of hypotheses rejected. So, if we want a false discovery proportion, γ, of 0.05, we need $k \sim 0.05 \times \text{number of hypotheses rejected}$.

Romano and Wolf also derived an algorithm to do just that. The algorithm is as follows. Start with $k = 1$. Apply the k-FWER procedure, and note the total number of hypotheses rejected, N. If $N < \frac{k}{\gamma} - 1$, stop and you have identified k. Otherwise, increase k by 1, and repeat[2]. The way you determine k-FWER is in eq. 13, instead of taking the maximum value in each row, you take the kth largest value. Finally, to get median control of the false discovery proportion, $\alpha = 0.5$. This means that 50% of the time you will get a value greater than γ, and 50% of the time you will get a value less than γ, leading to control of the median[2]. This is a similar decision to what is done when calculating the false discovery rate with Benjamini-Hochberg or Benjamini-Yekutieli, except the false discovery rate controls the mean of the false discovery proportion instead of the median.

Power estimation for CBAS. To estimate the statistical power of CBAS for a given sample size we resampled the dataset without replacement and ran a CBAS to determine the number of significant sequences. For each sample size we performed 20 repeats (Fig S1b, S2c, and S3b). We also ran a CBAS comparing each group to itself (for the fly and rat datasets) (Fig S1b and S3b), with 20 repeats for each group; or correlating the sequence counts with a randomly generated set of factor 2 scores drawn from the same distribution as the actual factor 2 scores (for the human dataset) (Fig S2c), with 40 repeats. Then for each sample size, the power is estimated by identifying the largest 20 values of significant sequences, and determining the fraction of those values that were generated by the comparison of the two groups or the correlation with the data with the actual factor 2 scores.

Spatial alternation category fraction determination of significance. There are a total of 20 different sets of three arms, and 1 set of three arms with all choices being rewarded. For each of the six contingencies the fraction of sequences that exclusively contains the set of three arms that are significant is calculated separately for significant sequences greater in the WT and greater in the *Scn2a*$^{+/-}$. That means that there a total of $21 \times 2 \times 6 = 252$ fractions across all contingencies. To determine whether the fraction is significantly larger than expected by chance we proceeded as follows: Within a contingency, we determined the sequences that belonged to any of the 21 categories. Then, separately for sequences greater in WT and sequences greater in *Scn2a*$^{+/-}$, we permuted the association between those sequences and whether or not they were significant. With that permuted association we recalculated the fraction of sequences in each category that were significant. We repeated that process 25,000 times and determined significance by calculating the number of permuted fractions that were greater than or equal to the actual fraction value for each category and divided that by 25,001 (as with eq. 15). That number was then corrected for multiple comparisons using the Bonferroni method and multiplied by 252, the number of tests being performed. Any category whose corrected p-value was < 0.05, was determined to be significantly larger than expected.

Data availability. Data will be made available upon reasonable request to the lead author.

Code availability. Code used to calculate CBAS will be posted to Github upon publication.

Acknowledgements. We thank Claire Gillan and Benjamin de Bivort for making their data publicly available. We thank Thomas Akam and Adam Frank for helpful comments on the manuscript. We thank the Simons Foundation Autism Research Initiative for supporting the generation of the *Scn2a*$^{+/-}$ rat line. This work was supported by grants from the Jane Coffin Childs Memorial Fund for Medical Research (D.B.K.), the UCSF Physician Scientist Scholars Program (D.B.K.), an NIH R25 (R25MH060482) (D.B.K), a Simons Foundation Autism Research Initiative grant (899599) (D.B.K.), the Max Planck Society (P.D) and the Humboldt Foundation (P.D.). Some of the data used in this study were collected in the laboratory of Loren Frank at the University of California, San Francisco.

Author Contributions. D.B.K. and P.D. designed the study, D.B.K., J.P.R., and P.D. developed the analysis method, D.B.K., G.W., and C.H. collected the data, D.B.K. analyzed the data, D.B.K and P.D. wrote the manuscript, and D.B.K., J.P.R. and P.D. edited the manuscript.

Declaration of interests. The authors declare no competing interests.

1. Romano, J. P. & Wolf, M. Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. *J Am Stat Assoc* **100**, 94–108 (2005).

2. Romano, J. P. & Wolf, M. Control of generalized error rates in multiple testing. *Ann Statistics* **35**, 1378–1408 (2007).

3. DiCiccio, C. J. & Romano, J. P. Robust Permutation Tests For Correlation And Regression Coefficients. *J. Am. Stat. Assoc.* **112**, 1211–1220 (2017).

4. Buchanan, S. M., Kain, J. S. & Bivort, B. L. de. Neuronal control of locomotor handedness in Drosophila. *Proc National Acad Sci* **112**, 6700–6705 (2015).

5. Kastner, D. B. *et al.* Spatial preferences account for inter-animal variability during the continual learning of a dynamic cognitive task. *Cell Reports* **39**, 110708–110708 (2022).

6. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife* **5**, e11305 (2016).

7. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron* **93**, 480–490 (2017).

8. Bialek, W. On the dimensionality of behavior. *P Natl Acad Sci Usa* **119**, e2021860119 (2022).

9. Meister, M. Learning, fast and slow. *Curr Opin Neurobiol* **75**, 102555 (2022).

10. Gomez-Marin, A. & Ghazanfar, A. A. The Life of Behavior. *Neuron* **104**, 25–36 (2019).

11. Trillmich, F., Günther, A., Müller, C., Reinhold, K. & Sachser, N. New perspectives in behavioural development: adaptive shaping of behaviour over a lifetime? *Front Zool* **12**, S1 (2015).

12. Dayan, P., Roiser, J. P. & Viding, E. The first steps on long marches: the costs of active observation. in *Rethinking Biopsychosocial Psychiatry.* (eds. Davies, W., Roache, R. & Savulescu, J.) (Rethinking Biopsychosocial Psychiatry., 2018).

13. Hession, L. E., Sabnis, G. S., Churchill, G. A. & Kumar, V. A machine-vision-based frailty index for mice. *Nat. Aging* **2**, 756–766 (2022).

14. Geuther, B. Q. *et al.* Action detection using a neural network elucidates the genetics of mouse grooming behavior. *eLife* **10**, e63207 (2021).

