brain sciences

MDPI

Article

Using Machine Learning to Discover Latent Social Phenotypes in Free-Ranging Macaques

Seth Madlon-Kay [1,*], Lauren J. N. Brent [2], Michael J. Montague [1], Katherine A. Heller [3] and Michael L. Platt [1,4,5]

[1] Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; montag@mail.med.upenn.edu (M.J.M); mplatt@mail.med.upenn.edu (M.L.P.)

[2] Center for Research in Animal Behaviour, University of Exeter, Exeter EX4 4QG, UK; L.J.N.Brent@exeter.ac.uk

[3] Department of Statistical Science, Duke University, Durham, NC 27708, USA; kheller@stat.duke.edu

[4] Department of Psychology, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA

[5] Marketing Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

* Correspondence: madseth@mail.med.upenn.edu; Tel.: +1-215-898-1579

Academic Editor: Stephanie Cacioppo
Received: 29 March 2017; Accepted: 16 July 2017; Published: 21 July 2017

Abstract: Investigating the biological bases of social phenotypes is challenging because social behavior is both high-dimensional and richly structured, and biological factors are more likely to influence complex patterns of behavior rather than any single behavior in isolation. The space of all possible patterns of interactions among behaviors is too large to investigate using conventional statistical methods. In order to quantitatively define social phenotypes from natural behavior, we developed a machine learning model to identify and measure patterns of behavior in naturalistic observational data, as well as their relationships to biological, environmental, and demographic sources of variation. We applied this model to extensive observations of natural behavior in free-ranging rhesus macaques, and identified behavioral states that appeared to capture periods of social isolation, competition over food, conflicts among groups, and affiliative coexistence. Phenotypes, represented as the rate of being in each state for a particular animal, were strongly and broadly influenced by dominance rank, sex, and social group membership. We also identified two states for which variation in rates had a substantial genetic component. We discuss how this model can be extended to identify the contributions to social phenotypes of particular genetic pathways.

Keywords: machine learning; behavioral genetics; social neuroscience; animal models

1. Introduction

Understanding how biological and environmental factors shape human social phenotypes—the ways in which we interact with others and in which others shape our own behavior—is a topic of intense interest in neuroscience. While neuroscientific research has made impressive strides in identifying genetic pathways and brain areas that contribute to specific social phenotypes, traditional approaches are limited in their ability to assess how biological and environmental factors contribute to naturalistic social behaviors in real-world environments. Animals used in laboratory research may differ from humans in the brain areas associated with social cognition, and live in laboratory environments with much less varied and dynamic social interactions than natural environments [1]. Research in humans faces similar limitations, as comprehensive, in-depth observations of human social behavior in natural environments is both logistically and ethically problematic, while controlled laboratory tasks take place in restricted and low-stakes environments which may leave the true social environment unobserved.

An alternative approach to understanding biological contributions to human social behavior is to study free-ranging animals whose biology and social behavior more closely approximate our own. Free-ranging non-human primates (NHPs) provide naturalistic and human-relevant variability in social behavior that cannot be modeled in the laboratory or easily (or ethically) collected in humans. With free-ranging NHPs, it is possible to collect comprehensive data on many different aspects of social behavior, as well as both noninvasive and invasive biological samples and life history information that can be used to disentangle genetic and environmental influences. However, natural social behavior is both noisy and high-dimensional, presenting severe challenges for traditional methods of data analysis. Standard methods for relating genetic variation to traits take an atomistic view by measuring genetic influences on each measured variable individually. While this approach makes sense for specific, well-defined phenotypes of interest such as educational attainment [2] or height [3], it is less useful for high-dimensional data where any individual dimension is of limited or unknown significance.

In the case of natural behavior, continuous streams of activity are necessarily segmented into many discrete behavioral categories during measurement, in many cases with no specific behavior being of special interest relative to the whole. The challenge is thus to develop a rigorous and biologically meaningful approach for reconstructing a "whole" behavioral state based on discrete, atomized behaviors. Such a challenge can be viewed as a type of latent structure learning, in that there is some underlying structure that is not directly observed but that ties together many different individual pieces of data. Models of this sort are increasingly popular in the field of probabilistic machine learning [4]. Though machine learning models are most frequently used for prediction, here we adapt methods from the machine learning literature to infer a meaningful and useful structure in observational data sets.

Any source of genetic variation is unlikely to affect only a single discrete social behavior. A genetic variant associated with grooming behavior, for instance, may also be associated with increased time spent in the proximity of other animals in general. Genetic influences are more likely distributed across many different individual social behaviors than localized to specific types of actions [5]. Furthermore, in the case of observations of natural social behavior, it is important to recognize that each type of behavior does not occur in isolation. Rather, the set of behaviors that occur during a given observational session may reflect the current situation or behavioral state the animal is in. The relationship between any given behavior and an underlying social phenotype thus may vary with context. For example, adjusting a watchband while speaking to a group of people may be a nervous tick reflecting social anxiety, while adjusting a watchband in one's bedroom might reflect only an ill-fitting watchband.

Thus, it is critical for any useful model of observed social behavior to reflect the complex interrelatedness of individual behaviors, while retaining identifiability and scalability to large data sets with dozens of individual behaviors and thousands of observations. To achieve this goal, we here adapt a family of models from machine learning known as topic models. Topic models are tools for describing the topics that are covered by a corpus of documents and to what degree each document is focused on each topic [6,7]. In our analysis, each animal is analogous to a document, while the animal's social phenotype, or the weights assigned to different behavioral states, correspond to the weights a document assigns to each of the different topics. Under a topic model, each word in a document belongs to a topic, just as in the current model each observational session belongs to a behavioral state. Of the many topic modeling variants that exist in the literature, the current model is most closely related to the logistic normal topic model [8,9]. Our model differs from the standard logistic normal topic model in that we do not explicitly model correlations between topic weights (here, weights on behavioral states); instead, our model uses a hierarchical regression layer to incorporate outside information about documents (here, animals) in predicting their phenotypes. To the best of our knowledge, no previous topic model has included hierarchical regression at the level of topic weights (but see [10] for a related model).

Recent research on behavioral phenotypes in NHPs has used principle components analysis (PCA) to identify latent structure in natural behavior [5,11], as well as in the study of animal behavior generally [12]. However, that research relied on rates of behaviors averaged across many different observational periods on the same animal. This means that the relationships among behaviors and the patterns in which they co-occur within a single observation are lost. That is, PCA captures variation among animals, but at the cost of losing all information on variability at the level of individual observational periods. By contrast, the hierarchical nature of topic models allows us to explicitly capture variability at both the level of individual animals and the level of observational sessions, in the form of the phenotypes used to represent individual animals and the behavioral states used to characterize observations, respectively.

Our behavioral data is a large set of animal observations taken from the free-ranging rhesus macaque (*Macaca mulatta*) colony on Cayo Santiago Island off the coast of Puerto Rico. Rhesus macaques provide an excellent model species for studying biological and environmental influences on human social behavior due to their extensive use in lab and field research; neural circuitry that is homologous to that of humans; and complex social behaviors that are critical to their biological success [13–16]. Because demographic, pedigree, and genetic data are also available for the colony [17,18], this detailed record of natural social behavior provides a unique opportunity for modeling genetic, demographic, and environmental influences on complex social phenotypes.

2. Materials and Methods

2.1. Study Site and Subjects

The studied population is a colony of rhesus macaques (*Macaca mulatta*) living on the island of Cayo Santiago, a 37-acre island 1 km off the southeastern coast of Puerto Rico. This is a free-ranging, freely-breeding population with known pedigrees, rich data on life histories and fitness, and extensive genetic and observational data on behavior. The colony was founded in 1938 with a population of 409 Indian-origin rhesus macaques and is currently maintained by the University of Puerto Rico Medical Sciences Campus [19]. The current population numbers approximately 1600 animals self-organized into six different social groups. Approximately 600 of the animals are adults of age six or above, and 600 are juveniles between the ages of one and five. Researcher and caretaker intervention in the population is minimal. Animals in the colony are provided commercially available monkey chow daily and unlimited access to water. Animals are handled only during designated annual trapping periods, during which infants are tagged for identification and population control can be implemented via the removal of small numbers of animals.

The data used in this study were 8205 ten-minute focal observations collected from adult (age > 6 years) male and female macaques from three social groups (F, HH, and R). Observational data was collected from group F during 2012 and 2013 and from group HH and R during years 2014 and 2015, respectively. A total of 227 macaques (71 male) were represented in this data set, with 106 from group F (41 male), 41 from group HH (11 male), and 80 from group R (19 male). The number of focal observations per animal ranged from 80 (approximately 13.33 h) to 11 (approximately 1.83 h), with an average of 36.15 focal observations per animal (approximately 6 h).

2.2. Observational Data

The data is comprised of ten-minute focal samples, wherein a trained researcher tracks an individual monkey (the focal animal) for a ten-minute period while comprehensively recording the animal's behaviors [17,18]. Once each ten-minute session is complete, the observer moves on to a new focal animal. Observers recorded the times at which the monkey engaged in any behaviors specified by a predetermined ethogram, as well as the identities of nearby animals during the focal observation [17]. The order in which animals were observed was semi-randomized to equalize the times of day and year that animals were observed. The ethogram used to guide data collection consisted of both social and non-social behaviors, both of which were classified as either activities, which have durations, or events, which occur over effectively zero duration. Social behaviors were further classified into affiliative and agonistic behaviors. Each social behavior was further divided based on whether the behavior was initiated by another animal and directed towards the focal animal, or initiated by the focal animal towards another animal. A total of 30 behaviors were used in the present study, including behaviors divided by giving versus receiving. For a description of each behavior included in the model, see the Appendix. Only social behaviors involving another adult macaque were used in this study; behaviors directed at infant or juvenile macaques or human researchers were not included.

This research complied with protocols approved by the Institutional Animal Care and Use Committee of the University of Puerto Rico (protocol #A6850108) and adhered to the legal requirements of the United States of America.

2.3. Genetic Data

We obtained animal pedigrees from a long-term database maintained by the Caribbean Primate Research Center (CPRC). From the founding of the population up through 1992, maternal identity was ascertained by behavioral observations, such as nurturing behaviors and lactation. For most macaques born after 1992, both maternal and paternal identity were ascertained genetically through the analysis of 29 microsatellite markers. Previous studies revealed a 97.4% agreement rate between maternity determined by behavior and by genetics [5]. The population pedigree was used to generate a kinship matrix for the animals in the present study via the R package kinship2 [20]. The kinship matrix was then element-wise multiplied by two to create the genetic covariance matrix used to estimate the heritability of behavioral states.

