Identification of breed contributions in crossbred dogs

Orlando Döhring

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University College London.

UCL Genetics Institute

Research Department of Genetics, Evolution and Environment

Division of Biosciences

Faculty of Life Sciences

School of Life and Medical Sciences

University College London

University of London

March 29, 2015

Statement of originality

I, Orlando Döhring, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

There has been a strong public interest recently in the interrogation of canine ancestries using direct-to-consumer (DTC) genetic ancestry inference tools. Our goal is to improve the accuracy of the associated computational tools, by developing superior algorithms for identifying the breed composition of mixed-breed dogs. Genetic test data has been provided by Mars Veterinary, using SNP markers.

We approach this ancestry inference problem from two main directions. The first approach is optimized for datasets composed of a small number of ancestry informative markers (AIM).

Firstly, we compute haplotype frequencies from purebred ancestral panels which characterize genetic variation within breeds and are utilized to predict breed compositions. Due to a large number of possible breed combinations in admixed dogs we approximately sample this search space with a Metropolis-Hastings algorithm. As proposal density we either uniformly sample new breeds for the lineage, or we bias the Markov Chain so that breeds in the lineage are more likely to be replaced by similar breeds.

The second direction we explore is dominated by HMM approaches which view genotypes as realizations of latent variable sequences corresponding to breeds. In this approach an admixed canine sample is viewed as a linear combination of segments from dogs in the ancestral panel.

Results were evaluated using two different performance measures. Firstly, we looked at a generalization of binary ROC-curves to multi-class classification problems. Secondly, to more accurately judge breed contribution approximations we computed the difference between expected and predicted breed contributions.

Experimental results on a synthetic, admixed test dataset using AIMs showed that the MCMC approach successfully predicts breed proportions for a variety of lineage complexities. Furthermore, due to exploration in the MCMC algorithm true breed contributions are underestimated. The HMM approach performed less well which is presumably due to using less information of the dataset.

Acknowledgements

I would like to thank my supervisor David Balding for being an excellent and dedicated supervisor. I have been fortunate to have had the opportunity to work with David for the last four years. Thank you for taking me on as a PhD student. David has been a great inspiration and I have benefited greatly from his ideas, time and feedback. Thereby, I also would like to acknowledge his support and patience in challenging times.

I am grateful to the staff of Mars Veterinary: Research director and secondary supervisor Neale Fretwell helped a lot to setup this project, provided crucial discussions and comments at project meetings. Stephen Davison, Bioinformatics technical leader at Mars Veterinary, was very helpful to supply me with background information on the project in general, Mars' proprietary code and improvements about my own Java implementation. Furthermore, I would like to thank Alan Martin, data analyst and research developer, for offering answers to some technical questions about their original implementation. Furthermore, I also appreciate the generous studentship of Mars Veterinary to support this exciting research resolving the ancestries of our beloved dogs. Many thanks also to my tertiary supervisor Mark Thomas from UCL who examined my transfer with Neale. Finally, I am also grateful to my PhD examiners, Ian Wilson and Maria De Iorio, to agree to read and provide feedback on my thesis.

I will greatly miss the great atmosphere, seminars and interesting discussions at the UCL Genetics Institute. In particular, Garrett Hellenthal provided helpful answers to questions about the ChromoPainter software. Furthermore, I acknowledge support of Adam Powell, Marco Scutari and Doug Speed who gave me advice on many programming and statistical problems. There are many other members I interacted with, including Andres Ruiz-Linares, Claudia Giambartolomei, Chris Steele, Cian Murphy, Dace Ruklisa, Delilah Zabaneh, Jon White, Julie Bertrand, Kaustubh Adhikari, Sofia Morfopoulou, Sonia Shah, Valentina Cipriani, Vincent Plagnol and Warren Emmett. I am also grateful to our administrative staff members Shush Datta, Claire Glen, Elvira Mambetisaeva, Simona Wade who run things smoothly so we can focus on our daily research activities.

There are several researchers who kindly offered answers to my questions: Claire Churchhouse, Lucie Gattepaille, Colin de La Higuera, Mattias Jakobsson, Brian Keith Maples, Kerstin Lindblad-Toh, Jonathan Marchini, Jesse Rodriguez, Sam Oman, Massimiliano Pontil, Paul Scheet, Inge Svein Helland,

Elaine Ostrander, Matthew Stephens, Rolf Sundberg, Yiming Yang, Robert Wayne and Amy Williams.

I also thank the 'Centre for Computational Statistics and Machine Learning' to offer seminars which connect members from many different departments who are interested in statistics which initiated ample exchange with Bryan Feeney, Samuel Livingstone and Stephen Pasteris.

I acknowledge the use of the UCL Legion High Performance Computing Facility (Legion@UCL), and associated support services, in the completion of this work. I also would like to thank Vincent Plagnol to setup the UGI cluster environments as well as Tristan Clark for support using the Computer Science Cluster.

Finally, and most importantly, I would like to thank my family and my parents.