15. Geuther, B. *et al.* High-throughput visual assessment of sleep stages in mice using machine learning. *SLEEP* **45**, zsab260 (2021).

560 16. Sheppard, K. *et al.* Stride-level analysis of mouse open field behavior using deep-learning-
561 based pose estimation. *Cell Rep.* **38**, 110231 (2022).

562 17. Weinreb, C. *et al.* Keypoint-MoSeq: parsing behavior by linking point tracking to pose
563 dynamics. *bioRxiv* 2023.03.16.532307 (2023) doi:10.1101/2023.03.16.532307.

564 18. Wiltschko, A. B. *et al.* Mapping Sub-Second Structure in Mouse Behavior. *Neuron* **88**, 1121–
565 1135 (2015).

566 19. Pereira, T. D. *et al.* Fast animal pose estimation using deep neural networks. *Nat. Methods*
567 **16**, 117–125 (2019).

568 20. Pereira, T. D. *et al.* SLEAP: A deep learning system for multi-animal pose tracking. *Nat.*
569 *Methods* **19**, 486–495 (2022).

570 21. Marshall, J. D. *et al.* Continuous Whole-Body 3D Kinematic Recordings across the Rodent
571 Behavioral Repertoire. *Neuron* **109**, 420-437.e8 (2020).

572 22. Dunn, T. W. *et al.* Geometric deep learning enables 3D kinematic profiling across species
573 and environments. *Nat. Methods* **18**, 564–573 (2021).

574 23. Huys, Q. J. M., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from
575 neuroscience to clinical applications. *Nat Neurosci* **19**, 404–413 (2016).

576 24. Kumar, P., Dayan, P. & Wolfers, T. From Complexity to Precision—Charting Decision-Making
577 Through Normative Modeling. *JAMA Psychiatry* **81**, (2024).

578 25. Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev Gen*
579 *Psychol* **2**, 175–220 (1998).

580 26. Kerr, N. L. HARKing: Hypothesizing After the Results are Known. *Pers Soc Psychol Rev* **2**,
581 196–217 (1998).

582 27. Dezfouli, A., Griffiths, K., Ramos, F., Dayan, P. & Balleine, B. W. Models that learn how
583 humans learn: The case of decision-making and its disorders. *PLoS Comput. Biol.* **15**, e1006903
584 (2019).

585 28. Eckstein, M. K., Summerfield, C., Daw, N. D. & Miller, K. J. Predictive and Interpretable:
586 Combining Artificial Neural Networks and Classic Cognitive Models to Understand Human
587 Learning and Decision Making. *bioRxiv* 2023.05.17.541226 (2023)
588 doi:10.1101/2023.05.17.541226.

589 29. Wiltschko, A. B. *et al.* Revealing the structure of pharmacobehavioral space through Motion
590 Sequencing. *Nat. Neurosci.* **23**, 1433–1443 (2020).

591 30. Gschwind, T. *et al.* Hidden behavioral fingerprints in epilepsy. *Neuron* **111**, 1440-1452.e5
592 (2023).

593 31. Uffelmann, E. *et al.* Genome-wide association studies. *Nat Rev Methods Primers* **1**, 59
594 (2021).

595 32. Farrell, M. S. *et al.* Evaluating historical candidate genes for schizophrenia. *Mol Psychiatr* **20**,
596 555–562 (2015).

597 33. Border, R. *et al.* No Support for Historical Candidate Gene or Candidate Gene-by-Interaction
598 Hypotheses for Major Depression Across Multiple Large Samples. *Am J Psychiat* **176**, 376–387
599 (2019).

600 34. Duncan, L. E., Ostacher, M. & Ballon, J. How genome-wide association studies (GWAS) made
601 traditional candidate gene studies obsolete. *Neuropsychopharmacol* **44**, 1518–1523 (2019).

602 35. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *Plos Med* **2**, e124 (2005).

603 36. Markowitz, J. E. *et al.* Spontaneous behaviour is structured by reinforcement without
604 explicit reward. *Nature* **614**, 108–117 (2023).

605 37. Cui, T. *et al.* Gene–gene interaction detection with deep learning. *Commun. Biol.* **5**, 1238
606 (2022).

607 38. Fang, G. *et al.* Discovering genetic interactions bridging pathways in genome-wide
608 association studies. *Nat. Commun.* **10**, 4274 (2019).

609 39. Skutt-Kakaria, K., Reimers, P., Currier, T. A., Werkhoven, Z. & Bivort, B. L. de. A neural circuit
610 basis for context-modulation of individual locomotor behavior. *bioRxiv* 797126 (2019)
611 doi:10.1101/797126.

612 40. Ayroles, J. F. *et al.* Behavioral idiosyncrasy reveals genetic control of phenotypic variability.
613 *Proc National Acad Sci* **112**, 6706–6711 (2015).

614 41. Werkhoven, Z. *et al.* The structure of behavioral variation within a genotype. *eLife* **10**,
615 e64988 (2021).

616 42. Akhund-Zade, J., Ho, S., O'Leary, C. & Bivort, B. de. The effect of environmental enrichment
617 on behavioral variability depends on genotype, behavior, and type of enrichment. *J. Exp. Biol.*
618 **222**, jeb202234 (2019).

619 43. Bivort, B. de *et al.* Precise Quantification of Behavioral Individuality From 80 Million
620 Decisions Across 183,000 Flies. *Front. Behav. Neurosci.* **16**, 836626 (2022).

621 44. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-Based Influences on
622 Humans' Choices and Striatal Prediction Errors. *Neuron* **69**, 1204–1215 (2011).

623 45. Akam, T. *et al.* The Anterior Cingulate Cortex Predicts Future States to Mediate Model-
624 Based Action Selection. *Neuron* **109**, 149-163.e7 (2020).

625 46. Silva, C. F. da & Hare, T. A. Humans primarily use model-based inference in the two-stage
626 task. *Nat. Hum. Behav.* **4**, 1053–1066 (2020).

627 47. Kim, S. M. & Frank, L. M. Hippocampal Lesions Impair Rapid Learning of a Continuous Spatial
628 Alternation Task. *Plos One* **4**, e5494 (2009).

629 48. Singer, A. C., Carr, M. F., Karlsson, M. P. & Frank, L. M. Hippocampal SWR Activity Predicts
630 Correct Decisions during the Initial Learning of an Alternation Task. *Neuron* **77**, 1163–1173
631 (2013).

