Genet. Sel. Evol. 39 (2007) 447–463
© INRA, EDP Sciences, 2007
DOI: 10.1051/gse:2007014

Available online at:
www.gse-journal.org

Original article

A heuristic two-dimensional presentation of microsatellite-based data applied to dogs and wolves

Claudia E. VEIT-KENSCH[a 1], Ivica MEDUGORAC[a 1*], Włodzimierz JEDRZEJEWSKI[b], Aleksei N. BUNEVICH[c], Martin FOERSTER[a]

[a] Institute for Animal Breeding, Faculty of Veterinary Medicine, The Ludwig-Maximilians-University Munich, Veterinaerstr. 13, 80539 Munich, Germany
[b] Mammal Research Institute, Polish Academy of Sciences, 17-230 Bialowieza, Poland
[c] State National Park Belovezhskaya Pushcha, Brest Oblast, Kamenec Raion, 225063 Kamenyuki, Belarus Republic

(Received 10 February 2006; accepted 14 February 2007)

Abstract – Methods based on genetic distance matrices usually lose information during the process of tree-building by converting a multi-dimensional matrix into a phylogenetic tree. We applied a heuristic method of two-dimensional presentation to achieve a better resolution of the relationship between breeds and individuals investigated. Four hundred and nine individuals from nine German dog breed populations and one free-living wolf population were analysed with a marker set of 23 microsatellites. The result of the two-dimensional presentation was partly comparable with and complemented a model-based analysis that uses genotype patterns. The assignment test and the neighbour-joining tree based on allele sharing estimate allocated 99% and 97% of the individuals according to their breed, respectively. The application of the two-dimensional presentation to distances on the basis of the proportion of shared alleles resulted in comparable and further complementary insight into inferred population structure by multilocus genotype data. We expect that the inference of population structure in domesticated species with complex breeding histories can be strongly supported by the two-dimensional presentation based on the described heuristic method.

dog / microsatellite / genetic distance / two-dimensional presentation

1. INTRODUCTION

While genetic distance methods based on a sum over loci such as the Nei D_A-distance [19] provide valuable insight into the phylogenetic relationship between breeds of several domestic species, they have often failed to support

* Corresponding author: ivica.medjugorac@gen.vetmed.uni-muenchen.de
[1] Both authors contributed equally to this work.

the analysis of dog breeds [12,15]. It is well accepted that the true evolutionary history of dog breeds is not sufficiently represented by a bifurcating tree since individuals from existing breeds are arbitrarily chosen to be founders of new breeds [22]. Since each reduction of information loss during the process of converting multidimensional genetic distance matrices into graphical presentations facilitates the interpretation of phylogenetic results, we were interested in methods for a two-dimensional (2D) illustration of genetic distances. As with phylogenetic trees, we also produced a consensus 2D graph to demonstrate the stability of the presentation of each particular population as well as the stability of the complete consensus graph. We used the cophenetic correlation coefficient [27] to analyse to which extent a tree or a 2D illustration (2DI) represents the multi-dimensional relationships within genetic distance data. To further evaluate the explanatory power of distance-based 2DI, we performed a model-based cluster analyses with the *Structure* programme [7, 23] that uses multilocus genotypes instead of distances. The individual distances based on the proportion of shared alleles (D_{PS} [1]) also use multilocus genotypes and avoid averaging over individuals. Therefore, the comparison of 2DI based on allele sharing distances with the results of the methods implemented in the *Structure* programme should give an appropriate insight into the usefulness of the heuristic algorithms developed in this work. To demonstrate the application of 2DI we analysed the biodiversity in a data set of nine dog breeds sampled in Germany and one free-living wolf population from the border of Poland and Belarus.

2. MATERIALS AND METHODS

2.1. Animals

Nine dog breeds, the Pyrenean shepherd dog (PS, $n = 33$), German shepherd dog (SH, $n = 28$), Saarloos Wolfhound (WH, $n = 30$), Bernese mountain dog (BS, $n = 31$), Entlebuch mountain dog (ES, $n = 29$), Rottweiler (RW, $n = 29$), Yorkshire Terrier (YT, $n = 23$), Beagle (BEA, $n = 142$) and Golden Retriever (GR, $n = 32$), and one free-living wolf population (PW, $n = 33$) from the Bialowieza Primeval Forest in Poland and Belarus were sampled. The choice of breeds was restricted by availability but tried to comprise some of the most common in Germany. The Beagle samples represent the status of a laboratory breeding population comprised of 142 animals that was completely blood sampled in 1996. All individuals were used to test the reliability of the applied marker set since these individuals were related in a complex manner.

Thorough revision of the pedigree revealed twelve unrelated individuals that were founders or unrelated Beagles brought into the population from other laboratories. Only these twelve unrelated individuals were included in statistical analyses. The Pyrenean shepherd dogs, the Entlebuch mountain dogs as well as the Saarloos Wolfhound were chosen from tissue banks exclusively established for those breeds. The tissue banks of the Entlebuch mountain dogs and the Saarloos Wolfhounds were established at the Institute for Animal Breeding, University Gießen (Germany). The blood bank of the Pyrenean shepherd dogs is based at the Institute for Animal Breeding, University Munich (Germany). Care was taken to be sure that the individuals were not related. All other breeds were sampled during the period of 1996 to 2000 and are derived from patients of the Small Animal Clinic for Surgery of the Ludwig-Maximilians-University Munich.

2.2. Microsatellite markers

The DNA analysis was based on a set of 23 microsatellite markers of seven dinucleotide (CPH02, CPH03, CPH04, CPH06, CPH07, CPH08, CPH17 [11]) and 16 tetranucleotide markers (2001, 2010, 2016, 2054, 2097, 2109, 2130, 2132, 2137, 2140, 2142, 2161, 2164, 2168, 2175, 2201 [10]). All markers were tested for use in a parentage test kit at our institute. Thus, we chose markers with a high PIC-value according to the authors mentioned above. We first genotyped two complex families of the Beagle population with known relationships ($n = 142$). The results assisted in assembling effective marker multiplex sets and served as a standard scale for the genotyping procedure. According to the results of the quality control (reproducibility and Mendelian segregation) we excluded three markers from all further analysis. Marker 2132 appeared extremely polymorphic with 31 alleles and presented null alleles; markers 2130 and 2142 were not able to generate reproducible PCR results.