Contents

1 Background ancestry testing and canine research **19**

1.1 Chapter outline and project motivation . . . 19
1.2 Introduction to DTC kits . . . 21
1.3 Data sources, applications and limitations of genetic testing . . . 22
1.3.1 Data sources . . . 22
1.3.2 Local and global ancestry estimation . . . 25
1.3.3 Applications . . . 25
1.3.4 Limitations . . . 26
1.4 Review canine research . . . 29
1.4.1 Canine history . . . 29
1.4.2 SNP marker data . . . 30
1.4.3 Linkage disequilibrium . . . 31
1.4.4 Breed clustering and diversity . . . 33
1.5 Datasets and marker profile . . . 35
1.5.1 Marker profile . . . 35
1.5.2 Datasets . . . 35
1.5.3 Purebred reference dataset . . . 36
1.5.4 Test Sample of synthetic dogs . . . 37
1.5.5 Recombination test sample of synthetic dogs . . . 39
1.6 Thesis contributions and further chapters' outline . . . 41
1.7 Conclusions . . . 44

2 Breed similarity in dogs **45**

2.1 Motivation to infer interbreed relationships . . . 45
2.2 Introduction to similarity measures . . . 46
2.3 Measures of proximity . . . 47
2.3.1 Proximity measures for continuous data . . . 48
2.4 Graphical methods for proximity matrices . . . 50
2.5 Breed similarity results using genotypes and haplotypes . . . 51
2.5.1 Selection of subset of breeds for literature comparison . . . 51
2.5.2 Discarding correlation structure in genotypes . . . 52
2.5.3 Haplotype analysis discarding LD . . . 59
2.5.4 Haplotype analysis . . . 63
2.5.5 Summary of results . . . 63
2.6 Conclusions . . . 64

3 Methodology for pure breed identification 65
3.1 Haplotype frequency model . . . 65
3.1.1 Inference in purebred haplotype frequency model . . . 67
3.2 Computation of chromosomal likelihood . . . 67
3.2.1 Mars approach: . . . 67
3.2.2 DBAncestry . . . 68
3.3 Performance measures for pure breed identification . . . 75
3.4 Results for DBAncestry . . . 76
3.4.1 SmallPure case . . . 76
3.4.2 BigPure case . . . 77
3.5 Conclusions . . . 77

4 MCMC methodology for mixed breed identification 79
4.1 Ancestry inference for crossbred dogs using MCMC . . . 80
4.2 Inference in crossbred haplotype frequency model . . . 81
4.2.1 Mars approach . . . 81
4.2.2 DBAncestry . . . 81
4.3 Breed proportion estimates using MCMC . . . 84
4.4 Extensions . . . 85
4.5 Performance measures . . . 86
4.6 MCMC run length . . . 87
4.7 Experimental results . . . 87
4.7.1 Short Run . . . 88
4.7.2 Long Run . . . 88
4.8 Conclusions . . . 93

5 Alternative ancestry inference models 95
5.1 Review ancestry inference techniques . . . 95
5.1.1 Regression . . . 95
5.1.2 Global model-based clustering . . . 100
5.1.3 Local window-based approaches . . . 101
5.1.4 Local HMMs . . . 102
5.1.5 Non-Parametric Bayesian approaches . . . 108
5.1.6 PCA-based approaches . . . 109
5.1.7 Machine Learning based approaches . . . 110
5.2 ChromoPainter data analysis . . . 112
5.2.1 Algorithm selection for a comparison with DBAncestry . . . 112
5.2.2 Computation of haplotype representation . . . 115
5.2.3 Recombination and mutation rate . . . 115
5.2.4 Computation of breed fraction estimates . . . 117
5.2.5 ChromoPainter results . . . 117
5.3 Summary of ancestry inference techniques and conclusions . . . 120

6 Conclusions and future work 123
6.1 Conclusions . . . 123
6.2 Future work . . . 125

A Appendix **128**
A.1 Overview of breeds and analysis pure breed dogs 128
A.2 Enumerating haplotype pairs 133
A.3 Results for pure breed identification 134
A.4 HMM Forward Algorithm 142
A.4.1 Toy example 144
A.5 Mars algorithm as pseudo code 145
A.6 Multiclass ROC Curves 147
A.7 DBAncestry Results 148
A.7.1 Breed estimation view 148
A.7.2 Breed classification view 150
A.7.3 Inferring ancestral boundaries: quantile view 154
A.8 ChromoPainter Results: classification view 154

Bibliography **157**

List of Figures

1.1 This figure shows the median distance and standard deviation in base pairs between consecutive SNP markers. Each of the points represents one chromosome. The data suggests that a few SNP have strong linkage while other can be considered unlinked given their distance of more than a million base pairs. 36

1.2 This figure shows the way Mars arranged lineage trees Φ by complexity. Each ggp leaf is encoded by colour according to the breed it represents. The most complex case is shown in tree 11 which assumes eight distinct pure breed ggps. Although these ggps may not pure breed ancestors I make this assumption because it is unlikely to detect ancestry on a finer scale than this. All other trees form special cases, e.g. tree 1 represents purebred dogs assuming that all eight ggps are from the same breed while tree 2 shows designer dogs which are a cross of two purebred dogs. The trees are roughly sorted in order of complexity although tree 6 (1 parent, 4 ggps) is more complex than tree 7 (3 gp, 1 ggp). These figures represent a classification perspective where time flows upwards (current individual at top) which is contrary to coalescent genealogical trees (Felsenstein, 2013) where the most recent individual is placed at the bottom of the figure. 37

1.3 The two sequences correspond to two examples for haplotypes on chromosome c where either the maternal or paternal part of the parental haplotype is copied first, respectively. Each haplotype is a combination of maternal **(M)** breed and paternal **(P)** breed haplotype subsequence sampled according to haplotype frequency estimates, such that with 50 percent chance either parent forms the first part of the sequence. Without loss of generality, I assume that the first sequence is related to the maternal haplotype: then, the first r alleles are copied from the maternal haplotype H^c_M, i.e. $H^c_M[1:r]$ while the remaining alleles from $r+1$ to n_c (chromosome length) are copied from the paternal haplotype H^c_P, i.e. $H^c_P[(r+1):n_c]$. 40