632 49. Kay, K. *et al.* Constant Sub-second Cycling between Representations of Possible Futures in
633 the Hippocampus. *Cell* **180**, 552-567.e25 (2020).

634 50. Fernández-Ruiz, A. *et al.* Long-duration hippocampal sharp wave ripples improve memory.
635 *Science* **364**, 1082–1086 (2019).

636 51. Sigurdsson, T., Stark, K. L., Karayiorgou, M., Gogos, J. A. & Gordon, J. A. Impaired
637 hippocampal–prefrontal synchrony in a genetic mouse model of schizophrenia. *Nature* **464**,
638 763–767 (2010).

639 52. Kastner, D. B., Gillespie, A. K., Dayan, P. & Frank, L. M. Memory alone does not account for
640 the speed of learning of a simple spatial alternation task in rats. *J Neurosci* **40**, JN-RM-0972-20
641 (2020).

642 53. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly
643 associated with autism. *Nature* **485**, 237–241 (2012).

644 54. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both
645 Developmental and Functional Changes in the Neurobiology of Autism. *Cell* **180**, 568-584.e23
646 (2020).

647 55. Lord, C. *et al.* Autism spectrum disorder. *Nat. Rev. Dis. Prim.* **6**, 5 (2020).

648 56. Clarke, D., Romano, J. P. & Wolf, M. The Romano-Wolf Multiple Hypothesis Correction in
649 Stata. *Ssrn Electron J* (2020) doi:10.2139/ssrn.3513687.

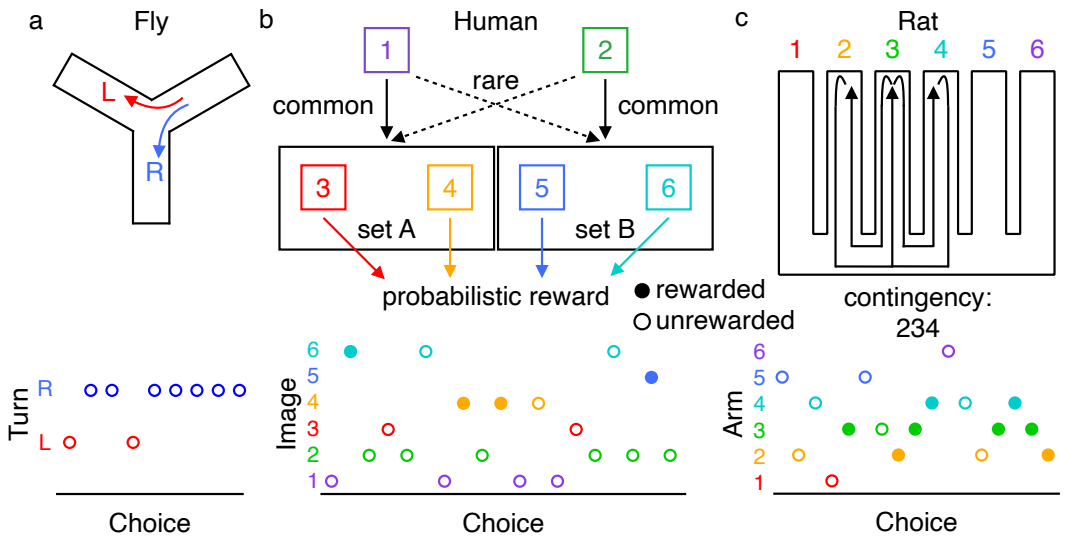


Figure 1. Choice is a common discretization for behavior. In the three behavioral tasks under consideration in this work, the actions of the subjects are broken down into a series of choices. (**a**) Fly y-maze (top), and example set of choices of left (red) and right (blue) turns (bottom) of an individual fly. Data come from Buchanan et al.[4]. (**b**) Two-step task, performed by human subjects (top), and an example set of choices of an individual subject (bottom). The colors correspond to the different objects chosen (bottom). Data come from Gillan et al.[6]. (**c**) Spatial alternation behavior, performed by rats (top), and an example set of choices of an individual rat (bottom). The contingency is defined as the 3 arms that can be rewarded, and the 6-arm track enables different sets of three arms to be rewarded. In **b** and **c,** filled in circles indicate that reward was received.