2.3. Laboratory analysis

Samples of Saarloos Wolfhounds and Entlebuch mountain dogs were supplied as DNA samples. The tissue samples and hair roots of wolves were stored at −20 °C upon collection and were analysed several months later. All other dog samples consisted of EDTA-blood. Genomic DNA was prepared from peripheral blood, hair roots, and tissue samples using standard methods.

Multiplex-PCR was carried out in 15 μL reactions using approximately 100 ng genomic DNA in 1.5 mM $MgCl_2$ (Sigma), 200 mM dNTP (Peqlab),

1 X buffer (Sigma), and 0.5 U Taq polymerase (Sigma). The forward primer of each microsatellite marker was synthesised with an additional tail of M13MP18 phage (5′ CGT TGT AAA ACG ACG GCC AGT 3′). The complementary primer to this tail was labelled with TET, FAM or HEX fluorescent dyes. We used M13MP18-tailing for primers to combine various four to six markers into multiplex sets labelled with one of three fluorescent dyes and to be able to exchange individual markers as necessary. The PCR conditions were as follows: initial denaturation for 4 min at 94 °C; 10 cycles consisting of denaturation at 94 °C, 1 min; annealing at 60 °C, 1 min and extension at 72 °C, 1 min. For the next 30 cycles, the annealing temperature was changed to 55 °C. The PCR ended with a final extension step at 72 °C for 7 min. PCR products of three marker sets, each labelled with a different dye, were mixed together for fragment analysis with an ABI 310 Sequencer (Perkin Elmer) using an internal TAMRA-labelled standard. For each run, internal and external standards were used to determine allele lengths. External standards corresponded to samples analysed in previous runs with excellent quality.

2.4. Statistical analysis

For the statistical analyses, we chose 267 samples with ten and more reliable genotypes including a sub-sample of twelve unrelated Beagles. We excluded 5.5% of samples from the analysis because they were of lower quality and resulted in less than 10 genotypes. Unbiased estimates of heterozygosity were calculated according to Nei [18]. For the measurement of population subdivision, we used G_{ST} [17]. Wright's formulation of fixation indices was developed for two alleles. For this reason, F_{ST} is often denoted as G_{ST} when defined in the context of multiple alleles. We used G_{ST} as a statistic measure that estimates F_{ST} and further to measure the average number of migrants per generation, *Nm*, as suggested by Slatkin and Barton [26].

The Nei unbiased D_A-distance [19] was calculated based on microsatellite frequencies while the individual distances were based on the proportion of shared alleles $D_{PS} = -\ln(PS)$ [1]. The phylogenetic trees of the D_A-distance and the individual D_{PS}-distances were calculated by the NEIGHBOR programme from the PHYLIP programme package [9] and plotted by the TreeView programme [20]. To test the stability of the D_A-distance tree, 1000 distance-matrices were produced by bootstrapping over loci [8]. The resulting consensus tree was generated using the CONSENSUS programme from the PHYLIP programme package [9]. The cophenetic correlation coefficient [27] was calculated using an Excel sheet developed by Dighe *et al.* [4].

To present the genetic distance matrix in the 2D space we applied a novel heuristic approach. In a 2D graph, each of *nP* populations (D_A-distances) or *nI* individuals (D_{PS}) is presented by a point in the Euclidean space. The spatial distances between points on the Euclidean plane are summarised in the Euclidean dimensional matrix, which should reflect the genetic distance matrix between the units. We maximised the correlation between multidimensional genetic distance matrix and the Euclidean two-dimensional matrix using the modified great deluge algorithm (GDA) of Dueck [5]. The GDA method is formally similar to simulated annealing [13] but easier to implement. In the first iteration of the GDA procedure, we chose a random distribution of *nP* (*nI*) points in the plane as the initial configuration. Then the spatial distances between these points were generated and the correlation between the two-dimensional Euclidean matrix and multidimensional genetic distance matrix, *r*2*D*, was calculated. The initial quality level (water level, hence great deluge) was set to *QL* = *r*2*D*, *i.e.* correlation between random 2D configuration and true multi-dimensional configuration. In the second iteration, a small stochastic perturbation (mutation) of the initial configuration was produced. One randomly chosen population or individual (random number from uniform distribution) was shifted by a random vector (two random numbers from normal distribution) in the plane. The quality of this new configuration was computed as the *r*2*D*new. The new configuration was rejected if *r*2*D*new < *QL* and accepted if *r*2*D*new > *QL*. We increased the quality level by *RS* (rain speed) only for a new configuration above the actual quality level. *RS* is calculated as max((*r*2*D*new–*QL*)/20, 0.000001) and *newQL* as *QL*+*RS*. If accepted, the new configuration served as the initial configuration for the next stochastic perturbation. The GDA procedure accepts all new configurations with a quality above the slowly increasing quality level (*QL*), *i.e.* also configurations with a lower quality than the previous one are accepted. The iterations stop when the number of iterations exceeds a user-defined maximum or when no further increase in quality for nE_i iterations is achieved (see below). The default maximal number of iterations is set to 100 000 for populations and 1 000 000 for individuals. To avoid getting arrested in a local optimum, we modified the original great deluge algorithm (GDA, [5]) by use of ten alternate "ebb" (*E*) and "floods" (*F*). If there was no increase in the quality of the current configuration after $nE_i = (2E_i+1)nP$ or $nE_i = (2E_i+1)nI$ iteration steps, E_i is the current ebb step (E_i = 1,...,10), quality (water) level was decreased by 20%, *newQL* = *QL**0.8. Since the GDA accepts all new configurations with a quality above the *newQL,* the stochastic perturbation partly destroys optimal or suboptimal configurations reached before the ebb and then re-optimises the current

configuration. After ten alternating ebbs and floods, the best of ten stored configurations is determined and re-optimised by 5000 additional iterations.