2.4. A Model for Social Phenotypes

At the heart of the model is a finite set of latent behavioral states that determines the rates at which different behaviors occur. An animal's social phenotype is represented in the model as the rate at which that animal experiences the different behavioral states. We describe each component of the model in detail below. In this section, we first present a conceptual description of the model, and the description of the mathematical implementation follows in Section 2.6.

Each ten-minute focal sample is assumed to belong to one and only one behavioral state; that is, during a given focal observation, the focal animal is assumed to be in a particular behavioral state. Each state is defined by a set of parameters that determines the probabilities that each of the discrete behaviors will occur, and, should they occur, how frequently or in what amount they will occur during a focal observation. In the data analyzed here, we used a categorical distribution over various levels of behavioral quantities, but the algorithm can trivially be modified to use different data likelihoods, such as a Poisson distribution for data in count form, a normal distribution for continuous data, and so on. For details on the coding of discrete behaviors in the current study, see Section 2.5 below.

Every animal has a different probability distribution over behavioral states, which describes how often it finds itself in each of those states. This probability is what we call the animal's social phenotype. An intuitive interpretation of this aspect of the model is that at the beginning of a focal observation, the focal animal flips a K-sided die where K is the number of states in the model. Whichever side the die lands on determines which behavioral state the animal will be in for the duration of the focal observation. While the K behavioral states are common across all animals, each animal has a unique die with different probabilities of landing on any given side compared to any other animal.

Ultimately, we are interested in not just describing social phenotypes and behavioral states, but also understanding how they are associated with other variables of scientific interest, be they morphological, demographic or genetic. Accordingly, our model allows the phenotypes of each animal to be influenced by covariates. We accomplish this by having the probabilities in the animals' phenotype be determined by a mixed-effects multinomial logistic regression. The regression model incorporates the influence of covariates as fixed effects. Two random effects components are also included in the model. The first describes the intercepts of each of the K states, which determine how typical each state is across the population as a whole. The second describes animal-specific error terms for each state, which capture how each animal deviates from the predicted phenotypes based on its covariates and the population average. This mixed effect formulation allows the model to refine the predicted phenotypes for each individual animal by pooling information across the entire population, as represented in the data [21].

In addition to the random effects terms described above, the model can also incorporate a third random effects component based on groupings or relationships among animals. Here we use this term to incorporate genetic effects into the model so that the heritability of phenotypes can be calculated as in the popular "animal model" in behavioral ecology [22].

A major constraint in the model is that during a ten-minute focal sample, the focal animal is confined to only one behavioral state. This constraint is equivalent to assuming that the rate of switching between behavioral states is low enough that the probability of changing states during a consecutive ten-minute period is effectively zero. Though this constraint is not particularly realistic, attempting to infer state shifts within focal samples would dramatically increase the computational complexity of the model. Furthermore, because many individual behaviors are quite sparse—occurring on average at a rate of less than once per focal observation—state transitions within focal samples are likely to be very difficult to detect. We therefore justify this simplifying assumption for the computational efficiency and scalability it permits, though it can be relaxed in the future or with different data.

2.5. Data Processing and Likelihood

To construct the input to the model, we calculated the total amount of each behavior present in each observation, and converted those amounts into ordered levels. The exact procedure is described below:

1. Construct a data matrix with a row for each focal observation, and a column for each behavior in the ethogram.
2. For each focal observation:
   (a) For each "event" behavior, count the number of times that behavior occurred during the observation.
   (b) For each "activity" behavior, calculate the total proportion of the focal observation spent engaged in that behavior.
   (c) Populate the associated row in the data matrix with these values.
3. For each behavior:
   (a) Calculate quintiles, e.g., 20th percentile, 40th percentile, etc., of the values in that behavior's associated column in the data matrix.
   (b) Also calculate the 1st and 99th percentile of the behavior to make high and low outliers.
   (c) Bin using the quantiles calculated above as cutpoints, e.g., values $\leq$ 1st percentile being 1, $>$ 1st and $\leq$ 20th percentiles being 2, etc.

This procedure divides each behavior into up to seven levels, with 1 being the smallest amount of a behavior and 7 being the most. However, in practice many behaviors occurred so infrequently that the bottom 50% or more of the data were all zeros. The large number of zeros in most behaviors can be seen in Supplementary Figure S1. This meant that 19 of 30 behaviors were split into three levels representing 0, 1, and >1 occurrences, and no behavior was divided into more than 6 levels.

A behavioral state consists of a set of categorical distributions, one for each behavior. Each categorical distribution has as many parameters as the behavior has levels that determine the probability of each level of the behavior occurring under that state. This allows each behavioral state to be extremely flexible; a state need not specify a narrow range of values or a specific distributional shape for every behavior. A state may be associated with a very specific amount of one behavior, but also consistent with a wide range of amounts of a different behavior.

The categorical distribution we used as the data likelihood can be easily changed to any distribution with a conjugate prior. For example, one might use the Poisson distribution for event behaviors and the zero-truncated normal distribution for activity behaviors. We chose the categorical distribution here because, as can be seen in Figure S1, many behaviors were zero-inflated with long right tails, which cannot be captured by Poisson and normal distributions.

2.6. Mathematical Description of the Model

The content of a focal observation is represented by $y^{(i,f)}$, a vector of length B, where B is the number of individual behaviors considered in the model. Element $y_b^{(i,f)}$ is the amount of behavior b that occurred in that focal observation. Each focal observation in the data belongs to a single behavioral state. Formally, this means that each observation f of animal i is associated with a latent variable $z_{i,f}$ which can take on values 1 through K, where K is the number of behavioral states in the model. This value denotes which of the K behavioral states to which the focal observation belongs. Given a value for behavioral state $z_{i,f}$ we can write the data likelihood of the focal observation using the parameters defining the behavioral state:

$$\mathrm{p}(y^{(i,f)}|z_{i,f} = k, \theta^{(k)}) = \prod_{b \in 1:B} \theta^{(k,b)}_{y_b^{(i,f)}} \tag{1}$$

Here $\theta^{(k)}$ is the set of all parameters associated with state k, $\theta^{(k,b)}$ is the vector of probabilities for different levels of behavior b in state k, and $\theta_l^{(k,b)}$ is the probability that behavior b occurs at level l in state k. Note that this data likelihood is simply a product of categorical likelihoods, one for each behavior. Because each focal observation is statistically independent conditional on the state assignments, we can write the complete likelihood for animal i as simply the product of the likelihoods of the focal observations:

$$\mathrm{p}(y^{(i)}|z_i, \theta) = \prod_{f \in 1:n_i} \mathrm{p}(y^{(i,f)}|z_{i,f}, \theta^{(z_{i,f})}) \tag{2}$$

Here n_i is the number of focal observations for animal i.

We now turn our attention from the likelihood to the phenotype, which is the prior probability of a focal observation f of animal i belonging to each of the k behavioral states. We can write this prior probability as,

$$\begin{aligned} \mathrm{p}(z_i|\pi_i) &= \prod_{k \in 1:K} \pi_{i,k}^{n_{i,k}} \\ n_{i,k} &= \sum_{f \in 1:n_i} \mathbf{1}(z_{i,f} = k) \end{aligned} \tag{3}$$

Here $\pi_{i,k}$ is the prior probability that animal i finds itself in state k and $n_{i,k}$ is the total number of focal observations of animal i belonging to state k. The probabilities $\pi_{i,k}$ are themselves determined by via multinomial logistic regression:

$$\begin{aligned} \pi_{i,k} &= \frac{\exp(\eta_{i,k})}{\sum_{k \in 1:K} \exp(\eta_{i,k})} \\ \eta_{i,k} &= \alpha_k + X_i^T \beta_k + u_{i,k} + \epsilon_{i,k} \end{aligned} \tag{4}$$

Here η_i is a vector of unbounded propensities for animal i to fall into each behavioral state, which are transformed by the softmax function into the probability distribution over states, π_i. X_i is a vector of animal-specific covariates, with the state-specific fixed effect regression coefficient vectors β_k. The quantities α, u, and ϵ are random effects terms representing baseline state propensities, genetic effects, and individual animal effects, respectively. These random effects are given normal distributions with the variances as free parameters to be estimated:

$$\begin{aligned} \epsilon_{i,k} &\sim \mathrm{N}(0, \sigma_k^2) \\ \alpha_k &\sim \mathrm{N}(0, \tau\sigma_k^2) \\ u_{\cdot,k} &\sim \mathrm{N}(0, \gamma_k \sigma_k^2 A) \end{aligned} \tag{5}$$

The parameters σ_k^2 determine how much variability exists across animals in the propensity for state k (outside of variability accounted for by the covariates X_i). The parameter τ controls the extent to which the states themselves vary in their average propensities across animals. That is to say, in a model where the σ_k^2 are large and τ is small, no state will consistently have high or low probability across the population, whereas when σ_k^2 are small and τ is large, states will have similar probabilities across animals but some states will be consistently high probability and others low. Finally, the covariance matrix A is the relatedness matrix of the animals in the study population. The parameters γ_k determine how much variation in each state's propensities is accounted for by genetic effects. Further, note that the variance component parameters τ and γ_k are being scaled by the "global" variances σ_k^2. This causes the linear regression in Equation (4) to be fully conjugated, such that the sampler can be "partially collapse" during inference by integrating out the components of the linear regression [23]. See [24] for an explanation of partially collapsed samples in mixed effects regression.

As we are using a fully Bayesian approach, we must specify priors for the free parameters. We use standard conjugate uninformative or weakly informative priors in all cases:

$$\begin{aligned} \theta^{(k,b)} &\sim \mathrm{Dirichlet}(1) \\ \beta_k &\sim \mathrm{N}(0, 1) \\ \sigma_k^2 &\sim \mathrm{InvGamma}(0.005, 0.0005) \\ \tau &\sim \mathrm{InvGamma}(0.005, 0.0005) \\ \gamma_k &\sim \mathrm{InvGamma}(0.165, 0.0165) \end{aligned} \tag{6}$$

See [25] for a discussion on weakly informative priors, though we do not use their exact priors. Note that the seemingly weakly informative prior on the γ_k parameters was chosen in order to achieve an uninformative (high variance) prior on heritability, which in this model is approximately $\gamma_k/(1+\gamma_k)$. An uninformative prior such as $\gamma_k \sim \mathrm{InvGamma}(0.005, 0.0005)$ would actually place high prior probability mass on values of $\gamma_k/(1+\gamma_k)$ near 1. The chosen prior is roughly symmetric on $\gamma_k/(1+\gamma_k)$ and places much of the probability mass near both 0 and 1.

