1 Supplementary information

2 Title: Parent-offspring facial resemblance increases with age in rhesus macaques

3 Authors: Anahita J.N. Kazem, Yvonne Barth, Dana Pfefferle, Lars Kulik and Anja Widdig

4 Journal: Proceedings of the Royal Society of London B

5 doi: 10.1098/rspb.2018.1208

6

7 Supplemental Experimental Procedures

Participants. Data from two studies, A and B, have been combined (total N = 109 adult raters; 69 females), in order to include trials encompassing the widest possible range of macaque ages (offspring age range in experiment A: 3 months - 6 years, and in B: 4 - 12 years). To ensure that participants were comparable, only the subset of 35 subjects which possessed experience working with nonhuman primates was utilised from study B, as that study had included a comparison between naïve and experienced raters [1]. Potential effects of the age of the stimulus animal had not been examined in this previous dataset, which included only sexually mature macaques.

Approximately half the participants (47%) for experiment A were tested at the Max-Planck-Institute for Evolutionary Anthropology, Leipzig, Germany, and the remainder at the German Primate Centre in Göttingen, Germany. Experiment B was conducted entirely at the Max-Planck-Institute for Evolutionary Anthropology. Identical testing conditions were used at both locations (e.g. size and resolution of computer monitors used to present the task).

20 Experimental design. Subjects completed a brief background questionnaire followed by the 21 experimental task, at individual computer terminals. A single practise trial was provided before the 22 main task, which was split into two blocks with a break of 1-2 minutes halfway through. See main 23 text for a description of the kin discrimination (KD) versus individual discrimination (ID) trials. Each 24 participant received trials in a unique order, with the four types of KD trial (mother-daughter, 25 mother-son, father-daughter, father-son), and two types of ID trial (male-male, female-female), 26 interleaved evenly across the session. The experiment was controlled via a custom-designed 27 online presentation system, linked to a MySQL database. Participants were allowed as much time 28 as desired to make decisions.

29 Within an image triad, the offspring and decoy individuals for kin discrimination trials (or self 30 and decoy for individual discrimination trials) were matched in age as closely as possible. The 31 mean ± SE absolute age disparity was 0.38 ± 0.07 years for KD trials, and 1.14 ± 0.34 years for ID 32 trials; care was taken to match the youngest age classes particularly closely (e.g. the mean disparity was only 25.9 and 18.7 days for the infant pairs in KD and ID trials, respectively). In kin 33 34 discrimination trials, the age distribution of offspring and the distribution of age disparities 35 between parents and offspring were evenly distributed across the four experimental treatments. 36 As offspring age (days) and parent-offspring age disparity (days) were not significantly correlated 37 (Pearson r = -0.15, N = 92, P = 0.154), both variables were included in statistical models. For 38 individual discrimination trials, the age distribution of focal individuals was chosen to encompass 39 the full range of offspring ages used in the kin discrimination trials. Twelve ID trials were used in 40 experiment B, involving equal numbers of male versus female trials; within each sex in half the 41 trials the pair of 3/4-view images were facing right and in half facing left. Nine such trials were 42 presented in experiment A (restricted to the youngest and potentially most difficult age group: 0 -43 3 years).

44 The duration (months) of a participant's experience working with nonhuman primates was 45 initially recorded using six categories: ≤ 1 , >1 and ≤ 3 , >3 and ≤ 6 , >6 and ≤ 12 , >12 and ≤ 24 , or >2446 months). For analysis purposes we simplified this to three (≤ 6 , >6 and ≤ 24 , or >24 months), as 47 using fewer factor levels reduces the complexity of the statistical model, while retaining three 48 levels still allows detection of a potentially non-linear effect. This subdivision was chosen because 49 it was the first re-categorization which also ensured that the number of participants would 50 become more evenly distributed across levels of the factor (dividing the sample into terciles was 51 not possible, as the break points would fall within two of the original categories). There was only 52 one alternative subdivision of experience into three categories (≤ 12 , >12 and ≤ 24 , or >24 months) 53 which could also achieve a more even distribution; we subsequently checked that the particular 54 re-categorization we used was not responsible for our results, by re-running the main models 55 using this alternative subdivision. The pattern of results and statistical significance was the same 56 as in the original analysis, in all cases (Gaussian and binomial versions of the full- and, where 57 applicable, reduced models; for both the kin- and individual-discrimination tasks).