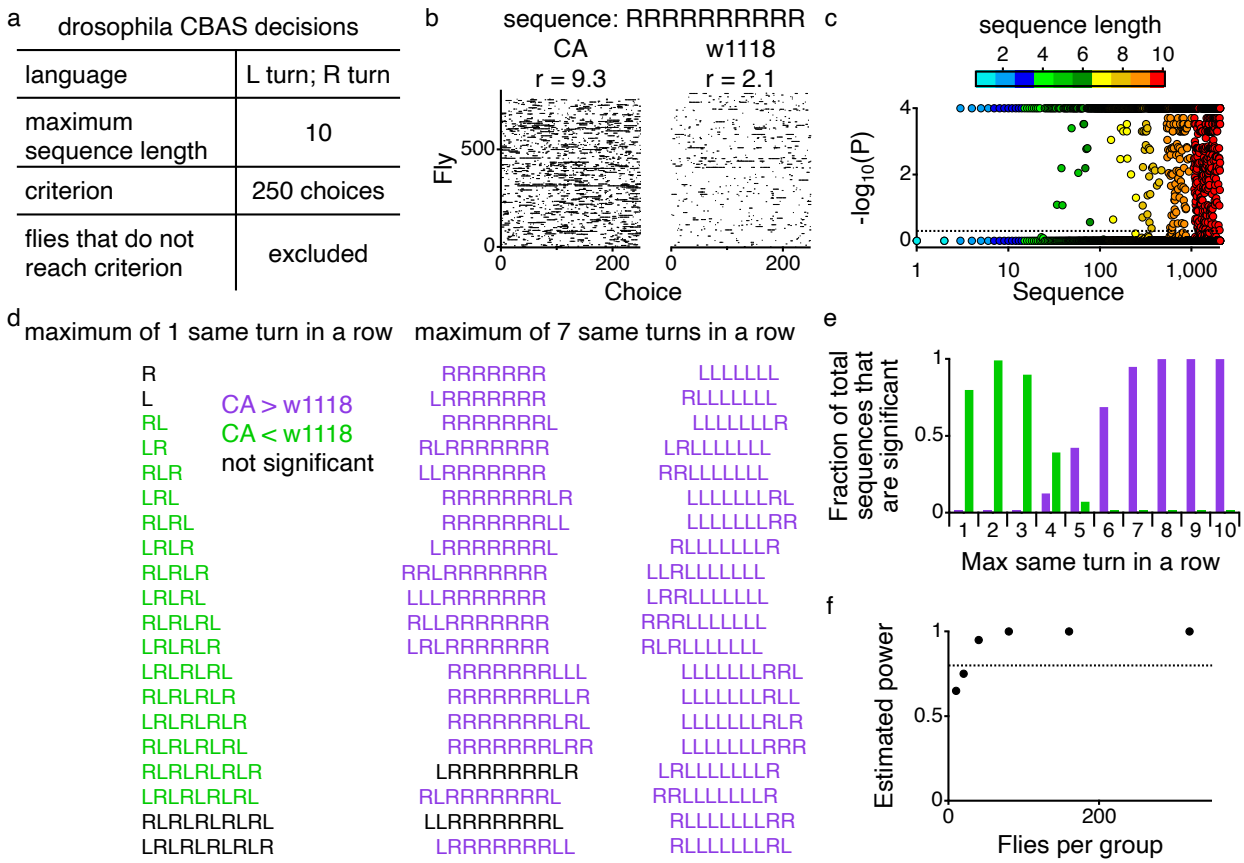


Figure 2. CBAS applied to y-maze produces interpretable differences between fly strains. CBAS was applied to flies tracked on a y-maze (Fig 1a) (**a**) This table demarcates the decisions made to perform the CBAS on this dataset. There was a total of 1,225 Cambridge-A (CA) and 1,372 w1118 flies in the dataset, 466 CA and 565 w1118 were excluded from analysis due to not reaching a total of at least 250 turns. (**b**) The occurrence of a single sequence of turns (10 right turns in a row) in the two strains during the first 250 turns. The rows show the occurrence of that sequence of turns for all individual flies from each strain. The CA strain has a rate of 9.3, meaning that, on average, each fly utilizes this sequence 9.3 times. The W1118 strain has a rate of 2.1, meaning that, on average, each fly utilizes this sequence 2.1 times. (**c**) CBAS Manhattan plot displays the p-value for each sequence. Each sequence has two p-values on this plot, one for CA > w1118, and the other for CA < w1118. The sequences are ordered based on the number of choices in the sequence, and they are displayed on a log scale to make all sequence lengths visible. Within a given sequence length, the sequences are ordered based on frequency of occurrence in the entire dataset. The horizontal dotted line indicates the significance threshold of 5% control of the median false discovery proportion. A total of 10,000 resamplings were used to calculate the p-values making the maximum value on the plot 4.00004 ($-\log\left(\frac{1}{10{,}001}\right)$). (**d**) Every sequence that either has a maximum of 1 turn in a direction in a row (left) or a maximum of 7 turns in the same direction in a row (right). Sequences are colored based on whether CBAS identifies them as occurring significantly more in the CA strain (purple), w1118 strain (green) or as not significantly different (black). The sequences on the right are aligned to the 7 right or left turns. (**e**) The maximum number of the same turns in a row was

681 calculated for every sequence in the dataset (2,046 total sequences), and for every number of
682 maximum turns in a row, the fraction of sequences that whose prevalence was significantly
683 greater in the CA strain (purple) or the w1118 strain (green) is plotted. The CA and w1118
684 strains showed separation with other related metrics even for the max turns that show both
685 significance for both direction of comparison (e.g. max same turn in a row of 4) (Fig S1a). (**f**)
686 Power estimate for different sized groups of flies in each strain (see methods). Horizontal
687 dotted line shows a value of 80%.