2.1 This figure shows a heatmap visualization of the Manhattan breed distance matrix using genotype data of 125 breeds. The columns are re-ordered according to the dendrogram of the hierarchical clustering of the distance matrix with complete linkage. Breed pairs which tend towards red have small distance (high similarity) while breed pairs going to the yellow-white spectrum are more distant breeds. Breeds have considerable distance among each other which can be seen from mostly yellow-colored matrix entries and mostly long dendrogram leaf branches. Furthermore, the dendrogram shows a flat cluster structure which implies limit subpopulation structure. 53

2.2 This figure shows the MDS plot for the **ChromoPainter** coancestry similarity matrix which has been converted to a distance matrix. This measure visually discriminates Retrievers and Scent hounds while Small Terriers are partially separated out. 54

2.3 In Figures (a,c,e,g) each line corresponds to a breed. For each of these breeds I show the decreasingly ordered breed genotype similarities which have been exponentially transformed. In Figures (b,d,f,h) I show the corresponding breed transition probabilities from closest (most similar, left) to furthest breed (right). Figure (i) shows the distribution of distance values according to heatmap in Figure 2.1. Finally, Figure (j) shows the proposed transition probability based on rank in distance sorted breeds. 55

2.4 These Figures show a heatmap either based on the distance (left column) or similarity matrix (right column). In the left column red denotes low distance (high similarity) and white refers to high distance (small similarity) while in the right column it is the reverse case: red denotes little similarity and white high similarity. The first row shows the heatmap based on the original distance matrix while in the second row the genotype distances have been transformed by function f from Figure 2.3 (j), such that most breeds are very far away from a given breed (whitish colour) and only few breeds very close (red). 57

2.5 These figures show the hierarchical clustering results using complete linkage for Manhattan distance and Pearson correlation based on the **SmallHap haplotype data**. Breeds from the same group have their branches shown in the same colour. Although breeds from the same breed group tend to be adjacent I notice that breeds from a given group are not distributed homogeneously, i.e. not all breeds from the same group are in the same cluster. As before there is a flat cluster structure confirming limited population substructure. Furthermore, most breeds are associated with long leaf branches suggesting strong differences in breed. Compared with the genotype data dendrogram there are fewer short branches which suggests less strong breed discrimination. 58

2.6 These figures shows the MDS plot for two distance, Manhattan distance (ManSmallHap) and Hellinger (HellSmallHap), and two correlation measures, Pearson (PearSmallHap) and Bhattacharyya (BhattSmallHap) based on the **SmallHap haplotype data**. Breeds from the same group are highlighted by transparent polygons. I see that all four proximity measures separate out the Retriever group and BhattSmallHap also separates out Mastiff-like breeds. Furthermore, some breeds are partially visually discriminated, such as (Herding dogs, Toy breeds) for ManSmallHap, (Herding dogs, Mastiff-like breeds, Toy breeds) for HellSmallHap, (Herding dogs, Toy breeds, Working dogs) for PearSmallHap and (Small terriers, Spaniels, Toy breeds) for BhattSmallHap. 60

2.7 This figure shows MDS plots where breeds from the same group are highlighted by transparent polygons. The first row shows the original Manhattan distance (MDG) and the rank Manhattan distance based on **genotype data** while the second row illustrates the Manhattan distance (ManBigHap) and Pearson correlation (PearBigHap) based on the **BigHap haplotype data**. The MDG measure separates out Ancient/Spitz dog, Mastiff-like breeds, Retrievers, Toy breeds and partially discriminates Small terriers. Both, ManBigHap and PearBigHap visually discriminate Retrievers and partially out either Herding dogs, Toy breeds or Herding dogs, Scent hounds, Toy breeds, respectively. 61

2.8 This figure shows that the number of haplotypes with frequency > 0.01 for any of the breeds grows exponentially with the number of SNPs on the chromosome. Each of the points represents one chromosome. Note the y-axis is shown in $\log_{10}-$scale. 62

3.1 Breed data structure Φ : The left column shows the maternal lineage painting and the right column the paternal lineage painting while the centre block refers to the genome painting. 65

3.2 This table shows the enumerated haplotype pairs for genotype $X_k^j = [0, ?, 0, 1, 2, ?]$. The first columns list the haplotype pair ID i followed by a dot and either a 1, first part of haplotype pair, or 2, second part of haplotype pair. In particular, in this example there are $m_{jk} = 32$ haplotype pairs. Details for the recursive enumeration are listed in Appendix A.2: the other columns shows for $t \in \{1, \ldots, 6\}$ which pattern has been used for haplotye pair $[H_1^i(j)[t], H_2^i(j)[t]]$. 73

4.1 **CCBC breed proportion estimates for the ULP update proposals within the long MCMC run:** these figures show true ancestral proportion as bold red line while the boxplots offers information on the breed proportion estimates, i.e. the median, lower and upper quartiles, whiskers which extend to the most extreme data point less than 1.5 IQR from the box, and outliers which are drawn as small circles. These plots represent a subset of the 11 lineage trees: in particular the top two rows correspond to the simplest trees while the bottom two columns match the most complex lineage trees. 89

4.2 This figure shows breed proportion estimates averaged across lineage trees. More in detail, CCBC breed proportion estimates for the ULP update proposal within long MCMC run are visualized as boxplot showing median, lower and upper quartiles of all ancestral levels. Numerical values for the median and quartiles are also given in Table 4.2. 90

4.3 This figure shows a density plot (gaussian kernel smoother with bandwidths 0.01-0.02) for global estimated breed contributions (CCBC, ULP, Long). Each colored transparent polygon corresponds to one of the ancestral levels, i.e. purebred, parent, gp, ggp and non-ancestry. Furthermore, vertical lines colored according to the TAP level illustrate the true breed contributions. Vertical grey dashed lines show the inferred breed contribution cut-offs between consecutive TAP levels, i.e. the cut-off between parental and purebred level is about 0.5258. 91