58

59 Macaque facial images. Cayo Santiago (18°09' N, 65°44' W), Puerto Rico, is an island colony of 60 >1000 free-ranging rhesus macaques, residing in 6-8 naturally formed social groups. All animals in 61 this population are individually identifiable via unique tattoos and ear notch combinations. Colour 62 images were taken in two orientations, frontal (both face and gaze directed at the camera) and ³/₄-63 view (left side of face oriented approximately 45° away from camera and gaze straight ahead of 64 face), with neutral expression. We obtained the majority of images at 1.5 - 3.0 m from the animal 65 and under even lighting conditions (open shade). The image background and ears were masked 66 using Adobe Photoshop (CS3 v.10.0 and CS4 v.11.0), head height standardized to 400 pixels, and 67 then centred on a black background. The masking procedure was performed blind with respect to 68 the kinship category of an image. When selecting images for the experiment, the 'match' and 69 'decoy' images used in a given trial were matched for general lighting conditions.

70 A total of 256 macaques, drawn from all social groups in the population, were used as stimuli in 71 the KD task. Due to the limited number of genetically assigned individuals possessing high quality 72 images, 11 individuals (images) were re-used once within the KD task (always in a different 73 experimental condition), and in nine further cases an individual was used across experiments A 74 and B (using two images, taken in different years). The ID task utilised a total of 42 macaques, 75 seven of whom had been used in the KD task (although not within the same experiment, always in 76 a different role, and typically using two different images). Image triad identity, and the identity of 77 the images used as target, left match and right match in each case, were therefore included as 78 random effects in statistical analyses.

Determination of kinship. The comprehensive genetic database for this population includes 4641
animals, genotyped at an average of 27.6 ± 1.63 microsatellite markers, with maternity genetically
confirmed and paternity assigned for >98% of animals genotyped [2]. The 291 individuals used in
this study were genotyped at an average of 27.8 ± 0.83 loci (Masterfile 2015-07-02).

Calculation of relatedness was based on complete pedigrees covering a minimum of three generations. Relatedness was genetically confirmed for all parent-offspring (i.e. target-match) pairs used in the KD task. Maternity as observed in the field was genetically confirmed for all 46 mother-offspring dyads. Paternity in all 46 father-offspring dyads was assigned using a strict exclusion rule and likelihood method, and confirmed at the 95% confidence level (for details, see Supplementary Information in [2]). Later additions to the wider pedigree revealed an updated estimate of r = 0.313 for one dyad that had originally been classified as sire-daughter ($r \approx 0.5$). 90 Given that the two individuals were nevertheless related through the paternal line and their r-91 value was substantially greater than between the target and decoy individuals (nonkin, r < 0.001), 92 the triad was retained in the analysis. To be selected as the unrelated decoy in a KD trial, our aim 93 was that the individual was unrelated for a minimum of two generations to both the target and offspring individuals, i.e. possess no parents or grandparents in common with either animal (r < 94 95 0.0625). This was achieved for all except one target-decoy and one offspring-decoy pair (from 96 different triads), in which the two individuals shared one of their 12 ancestors up to the 97 grandparental generation, hence $r \approx 0.0625$.

98 For the ID task we confirmed that the match (i.e. self) and respective decoy individuals were 99 unrelated in all except one of 21 trials, where a later update of the genetic database revealed that 100 the two individuals were in fact paternal half-siblings, thus making our test slightly conservative.

101 In cases where the sire or grandsire of an individual was unknown (e.g. due to lack of genetic 102 samples), an exclusion rule was used. We first identified potential sires for an individual based on 103 the reproductive males present in the population at the time of its conception, reduced by all 104 males already excluded as the sire based on genotypic data. The identities of the non-excluded 105 potential sires of individual A were then compared with the identities of the assigned sire and 106 grandsires of individual B, and if no overlap occurred the dyad was considered to be unrelated (r <107 0.0625). If an overlap occurred, an expected relatedness was calculated based on the probability 108 of sharing one or several ancestors whilst accounting for the empirical level of inbreeding 109 avoidance in this population (R script by R Mundry & L Kulik, unpublished data). In all cases, the 110 expected *r*-value obtained for these nonkin dyads was << 0.001.

111 Statistical analysis. Participants' responses were first assigned values from -3.5 to +3.5 (in 112 increments of 1.0), where positive values indicate a preference for the correct image in a trial and 113 negative values the reverse. All covariates were z-transformed to a mean of 0 and standard 114 deviation of 1 before use. To allow for Likelihood Ratio tests, models were fitted using Maximum 115 Likelihood (rather than Restricted Maximum Likelihood; [3]). Models incorporated a control object 116 specifying the number of iterations as 100000 (with optimiser "nloptwrap" and argument 117 'calc.derivs' set to FALSE). More accurate P-values for individual fixed effects were obtained using 118 the function drop1 [4] with argument 'test' set to "Chisq". For each full model, we checked 119 normality and homogeneity of residuals by visually inspecting a gg-plot and the residuals plotted 120 against fitted values. We also checked the stability of the full model by i) excluding data points and