To assess the possible benefits of a 2DI over a phylogenetic tree we generated jackknife [6] series of *nP* trees and the appropriate 2DI based on D_{PS} distances. For each replication we omit all individuals of one breed, *i.e.* jackknife over populations. For each of these trees and 2DI, we calculated the cophenetic correlation coefficient and maximised *r*2*D*. We used both correlations as a measure to which degree a tree or a 2DI represents the multi-dimensional relationships within the genetic distance data.

As for phylogenetic trees, we aimed to generate a consensus 2D presentation which demonstrates the stability of the presentation of each particular population as well as the stability of the complete consensus presentation. This was achieved by bootstrapping and subsequent 2DI of all genetic distance matrices. We used 200 bootstrap distance matrices that resulted in 200 points per population anywhere in the Euclidean space. The consensus 2DI simultaneously formed a scatter plot for each population and maximised the spatial distances between the population clouds. Thus, the F-value was maximised, *i.e.* minimisation of the presentation variance within the populations while maximising the presentation variance between the populations. We maximised the F-value using the GDA again. First, we standardised the size and position for all 200 2DI. The size was standardised by equating the sum of Euclidean distances with the sum of the appropriate genetic distances. The position was standardised by placing one population to the coordinate origin and rotating the 2DI to set the second one on the diagonal in quadrant II. We then rotated all 2DI around this diagonal (quadrant II and IV), and accepted or rejected a rotation depending on the quality (F-value) of the new configuration. The series of rotations around the diagonal were done for $nP(nP-1)$ re-positioning of populations onto the origin and diagonal of the coordinate system. Thus, the standardised configuration with the highest F-value served as the initial configuration for the next GDA to optimise the consensus 2DI. The GDA procedure is similar as presented for maximisation of *r*2*D* above. Randomly chosen 2DI (uniform distribution) were shifted by a random vector (normal distribution) and rotated by a random angle (normal distribution). The quality of the new configuration was calculated, and the new configuration was accepted if the quality was above the consistently increasing quality level. The *RS* parameter and alternate use of "ebbs" and "floods" was performed as described and defined above.

The confidence interval for the consensus position of each particular population in the final configuration can be demonstrated by a circle around the

consensus position of each population. The radius (R) is defined by minimum significant difference (MSD [28]). The application "*PhyloGen*" of this heuristic algorithm for presentation of phylogenetic results with the statistical background and definitions is described in more detail by Medugorac [16] and can be found and downloaded on the following website: http://www.vetmed.uni-muenchen.de/gen/forschung/PhyloGen.html. The plot of the Euclidean distances in the 2D space was drawn with Microsoft Powerpoint software.

An assignment test was carried out with the *Doh* programme [2]. The *Doh* programme implements the multilocus genotype based assignment index procedure first described by Paetkau *et al.* [22].

To infer genetic ancestry of individual dogs from distinct breeds and to identify subgroups that have distinctive genotype patterns, we analysed multilocus genotypes by the model-based clustering algorithm, implemented in the computer programme *Structure* [7, 23]. Ten runs of *Structure* were performed with K equal to the total number of breeds (K = 10) and subsequently with K = 2 to K = 9, with twenty runs at each K. We ran *Structure* for 1 000 000 iterations of the Gibbs sampler after a burn-in of 100 000 iterations. The correlated allele frequency model was used allowing for admixture. The similarity coefficient across runs of *Structure* was computed as described in Rosenberg *et al.* [24].

3. RESULTS

Only Saarloos Wolfhound fell below a heterozygosity of 0.500 (0.454). Yorkshire Terrier and Polish wolves showed the highest heterozygosity with values of 0.748 and 0.736 respectively.

To investigate the population subdivision and the average number of migrants per generation we estimated the G_{ST} and *Nm* values respectively. Values for G_{ST} varied from 0.12 to 0.42, being 0.23 as mean over loci. The average number of migrants per generation, *Nm,* varied from 0.34 to 1.77, with 0.83 as the mean over loci.

The consensus tree of the Nei D_A-distance was unstable with the highest bootstrap value being 61% for the BS-PS cluster and only 23–41% for the others (graph not shown). The consensus two-dimensional diagram (Fig. 1) demonstrates the existence of two main clusters consisting of several breeds partly overlapping and four breeds laying separately. The first cluster comprises the Golden Retriever and Entlebuch mountain dog [GR-ES], and the second Rottweiler, Bernese mountain dog, Pyrenean shepherd dog and Yorkshire Terrier [RW-(PS-BS)-YT]. The Saarloos Wolfhound [WH], German shepherd [SH], Polish wolves [PW] and the Beagles [BEA] are clearly separated from