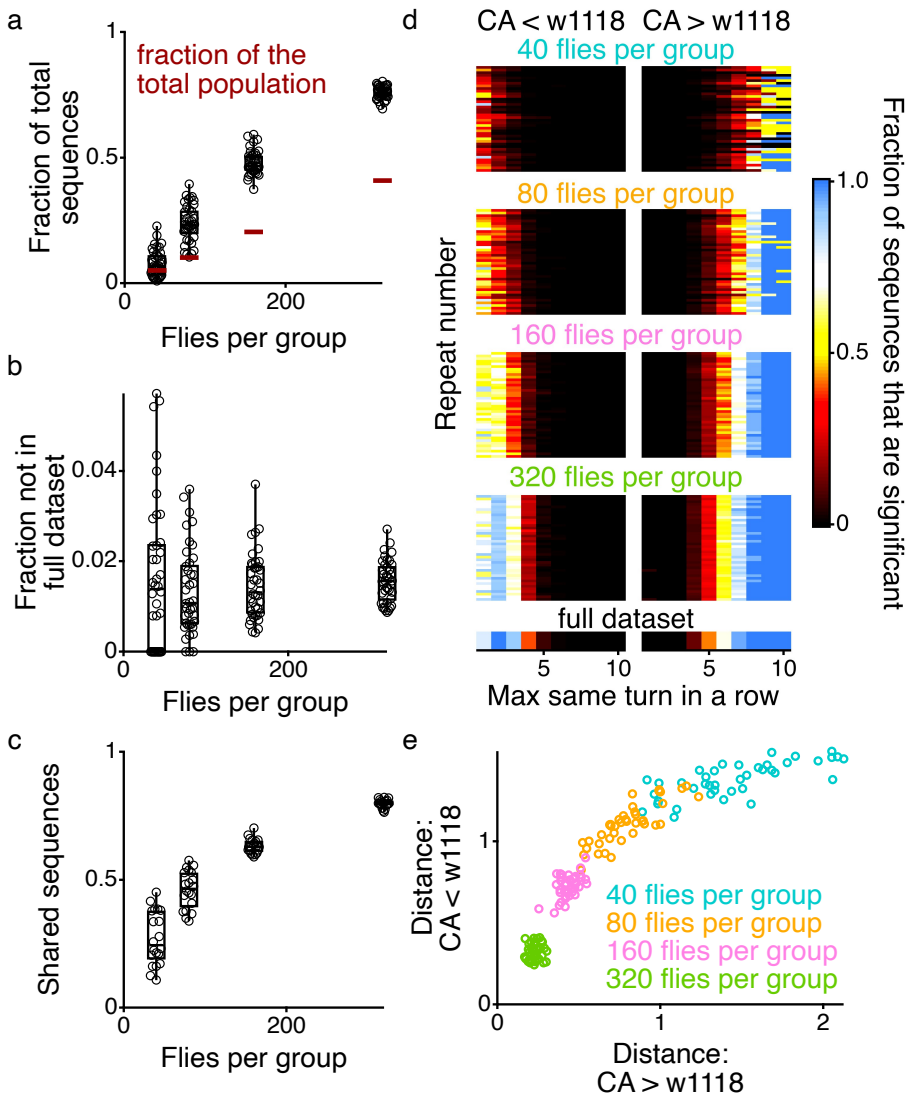


Figure 3. Common, but degraded output with CBAS on smaller sample sizes. For all panels in this figure, many repeats of smaller sample sizes were generated from the fly data used in Fig 2 by resampling subjects (without replacement within a group) from the full dataset. (**a**) Each CBAS run on the smaller sample size identified some number of significant sequences. Those sequences were compared to the sequences identified in the CBAS on the full dataset, and the graph shows the ratios of the number of sequences identified by the smaller sample size that were also in the full dataset to the total number of sequences identified in the full dataset. As a comparison, the ratio of the number of flies in the smaller CBAS to the total number of flies in the dataset is plotted in the maroon horizontal lines. (**b**) The number of significant sequences identified in the smaller sample size that were not also identified in the full dataset is plotted over the total number of significant sequences identified in the smaller sample size. (**c**) In creating the smaller samples sizes, 20 paired sets of animals were generated that had no overlapping individuals. The number of the same significant sequences that were identified with these nonoverlapping sets of flies is plotted over the average number of significant sequences in the pair. (**d**) For each repeat of each sample size, all of the sequences were

categorized based on the maximum number of turns in the same direction and the fraction of significant sequences within those categories are plotted (as in Fig 2e). The bottom row of this plot is from the full dataset and is identical to Fig 2e. (**e**) The Euclidean distance between each row from panel **d** and the full dataset row is plotted. Colors correspond to the sample sizes as shown in panel **d**. For panels **a** – **c** data points are overlayed by box and whisker plots. The center line of the box displays the median of the data, the top and bottom lines of the box show the 25th and 75th quartiles, respectively, and the end of the whiskers show the full range of the data.