- 121 ii) levels of the random effects one by one from the dataset and comparing the estimates derived
- 122 with those from the full model. None of these tests showed any indication for the existence of
- 123 influential cases. Variance Inflation Factors (VIF) were calculated using the function vif of the R-
- 124 package car [5] applied to a standard linear model excluding the random effects, which revealed
- no issue with collinearity (KD model: all VIFs < 1.43, ID model: all VIFs < 1.05). Confidence intervals
- 126 were derived using the function bootMer of the package lme4 [6], using 1000 parametric
- 127 bootstraps and bootstrapping over the random effects too (argument 'use.u' set to TRUE), in R
- version 3.3.1 [7]. Statistical tests were two-tailed, with the alpha level set at *P* = 0.05.
- 129 We initially checked whether a nonlinear term for offspring age was a more appropriate fit for
- 130 the kin discrimination data, by comparing full models based on age (days) versus log age.
- 131 Comparison of AIC values suggested the linear model was a better fit (age: 25298, log age: 25301),
- 132 hence age was used in subsequent analyses.
- 133 Example formulae used for the full models:
- 134 Kin discrimination, Gaussian model
- 135 Imer(final_rating ~ z_age*line + z_age_disparity + sex_same + experience + z_trial_posn +
- 136 (1+z_age+line_d+z_age_disparity+sex_same_d+z_trial_posn+z_age.line_d|idnmr) +
- 137 (1+experience_d1+experience_d2+z_trial_posn|triad) +
- 138 (1|target_id) + (1|match_id) + (1|decoy_id), data = KD_data, REML = F, control = contr_gauss)
- 139 Kin discrimination, binomial model
- 140 The model parallels the Gaussian version, but uses the function glmer rather than Imer, and
- 141 response_correct as the response variable. The argument family = binomial replaces REML = F, and
- 142 the control object is contr_binom.

143 Individual discrimination, Gaussian model

- 144 Imer(final_rating ~ z_age + triad_sex + experience + z_trial_posn +
- 145 (1+z_age+triad_sex_d+z_trial_posn|idnmr) + (1+experience_d1+experience_d2+z_trial_posn|triad) +
- 146 (1|target_id) + (1|match_id) + (1|decoy_id), data = ID_data, REML = F, control = contr_gauss)

147

148 Individual discrimination, binomial model

149 The model parallels the Gaussian version, but uses the function glmer rather than Imer, and

- 150 response_correct as the response variable. The argument family = binomial replaces REML = F, and
- 151 the control object is contr_binom.
- 152 Specification of control objects
- 153 contr_gauss <- ImerControl(optimizer = "nloptwrap", optCtrl = list(maxfun=100000), calc.derivs = F)

154 contr_binom <- glmerControl(optimizer = "nloptwrap", optCtrl = list(maxfun=100000), calc.derivs = F)

- 155 Variable names incorporating 'z_' were z-transformed, and those incorporating '_d', '_d1' or '_d2'
- 156 refer to factors that were manually dummy coded when specifying the random slopes.

157 Supplemental Results

158 Kin discrimination performance.

159 As there was slight patterning in the residuals of the Gaussian (LMM) full models, the results were 160 confirmed by conducting generalized linear mixed models (GLMM) with binomial error structure 161 and logit link, a more conservative approach. In the binomial version of the initial full- versus null-162 model check, there was a nonsignificant trend for a combined effect of the main predictors impacting upon subjects' performance (full versus null model: $X^2 = 6.98$, df = 3, P = 0.072). In the 163 full model itself, the interaction between offspring age and parental line was not significant 164 (estimate = $-0.11 \pm SE 0.205$, $X^2 = 0.23$, df = 1, P = 0.631), and in the reduced model there was a 165 significant positive effect of age upon participant's success in the task (estimate = 0.26 ± SE 0.104, 166 $X^2 = 6.06$, df = 1, P = 0.014), but not of parental line (estimate = 0.17 ± SE 0.231, $X^2 = 0.51$, df = 1, P 167 = 0.473). In addition, trial position was significant (estimate = $0.07 \pm SE 0.033$, $X^2 = 4.66$, df = 1, P = 168 169 0.031); participants' success rates increased slightly across the experimental session. Restricting 170 the GLMM to the 16 raters who had experienced the full range of offspring ages demonstrated a 171 similar pattern: there was no significant age*parental line interaction in the full model (estimate = 0.06 ± SE 0.288, X^2 = 0.04, df = 1, P = 0.838), but a significant effect of offspring age in the final 172 model (estimate = $0.33 \pm SE 0.148$, $X^2 = 4.71$, df = 1, P = 0.030). In this smaller sample trial position 173 174 was no longer significant and all remaining *P*-values were between 0.152 – 0.891.