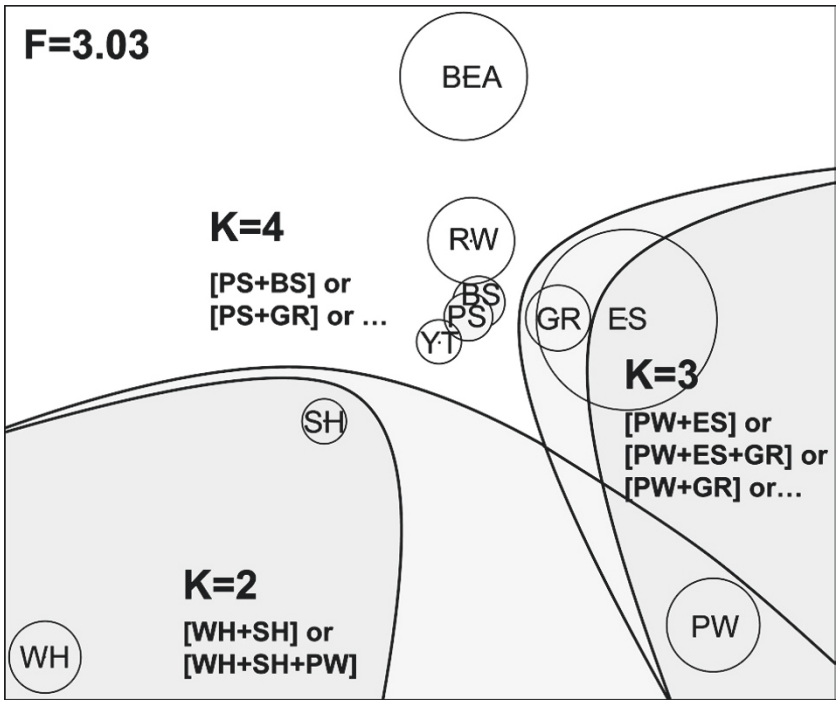

Figure 1. Consensus two-dimensional presentation based on 200 bootstrap genetic distance matrices. The F-value of the presentation variance is maximised by the GDA procedure. The circle around the consensus position of each population demonstrates the 95% confidence interval. The radius is defined by the minimum significant difference (MSD [28]). Grey underlined areas highlight consistency of this 2DI with results of structure analyses. Abbreviations of populations are as follows: WH Saarloos wolfhound, SH German shepherd, PW Polish wolves, YT Yorkshire Terrier, PS Pyrenean shepherd, BS Bernese mountain dog, RW Rottweiler, GR Golden retriever, ES Entlebuch mountain dog, BEA Beagle.

both clusters with the [WH] cluster being the farthest from all other breeds. The first neighbouring cluster is [SH]. Wolves show the largest distance to [WH], then [BEA], [SH], [RW-(PS-BS)-YT] and the smallest to the cluster of [GR-ES]. The neighbour joining tree for individual D_{PS} distances estimated by proportion of shared alleles is shown in Figure 2. Eight out of 267 individuals (3.0%) were found in "wrong" clusters. We generated the 2DI of the D_{PS} distances simultaneously to the phylogenetic tree (Fig. 3) and calculated the cophenetic correlation coefficients for both, tree and 2DI. By using *nP* jackknife replicates, we showed that the average cophenetic correlation for the

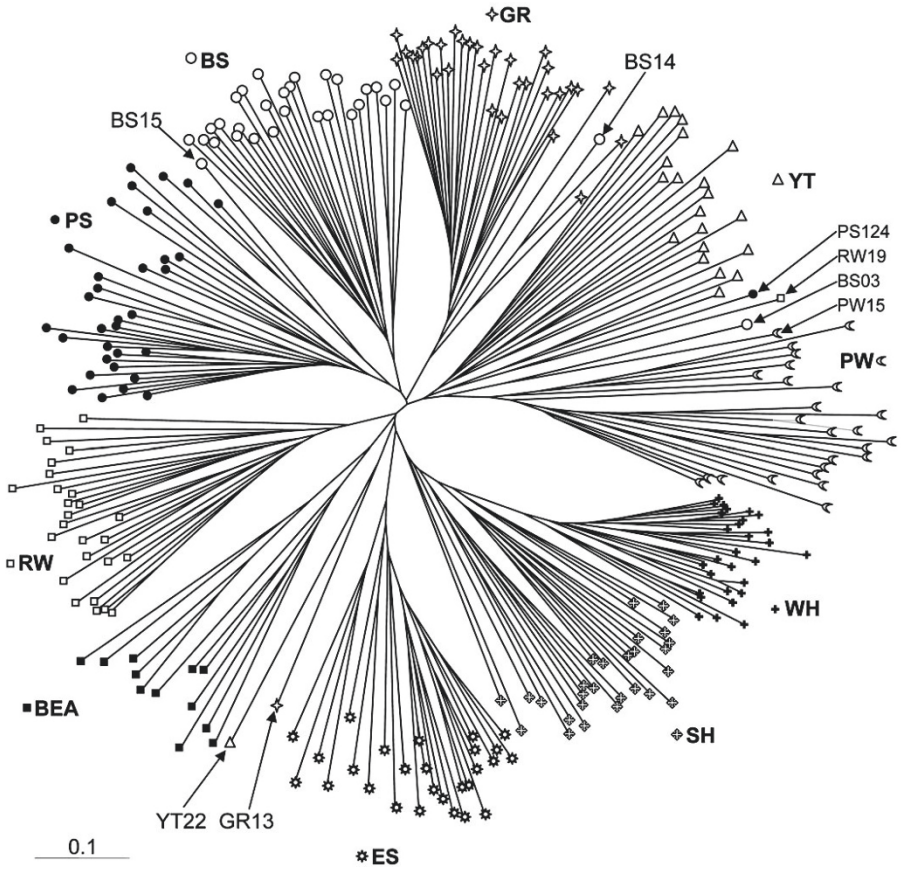

Figure 2. The neighbour-joining tree of individual allele sharing distance. Individuals being found in "wrong" clusters are marked with an arrow and the corresponding animal ID. The abbreviations are as follows: WH Saarloos wolfhound, SH German shepherd, PW Polish wolves, YT Yorkshire Terrier, PS Pyrenean shepherd, BS Bernese mountain dog, RW Rottweiler, GR Golden retriever, ES Entlebuch mountain dog, BEA Beagle.

phylogenetic tree (0.604) is significantly lower ($P < 0.00001$) than the average cophenetic correlation for 2DI (0.695).

Figure 3 shows the 2DI of 267 individual dogs and wolves optimised by the GDA method ($r2D = 0.687$).