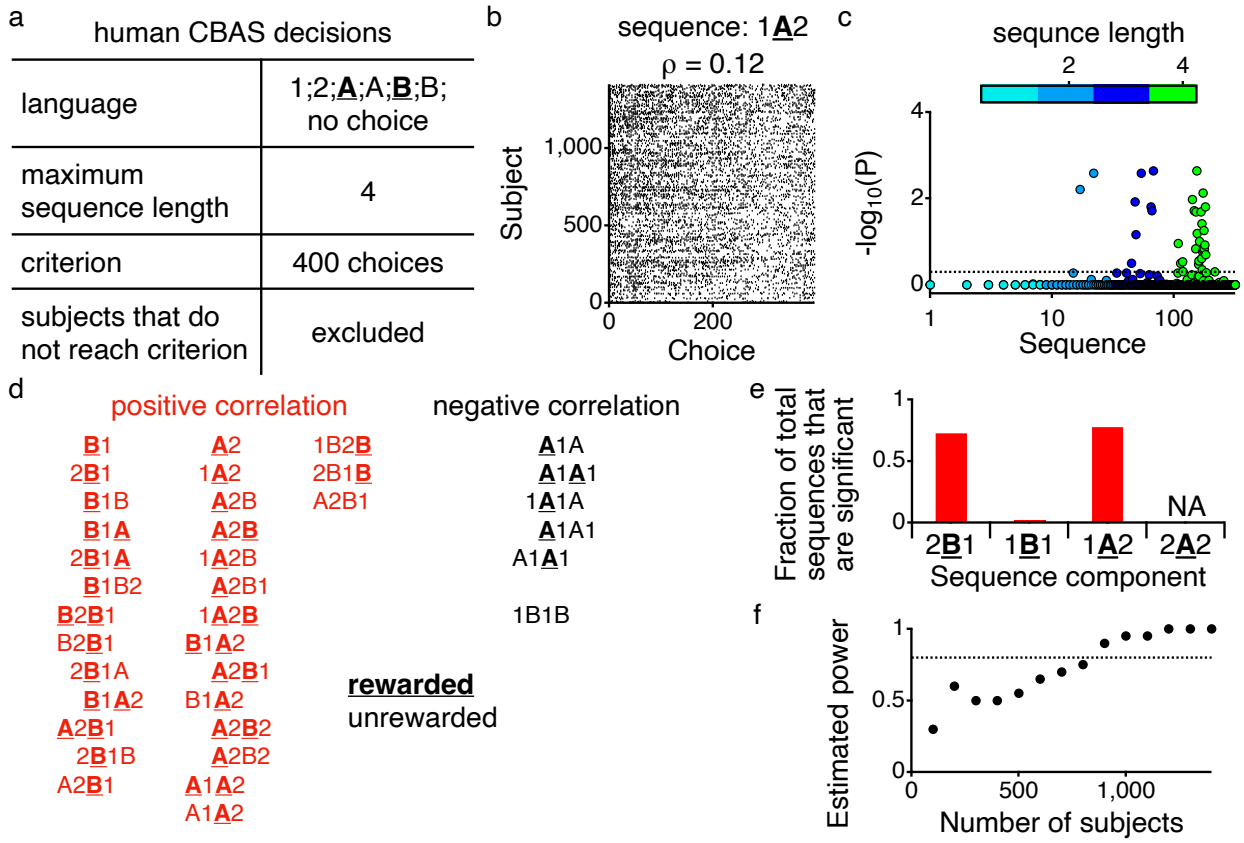


Figure 4. CBAS identifies unexpected sequences in human dataset. CBAS was applied to humans performing the two-step task (Fig 1b) (**a**) This table demarcates the decisions made to perform the CBAS on this dataset. There was a total of 1,413 human subjects in the dataset. (**b**) The occurrence of a single sequence of choices ('1**A**2', meaning choosing object 1, then getting rewarded for a choice in set A, then choosing object 2) across all subjects in the dataset. Subjects are ordered based on their factor 2 ("intrusive thoughts and compulsive behaviors") score. This sequence shows a correlation (ρ) of 0.12 with factor 2 score from the questionnaire factor analysis (see S2a). (**c**) CBAS Manhattan plot displays the p-value for each sequence. Each sequence has two p-values on this plot, one for positive correlation and one for negative correlation. The sequences are ordered based on the number of choices in the sequence and are displayed on a log scale to make all sequence lengths visible. Within a given sequence length, the sequences are ordered based on frequency of occurrence in the entire dataset. The horizontal dotted line indicates the significance threshold of 5% control of the median false discovery proportion. A total of 10,000 resamplings were used to calculate the p-values making the maximum value on the plot 4.00004 ($-\log\left(\frac{1}{10{,}001}\right)$). (**d**) Every sequence that was significantly positively (left) or negatively (right) correlated with factor 2 score. The sequences on the left are aligned to **B**1 or **A**2, and sequences on the right are aligned to **A**1. (**e**) For the categories listed (2**B**1, 1**B**1, 1**A**2, 2**A**2), the number of significant sequences that contain the category is plotted over total number of sequences that exist in the dataset that contain the category. For 2**A**2 no subject in the dataset performed that sequence. (**f**) Power estimate for different sized groups of subjects (see methods). Horizontal dotted line shows a value of 80%.

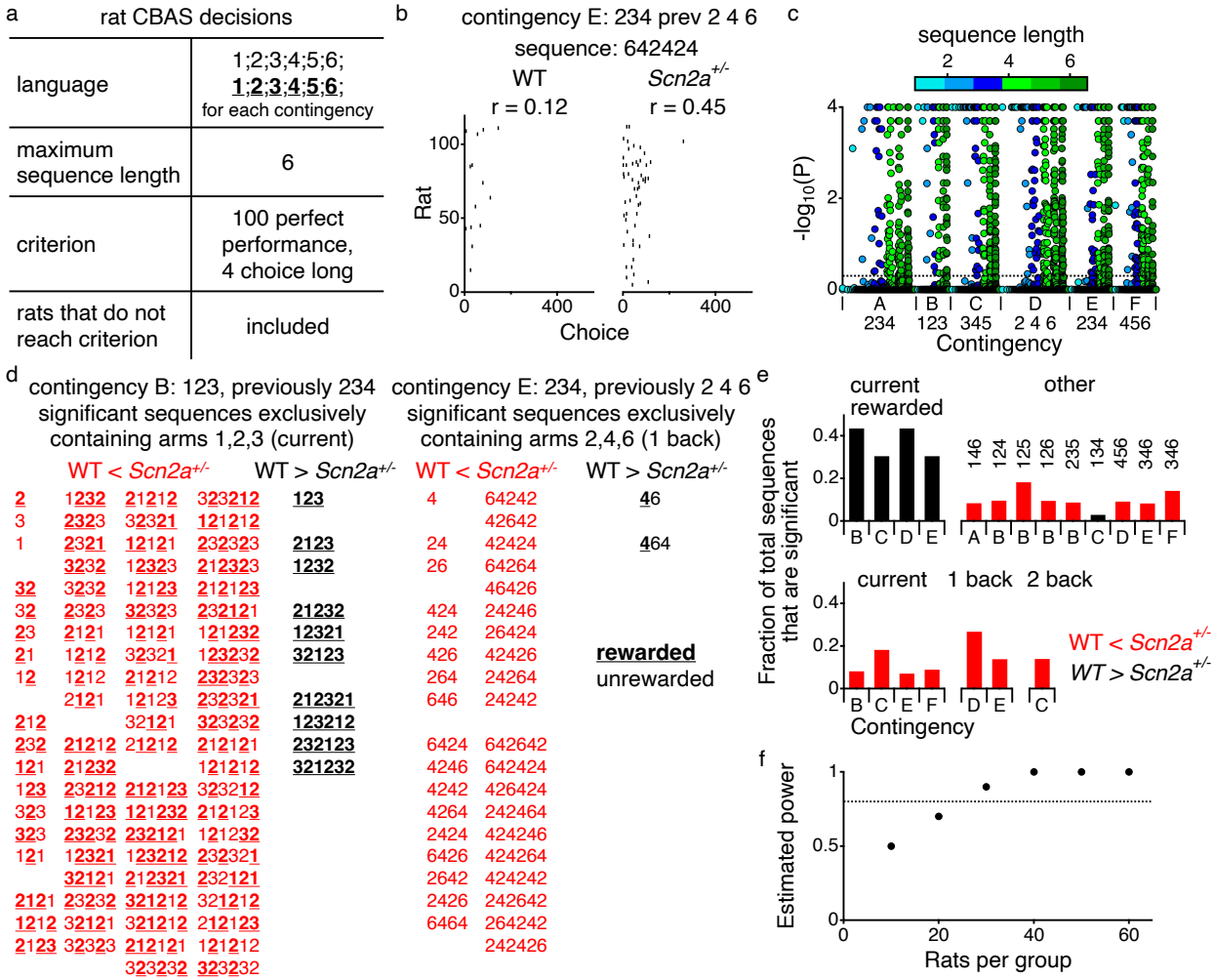
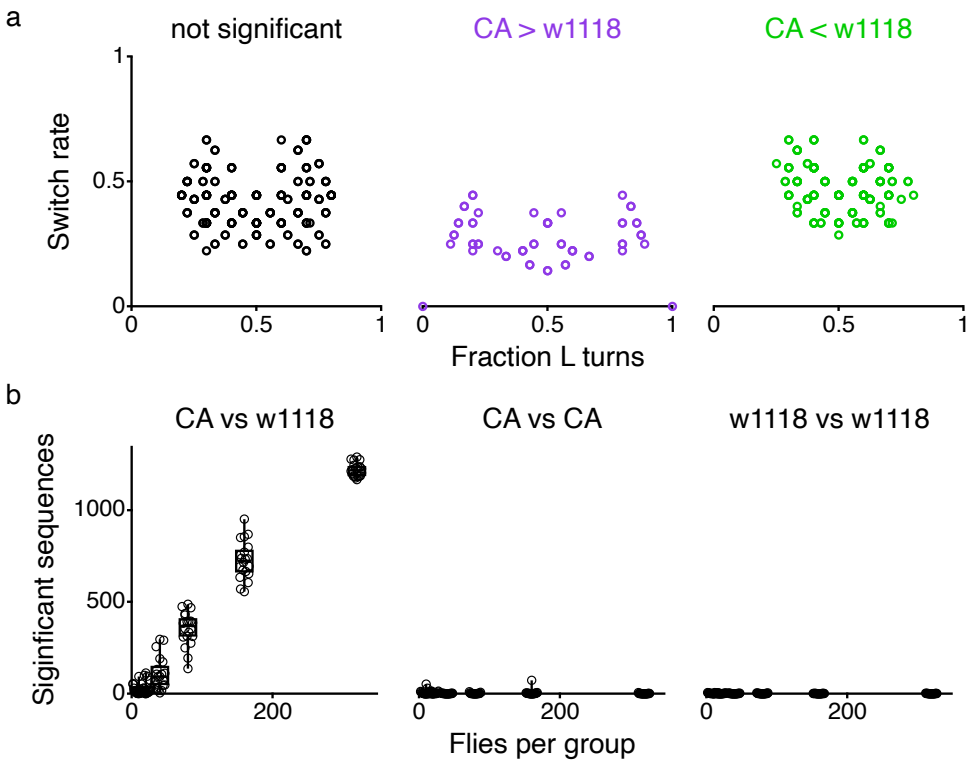
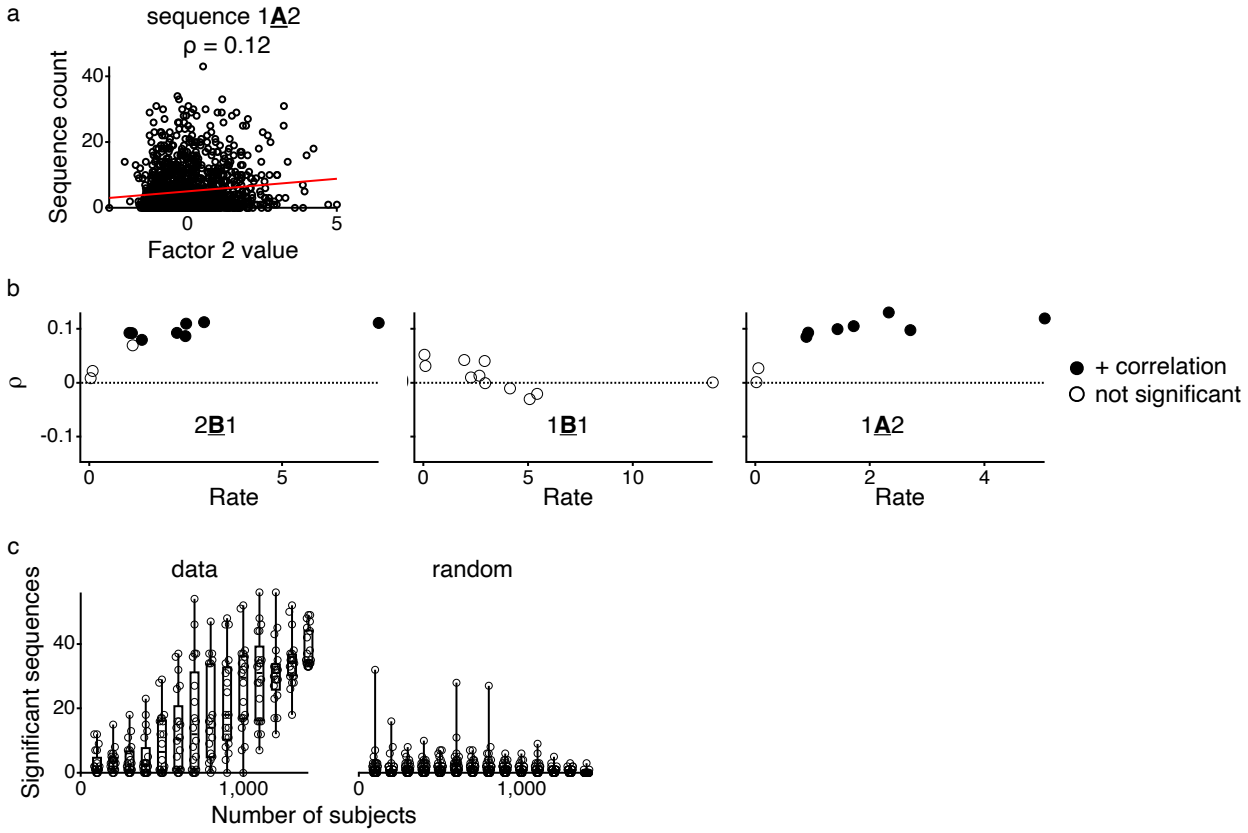