2.7. Model Fitting

We fit the model using a custom Gibbs sampler implemented in the Julia technical computing language v0.5.0 [26]. As standard logistic regression representation is non-conjugate and therefore cannot be sampled from using Gibbs, we use the fully conjugate latent variable formulation of logistic regression described in [27] and previously applied to topic models in [9]. In the results reported below we ran two chains of 100,000 samples each, which were thinned to 1000 samples, with the first 100 of those discarded as burn-in.

Our model involves unsupervised classification with several hundred parameters, which means the posterior is likely to be highly multimodal. Gibbs sampling with naively chosen random starting points can often get stuck around suboptimal and idiosyncratic local maxima, which leads to both poor inference and results that are unpredictable and unreproducible between runs. Tests with random starting points indicated that even in simple synthetic data sets, where observations fall into distant, non-overlapping clusters, Gibbs sampling alone very frequently yielded incorrect clustering whenever more than three or four behavioral states were used. In many applications of topic modeling, reproducibility and interpretability of model outputs are of secondary concern to pure predictive performance, but for scientific inference they are paramount.

In order to improve the quality and reliability of inference we first fit a simpler version of the model which we then used to generate starting points for the full model. Specifically, we fit a "flattened" version of the full model which is equivalent to assuming that all observations came from the same animal, thus discarding all information about differences between animals and population level covariates. This model is in effect a naive Bayes classifier, where each possible classification is analogous to a behavioral state. This flat model was fit by maximum likelihood (ML) using the Stan software package v2.14.0 [28]. The initial points for this optimization step were generated by first taking the ML solution for a model with a single behavioral state, then adding independent Gaussian noise to each parameter to generate starting points for each behavioral state. Multiple fits with independent initializations were generated and the fit with the highest posterior density was used to initialize the Gibbs sampler. Specifically, for each focal observation, an assignment to a behavioral state was randomly drawn from the posterior distribution of behavioral state memberships under the ML fit, and this mapping from observations to behavioral state was used as the starting point for the Gibbs sampler. We found that in practice this procedure very reliably recovered the true parameters when applied to simulated data and provided consistent results using real data.

In order to choose an appropriate number of behavioral states, *K*, we calculated the widely-applicable information criterion (WAIC) [29] for models (without a genetic component) with 5, 10, 15, and 20 states, and used the number of states associated with the lowest WAIC. See Supplementary Figure S2 for the WAIC results.

2.8. Repeatability Analysis

Seventy-seven animals from social group F were observed for two consecutive years (2012 and 2013), and we used these animals to assess the stability of the phenotypes discovered by our model. To accomplish this objective, we ran a separate model in which all animals from the previous analysis were included, but animals with dual observation years were permitted to have independent phenotypes for each year. We then examined the correlation between the posterior means of the phenotypes in 2012 and 2013. No heritability component was included in this model. Note that this version of the model was not used for any analysis other than heritability, as the artificial inflation of the number of animals might lead to unwarranted confidence in population-level inferences.

2.9. Simulations

To verify that the behavioral state model could correctly recover both state contents and individual phenotypes, we tested the model on synthetic data in series of five simulations using 5, 10, 20, 40, and 80 states. For each simulation, we simulate 200 individuals with 100 focal observations each, using an ethogram consisting of 30 behaviors. Each behavior was represented as a categorical distribution with 3 levels. States were generated by sampling from the following prior:

$$\begin{aligned} \lambda_l^{(k,b)} &\sim \mathrm{N}(0, 1.0) \\ \theta_l^{(k,b)} &= \frac{\exp(\lambda_l^{(k,b)})}{\sum_{l' \in 1:3} \exp(\lambda_{l'}^{(k,b)})} \end{aligned} \tag{7}$$

Similarly, individual phenotypes were generated by sampling from the following distribution:

$$\begin{aligned} \eta_{i,k} &\sim \mathrm{N}(\alpha_k, 0.0625) \\ \alpha_k &\sim \mathrm{N}(0, 0.25) \end{aligned} \tag{8}$$

These values are converted into individual probability distributions over states. Each individual's probabilities are then used to sample a state membership for each of that individual's 100 focal observations.

Because states have no intrinsic ordering, some method is required for associating states in the simulated data with a matching state in the model output before it is possible to determine whether a state or phenotype has been accurately estimated. To accomplish this, we used a greedy matching algorithm with the folowing steps:

1. Pick an output state k' and calculate the posterior mean for each of its parameters, $\hat{\theta}^{(k')}$.
2. For each simulated state k that is not already matched with an output state, calculate the correlation between $\theta^{(k)}$ and $\hat{\theta}^{(k')}$.
3. Pick the simulated state with the highest correlation as the match for the output state k'.
4. Repeat for each k'.

After the matching procedure, we assessed how well the model recovered individual phenotypes by calculating for each individual the correlation between that individual's simulated probability of being in each state and the probabilities of being in the matching states of under the fitted model. We also assessed our ability to recover the true number of behavioral states in the 5-state simulation by comparing WAIC scores of the model with 5 states to models with 3, 4, 6, and 7 states fitted to the same data.

The model was fit to the simulated data using the same fitting procedure as used with the Cayo data. However, no regression coefficients or heritabilities were calculated.

2.10. Comparisons with Factor Analysis

We fit factor analysis models to the Cayo Santiago data set to compare with the behavioral state model presented here. As factor analysis and related models have no explicit hierarchical structure, they cannot separate observation-level relationships among behaviors from organism-level relationships; they can either examine how behaviors co-occur within observations, or how average rates of behaviors can co-occur between macaques, but not both simultaneously. Therefore we fit two models. The factor model 1 used the focal observations themselves as independent data points. For factor model 2, we calculated average rates of each behavior for each macaque and used macaques as independent data points. To set the number of factors used, we used 5-fold cross-validation to choose amongst 5, 10, 15, and 20 factors. For both versions 1 and 2 of the model, models with 10 factors or more performed similarly, so we used 10 factors for ease of comparison with the state model. We defined the phenotypes infered by the factor models as the factor scores of the individual animals. In the case of factor model 1, scores are associated with individual focal observations rather than individual animals, so the phenotypes were the average of the scores across each macaque's focal observations.

To estimate heritability of phenotypes under both models, we fit an animal model to the inferred phenotypes. These models used the same covariates as the state model.

To calculate conditional means for rates of behaviors under factor model 1, we sampled focal observations from a multivariate normal distribution with the means and covariance matrix fit by the factor model. Since the true data is discrete rather than continuous, for each of the sampled observations we rounded the behavior amounts to the nearest whole number occurring in the true data.

All factor analyses were carried out using the (factanal) function in R 3.4.0 using "regression" scores. [30]. Animal models were written and fit using the Stan software package [28].

2.11. Assessing Genetic and Covariate Influences on Social Phenotypes

In multinomial logistic regression, regression coefficients associated with a state are only interpretable relative to a baseline state [31]. This is due to the fact that a probability distribution over K states has only $K - 1$ free variables, as the probability of a single state is determined completely by that of the remaining states. If a baseline state is not specified, parameters will be unidentifiable with respect to the data and will be constrained only by the prior. Regression coefficients and genetic effects, with the coefficients and genetic effects of the baseline state subtracted, therefore reflect an influence on the relative probabilities of a each state occurring versus the baseline state.

A related issue is that the state probabilities, here the phenotypes of interest, are a nonlinear function of the underlying model parameters and covariates. The impact on the phenotypes of any given component of the model depends on the value of every other covariate and parameter in the model. Due to this nonlinearity, and because values depend on the baseline selected, parameter estimates themselves can be very difficult to interpret. However, an advantage of our Bayesian approach is that we can easily derive estimates and central credible intervals (CIs) for the influence of specific components of the model on the phenotypes of interest. We accomplished this by generating predicted phenotypes from posterior samples under different assumptions to assess the contributions of different components of the model. To assess the impact of a specific covariate on phenotypes, we generated predicted state probabilities at varying levels of that covariate while holding every other covariate fixed at its population average value (or, in the case of discrete covariates, the modal level). This process was repeated for every posterior sample of the model parameters, yielding a full posterior distribution of predicted phenotypes.

To assess which, if any, behavioral states were strongly influenced by genetic effects, we calculated a pseudo-h^2 metric to describe the amount of variance in state probabilities that was explained by genetic effects and covariates combined relative to the variance captured by covariates alone. Formally, for state k the total variance in state probabilities is $\mathrm{var}(\pi_{\cdot,k})$, where $\pi_{\cdot,k}$ is the vector of state probabilities as defined above (Equation (4)). We can also define partial variances based on state probabilities estimated when certain components of the model are omitted:

$$\begin{aligned} \hat{\pi}_{i,k}^{(B)} &= \frac{\exp(\hat{\eta}_{i,k}^{(B)})}{\sum_{k \in 1:K} \exp(\hat{\eta}_{i,k}^{(B)})} \\ \hat{\eta}_{i,k}^{(B)} &= \alpha_k + X_i^T \beta_k \end{aligned} \tag{9}$$

$$\begin{aligned} \hat{\pi}_{i,k}^{(u)} &= \frac{\exp(\hat{\eta}_{i,k}^{(u)})}{\sum_{k' \in 1:K} \exp(\hat{\eta}_{i,k'}^{(u)})} \\ \hat{\eta}_{i,k}^{(u)} &= \alpha_k + X_i^T \beta_k + u_{i,k} \end{aligned} \tag{10}$$

We then define pseudo-h^2 for behavioral state k as,

$$1 - \frac{\mathrm{var}(\pi_{\cdot,k} - \hat{\pi}_{\cdot,k}^{(u)})}{\mathrm{var}(\pi_{\cdot,k} - \hat{\pi}_{\cdot,k}^{(B)})} \tag{11}$$

This is the proportion of variance explained by adding genetic effects back into a model, out of the residual variance left over by a model with covariates alone. This is both closely related to a partial R^2, and to the traditional h^2 measure used in animal models, where the latter is commonly defined as the proportion of variance explained by genetic effects out of the residual variance after the effects of covariates are removed [32]. As with covariates, we calculated pseudo-h^2 for every posterior sample. Finally, we note that unlike standard h^2 and R^2, pseudo-h^2 can be negative, indicating that including genetic components reduces predictive accuracy.

3. Results

3.1. Simulation Tests

We first verified that our model could accurately recover both the contents of behavioral states and phenotypes of individual animals by fitting the model to simulated data sets and comparing the model outputs to the ground truth. Figure 1 shows that for data sets of similar size as the Cayo Santiago data, the model accurately recovered all behavioral states so long as the number of states did not exceed 20. In each simulated data set with up to 20 states, every simulated state had a unique state in the fit model for which their state parameters $\theta^{(k)}$ had a correlation coefficient above 0.95. When the number of states exceeded 20, the majority of simulated states had a close match in the model's estimated states, but several simulated states had no match with correlations above 0.5. This suggests that the model was unable to find some behavioral states.