4.4 Breed proportion proportion estimates of lineage tree ABCDEFGH for ULP update proposal within long MCMC run based on own simulation data discussed in Section 1.5.5. The red vertical line represents the true non-ancestry and great-grandparent levels while the bold black line corresponds to the median estimate in the boxplot. Numerical values for the median estimates and quartiles are also given in Table 4.4. 93

5.1 This Figure is due to Li and Stephens (2003) and shows the fourth haplotypes are coloured according to an imperfect mosaic of the previous ones, with a probability I denote as $\hat{\pi}(h_{4\{A,B\}}|h_1, h_2, h_3)$. This example captures the copying process for two examples of the 4th haplotype where h_{4A} has a lower switch rate than h_{4B}. The shading for those next haplotypes illustrates which previous haplotypes I copy from. And jumps in the shading indicates switches in the ancestral copying. Given that shadings are unavailable in actual datasets I need to sum over all possible ones. Each circle represents a biallelic SNP marker along the chromosome with white/black color coding according to allelic variation. The third SNP marker illustrates the effect of mutations in the imperfect copying process where both $h_{4\{A,B\}}$ have the black allele but still copy from the second haplotype which has the white allele. 107

5.2 These figures show the distribution of the ChromoPainter estimates for recombination scaling and mutation rate either separately for each pure breed individual or aggregated by breed. 116

5.3 **CCBC breed proportion estimates for the ChromoPainter standard and NNLS variant:** These figures show true ancestral proportion as bold red line while the boxplots offers information on the breed proportion estimates, i.e. the median, lower and upper quartiles, whiskers which extend to the most extreme data point less than 1.5 IQR from the box, and outliers which are drawn as small circles. These plots represent a subset of the 11 lineage trees: In particular the top two rows correspond to the simplest trees while the bottom two rows match the most complex lineage trees. Furthermore, the left column shows the standard variant wile the right column corresponds to the NNLS variant. These figures show that NNLS has a much larger IQR than the standard variant. 118

5.4 This figure shows breed proportion proportion estimates averaged across lineage trees. More in detail, CCBC breed proportion estimates for the **NNLS** variant are visualized as boxplot showing median, lower and upper quartiles of all ancestral levels. Numerical values for the median and quartiles are also given in Table 5.2. 119

5.5 This figure shows breed proportion proportion estimates averaged across lineage trees. More in detail, CCBC breed proportion estimates for the **standard** variant are visualized as boxplot showing median, lower and upper quartiles of all ancestral levels. Numerical values for the median and quartiles are also given in Table 5.2. 119

A.1 HMM with two hidden markov chains unfolds in time as a lattice 142

A.2 HMM hidden path illustration for genotype [1,2,0] for three time steps 142

A.3 This figure shows CCBC-based results for the area under the curve and the F_1 criterion using the long MCMC run with ULP update proposal for all 11 lineage trees from Figure 1.2. A visualization of the actual performance plots are shown in Figure A.4. Firstly, it can be seen that both criteria are very correlated, i.e. have a correlation of 0.97. For the simplest tree both performance criteria show almost perfect performance. After that, both criteria decrease quickly until intermediate tree complexity, i.e. around 0.86 for AUC and 0.8 for F_1 score, and then continue to be reduced moderately at a level of 0.81 (AUC) and 0.78 (F_1) for the most complex tree, respectively. 150

A.4 **DBAncestry classification performance:** For all 11 lineage trees these two plots show PPV vs. sensitivity and F_1 score as a function of breed proportion, respectively. These results are based on CCBC-based breed proportion estimates from the long MCMC run with a ULP update proposal. According to the AUC-plot it can be seen that the AUC score is near perfect for trees AAAAAAAA (tree 1) and AAAABBBB (tree 2) and very high for AAAABBCC (tree 3) and AABBCCDD (tree 4). Furthermore, tree 5-11 have similar AUC scores at a high level. The F_1 plot suggests that almost independently of the chosen breed proportion cut-off there is a very high F_1 score for tree 1. Even for tree 2 the F_1 score only starts to decrease sharply at around 0.4. All other trees typically have optimal cut-off values of less than 0.2. While the simplest three trees achieve almost perfect F_1 score, with increasing amount of complexity the best F_1 score goes down to 0.78 for the most complex tree ABCDEFGH. 151

A.5 Estimated global breed contributions (CCBC, ULP, Long) percentiles are shown for each TAP. Each coloured curve corresponds to one of the ancestral levels, i.e. purebred, parent, gp, ggp and non-ancestry. Furthermore, horizontal yellow lines illustrate the true breed contributions for the different TAP levels. Vertical grey dashed lines correspond to the inferred breed contribution cut-offs between consecutive TAP levels. Related to that horizontal grey dashed lines show quantiles associated to those breed cut-offs. 154

A.6 **ChromoPainter NNLS classification performance:** This figure shows the AUC score and F_1 criterion for all 11 lineage trees. Prediction quality for pure breed tree AAAAAAAA is reasonably good at 0.83 (AUC) and 0.74 (F_1). Then, for more complex lineage trees classification performance drops sharply to [0.37,0,51] (AUC) and [0.44,0.55] (F_1) which suggests ChromoPainter does not deal well with lineages involving multiple breeds. 155

A.7 **ChromoPainter NNLS classification performance:** For all 11 lineage trees these two plots show PPV vs. sensitivity and F_1 score as a function of breed proportion, respectively. These results are based on CCBC-based breed proportion estimates using ChromoPainter with the NNLS variant. According to the AUC-plot it can be seen that the AUC score is high for tree AAAAAAAA (tree 1) with a value of 0.81, i.e. the algorithm performs reasonably well for pure breed synthetic test dogs. For more complex trees there sharp decline in performance to an AUC: while tree AAAABBBB (tree 2) has an AUC value of 0.51 all other lineage trees have an AUC score between 0.37 and 0.43. The same behaviour can be observed for the F_1 score plots: in particular this figure shows that tree AAAAAAAA has a much higher F_1 score value of 0.75 compared to the next best value of 0.45 for AAAABBBB (tree 2). 156