We used the direct assignment method described by Paetkau *et al.* [22] to assess the capability of the used marker set to assign the individual dogs to their breed on the basis of genotype data alone. The direct assignment method with a leave-one-out analysis was able to correctly assign 99% of the individual

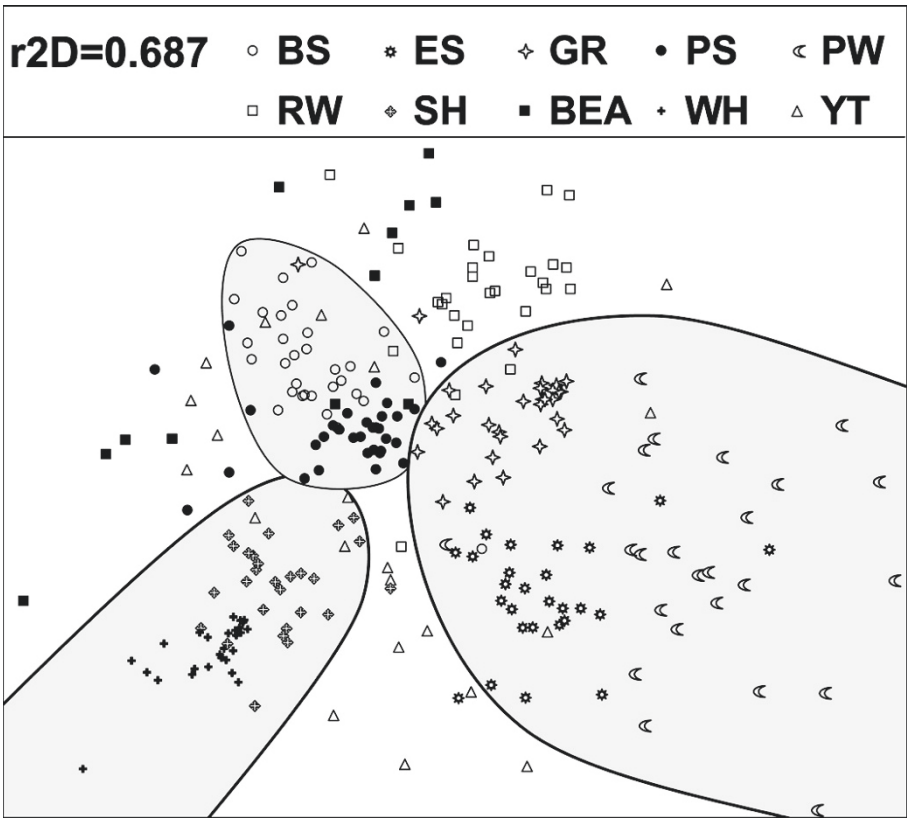

Figure 3. Two-dimensional illustration (2DI) of individual distances based on the proportion of shared alleles (D_{PS} [1]). The *r2D* is the cophenetic correlation coefficient between Euclidean and genetic distance matrix maximised by the GDA procedure. Grey underlined areas highlight the consistency of this 2DI with the results of the *Structure* analyses (*e.g.* K = 4F). The abbreviations are as follows: WH Saarloos wolfhound, SH German shepherd, PW Polish wolves, YT Yorkshire Terrier, PS Pyrenean shepherd, BS Bernese mountain dog, RW Rottweiler, GR Golden retriever, ES Entlebuch mountain dog, BEA Beagle.

dogs to their corresponding breeds. Only three out of 267 individuals were assigned incorrectly: one Bernese mountain as a Pyrenean shepherd, one Golden Retriever as a Rottweiler, and one German shepherd as a Yorkshire Terrier.

In Figure 4, the results of the *Structure* based analysis are demonstrated. Assuming 10 clusters (K = 10; Fig. 4K), the *Structure* programme assigned almost all individual dogs to each pre-defined population. On average, the proportion of membership of individuals in each of the 10 pre-defined breeds was in the range from 0.87 (PS) to 0.97 (WH). In 20 independent *Structure* runs,