Similarly, the ability of the model to reliably recover phenotypes declined as the number of states increased. With 5 and 10 states, the correlation between fitted and simulated phenotypes was above 0.75 for most individuals and above 0.5 for nearly all of them, while in the data sets with 40 and 80 states most individuals had correlations below 0.5. This is not surprising, as larger numbers of states means more parameters must be estimated for each individual using the same number of observations.

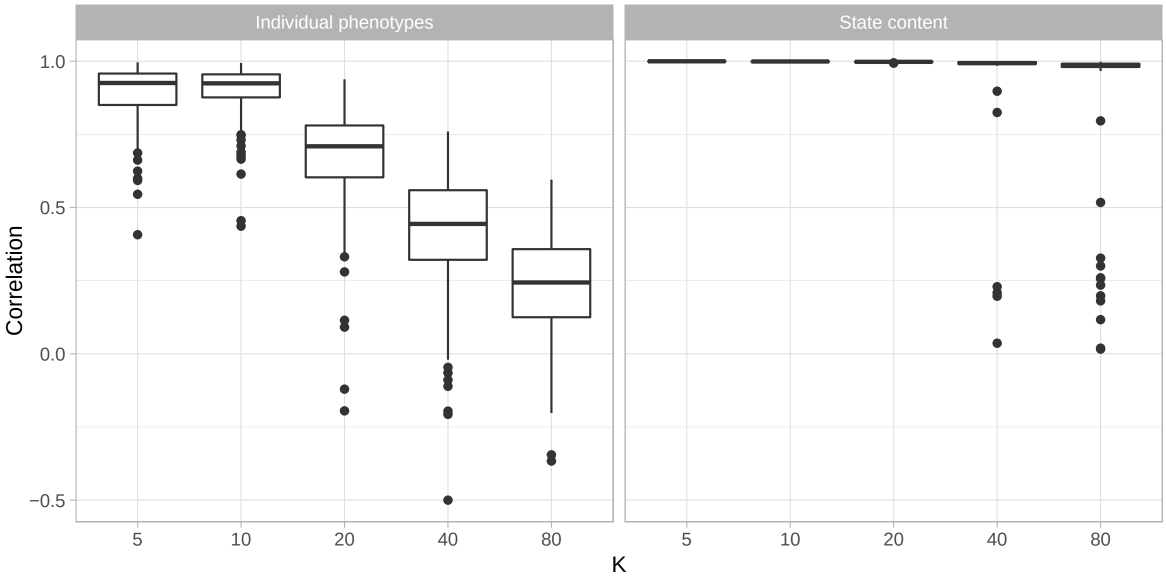

Figure 1. Correlations between simulated and fit phenotypes (**left**) and state contents (**right**). Both panels show boxplots, though for the state contents the values are concentrated enough that the hinges of the plots are not distinguishable. For phenotypes, each data point is an individual, while for states, each data point is a state.

We also found that for the data set with five states were were able to identify the model with five behavioral states as the optimal model using WAIC goodness-of-fit metric. See Supplementary Figure S3 for the WAIC results.

3.2. Phenotype Distributions and Behavioral State Content

We fit the model allowing ten behavioral states (referred to hereon as S1–S10). Figure 2 shows the distribution of phenotypes across the population, in the form of the rate at which each animal exhibited each behavioral state. We ordered the states in terms of how frequently each state occurred in the population, with S1 having the highest average rate of occurrence and S10 having the lowest. S1–S3 showed both high overall rates, with animals spending on average 55% of their focal observations in these states, as well as high variability in their rates across animals, with substantial numbers of animals spending more than 30% of their focal observations in a single one of these three states. S4–S9, on the other hand, all had median rates falling in a narrow range between 7% and 5.6%, with only one animal displaying a rate greater than 20% in any one of those states. Finally, S10 has a median rate of only 1.4%.

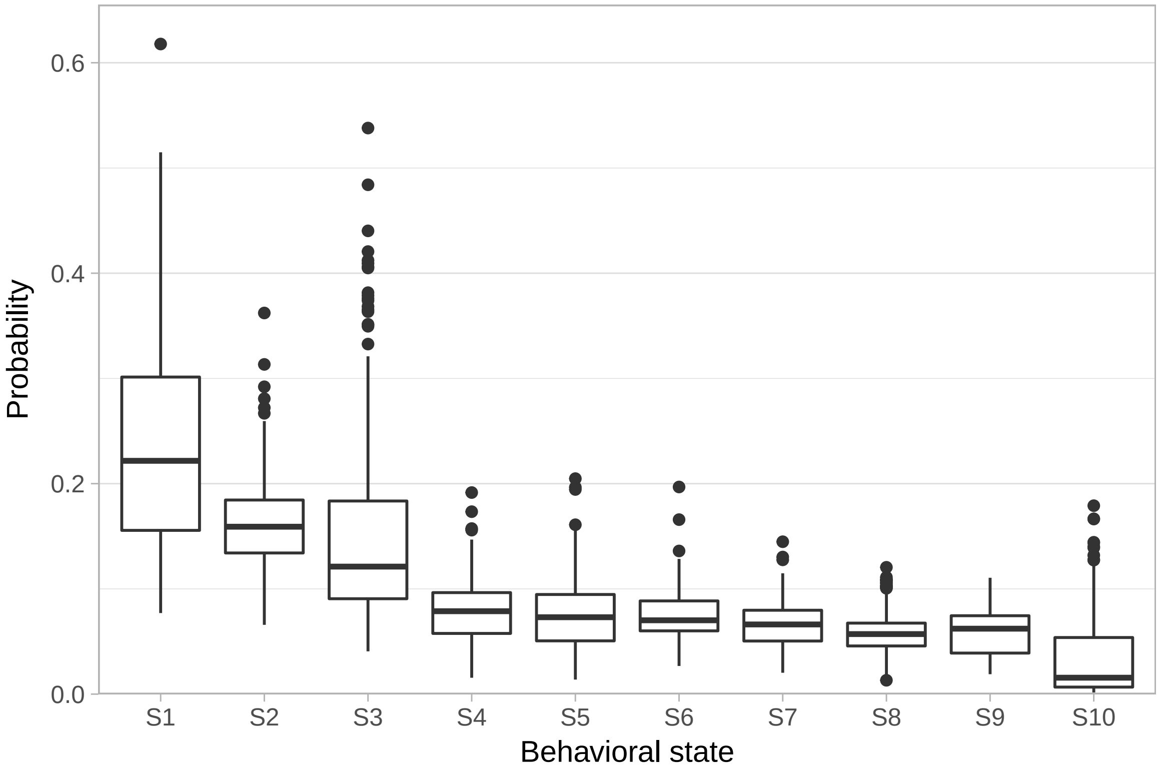

Figure 2. The distribution of phenotypes of the studied population. The box-and-whisker plots display the distribution of posterior mean probabilities of being in each state across all studied individuals. States are ordered by descending mean probability across the population. "Hinges" of the boxes represent 25%, 50%, and 75% quartiles.

Next we examined the contents of these behavioral states. Figure 3 visualizes the typical behaviors of each state in terms of their relative frequency, with very rare behaviors omitted for legibility. Such "at a glance" summaries distinguish particular features of each behavioral state. Immediately apparent is the distinction between the more frequent S1–S3 and the infrequent S4–S9. S1–S3 paint a rather prosaic portrait of macaque life, consisting largely of self-directed behaviors, such as scratching oneself, eating, and walking. Indeed, S2 entailed doing little other than resting. The social interactions that did occur in these three states were primarily agonistic. In particular, both overt non-contact aggression and less overt agonistic actions, such as fear grimaces and avoidances, were quite common in S1. The middle infrequent S4–S9 on the other hand, displayed mixtures in varying proportions of agonistic and affiliative actions. Incidental affiliative behaviors such as approaching other animals occurred commonly throughout all of these infrequent states, while grooming, a more significant sign of affiliation, appeared concentrated in S4, S5, and S8.

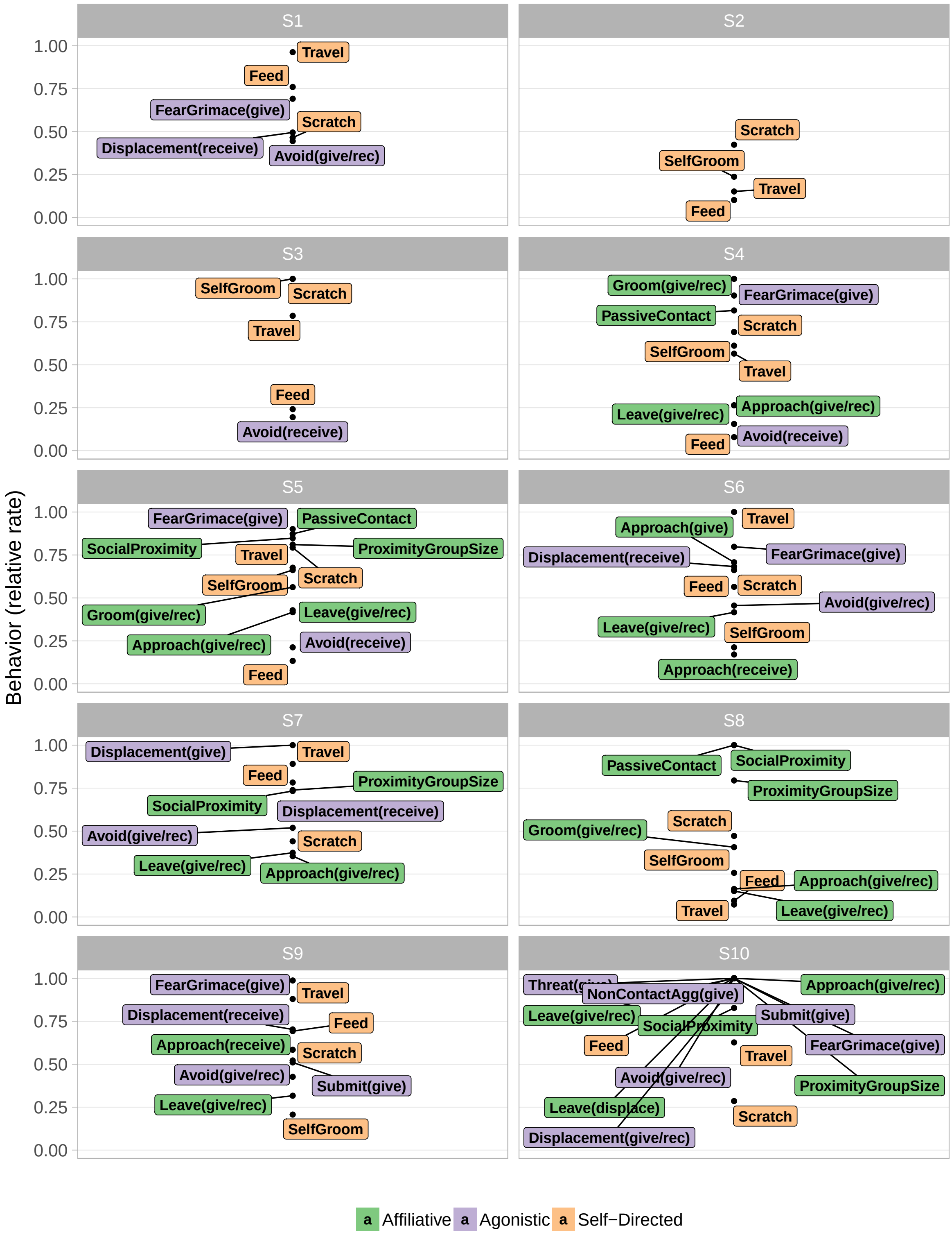

Figure 3. The content of behavioral states fit by the model. Relative rates for each behavior are calculated by dividing the posterior mean rates across states by the mean rate of the highest state, such that 1 represents the highest mean rate across states. For visual clarity, behaviors with relative rates below 0.05 are omitted, and behaviors for which the difference between the “give” and “receive” variants is less than 0.33 are concatenated into a single label (give/rec).