List of Tables

1.1 Observed allele distributions under LD. 31

1.2 This table shows values for the fixation index for a variety of marker types and dog datasets. 34

1.3 The distribution of the 320 SNPs across 25 of the 38 chromosomes of the dog genome. . 35

1.4 For each true ancestral proportion level I list which of the lineage trees have corresponding levels. Lineage tree IDs correspond to those defined in Figure 1.2. 37

1.5 For each lineage tree shown in Figure 1.2 I list the counts for the different true ancestry proportion levels. For example, lineage tree 3 is composed of one parent and two grandparents while tree 10 has 1 grandparent and 6 ggps. This table also implies that all trees have a unique combination of purebred, parents, gps and ggps except for lineage trees 8 and 9 which both have 2 gps and 4 ggps. 38

1.6 For OrgSyntheticRed I list number of test samples in each of 11 lineage trees. 38

2.1 The number of unique haplotypes in 1000s on a chromosome with non-zero population sample frequency each for at least one of the breeds. 62

3.1 For a subset of the chromosomes chromosome painting with associated breeds is shown. 66

3.2 Breed-specific SNP transition probabilities on chromosome 1 for breed Siberian Husky. There are 14 SNPs sequenced for this first chromosome. In each row those transition probabilities sharing the same old state sum to 1, i.e. $p(0 \to 0) + p(0 \to 1) = 1$, and $p(1 \to 0) + p(1 \to 1) = 1$. Table courtesy of Davison and Fretwell (2012). 67

3.3 Results for pure breed identification on **SmallPure** using performance criterion $\bar{r}_{1:B}$ defined in Equation 3.10. I selected the extreme values for $\bar{r}_{1:B}$ with its corresponding position d from Table A.6 according to different combinations of haplotype frequency estimation and evaluation of $p\left(X_c|\theta\right)$. Detailed results for combination *(PSEPARATE,DFPB)* are in Table A.8, for *(PPOOLED,DFPB)* are in Table A.9, for *(PSEPARATE,OneHap)* are in Table A.10, for *(PPOOLED,OneHap)* are in Table A.11, for *(PSEPARATE,PSEUDO)* are in Table A.12 and for *(PPOOLED,PSEUDO)* are in Table A.13. 76

3.4 Results for pure breed identification on dataset **BigPure**. The minima and maxima of performance criterion $\bar{r}_{1:B}$ according to Equation 3.10 for separate phasing with either DFPB or OneHap are shown in Table A.7. Detailed results $1 - q_b$ for each breed b are given in Table A.14 for *(PSEPARATE,DFPB)* and in Table A.15 for *(PSEPARATE,OneHap)*. 77

4.1 **Inference for short MCMC run:** a list of the median and lower/upper quartiles for the breed proportion estimates for all true underlying ancestral levels. The results show that the ULP proposal mechanism leads to less underestimation than for BSLP updates. Furthermore, re-scaling estimates from UCBC to CCBC shows there is a considerable exploration bias for the ancestral breeds, i.e. [3,15] percent for ULP and up to [2,17] percent for BSLP. 88

4.2 **Inference for short/long MCMC run:** a list of the median and lower/upper quartiles for the breed proportion estimates for all true underlying ancestral levels. The results show that a longer MCMC run further reduces underestimation by about to 1 to 3 percent. Furthermore, for the long MCMC run the BSLP update proposal yields almost identical estimates to the ULP update which suggests that the breed-biased update mechanism directed proposals over larger parts of the breed space in the additional iterations. However, due to limited hierarchical population structure the uniform updates are sufficient to explore the breed space. Full results for all combinations of MCMC run length, update proposals and breed proportion computations are shown in Table A.19. 89

4.3 The breed intervals for each TAP level for the quantile and density view are shown. For the quantile view I also provide the cumulative density threshold. For example, if a dog has a predicted breed contribution in the interval [0.17,0.29] I assign it to the TAP class grandparent. Furthermore, the upper bound 0.29 for the grandparent level is associated with the 86-th quantile and 0.14-th quantile for the cumulative density of grandparent and parent, respectively. Note the results for the quantile and density views are slightly different because for the quantile I chose a 200-quantile as approximation while for the density view I selected the default option of machine precision. 91

4.4 **Inference for ULP update proposal within long MCMC run:** simulation dataset OrgSynRecC1 yields similar median ggp estimates when compared to Mars' dataset OrgSyntheticRed. Furthermore, DBAncestry is robust towards recombination with roughly equal medians but slightly smaller lower quartile estimates. 92

5.1 A summary of the reviewed ancestry inference techniques is shown. In particular, listed types are based on regression in Section 5.1.1, global model-based clustering in Section 5.1.2, local window-based approaches in Section 5.1.3, hidden markov models in Section 5.1.4, non-Parametric Bayesian approaches in Section 5.1.5, PCA-based approaches in Section 5.1.6 and Machine Learning (ML) based approaches in Section 5.1.7. For each of these methods I show whether this technique is applicable for more than two populations. I also list whether a technique deals with correlation. In particular, for regression techniques I list whether correlation in predictor, response space or both is modelled. . . 96

5.2 **CCBC breed proportion estimates for the ChromoPainter standard and NNLS variant:** The median, lower and upper quartile estimates of the re-scaled (CCBC) breed proportion for all possible true ancestral levels are shown. Breed fraction estimates are computed taking predictions from all lineage trees into account. Further visualization and discussion of these estimates is given in Figure 5.3. 120