Figure 5. CBAS provide interpretable information to phenotype *Scn2a*$^{+/-}$ rats. CBAS was applied to rats performing a spatial alternation task (Fig 1c) (**a**) This table demarcates the decisions made to perform the CBAS on this dataset. There was a total of 121 WT and 120 *Scn2a*$^{+/-}$ rats for the analyses (see methods). (**b**) The occurrence of a single (unrewarded) sequence of arm choices (642424; all from contingency D which rewards 246) in the two rat genotypes during contingency E, which is rewarded at arms 234. The rows show the occurrence of that sequence of choices for all individual rats from each genotype. WT rats used this sequence at a rate of 0.12, meaning that on average about 1 out of every 10 rats used this sequence. *Scn2a*$^{+/-}$ rats used this sequence at a rate of 0.45, meaning that on average 1 out of every 2 rats used this sequence. (**c**) CBAS Manhattan plot displays the p-value for each sequence. Each sequence has two p-values on this plot, one for WT > *Scn2a*$^{+/-}$, and the other for WT > *Scn2a*$^{+/-}$. The sequences are ordered based on the number of choices in the sequence, and within each contingency are displayed on a log scale to make all sequence lengths visible. Within a given sequence length, the sequences are ordered based on frequency of occurrence in the entire dataset. The rewarded arms for each contingency are shown below the letter symbol for each contingency. The horizontal dotted line indicates the significance threshold of 5% control of the median false discovery proportion. A total of 10,000 resamplings were used to calculate the p-values making the maximum value on the plot 4.00004 ($-\log\left(\frac{1}{10{,}001}\right)$). (**d**)