3.3. Repeatability

As our goal is to identify phenotypes with underlying biological bases, it is important to show that phenotypes produced from the model are consistent within animals and relatively stable across time. To that end we compared estimated phenotypes from animals in group F from years 2012 and 2013. Phenotypes in 2012 were strongly correlated with phenotypes in 2013 for all behavioral states, with the lowest correlation coefficient being 0.64 for S1, and the largest being 0.88 for S8 ($p < 0.001$ for both).

3.4. Group, Rank, and Sex Effects on Social Phenotypes

While describing the content of behavioral states is useful, scientists are often more interested in identifying sources of behavioral variability. In rhesus macaques and other NHPs, it is important to understand how much variability in social phenotypes across animals can be predicted on the basis of demographic covariates, and how much is idiosyncratic to each individual. In the present model we included age, sex, dominance rank, social group, and age-by-sex and rank-by-sex as population-level predictors of phenotypes. Dominance rank was represented on an ordered categorical scale: low-ranking animals outranked less than 50% of their social group, medium-ranking animals outranked between 50% and 80%; and high-ranking animals outranked greater than 80%. Ranks were available annually, so for animals with multiple years of observations that changed ranks, their average rank was used. Figure 4 displays the regression coefficients for each covariate's influence on the probability of being in each behavioral state. Here we choose S2 as baseline as it provides a neutral default state with little in the way of positive or negative social interactions.

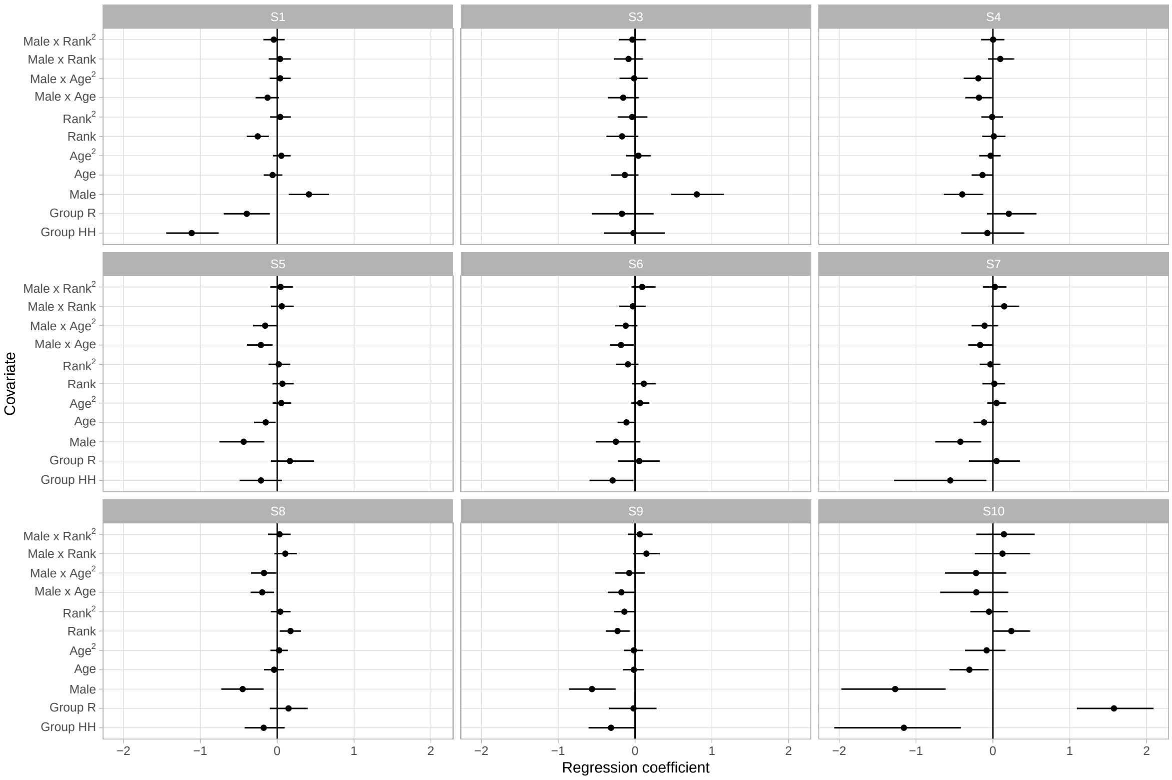

Figure 4. Regression coefficients with S2 as a baseline (see Methods). Points represent posterior means and lines, 95% central credible intervals.

The posterior distributions of the regression coefficients indicate that social group membership and sex had large influences on social phenotypes. Animals in groups R and HH had, for instance, lower relative rates of being in S1 versus S2 than animals in the largest social group, F, with regression coefficients of −0.52 ([−0.78, −0.27] 95% CI) and −0.88 ([−1.20,−0.58] 95% CI), respectively. As S1 had much higher rates of agonistic behavior than the unsocial S2 and typically involved very little affiliative behavior, this suggests that animals in group F were more likely to become involved in agonistic interactions. Sex and linear dominance rank were the most common predictors of relative behavioral state occurrence rates, with four and five of the nine non-baseline states respectively showing significant effects. These states overlap, with S1, S3, and S6 being more frequent in male macaques and in lower ranking monkeys, which we note were states with higher rates of agonistic behaviors and relatively lower rates of affiliative behaviors.

Though examining the posterior of regression coefficients predicting relative odds of states can be useful, it is often more intuitive to directly examine the influence of covariates on the absolute rates at which an animal experiences behavioral states. We used this method to examine how sex, group membership, and dominance rank influenced phenotype. Figure 5 displays the influence of social group membership on the social phenotype of male and female animals of median age and rank. This representation of the model demonstrates that males experienced S1 and S3 more frequently than females, at the expense of slightly lower rates of all of the less frequent, more social S4–S10. S1 entailed giving and receiving aggression while moving and feeding, while S3 consisted of self-directed grooming and scratching with some amount of agonistic actions that may have served to defuse aggression. Both featured low rates of affiliative social behavior. In contrast, the states males experienced less often contained higher relative rates of affiliative behaviors and peaceful sharing of space with conspecifics. This implies that males spent more of their time avoiding confrontations and competing for resources and less time experiencing social support.

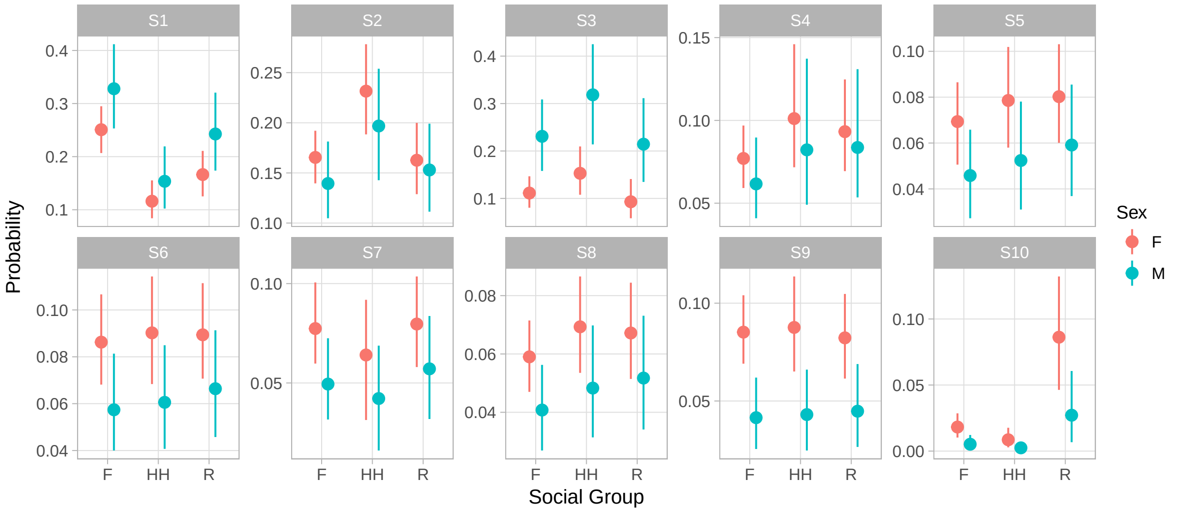

Figure 5. Expected probabilities of being in a state across different social groups for animals of average age and rank. Points represent posterior means and error bars, 95% central credible intervals of the expected probability.

It is worth noting that males experienced S6 at lower rates than females did, despite the fact that the S6 regression coefficient for ‘male’ has a positive posterior mean and little posterior mass below zero (0.13 posterior mean, [−0.03, 0.45] 95% CI). This illustrates the value of examining absolute state probabilities rather than regression coefficients themselves. In other words, because males spent so many focal observations in S1 and S3, the positive coefficient for males in S6 reduced the extent to which S1 and S3 crowded out S6 for males, but did not reverse the gap between sexes.

Finally, we note that S10, the least frequent of states overall, showed a dramatic difference between groups. Monkeys in group R experienced this state far more often than members of either F or HH. As before, this effect was not apparent from the raw regression coefficients, as the coefficients for

HH and R membership were similar in absolute value, with posterior means of −0.91 ([−1.80,−0.14] 95% CI) and 1.45 ([0.96, 1.95] 95% CI) respectively. S10 was characterized by very high rates of feeding while maintaining close proximity to large numbers of other animals. Focal animals in S10 also engaged in high rates of agonistic behaviors such as threats and non-contact aggression, suggesting this state captured episodes of food competition within or between groups.