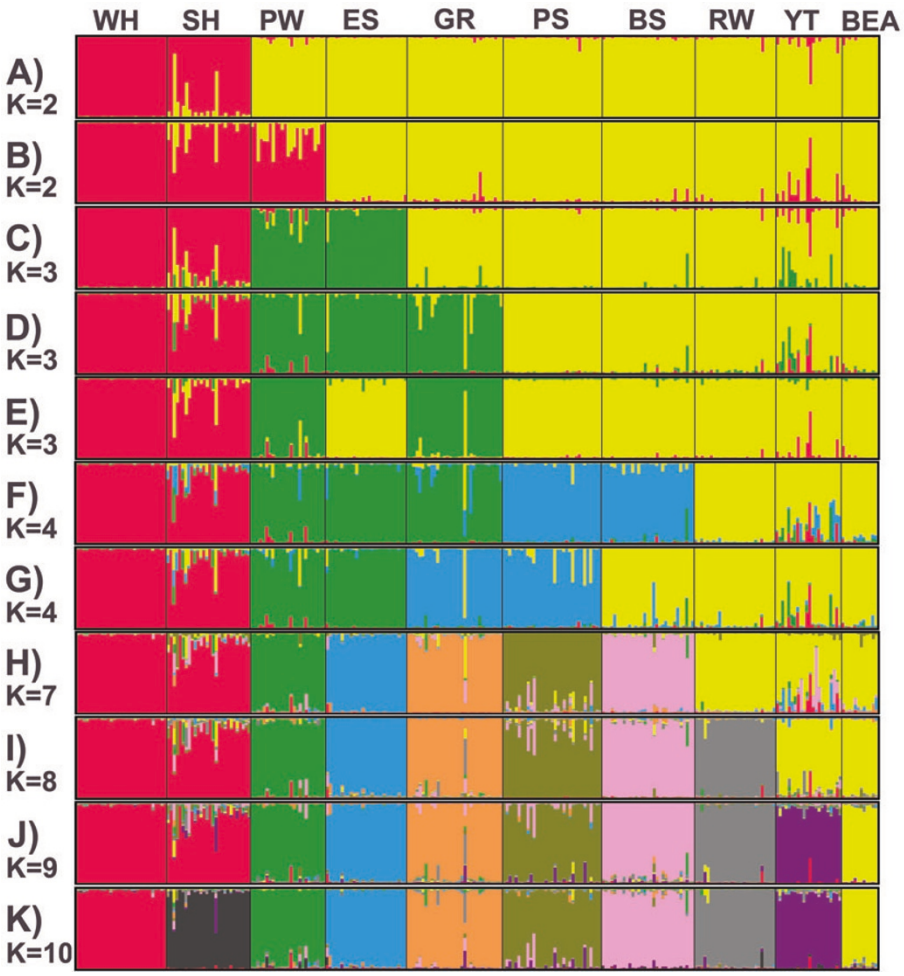

Figure 4. Clustering assignment of nine dog breeds and one wolf population. The given number of clusters (K) varied from two (**A**) to ten clusters (**K**). Individuals are represented by a single vertical column divided into K colours. Each colour represents one cluster, and the length of the coloured segment corresponds to the individual's estimated proportion of membership in that cluster. In **A**) and **B**), K = 2. **B**) Shows the more frequent pattern of 13 out of 20 runs and **A**) shows the remaining 7 runs. **C**), **D**) and **E**) represent the three most frequent patterns at K = 3. **F**) and **G**) represent the two most frequent patterns at K = 4. **H**) to **K**) represent the typical pattern at K = 7 to 10. The abbreviations are as follows: WH Saarloos wolfhound, SH German shepherd, PW Polish wolves, YT Yorkshire Terrier, PS Pyrenean shepherd, BS Bernese mountain dog, RW Rottweiler, GR Golden retriever, ES Entlebuch mountain dog, BEA Beagle.

individual dogs from the same breed shared similar membership coefficients within the inferred clusters. The similarity coefficient for 45 pairs of runs was 0.992. The 2DI of D_{PS} distances also clearly clustered the individuals from each pre-defined breed, only YT dogs spread over almost the entire parametric space. For 20 independent *Structure* runs with K = 2 to 9 the clustering results were not consistent during all 20 runs. The typical patterns are presented in Figures 4A to 4J. The similarity coefficient calculated within the typical patterns for specific K varied from 0.991 to 0.997. Figures 1 and 3 combine the results of the two-dimensional presentation and the *Structure* analyses with different numbers of a given cluster. Using increasing K, *Structure* first separated the most divergent groups into clusters. Using K = 2, *Structure* first built the [WH-SH-PW] cluster and second a cluster with all other breeds (Fig. 4, B)). Alternatively, at 35% of runs, the first cluster only includes WH and SH (Fig. 4, A)). Both alternative clustering matched very well with the consensus and individual 2DI (Figs. 1 and 3). For K = 3, five patterns were estimated and the three most frequent are cluster [WH-SH], with alternating [PW-ES] (35%) or [PW-ES-GR] (30%) or [PW-GR] (20%) as a second cluster and the remaining breeds in a third one (Fig. 4, C) to E)). This clustering is also shown by both 2DI. Given four clusters (K = 4), *Structure* again built [WH-SH] and alternatively [PW-ES] or [PW-GR] or [PW-ES-GR] clusters and then tried to distribute the remaining breeds into two clusters. The composition of the two remaining clusters depended on the alternative clustering of PW, ES and GR breeds. Therefore, the third cluster consisted of one to three breeds including all combinations of PS, BS and GR (Fig. 4, F) and G)). The fourth cluster included BEA, YT and RW in 90% of the replications and one or two additional breeds depending on the structure of the third cluster. The inconsistent clustering at K = 2, 3 and 4 by the *Structure* programme (most common clustering shown in Fig. 4, A) to G)) is visualised by the distribution of breeds (Fig. 1) and individual dogs (Fig. 3) across 2DI and thus can be better understood. Using K = 5 and K = 6 did not give consistent results across runs, but four clusters mentioned above built the most frequent groups again. At K = 7, WH and SH are grouped in one cluster in all runs and two breeds with wide spreading over individual 2DI, BEA and YT, always remained together in the second cluster (Fig. 4, H)). In 35% of cases, RW was found in this second cluster. Depending on the second cluster [BEA-YT] or [BEA-YT-RW], the remaining six or five breeds built five or four single or one additional cluster with various combinations of two closely related breeds. At K = 8, *Structure* inferred [WH-SH], [BEA-YT] and six remaining breeds as eight distinct clusters (Fig. 4, I)). For K = 9, in 80% of independent runs *Structure* first clustered WH and SH

and the other eight populations remained single (Fig. 4, J)). During the remaining 20% runs *Structure* first clustered PS and BS (closely related in 2DI) or BEA and YT (both with broad distribution over 2DI). Finally, assuming 10 clusters, *Structure* assigned almost all individual dogs to each pre-defined population (Fig. 4, K)).

4. DISCUSSION

The amount of genetic variation is similar to that found by other authors using at least partly the same markers [11, 12, 30]. The population differentiation found in this study (G_{ST} being 0.23) was similar to that found by other authors using different microsatellite marker sets in dog breeds: 0.233 in Koskinen and Bredbacka [15], 0.23 in Irion *et al.* [12], 0.27 in Parker *et al.* [21]. These observations confirmed that breeding barriers have led to a strong genetic isolation, $Nm < 1$. Variation among breeds in dogs is on the high end of the range reported for domestic livestock populations (typically in the range of eight to 16 percent; own data) or human populations (typically in the range of five to ten percent [24]). One reason to estimate Nm (the average number of migrants per generation) is that this combination of parameters indicates the relative strengths of gene flow and genetic drift. Thereby the force of gene flow is measured by the fraction of individuals that are immigrants (denoted by m) and the force of genetic drift is proportional to the inverse of the effective population size (N [25]). If Nm is less than 1, which is the case in our study with $Nm = 0.83$, the isolation between populations results in the genetic drift being the main force of genetic differentiation [26].