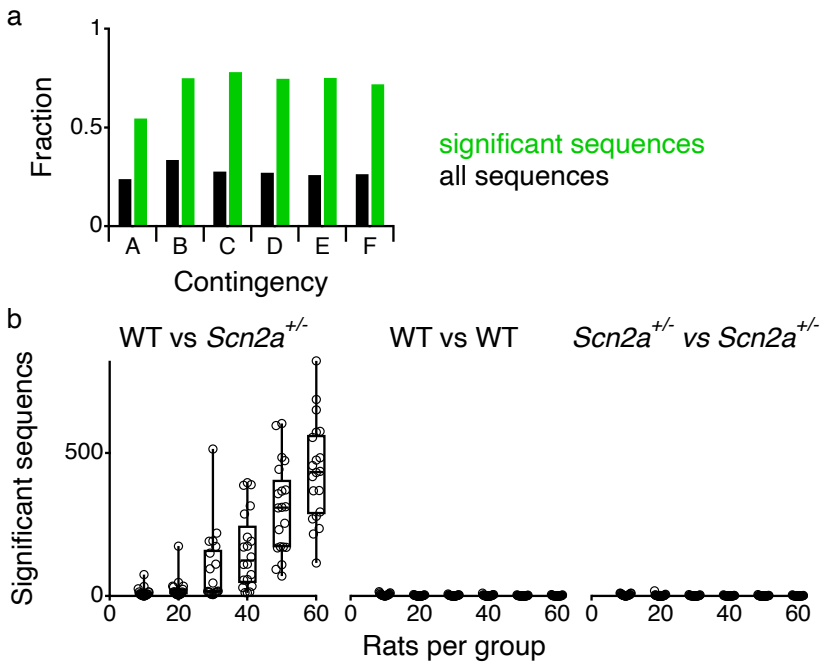
750 Every significant sequence that exclusively contains arms 1, 2, 3 (irrespective of reward) during
751 contingency B (left) or every significant sequence that exclusively contains arms 2, 4, 6
752 (irrespective of reward) during contingency E (right). Sequences that occur significantly more in
753 *Scn2a*$^{+/-}$ are shown in red, and those that occur significantly more in WT are shown in black. (**e**)
754 The sets of 3 arms which show a significantly greater fraction of significant sequences than
755 would be expected by chance (see methods). The categories above each group of bar plots
756 indicate the structure of the set of 3 arms. "Current" indicates the 3 arms from the current
757 contingency irrespective of reward, "current rewarded" indicates the 3 arms from the current
758 contingency all of which are rewarded, "1 back" indicates the 3 arms from the prior
759 contingency, "2 back" indicate the 3 arms from 2 contingencies prior, and "other" are
760 categories that show significant fractions that do not fit the other categories. Of note, the
761 sequences in the "current rewarded" category are a subset of the sequences in the "current"
762 category, as can be seen in panel **d**, left. (**f**) Power estimate for different sized groups of rats in
763 each genotype (see methods). Horizontal dotted line shows a value of 80%.



764 **Supplementary Figure 1.** (**a**) All sequences with a maximum length of 4 turns in the same
765 direction in a row (Fig 2e) were evaluated for the fraction of L turns in the sequence
766 (encompassing L = 4 – 8) and the rate of switching (number of changes in turn direction divided
767 by the total number of turns in the sequence). These two metrics show a separation in the
768 sequences that were significantly greater in the CA strain compared to those that were
769 significantly greater in the w1118 strain: the CA strain had sequences with more extreme
770 fraction of left turns and lower switch rate, consistent with more turns in the same direction.
771 (**b**) Left: the number of significant sequences when randomly resampling the populations
772 without replacement and calculating CBAS on the smaller sample sizes. Middle: the number of
773 significant sequences when comparing smaller samples sizes of the CA strain to itself, with
774 nonoverlapping individuals in each group. Right: the number of significant sequences when
775 comparing smaller samples sizes of the w1118 strain to itself, with nonoverlapping individuals
776 in each group. Data points are overlayed by box and whisker plots. The center line of the box
777 displays the median of the data, the top and bottom lines of the box show the 25th and 75th
778 quartiles, respectively, and the end of the whiskers show the full range of the data.



Supplementary Figure 2. (**a**) Correlation between the sequences count for each human subject and their factor 2 score for sequence: 1**A**2, which means choosing object 1, then making a choice in set A and getting rewarded, and then choosing object 2. Factor 2 reflects the intrusive thoughts and obsessive behavior loading on the psychiatric symptom questionnaires. Red line shows the linear fit to the data. (**b**) The correlation plotted as a function of the average sequence rate for all sequences containing the unit 2**B**1 (left), 1**B**1 (middle), or 1**A**2 (right). Filled in circles indicate those sequences that show a significant positive correlation. (**c**) Left: the number of significant sequences when randomly resampling the populations without replacement and calculating CBAS on the smaller sample sizes. Right: the number of significant sequence when randomly resampling the populations without replacement and comparing it to randomly generated factor 2 scores drawn from a distribution imputed from the original factor 2 scores and calculating CBAS on the smaller sample sizes. Data points are overlayed by box and whisker plots. The center line of the box displays the median of the data, the top and bottom lines of the box show the 25th and 75th quartiles, respectively, and the end of the whiskers show the full range of the data.



Supplementary Figure 3. (**a**) For the sequences that exclusively contain all sets of 3 arms (Fig 5d&e) the fraction of the total number of sequences is plotted in black for each contingency, and the fraction of significant sequences that are a part of the sequences that exclusively contain all sets of 3 arms compared to the total number of significant sequences in each contingency is plotted in green. (**b**) Left: the number of significant sequences when randomly resampling the populations without replacement and calculating CBAS on the smaller sample sizes. Middle: the number of significant sequence when comparing smaller samples sizes of the WT genotype to itself, with nonoverlapping individuals in each group. Right: the number of significant sequence when comparing smaller samples sizes of the *Scn2a*$^{+/-}$ genotype to itself, with nonoverlapping individuals in each group. Data points are overlayed by box and whisker plots. The center line of the box displays the median of the data, the top and bottom lines of the box show the 25th and 75th quartiles, respectively, and the end of the whiskers show the full range of the data.