Figure 6 shows the influence of dominance rank on the phenotype of an animal of median age in group F. Here we observed an overall pattern across behavioral states that was similar to the effect of gender, in that S1 and S3 were overall more frequent among low ranking animals than high, while S4–S10 were more frequent in high ranking animals. We also observed strong rank-by-sex interactions in S3 and S9, whereby higher rank had little effect on the rate of being in S3 for females but greatly decreased that rate for males, while the opposite was true for S9.

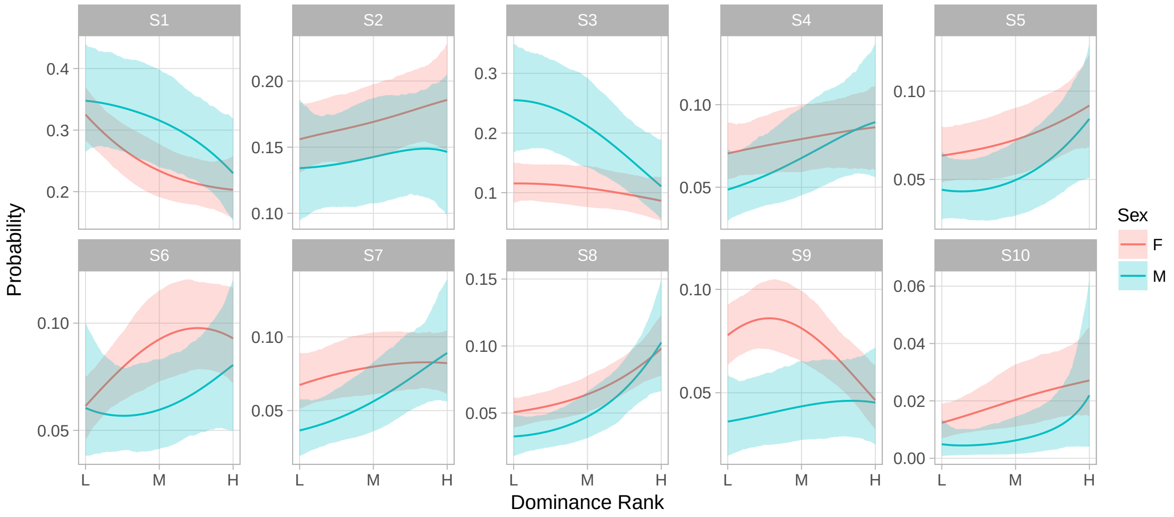

Figure 6. Relationship between dominance rank and the expected probability of being in a state for animals of average age in group F. Curves represent posterior means and shaded regions, 95% central credible intervals of the value of the curve at the corresponding rank.

3.5. Genetic Components of Social Phenotypes

Finally, we sought to clarify which aspects of social phenotypes were strongly influenced by genetic factors. To quantify how much variability in the probability of being in each behavioral state could be accounted for by genetic factors, beyond the variability explained by demographic covariates alone, we calculated pseudo-h^2 as described above for each behavioral state and displayed the posterior distributions in Figure 7. For most states, we found very wide 95% credible intervals for the variability explained with the lower bounds of eight of the ten states falling below 10%. For these states, we could not effectively rule out negligibly small contributions of genetics in determining behavioral state probabilities. However, S5 and S6 showed reliably high pseudo-h^2 of 0.69 ([0.43, 0.80] 95% CI) and 0.6 ([0.24, 0.76] 95% CI), respectively.

S5 was associated with high levels of affiliative behavior, involving high rates of friendly vocalizations, grooming, and passive physical contact, as well as being in close proximity to a high number of conspecifics for long periods of time (Proximity Group Size and Social Proximity, respectively, see Appendix for definitions). S5 also showed low rates of most agonistic interactions. Conversely, S6 was characterized by a more diverse selection of behaviors, entailing high rates of less confrontational forms of agonistic behavior (e.g., giving and receiving fear grimaces and displacements) and more actively confrontational behavior (e.g., giving threats and noncontact aggression, as well as submissive gestures). Nonetheless, animals in S6 were not deterred from affiliative approaches.

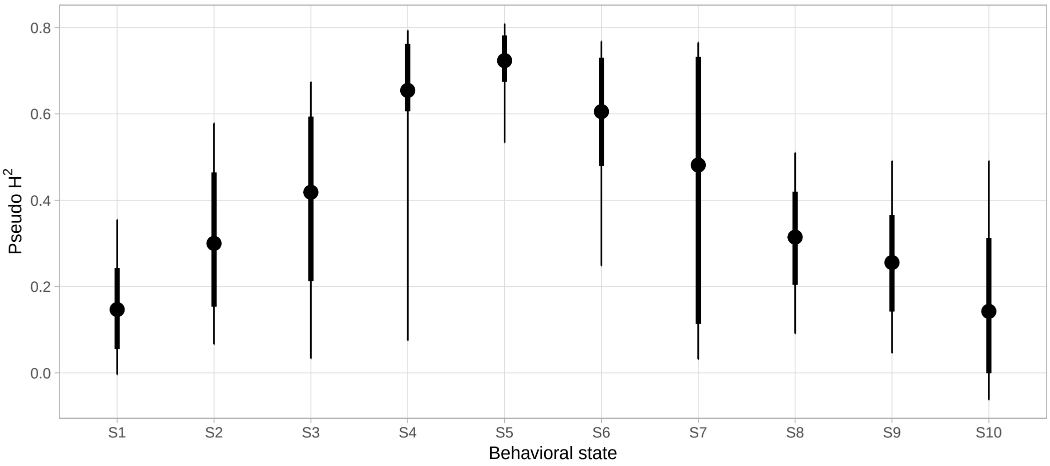

Figure 7. Genetic influences on the probability of being in a state in the studied population. Points represent posterior means, thick error bars, 66% central credible interval (roughly equivalent to 1 standard error), and thin error bars, 95% central credible intervals. See Methods for definition of the pseudo-h2 measure.

3.6. Comparison with Factor Analyis

One advantage of the state model over factor analysis and related methods is that it can capture a wider range of relationships among behaviors. Figure 8a gives an example of this using the behaviors Travel, Feed, and Noncontact Aggression (received). The relationship between traveling and feeding is nonlinear, with the rate of traveling increasing with the rate of feeding for low levels of feeding, but sharply decreasing at very high rates of feeding. Moreover, this relationship is moderated by noncontact aggression received. During focal observations when the animal is not feeding, animals travel much more when receiving aggression, but when feeding occurs received aggression has little impact on time spent traveling. The state model captures both the nonlinear relationship between travel and feeding as well as the interaction with noncontact aggression received, though it does notably underestimate how quickly time spent traveling decreases at high levels of feeding. Factor model 1, however, can only represent linear relationships [33] and so captures only the positive correlation between traveling and feeding at low levels of feeding. We did not assess factor model 2 in this way; because factor model 2 is a model of average behavior rates by individuals rather than a model of focal observations themselves, we saw no clear method for generating predictions for these observation-level relationships.

Finally, we also compared the state model to the factor models in the repeatability and heritability of the phenotypes generated by the model. As shown in Figure 8b, the phenotypes under the state model showed higher repeatability than those of either factor model. As shown in Figure 8c, the factor model yielded no phenotypes with heritability estimated at above 0.5, while the state model and factor model 1 phenotypes showed broadly similar distributions of heritabilities. However, given the high uncertainties associated with heritability estimates under all models, any conclusions about differences among models must be tentative at best.

The loadings associated with factor models 1 and 2 are shown in Supplementary Tables S1 and S2, respectively.

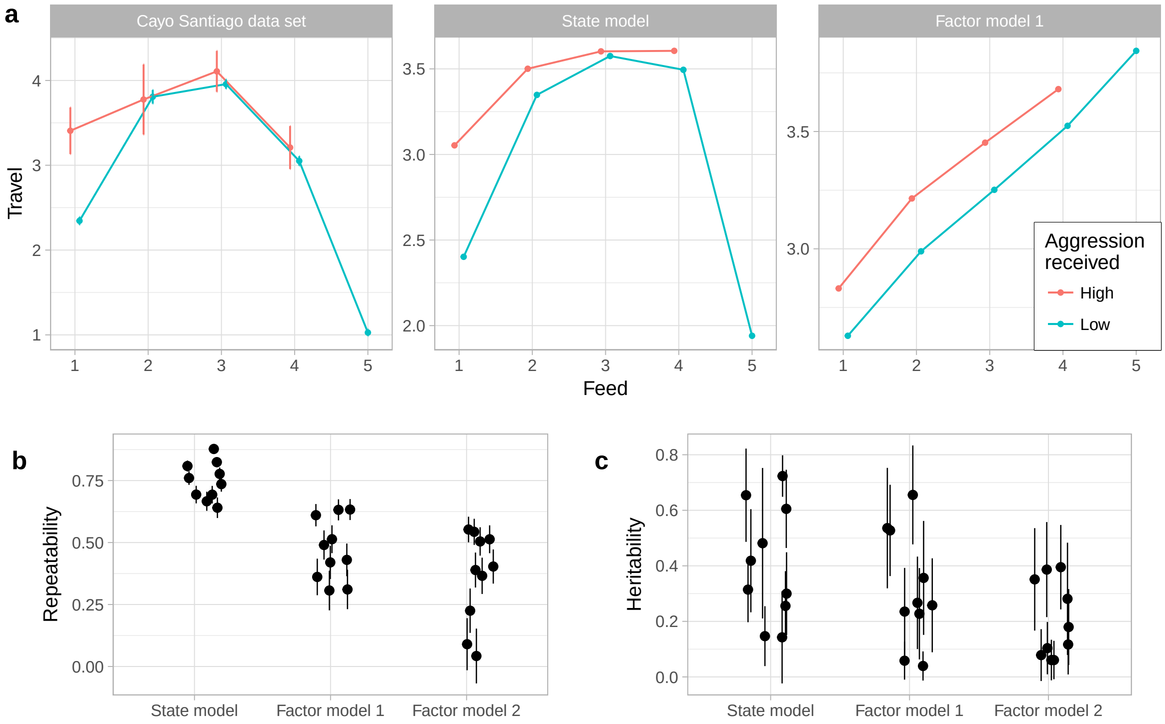

Figure 8. Comparisons between the state model and factor analysis models. (**a**) The leftmost panel shows observed means and standard errors of rates of Travel at varying levels of Feed and Noncontact Aggression (received); The right two panel shows the expected levels of Traveling under the state model and factor model 1; (**b**) Repeatability, as measured by the correlation between 2012 and 2013 phenotypes, for the three models; (**c**) Heritability estimates of phenotypes from the three models. Error bars in all panels represent one standard error.