176 Individual discrimination performance.

177 **Table S1.** Results of linear mixed model examining individual discrimination performance; initial

178 full model.

| Variable | Estimate | SE | Lower 95% Cl | Upper 95% Cl | <i>X</i> ² | df | P ^{1, 2} |
|---|--------------------|-------|-----------------|-----------------|-----------------------|----|-------------------|
| Intercept | 2.620 ³ | 0.309 | 2.006 | 3.198 | | | |
| Age ⁴ | -0.060 | 0.165 | -0.377 | 0.276 | 0.01 | 1 | 0.920 |
| Triad sex (female=0, male=1) | -0.330 | 0.343 | -1.099 | 0.408 | 0.85 | 1 | 0.357 |
| Experience (≤6 mths=0, >6 & ≤24 mths=1) | -0.231 | 0.188 | -0.580 | 0.129 | 1.66 | 2 | 0.437 |
| Experience (≤6 mths=0, >24 mths=1) | -0.105 | 0.172 | -0.461 | 0.216 | | | |
| Trial position ⁴ | 0.044 | 0.049 | -0.056 | 0.148 | 0.66 | 1 | 0.415 |

¹ *P*-values and test results from the drop1 function are calculated per predictor. ² P-value for the overall
 intercept does not have a meaningful interpretation and is therefore not shown. ³ Positive values indicate a
 preference for the correct image in a trial; zero indicates chance performance and negative values a
 preference for the incorrect image. ⁴ Covariates were z-transformed; mean ± SE of individual age was

183 2469.8 ± 501.15 days.

184

185 Again, the more conservative binomial (GLMM) models confirmed the main results of the

186 Gaussian (LMM) models. The initial full versus null model comparison was not significant, as there

187 was no significant effect of the main predictor, individual age (estimate = $0.33 \pm SE 0.331$, X^2 =

188 1.04, df = 1, P = 0.309). There was a nonsignificant trend for performance to be better on female-

189 female than male-male triads (triad sex: estimate = $-1.28 \pm SE 0.663$, $X^2 = 2.96$, df = 1, P = 0.085).

190 The remaining variables in the full model were nonsignificant (experience: $X^2 = 1.78$, df = 2, P =

191 0.410; trial position: $X^2 = 0.86$, df = 1, P = 0.354). Restricting the GLMM to the 16 participants who

192 had experienced the full range of offspring ages again showed no significant effect of individual

age (estimate = $0.29 \pm SE \ 0.820$, $X^2 = 0.16$, df = 1, P = 0.689). There was a nonsignificant trend for

194 performance to be better on female-female than male-male triads (triad sex: estimate = -3.74 ± SE

195 1.709, $X^2 = 3.27$, df = 1, P = 0.070); the remaining variables were not significant (experience: $X^2 =$

196 2.98, df = 2, P = 0.226; trial position: X^2 = 0.03, df = 1, P = 0.853).

197 Re-running the binomial model with all factors centred and covariates z-transformed, on data 198 from experiment B (the youngest triads; see main text for background), demonstrated that the 199 overall intercept was significantly greater than zero (estimate = $3.57 \pm SE 0.480$; lower and upper 200 95% CI = 2.74 and 25.41; *P* << 0.001). Therefore the absence of any effect of age was not due to 201 subjects failing to discriminate individual macaques.

202 Supplementary references

- Kazem AJN, Widdig A. 2013 Visual phenotype matching: cues to paternity are present in rhesus macaque faces. *PLoS ONE* 8, e55846. doi:10.1371/journal.pone.0055846.
- Widdig A, Muniz L, Minkner M, Barth Y, Bley S, Ruiz-Lambides A, Junge O, Mundry R, Kulik L.
 2017 Low incidence of inbreeding in a long-lived primate population isolated for 75 years.
 Behav. Ecol Sociobiol. 71, 18. doi:10.1007/s00265-016-2236-6.
- Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS. 2008
 Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.*, 24, 127-135. doi:10.1016/j.tree.2008.10.008.
- Barr DJ, Levy R, Scheepers C, Tily HJ. 2013 Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Memory Lang.* 68, 255-278.
 doi:10.1016/j.jml.2012.11.001.
- 5. Fox J, Weisberg S. 2011 An {R} companion to applied regression. 2nd edn. Sage, Thousand Oaks
 CA.
- Bates D, Mächler M, Bolker BM, Walker SC. 2015 Fitting linear mixed-effects models using
 Ime4. J. Stat. Software 67, 1-48. doi:10.18637/jss.v067.i01.
- R Development Core Team. 2016 *R: a language and environment for statistical computing*. R
 Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.Rproject.org/.