A.1 **Breed overview**: for the small and big pure breed dataset I show the used breeds. The big dataset covers all breeds while the small dataset covers those 34 breeds ranging from Afghan Hound to Bullmastiff. The breed identifier is shown in the first column. The fourth column has the total number of dogs n_b per breed. The second column 'Small' shows the number of training dogs for the small dataset, and column 'Big' the number of training dogs for the big dataset. 129

A.2 Breed Merging Part I: in the first two columns I show the IDs and names for the merged breed. Then, similarly in columns 5 and 6 I show the IDs and names of the pure breeds. In the centre columns 3 and 4 I show which pure breeds correspond to the merged breeds in the respective table row. 130

A.3 Breed Merging Part II: continued, see Table A.2 for a discussion. 131

A.4 Breed Merging Part III: continued, see Table A.2 for a discussion. 132

A.5 Values of repeats ρ, cycles π different pattern structures based on the ambiguous marker type based on $X_k^j = [0, ?, 0, 1, 2, ?]$. 133

A.6 Each cell in the table refers to $1 - \bar{q}_{1:B}$, i.e. **one minus the breed-averaged** mean posterior probability assigned to the correct breed for the small dataset based on **separate** and **pooled phasing**. I use **DFPB, PSEUDO and OneHap** to define $p(X_c|\theta)$. I use $B' = 34$ breeds. 135

A.7 Summary of $1 - \bar{q}_{1:B}$, i.e. **one minus the breed-averaged mean posterior probability assigned to the correct breed**, and the number of misclassifications for the **big dataset** based on **separate phasing**. Prediction misclassification is out of the total number of test dogs which is 1302. I use **DFPB and OneHap** to define $p(X_c|\theta)$. use $B = 149$ breeds. 135

A.8 These are the results for the **small pure breed dataset** based on **separate phasing**. I use **DFPB** to define $p(X_c|\theta)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d. 136

A.9 These are the results for the **small pure breed dataset** based on **pooled phasing**. I use **DFPB** to define $p(X_c|\theta)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d. 136

A.10 These are the results for the **small pure breed dataset** based on **separate phasing**. I use **OneHap** to define $p(X_c|\theta)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. We only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d. 137

A.11 These are the results for the **small pure breed dataset** based on **pooled phasing**. I use **OneHap** to define $p(X_c|\theta)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d. 137

A.12 These are the results for the **small pure breed dataset** based on **separate phasing**. I use **PSEUDO** to define $p(X_c|\theta)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d. 138

A.13 These are the results for the **small pure breed dataset** based on **pooled phasing**. I use **PSEUDO** to define $p(X_c|\theta)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d. 139

A.14 These are the results for the **big pure breed dataset** based on **separate phasing**. I use **DFPB** to define $p(X_c|\theta)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B = 149$ breeds (including those ones not shown) for a fixed value of d. 140

A.15 These are the results for the **big pure breed dataset** based on **separate phasing**. I use **OneHap** to define $p(X_c|\theta)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B = 149$ breeds (including those ones not shown) for a fixed value of d. 140

A.16 Breed proportions of all three quartiles for all 11 family trees are shown. The first 5 columns correspond to uncorrected while the last columns represent CCBC breed proportions. 141

A.17 HMM example forward probability for genotype [1,2,0] 144

A.18 Confusion matrix is given as contingency table and shows the relationship between true and predicted class. 147

A.19 **Breed proportion estimates:** Median, lower and upper quartile estimates of the raw (UCBC) and re-scaled (CCBC) breed proportion for all possible true ancestral levels are shown. In this experiment I look at combinations of short/long MCMC run-length and uniform (ULP) vs. breed proposal biased breed (BSLP) proposals. Breed fraction estimates are computed taking predictions from all lineage trees into account. 149

A.20 **Classification view criterion area under the curve (AUC):** TAUC values which were obtained from numerical integration are shown. AUC values evaluated for each lineage tree using UCBC and CCBC breed proportion estimates. Although according to Table 4.2 the long run only yields modest improvements in breed proportion estimates there is a considerable performance enhancement to correctly classify breeds. In particular, for complex lineage trees 5 to 11 there is an improvement from 0.05 to 0.1 for the AUC criterion. Furthermore, the finding from Table 4.2 are confirmed that for the long run performance of the ULP and BSLP updates are almost identical. Full AUC results listing the first four leading digits are shown in Table A.21 for all combinations of MCMC run length, update proposals and breed proportion computations. 152

A.21 **Area under the curve (AUC):** AUC values which were obtained from numerical integration are shown. AUC values evaluated for each lineage tree using raw (UCBC) and re-scaled (CCBC) breed proportion estimates. In this experiment I look at combinations of short/long MCMC run-length and uniform (ULP) vs. breed proposal biased breed (BSLP) proposals. 153

A.22 **Maximum value for F_1 score (maxF1):** Values for maxF1 values are shown which were obtained from numerical integration. The maxF1 values evaluated for each lineage tree using raw (UCBC) and re-scaled (CCBC) breed proportion estimates. In this experiment I look at combinations of short/long MCMC run-length and uniform (ULP) vs. breed proposal biased breed (BSLP) proposals. 153

Chapter 1

Background ancestry testing and canine research

This chapter forms the foundation for the next chapters by providing background about canine research and ancestry testing. With respect to ancestry testing we do not entirely limit the focus on dogs but also provide some information as it relates to human testing, mainly due to the following reasons: most of the initial work on genetic ancestry testing was developed with humans in mind, and a survey of the literature reveals that at present still most of the research is centered around human genetic testing. Furthermore, inspection of human genetic testing illuminates potential development opportunities for canine testing but also puts a spotlight on their similarities as well as differences.