4. Discussion

The influence of typical biological and environmental factors on social behavior is likely to be many-to-many, with any given factor impacting many different behaviors together rather than any specific behavior in isolation [34–37]. Dealing with high-dimensional, structured behavioral data will therefore be important for understanding the biological bases of behavior. The field of primatology has decades of experience collecting detailed, high-dimensional data on social behavior in species with complex, human-like social behaviors, and this data may prove invaluable for linking fundamental biology to complex social behavior. However, extracting useful structure from natural, high-dimensional and relatively noisy behavioral data sets remains a major challenge. In this paper we presented a model for identifying patterns of social behavior and relating them to environmental and genetic factors, as inspired by machine learning methods for uncovering underlying topics in corpora of documents.

The behavioral states our model discovered and their relationships with demographic covariates were consistent with our knowledge of rhesus macaque behavior generally. Macaques on average spent about 50% of their time in S1–S3, which involved little social interaction beyond receiving non-contact aggression and other agonistic encounters in S1. This pattern may reflect conflict over food or space given that state's high levels of feeding and traveling. The other 50% of the observed time was evenly distributed across the remaining states, S4–S10, which consisted of idiosyncratic combinations of agonistic and affiliative behaviors, reflecting the long-known importance of sociality in rhesus macaque life [15,19]. Among both males and females, higher dominance rank was associated with a shift away from the frequent S1 and S3, which involved mostly self-directed behavior amidst some agonistic interactions, and towards S4–S10 which had higher rates of many different social interactions overall as well as high rates of affiliative interactions. This is consistent with the many previous findings that high ranking macaques, and other primates as well, receive higher rates of affiliative behaviors from others in their social group, while low ranking individuals are more likely to experience social isolation [13,15,16,38–40]. The effect of sex on predicted phenotypes was similar to the effect of dominance rank, with males, like low-ranking animals, having higher probabilities of assignment to S1 and S3 at the expense of time spent in S4–S9 (but not S10, notably). This may reflect the inheritance of female rank in rhesus societies, with females remaining in their natal groups and acquiring a rank just below their mother. By contrast, males disperse into new social groups at adulthood and must earn their ranks there. This results in dominance hierarchies generally being more stable, and therefore social relationships more peaceful, among females than males [41,42].

Our model improves upon the commonly used factor analysis and PCA methods in that it captures information regarding the way individual behaviors co-occur. First, factor analysis and related methods implicitly assumes that data is normally distributed [33]. This assumption is strongly violated by observations of natural behaviors, which are typically represented as counts. Furthermore, many behaviors of interest are relatively infrequent (see Section 2.6 on data processing and likelihood), leading to highly skewed, zero-inflated distributions for individual behaviors. Our model uses instead a flexible discrete distribution to avoid mismatches between the assumed data distribution and the data itself (though any distribution with conjugate priors can be substituted with minimal effort). Second, and more because of the multivariate normal assumption, PCA and factor models can only represent relationships between two behaviors as correlations. The behavioral state model, being a type of mixture model, is more flexible and capable of capturing nonlinear relationships and interaction effects [43]. This means PCA cannot capture nonlinear relationships between pairs of behaviors, or relationships which are modulated by a third behavior. The seemingly quadratic relationship between time spent feeding and traveling in Cayo Santiago macaques, which itself depends on whether aggression is received, is an example of the kind of complex relationship among behaviors that that the behavioral state model is able to represent.

We are not the first to apply topic model-like methods to primate behavioral data. The popular program STRUCTURE, which was in fact developed prior to topic models, represents the genotypes of organisms as distributions over distinct genetic populations [44], just as topic models treat documents as distributions over topics and just as our current model describes the phenotype of an animal as a distribution over behavioral states. This model has been applied to distinguish Indian and Chinese rhesus macaque genotypes [45]. The current model is, to the best of our knowledge, the first attempt to apply topic models to natural animal behavior.

In this paper we used a pedigree to estimate additive genetic influences on behavioral phenotypes, as per classical heritability analyses. An alternative approach is to include individual genetic variants as predictors in the model as in genome-wide association analyses (GWAS). Our model's regression layer makes GWAS very straightforward to implement, as variant data such as minor allele count can be included directly in the model as a per-animal covariate. However, while this approach would yield much more precise information on how different genetic variants influence phenotypes, it may be difficult to implement for two reasons. First, the traditional GWAS methodology involves fitting a separate model for many common genetic variants [46]. While this is tractable when the model being fit is a standard linear regression, fitting a complex hierarchical model thousands of times is computationally infeasible. Second, the sample sizes available in observational data sets of wild or free-ranging animals reach the hundreds or low thousands at best, which are orders of magnitude too small to achieve adequate power to detect the generally small effects of common genetic variants [47–49]. However, inference on the effects of rare or de novo mutations, which can have larger effect sizes, may be more feasible [37,50,51]. Gene sets [52] or genetic risk scores, which aggregate known genetic effects on biological pathways, provide alternative approaches.

Another possibility is to use realized relatedness based on measured genetic similarity to estimate genetic influences on phenotypes. This method assumes that many genetic variants have some small effect and aggregates information across all of them rather than attempting to identify specific variants with large effects [34,53]. This assumption is highly appropriate for complex behavioral phenotypes such as social behaviors, which are known to be polygenic in humans [35,36]. While aggregating across many variants means individual variants cannot be singled out as important for a given phenotype, recent methodological advances suggest that analyses based on realized relatedness can be used to decompose overall genetic influences into the influences of specific genomic regions of interest, thus retaining some specificity and allowing insight into the roles of different genetic pathways [54–56]. Several groups have begun investigating the use of realized relatedness to examine genetic contributions to phenotypes in natural populations [57,58], though to the best of our knowledge no groups have yet examined the contributions of different genomic regions or pathways. This may present a path forward for understanding the genetics of natural behavior, when combined with models of behavior such as that developed here that effectively and efficiently aggregate information both across behaviors and animals.

5. Conclusions

As neuroscientists and biologists investigate the biological determinants of natural social behavior and attempt to translate laboratory findings into more realistic, unconstrained environments, they face the perennial challenge of quantifying natural social behavior. Because natural social behavior is high-dimensional, highly variable, and yet highly structured, straightforward measurement of its influences is difficult. We have worked to address this issue by developing a model for identifying patterns of social behavior and relating them to environmental and genetic factors. Based on recent advances in machine learning for identifying latent structure in sets of documents, this method captures a wider range of relationships among behaviors than is possible using popular methods such as factor analysis or PCA, and explicitly disentangles variability in an individual's social behavior from variability across the population. We hope that this model will aid researchers in quantifying social behaivior in a way that is rigorous and consistent, while also capturing the richness and complexity of social behavior.

Supplementary Materials: Supplementary material are available online at www.mdpi.com/2076-3425/7/7/91/s1. Figure S1: Distribution of behavior amounts across focal observations. Figure S2: Choosing number of states for the Cayo Santiago data using WAIC. Figure S3: Choosing number of states for simulated data using WAIC. Table S1: Factor loadings for factor model 1. Table S2: Factor loadings for factor model 2.

Acknowledgments: The authors would like to thank John Pearson, Sam Larson, Ashley Walker, Joel Glick, Josue Negron, and the CPRC staff for their feedback and research support. This research supported by NIH grant 5R01-MH096875-02. The CPRC is supported by grant 8-P40 OD012217-25 from the National Center for Research Resources and the Office of Research Infrastructure Programs of the National Institutes of Health.

Author Contributions: S.M., K.A.H., and M.L.P. conceived the project; L.J.N.B. provided the data; S.M., K.A.H., and L.J.N.B. designed the analyses, S.M. developed the algorithm and performed the analyses; S.M., M.J.M., and M.L.P. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CPRC	Carribean Primate Research Center
PCA	Principle components analysis
NHP	Non-human primate
CI	Central credible interval
WAIC	Widely-applicable information criterion

Appendix A

Note that all affiliative and agonistic behaviors can be either "given" or "received" depending on whether the focal animal directed the action at another or another directed the action towards the focal animal, with the exception of Passive Contact, Social Proximity, and Proximity Group Size. The "given" and "received" versions of behaviors are represented separately in the model. Further, all behaviors are measured in counts of the number of times the behavior occurs, except for Feed, Travel, Groom, and Passive Contact, which are measured in durations. For more information on the behaviors defined below, see [17].

Appendix A.1. Self-Directed Behaviors

- Scratch: Rapid and repeated movement of the nails of a hand or foot across the skin.
- Self groom: Running of hands or mouth through one's own hair for at least 5 s.
- Feed: Searching for, manipulating, holding and ingesting food items, including water, for at least 5 s.
- Travel: Movement from one location to another over a distance of at least 5 m.

Appendix A.2. Affiliative Behaviors

- Approach: One individual approaches another to within arms' reach (2 m) without physical contact, and remains within that distance for at least 5 s.
- Leave: Exiting a 2 m area around another without an agonistic interaction.
- Affiliative Vocalization (AffilVocal in figures): Emiting a friendly vocalization in the form of either a grunt, girney, vocal exchange, or lipsmack. Individuals will often emit many vocalizations in short succession.
- Groom: Running the hands or mouth through the hair of another monkey for at least 5 s.
- Passive Contact: Sitting or lying in physical contact with another animal without grooming.
- Social Proximity: Number of time points out of three at which the focal was observed to be within 2 m of at least one other animal. The time points were at the beginning, middle, and end of the focal observation.
- Proximity Group Size: Number of unique animals with whom the focal animal shared social proximity as defined above.

Appendix A.3. Agonistic Behaviors

- Threat: One individual threatens another with one or a combination of staring, barks, head bobs, and opening one's mouth with covered teeth.
- Avoid: Moving out of the way of another before they come within 2 m.
- Displacement: Similar to avoid, but within 2 m.
- Fear Grimace: Submissive facial expression wherein lips are retracted horizontally to expose teeth.
- Submit: Leaning away from another or crouching while raising hindquarters towards another.
- Noncontact Aggression: A lunge, charge, or chase that does not result in direct physical contact.
- Contact Aggression: Direct physical contact such as a bite, hit, push, or grab.

References

1. Morgan, K.N.; Tromborg, C.T.; Syaadah, O.; Norma, M.; Feldon, J.; Berckmans, D.; Hare, V.; Tepper, E.; Lindburg, D. Sources of stress in captivity. *Appl. Anim. Behav. Sci.* **2007**, *102*, 262–302.
2. Rietveld, C.A.; Medland, S.E.; Derringer, J.; Yang, J.; Esko, T.; Martin, N.G.N.W.; Westra, H.J.; Shakhbazov, K.; Abdellaoui, A.; Agrawal, A.; et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **2013**, *340*, 1467–1471.