Two previous studies using ten [14] and 96 [22] nuclear microsatellite loci showed that these could be used to accurately assign individual dogs to their breed of origin. Our results based on 20 microsatellite genotypes and the direct assignment method [21] confirm the high rate of correctly assigned dogs to breed of origin. Although the G_{ST}-value demonstrates that variation among breeds accounts for 23% of the total genetic variation and microsatellite genotypes assigned correctly 99% of individuals to their respective breeds, the Nei D_A distance appeared to be remarkably unstable in the consensus tree. Only one node was consistent in 61% of the replicates, others in less then 41% of the replicates (tree not shown).

By using similar material (*i.e.* similar number of markers, individuals and breeds) and similar conditions, we observed very stable UPGMA or Neighbour Joining consensus trees for other domesticated species such as horses, cattle, sheep and swine (data not shown). In these species, the distribution of alleles

over all breeds showed a regular pattern of two to three alleles per marker with higher frequencies and just the order of the alleles varied. Other alleles varied in rare frequencies or were only found in single populations. This pattern was completely different from the pattern detected in domestic dog breeds where most frequent alleles irregularly changed from breed to breed causing the single locus distances between breeds to change remarkably. Strong allele frequency differences between breeds with irregular patterns between loci resulted in a large population differentiation ($G_{ST} = 0.23$) but inaccurate clustering in the phylogenetic tree. Under the above circumstances, bootstrapping over loci inevitably results in unstable consensus trees.

This pattern of allele frequency distribution is partly explainable by the breeding history of dogs. Most of the breeds started with a few arbitrarily chosen individuals to favour a certain morphological and/or behavioural feature. The selected individuals often originated from different subpopulations. Breed standards were clearly defined (*e.g.* hair length, colour, skeletal shape) beginning in the 19th century. The subsequent artificial selection was strictly focussed on breed specific morphological traits, which are determined by a small number of quantitative trait loci (*e.g.* [3]). Neighbouring alleles inevitably underwent a strong selection by hitch-hiking effects [29] resulting in an irregular allele frequency pattern across dog breeds and consequently in instable consensus trees. For functional traits or loci, which are not the main objective of breed definition, balancing selection was practised, meaning that the extremes have been excluded from breeding.

The clustering algorithm implemented in the *Structure* programme was explicitly designed to overcome limitations through the allelic distribution pattern. Individual allele sharing distances (D_{PS}) are also based on multilocus genotype and not on an average over individuals and/or populations. Therefore, inference of population structure using the *Structure* programme and individual 2DI may complement each other. By using series of *nP* jackknife replications of individual D_{PS} distances, we could show that the 2DI extracts significantly more information from allele sharing distance matrices than the phylogenetic tree ($P < 0.00001$). This was due to the higher cophenetic correlation (0.695 *versus* 0.604). The comparative analyses of 2D clustering and the results of *Structure* with different number of given clusters (K = 2 to 10) suggest that the two methods can complement each other. The clustering pattern depicted in Figure 4 is comparable with the individual 2DI shown in Figure 3. Independent to K, the program *Structure* calculated a very high membership index for individuals of WH (0.97) and a low index for YT (0.90). We estimated the lowest (0.454) and the highest (0.748) heterozygosity for WH and YT,

respectively. Furthermore, WH individuals describe a compact population in a marginal position on the individual 2DI and YT individuals were distributed over almost the entire parametric space of the 2DI. On the contrary, the wolf population also appeared very heterozygous (0.736) showing the highest mean number of alleles per locus. The population was known to experience a lot of migration within their habitat at sampling time. This is most likely the reason for their scattered but still circumscribed marginal position in 2DI. Using K = 10, the *Structure* programme estimates a consistent membership coefficient for all individual dogs or wolves, *i.e.* the similarity coefficient between runs was 0.992. Four individual wolves consistently showed a relatively low membership coefficient to the wolf cluster 0.7–0.8. These four wolves were found in a central position, *i.e.* close to the neighbouring breeds in the 2DI.

By using K = 7 or K = 8 and 20 independent runs of the *Structure* programme, we observed alternative clustering within each of K. The most common clustering patterns are presented in Figure 4, H) and I). In Figure 4, there is no qualitative difference between [WH-SH] and [YT-BEA] cluster at K = 8 or between [WH-SH] and [RW-YT-BEA] cluster at K = 7. On the contrary, analysing the 2DI (Fig. 3) it is obvious, that the [WH-SH] is a compact cluster of two historical closely related breeds while the [YT-BEA] is a cluster of the "remaining" heterogeneous individuals widely spread over almost the entire parametric space. The same holds true for the [YT-BEA-RW] cluster at K = 7. And even using K = 9 (No of breeds – 1), the *Structure* programme tends to cluster individuals of the two most closely related breeds together (80% WH-SH and 5% PS-BS) or alternatively (15% cases) the remaining individuals spread all over the 2DI (YT and BEA). Therefore, the 2DI of individual genetic distances enables to differentiate between compact clusters of closely related individuals and scattered clusters of highly heterogeneous individuals (compare WH and YT as well as PW and YT, Fig. 3).

Although methods based on genotype patterns have proven to give deeper insight into the population structure of dogs [22], we could show that the method of two-dimensional plotting of distance matrices including consensus 2D graph and 2D graph of individual allele sharing distances give additional results. In addition, it can facilitate the interpretation of phylogenetic trees, parameters of population subdivision and *Structure* results. We present a powerful method to maximise the cophenetic correlation between estimated genetic distance and the Euclidean distance in a diagram in order to better exploit the information of a distance matrix. We could show that even in a species such as the dog having experienced an interlocking breeding history it provides a

suitable overview in comparison with and in addition to other phylogenetic methods.