1.1 Chapter outline and project motivation

PhD project sponsor Mars Veterinary chose UCL as host for this dog breed inference project due to its high-performance computing resources and statistical genetics expertise, e.g. Prof. David Balding is a leading driver in this field. In particular, Mars Veterinary sought the development of improved statistical genetics algorithms which infer the ancestry of crossbred dogs. These novel inference techniques should yield better approaches to estimate breed contributions based on genetic data than currently used by Mars Veterinary.

The main contributions of this thesis are outlined in Section 1.6: Firstly, extending work from Section 1.4.4 in Chapter 2 I perform an explanatory analysis of breed proximity using 2D visualizations to understand how different dogs breeds are genetically. Furthermore, I investigate which proximity measure should be incorporated into the novel ancestry inference algorithm which is developed in subsequent chapters. Secondly, in Chapter 3 it will be shown how my novel simulation-based algorithm for ancestry inference utilizes the correlation structure in the genetic data which Mars did not account for. Moreover, in Chapter 4 I will experiment with either uniform or breed proximity directed proposal mechanisms in my simulation-based algorithm while the original Mars algorithm only uses uniform update proposals. Finally, in Chapter 5 I will compare prediction results using my novel algorithm with an state-of-the-art human ancestry inference algorithm selected after a comprehensive literature review.

Currently, Mars commercially distributes canine direct-to-consumer (DTC) genetics inference tools for purebred and crossbred dogs. These DTC kits which are sold by Mars using the label Mars Wisdom Panel are described in more detail in Section 1.2. Mars was motivated to incorporate ideas and code from this research project into their DTC kit to obtain better canine ancestry predictions which then may lead to further customer satisfaction, fewer client complaints and better reputation in the DTC market.

Customer of Mars Wisdom Panel are typically private users who are either curious about the composition of their crossbred dog, would like to verify that their dog is indeed purebred or commercial customers involved in dog breeding (see Section 1.2 for further details). In Section 1.3.3 I will describe further application areas relevant to to Mars Wisdom Panel, such as population genetics, health and disease analysis, forensic applications and public safety concerns due to canine attacks on human which are relevant to local councils and insurance companies.

The next question reviewed in this chapter in Section 1.3.1 relates to the data sources which may be used for ancestry prediction. Data sources I will discuss include phenotype, language, isotope data and genetic data, such as uniparental inherited lineage-based haploid markers, autosomal DNA markers and short tandem repeats. After that in Section 1.3.2 I will outline why autosomal DNA from Mars Wisdom Panel was used in this project. In Section 1.3.4 I argue that it should not be expected that DTC kits yield a definite answer to ancestry inference because data sources only provide partial information from an individual's whole ancestry. Furthermore, quality of prediction also depends on the quality of the data (e.g. number of DNA markers and their informativeness for breed separation), the way samples were selected and how ancestry is defined. Briefly, I also will review how ancestry testing is related to the definition of identity and legal implications.

In Section 1.4.1 I briefly describe canine history starting from their original domestication, breed creation during the Victorian era up to the dog genome project. After that, in Section 1.4.2 I define haplotype and SNP marker data. Then, in Section 1.4.3 I discuss linkage disequilibrium (LD) which defines the correlation structure in the marker space. In my novel algorithm I incorporated LD information while the current Mars Wisdom algorithm does not directly account for LD in the SNP data. LD is a differentiator for ancestry inference algorithm given that high correlation in genome extends 10-100 fold longer in modern dogs than in humans. Although LD information could not be fully exploited in this study due to sparseness of the markers in the genetic dataset. Due to these comparably high LD values and consequently fairly inbred nature of dogs fewer SNPs are required in dogs to achieve similar levels of predictions as in humans.

In the final Section 1.4.4 of this review I look at breed separation to judge how well breeds are expected to be discriminated among each other. For this purpose, I look at the within- and between breed variability for lineage-based and autosomal markers. Furthermore, I discuss marker differentiation between specific breeds and whole population for a range of marker types and selection of breed groups.

The next section deals with the data from this ancestry inference project. Firstly, in Section 1.5.1 I discuss how markers were selected in this project and how they are distributed across chromosomes. Secondly, in Section 1.5.2 I discuss the datasets analyzed in this project. As first step in Section 1.5.3 I look at the purebred data collected by Mars. After that in Section 1.5.4 I show how Mars' synthetic test dogs were designed to measure prediction quality of the ancestry inference algorithms. In particular, test dogs are chosen from different ancestry complexities ranging from purebred parents to dogs which have eight different great-grandparents. However, these Mars' synthetic test dogs only model recombination up to the chromosome level. Therefore, in Section 1.5.5 I show how I simulated a more complex synthetic test dataset which incorporates recombination with rates matching real canine genomes.

Finally, as last part in this chapter in Section 1.6 I will explain the main novel contributions in this

thesis, provide outlines to the further chapters of this thesis and offer the chapter conclusions.

1.2 Introduction to DTC kits

In the last 15 years interest in human ancestry and genealogical research increased rapidly covering techniques to infer population histories, exploring genealogy and estimating individual ancestry proportions (Shriver and Kittles, 2004). More specifically, DTC genetics inference tools have been used to predict individual ancestral proportions, i.e. population groups, such as Sub-Saharan African, Scandinavian, Finnish or East Asian. Furthermore, DTC genetic tests are utilized to offer interpretation in the area of oncology, rare or undiagnosed disease and pharmacological drug response (The American Society of Human Genetics, 2008; Duster, 2009; Allison, 2012; Wagner and Weiss, 2012; Wagner et al., 2012). Recently, the DTC market matured due to two main factors (Lee et al., 2009; Bloss et al., 2011; Allison, 2012): firstly, DTC products were put under US governmental regulation by the Food Drug Authority (FDA) which led some of the test providers to cooperate with physicians to administer those tests. Secondly, due to dramatic price drop-off in whole-genome sequencing DTC testing is much more affordable and accessible to the average person.