3. Allen, H.L.; Estrada, K.; Lettre, G.; Berndt, S.I.; Weedon, M.N.; Rivadeneira, F.; Willer, C.J.; Jackson, A.U.; Vedantam, S.; Raychaudhuri, S.; et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **2010**, *467*, 832–838.
4. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; The MIT Press: Cambridge, MA, USA, 2012.
5. Brent, L.J.N.; Semple, S.; Maclarnon, A.; Ruiz-Lambides, A.; Gonzalez-Martinez, J.; Platt, M.L. Personality Traits in Rhesus Macaques (Macaca mulatta) Are Heritable but Do Not Predict Reproductive Output. *Int. J. Primatol.* **2014**, *35*, 188–209.
6. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
7. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77.
8. Lafferty, J.D.; Blei, D.M. Correlated Topic Models. In *Advances in Neural Information Processing Systems 18*; Weiss, Y., Schölkopf, P.B., Platt, J.C., Eds.; MIT Press: Cambridge, MA, USA, 2006; pp. 147–154.
9. Chen, J.; Zhu, J.; Wang, Z.; Zheng, X.; Zhang, B. Scalable Inference for Logistic-Normal Topic Models. In *Advances in Neural Information Processing Systems 26*; Burges, C.J.C., Bottou, L., Welling, M., Gharamani, Z., Weinberger, K.Q., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2013; pp. 2445–2453.
10. Rodríguez, A.; Dunson, D.B. Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.* **2011**, *6*, 145–177.
11. Seyfarth, R.M.; Silk, J.B.; Cheney, D.L. Variation in personality and fitness in wild female baboons. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 16980–16985.
12. Budaev, S.V. Using Principal Components and Factor Analysis in Animal Behaviour Research: Caveats and Guidelines. *Ethology* **2010**, *116*, 472–480.
13. Brent, L.J.N.; Heilbronner, S.R.; Horvath, J.E.; Gonzalez-Martinez, J.; Ruiz-Lambides, A.; Robinson, A.G.; Skene, J.H.P.; Platt, M.L. Genetic origins of social networks in rhesus macaques. *Sci. Rep.* **2013**, *3*, 1042.
14. Widdig, A.; Bercovitch, F.B.; Streich, W.J.; Sauermann, U.; Nürnberg, P.; Krawczak, M. A longitudinal analysis of reproductive skew in male rhesus macaques. *Proc. Biol. Sci.* **2004**, *271*, 819–826.
15. Carpenter, C.R. Characteristics of Social Behavior in Non-Human Primates. *Trans. N. Y. Acad. Sci.* **1942**, *4*, 248–258.
16. Bernstein, I.S.; Sharpe, L.G. Social Roles in a Rhesus Monkey Group. *Behaviour* **1966**, *26*, 91–104.
17. Brent, L.J.N. Investigating The Causes and Consequences of Sociality in Adult Female Rhesus Macaques Using a Social Network Approach. Ph.D. Thesis, University of Roehampton, London, UK, 2010.
18. Altmann, J. Observational Study of Behavior: Sampling Methods. *Behaviour* **1974**, *49*, 227–267.
19. Rawlins, R.G.; Kessler, M.J. *The Cayo Santiago Macaques: History, Behavior, and Biology*; State University of New York Press: Albany, NY, USA, 1986; p. 306.
20. Sinnwell, J.P.; Therneau, T.M.; Schaid, D.J. The kinship2 R package for pedigree data. *Hum. Hered.* **2014**, *78*, 91–93.
21. Gelman, A.; Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*; Analytical Methods for Social Research; Cambridge University Press: New York, NY, 2007.
22. Kruuk, L.E.B. Estimating genetic parameters in natural populations using the "animal model". *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2004**, *359*, 873–890.
23. Van Dyk, D.A.; Park, T. Partially Collapsed Gibbs Samplers: Theory and Methods. *J. Am. Stat. Assoc.* **2008**, *103*, 790–796.
24. Park, T.; Min, S. Partially Collapsed Gibbs Sampling for Linear Mixed-effects Models. *Commun. Stat. Simul. Comput.* **2014**, *45*, 165–180.
25. Gelman, A.; Jakulin, A.; Pittau, M.G.; Su, Y.S. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2008**, *2*, 1360–1383.
26. Bezanson, J.; Edelman, A.; Karpinski, S.; Shah, V.B. Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.* **2014**, *59*, 65–98.
27. Polson, N.G.; Scott, J.G.; Windle, J. Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *J. Am. Stat. Assoc.* **2013**, *108*, 1339–1349.
28. Stan Development Team. RStan: The R interface to Stan, R package version 2.14.1, 2016.
29. Vehtari, A.; Gelman, A.; Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **2017**, *27*, 1413–1432.
30. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.

31. Agresti, A. *Categorical Data Analysis*, 2nd ed.; Wiley: Hoboken, NY, USA, 2002.
32. Wilson, A.J.; Réale, D.; Clements, M.N.; Morrissey, M.M.; Postma, E.; Walling, C.A.; Kruuk, L.E.B.; Nussey, D.H. An ecologist's guide to the animal model. *J. Anim. Ecol.* **2010**, *79*, 13–26.
33. Roweis, S.; Ghahramani, Z. A Unifying Review of Linear Gaussian Models. *Neural Comput.* **1999**, *11*, 305–345.
34. Yang, J.; Benyamin, B.; McEvoy, B.P.; Gordon, S.; Henders, A.K.; Nyholt, D.R.; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Gen.* **2010**, *42*, 565–569.
35. Davies, G.; Tenesa, A.; Payton, A.; Yang, J.; Harris, S.E.; Liewald, D.; Ke, X.; Le Hellard, S.; Christoforou, A.; Luciano, M.; et al. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol. Psychiatry* **2011**, *16*, 996–1005.
36. Benjamin, D.J.; Cesarini, D.; Chabris, C.F.; Glaeser, E.L.; Laibson, D.I.; Guðnason, V.; Harris, T.B.; Launer, L.J.; Purcell, S.; Smith, A.V.; et al. The Promises and Pitfalls of Genoeconomics. *Annu. Rev. Econ.* **2012**, *4*, 627–662.
37. Neale, B.M.; Kou, Y.; Liu, L.; Ma'ayan, A.; Samocha, K.E.; Sabo, A.; Lin, C.F.; Stevens, C.; Wang, L.S.; Makarov, V.; et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **2012**, *485*, 242–245.
38. Brent, L.J.N. Friends of friends: Are indirect connections in social networks important to animal behaviour? *Anim. Behav.* **2015**, *103*, 211–222.
39. Seyfarth, R.M. The distribution of grooming and related behaviours among adult female vervet monkeys. *Anim. Behav.* **1980**, *28*, 798–813.
40. Seyfarth, R.M. A model of social grooming among adult female monkeys. *J. Theor. Biol.* **1977**, *65*, 671–698.
41. Colvin, J.D. Proximate Causes of Male Emigration at Puberty in Rhesus Monkeys. In *The Cayo Santiago Macaques: History, Behavior, and Biology*; Rawlins, R.G., Kessler, M.J., Eds.; State University of New York Press: Albany, NY, USA, 1986; Chapter 6.
42. Melnick, D.J.; Pearl, M.C.; Richard, A.F. Male migration and inbreeding avoidance in wild rhesus monkeys. *Am. J. Primatol.* **1984**, *7*, 229–243.
43. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2013.
44. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **2000**, *155*, 945–959.
45. Hernandez, R.D.; Hubisz, M.J.; Wheeler, D.A.; Smith, D.G.; Ferguson, B.; Rogers, J.; Nazareth, L.; Indap, A.; Bourquin, T.; McPherson, J.; et al. Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. *Science* **2007**, *316*, 240–243.
46. Balding, D.J. A tutorial on statistical methods for population association studies. *Nat. Rev. Gen.* **2006**, *7*, 781–791.
47. Chabris, C.F.; Hebert, B.M.; Benjamin, D.J.; Beauchamp, J.; Cesarini, D.; van der Loos, M.; Johannesson, M.; Magnusson, P.K.E.; Lichtenstein, P.; Atwood, C.S.; et al. Most reported genetic associations with general intelligence are probably false positives. *Psychol. Sci.* **2012**, *23*, 1314–1323.
48. Spencer, C.C.A.; Su, Z.; Donnelly, P.; Marchini, J. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLoS Gen.* **2009**, *5*, e1000477.
49. Willer, C.J.; Speliotes, E.K.; Loos, R.J.F.; Li, S.; Lindgren, C.M.; Heid, I.M.; Berndt, S.I.; Elliott, A.L.; Jackson, A.U.; Lamina, C.; et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Gen.* **2009**, *41*, 25–34.
50. Schaaf, C.P.; Sabo, A.; Sakai, Y.; Crosby, J.; Muzny, D.; Hawes, A.; Lewis, L.; Akbar, H.; Varghese, R.; Boerwinkle, E.; et al. Oligogenic heterozygosity in individuals with high-functioning autism spectrum disorders. *Hum. Mol. Gen.* **2011**, *20*, 3366–3375.
51. Gratten, J.; Wray, N.R.; Keller, M.C.; Visscher, P.M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.* **2014**, *17*, 782–790.
52. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550.
53. Lee, S.H.; Goddard, M.E.; Visscher, P.M.; van der Werf, J.H. Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. *Gen. Sel. Evolut.* **2010**, *42*, 22.

54. Gusev, A.; Lee, S.H.; Trynka, G.; Finucane, H.; Vilhjálmsson, B.J.; Xu, H.; Zang, C.; Ripke, S.; Bulik-Sullivan, B.; Stahl, E.; et al. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *Am. J. Hum. Gen.* **2014**, *95*, 535–552.
55. Kostem, E.; Eskin, E. Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *Am. J. Hum. Gen.* **2013**, *92*, 558–564.
56. Yang, J.; Lee, S.H.; Goddard, M.E.; Visscher, P.M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Gen.* **2011**, *88*, 76–82.
57. Gay, L.; Siol, M.; Ronfort, J. Pedigree-free estimates of heritability in the wild: Promising prospects for selfing populations. *PLoS ONE* **2013**, *8*, e66983.
58. Bérénos, C.; Ellis, P.A.; Pilkington, J.G.; Pemberton, J.M. Estimating quantitative genetic parameters in wild populations: A comparison of pedigree and genomic approaches. *Mol. Ecol.* **2014**, *23*, 3434–3451.