behaviours based on their relevance to chimpanzee socio-ecology and frequent expression in both captivity and the wild (van Hooff 1973; Nishida et al. 2010). However, the lack of sufficient data for all or the majority of the individuals forced me to exclude some behaviours (e.g. female submissive behaviour to other females; dominance displays) and combine others into larger categories (e.g. mild and severe aggression). Table 1 gives the final set of 16 variables sampled as potential personality behaviours.

Most of the behavioural variables were calculated as frequency per hour, corrected by the focal observation time per individual. Activity was calculated from time-budget data, derived from the focal sampling main activity per minute, and reported as the proportion of time spent not resting (i.e. walk, run, groom, play, forage, etc., but not rest or autogroom). The number of neighbours was the average number of individuals within 2 m of a focal subject at the beginning of the focal observation (sampled once/day). Grooming density was the proportion of grooming partners of all available partners (grooming partner was defined as a binary value of yes/no grooming given by a focal subject during the whole observation period; only adults and adolescents were considered). Grooming diversity index (GDI) was calculated with the Shannon–Wiener diversity index corrected for the group size effects (Cheney 1992; Di Bitetti 2000), as follows:

$$\textit{Grooming diversity index} = H/H_{\max}$$
$$H = -SUM(p_i x \ln(p_i))$$

in which p_i is the proportion of individual's grooming effort given to the ith individual

$$H_{\max} = \ln(N-1)$$

in which N is the number of individuals in the group. GDI results are in a value between 0 (representing perfect skew, i.e. all subject's grooming effort is directed to one individual) and 1 (representing perfect equality).

Data analyses

First, the data were divided into time periods to test repeatability (Lessells and Boag 1987). Data from AR covered nearly 3 years, allowing division into six periods based on the predominance of indoor and outdoor observations. The periods were: winter 1 (Oct 2002–April 2003, 74% observations indoors), summer 1 (May–September 2003, 100% observations outdoors), winter 2 (October 2003–January 2004, 91% observations indoors), summer 2 (June–October 2004, 100% observations outdoors), winter 3 (October 2004–April 2005, 56% observations indoors), summer 3 (May–August 2005, 98% observations outdoors). Data of CH and BB were divided into two periods: CH time 1 (June–mid-July 2008), time 2 (mid-July–September 2008); BB time 1 (May–mid-July 2009), time 2 (mid-July–September 2009). As the CH and BB studies were short-term studies, data could not be divided according to the indoor/outdoor observation context.

I calculated separately the individual behaviour scores for the 16 variables for each period. The repeatability of these individual behaviour scores was tested by intraclass correlation (ICC) with a two-way mixed model, with the period as fixed and the individual as a random factor (McGraw and Wong 1996). ICC analyses were run separately for each zoo. Repeatability analyses of the CH and BB data contained two behaviour scores per individual (i.e. time 1 and time 2) per behaviour. The individuals of BBa and BBb were pooled in the analysis. AR data allowed repeatability calculations of both short-term and long-term data. The short-term ICC was calculated on the individual behaviour scores from two consecutive winter observation periods (i.e. two scores per individual; winter 1 and winter 2), and the long-term ICC was calculated on the individual behaviour scores from all six periods (i.e. six scores per individual; from winter 1 to summer 3). Fewer individuals were included in the data set for the long-term ICC because four individuals were transferred to another zoo before the study was finished. Long- and short-term repeatability, respectively, were compared by testing the 16 variables' ICC values pairwise with Wilcoxon signed-rank test (given the non-normal distribution of the data). In addition, AR data allowed the comparison of the repeatability scores between different housing conditions; the ICC (model as above) of the variables was calculated for outdoor and indoor observations, respectively. The outdoor ICC consisted of individual behaviour scores from summer 1, summer 2 and summer 3 periods (i.e. three behaviour scores per individual). The procedure was repeated for the individual trait scores of winter 1, winter 2 and winter 3, representing indoor observations (again, three behaviour scores per individual). The indoor and outdoor ICC values were tested against each other with a Wilcoxon signed-rank test.

Secondly, I calculated the overall behaviour scores for each individual from the total observation time of those behaviours that were repeatable in at least two of the three zoos. There were two missing values of GDI (i.e. two individuals never groomed anyone during the study), which were replaced by the group mean GDI value. These scores were subjected to a Factor Analysis (FA). FA is a data reduction tool used to assess an unobservable latent construct that accounts for correlations amongst variables; it is preferable to principal components analysis when the aim is to interpret and label the emerging factors (Budaev 2010). The Kaiser–Meyer–Olkin Measure of Sampling Adequacy (KMO) and Bartlett's test of sphericity tests assured passable adequacy (KMO=0.66; Bartlett's sphericity