ACKNOWLEDGEMENTS

We are very grateful to M. Geyer and the staff of the Small Animal Clinic for Surgery of the Ludwig-Maximilians-University Munich for help in acquiring dog samples. We also thank W. Hecht of the Justus-Liebig-University Giessen for the DNA of Entlebuch shepherd and Saarloos Wolfhound. We would like to mention M. Schumm and M. Hagemann of the National Research Center for Environment and Health for providing the Beagle samples and pedigree. Last but not least we thank Dr. Petra Hanke from Roche Diagnostics GmbH Penzberg for comments on English grammar and style.

REFERENCES

[1] Bowcock A.M., Ruiz-Linares A., Tomfohrde J., Minch E., Kidd J.R., Cavalli-Sforza L.L., High resolution of human evolutionary trees with polymorphic microsatellites, Nature 368 (1994) 455–457.

[2] Brzustowski J., Doh assignment test calculator (2002) http://www2.biology.ualberta.ca/jbrzusto/Doh.php [consulted: 28 September 2006].

[3] Chase K., Carrier D.R., Adler F.R., Jarvik T., Ostrander E.A., Lorentzen T.D., Lark K.G., Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton, Proc. Natl. Acad. Sci. USA 99 (2002) 9930–9935.

[4] Dighe A.S., Jangid K., González J.M., Pidiyar V.J., Patole M.S., Ranade D.R., Shouche Y.S., Comparison of 16S rRNA gene sequences of genus Methanobrevibacter, BMC Microbiology 4 (2004) 20.

[5] Dueck G., New optimisation heuristics: the great deluge algorithm and record to record travel, J. Comp. Phys. 104 (1992) 86–92.

[6] Efron B., Gong G., A leisurely look at the bootstrap, the jackknife, and cross-validation, Am. Stat. 37 (1983) 36–48.

[7] Falush D., Stephens M., Pritchard J.K., Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, Genetics 164 (2003) 1567–1587.

[8] Felsenstein J., Confidence limits on phylogenies. An approach using the bootstrap, Evolution 35 (1985) 783–791.

[9] Felsenstein J., PHYLIP – Phylogeny Interference Package, Version 3.5, Department of Genetics, Washington University, Seattle, WA, 1993.

[10] Francisco L.V., Langston A.A., Mellersh C.S., Neal C.L., Ostrander E.A., A class of highly polymorphic tetranucleotide repeats for canine genetic mapping, Mamm. Genome 7 (1996) 359–362.

[11] Fredholm M., Wintero A.K., Variation of short tandem repeats within and between species belonging to the *Canidae* family, Mamm. Genome 6 (1995) 11–18.
[12] Irion D.N., Schaffer A.L., Famula T.R., Eggleston M.L., Hughes S.S., Pedersen N.C., Analysis of genetic variation in 28 dog breed populations with 100 microsatellite markers, J. Hered. 94 (2003) 81–87.
[13] Kirkpatrick S., Gelatt C.D. Jr., Vecchi M.P., Optimization by simulated annealing, Science 220 (1983) 671–680.
[14] Koskinen M.T., Individual assignment using microsatellite DNA reveals unambiguous breed identification in the domestic dog, Anim. Genet. 34 (2003) 297–301.
[15] Koskinen M.T., Bredbacka P., Assessment of the population structure of five Finnish dog breeds with microsatellites, Anim. Genet. 31 (2000) 310–317.
[16] Medugorac I., Genetischer Poymorphismus in Rinderrassen des Balkan und Phylogenie europäischer Rinder, Dissertation, Institute for Animal Breeding of Technical University of Munich, 1995.
[17] Nei M., Analysis of gene diversity in subdivided populations, Proc. Natl. Acad. Sci. USA 70 (1973) 3321–3323.
[18] Nei M., Estimation of average heterozygosity and genetic distance from a small number of individuals, Genetics 89 (1978) 583–590.
[19] Nei M., Tajima F., Tateno Y., Accuracy of estimated phylogenetic trees from molecular data, J. Mol. Evol. 19 (1983) 153–170.
[20] Page R.D., TreeView: an application to display phylogenetic trees on personal computers, Comput. Appl. Biosci. 12 (1996) 357–358.
[21] Paetkau D., Calvert W., Sterling I., Strobeck C., Microsatellite analysis of population structure in Canadian polar bears, Mol. Ecol. 4 (1995) 347–354.
[22] Parker H.G., Kim L.V., Sutter N.B., Carlson S., Lorentzen T.D., Malek T.B., Johnson G.S., DeFrance H.B., Ostrander E.A., Kruglyak L., Genetic structure of the purebred domestic dog, Science 304 (2004) 1160–1164.
[23] Pritchard J.K., Stephens M., Donnelly P., Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.
[24] Rosenberg N.A., Pritchard J.K., Weber J.L., Cann H.M., Kidd K.K., Zhivotovsky L.A., Feldman M.W., Genetic structure of human populations, Science 298 (2002) 2381–2385.
[25] Slatkin M., Gene flow and the geographic structure of natural populations, Science 236 (1987) 787–792.
[26] Slatkin M., Barton N.H., A comparison of three indirect methods for estimating average levels of gene flow, Evolution 43 (1989) 1349–1368.
[27] Sokal R.R., Rohlf F.J., The comparison of dendrograms by objective methods, Taxon 11 (1962) 33–40.
[28] Sokal R.R., Rohlf F.J., Biometry – the principles and practice of statistics in biological research, 3rd edn., W.H. Freeman and Company, San Francisco, 1998.
[29] Storz J.F., Using genome scans of DNA polymorphism to infer adaptive population divergence, Mol. Ecol. 14 (2005) 671–688.
[30] Zajc I., Mellersh C.S., Sampson J., Variability of canine microsatellites within and between different dog breeds, Mamm. Genome 8 (1997) 182–185.