In the last few years, DTC technology was adapted for the interrogation of canine ancestries to detect recent ancestry of a mixed-breed dog, i.e. to infer purebred contributions within a crossbred dog going back to the great-grandparent level (Mars Veterinary Wisdom Panel, 2013). In the US a large part of the canine population is formed by mixed breed dogs, also known as mongrel (American Veterinary Medical Association, 2002; Parker, 2012). For example, a crossbred dog could have a breed composition based on a three-way admixture of one parent and two grandparents with German Shepherd 50%, Bernese Mountain Dog 25% and Siberian Husky 25%. Some of the reasons for canine DTC testing are very similar to those in humans (Mars Veterinary Wisdom Panel, 2013; DNA MY DOG, 2013; Dog DNA, 2013): firstly, there are a large number of crossbred dogs with complex ancestries which are hard to trace back to purebred ancestors. Often, these crossbred dogs lack official recording of the sire, dam or grandparents. Therefore, those dog owners are interested in DTC tools which either infers the breed composition of their mixed bred dog, or validates that they have a purebred dog or designer dog which is formed by a 50/50 hybrid of two purebred parents. Secondly, the inferred individual ancestry proportions are a pointer to specific genetic health risks associated with these specific breeds (Sutter and Ostrander, 2004; Boyko, 2011).

DTC kits are utilized for the selection by breeders who are interested in optimizing certain parameters, such as puppy health, nutrition and canine veterinary drugs (Mars Veterinary Wisdom Panel, 2013). Even when your dog is still a puppy these DNA services offer prediction service to estimate adult values for weight, height and dietary nutrition volumes.

Patterson (2000); Calboli et al. (2008); Leroy and Baumung (2011); Leroy (2011); Shariflou et al. (2011); Bateson and Sargan (2012) showed that many breeds are dominated by popular sires because these dogs often have traits assumed to be desirable for the given breed which makes them a popular choice for frequent use in breeding. Therefore, some of these offspring may be paired with each other which increases inbreeding leading to less diversity. Inbreeding and small population sizes are relevant factors because they increases the chance of birth defects and other inherited health problems.

1.3 Data sources, applications and limitations of genetic testing

In this section I discuss data types which are predictive of ancestry, applications and limitations of ancestry inference tools.

1.3.1 Data sources

This section outlines data types which may be utilized to estimate human and canine ancestry proportions.

- **Phenotype**: absent other further information morphological characteristics have been used for immediate evaluation of dog breed composition or human origins. The following account shows that phenotypic self-assessment has limited accuracy.

    1. **Dogs**: previously, ancestor breeds of a given dog have been predicted based on phenotype. Dogs are of a wide range of physical appearance such size, weight, coat type (colour, length), skull type, leg length and tail shape. This morphological diversity may have been developed through multiple breeding options. Dogs who exhibit similar appearance might not be very similar for other traits, such as behaviour or certain chromosome regions. On the other hand, dogs with similar breed compositions may look quite different depending on which of the dominant and recessive genes got passed by its ancestors. Furthermore, some breeds show strong intra-breed variability for certain characteristics. Sacks et al. (2000) state that even among experts there is no agreement on the breed for a particular dog.

       Research by Voith et al. (2009) shows that phenotypic breed classification of purebred dogs carried out by shelter staff is unreliable when compared to the Mars product, Mars Wisdom MX, to infer breeds constituting the ancestry of a mixed breed dog. Poor human breed prediction may be due to limited feedback on the correct breed composition shelter staff obtain. Furthermore, mixed breeds are harder to predict because they exhibit more genetic variation than pure breed dogs, e.g. those mixed breed dogs show characteristics which may be common to a variety of pure breed dogs.

    2. **Humans**: description of ancestry in terms of race, ethnicity or parental origin countries can be either self-reported or be identified by an observer (Via et al., 2009). Burnett et al. (2006) performed a study on Caucasian sibling subjects which were independently asked to specify mother's and father's countries of origin. The study showed that self-reported ancestry has limited reliability due to major disagreement, especially when one of the parents had more than one country of origin (up to four answers). For example, only 68% of siblings provided an exact match for both parental countries of origin. Similar biased results hold for self-reportin in African populations (Yaeger et al., 2008) and native American ancestry in Hispanics (Klimentidis et al., 2009).

- **Language**: language analysis has been used by the UK border agency to distinguish members of different countries, such as the discrimination between likely asylum seekers from Somalia (who speak Somali and Arabic) and opportunistic economic advantage seeking migrants from Kenya (who speak Kiswahili and English) (Balding et al., 2010). However, based on the foreign language capabilities of an individual is not always straight forward to infer origin.

- **Genealogical records**: genealogical records form the foundation for the study of family history.

1. **Dogs**: genealogical analysis has been applied through pedigree analysis and some complete genealogies can be traced back to 1960s and 1970s (Baumung and Solkner, 2003; Leroy, 2011). But these records tend to be limited because dog lovers actually need to register their dogs, such as the Kennel clubs, to be added to the database. In this case I know its breed because breed records require parental breed registration certificates (Crowley and Adelman, 1998). At other times written records are limited because have not been transferred to computers while other breeds were just founded recently.
2. **Humans**: research into human family history and tracing of one's own lineage is becoming increasingly popular (Shriver and Kittles, 2004; Duster, 2006). Common genealogical records include family trees, trusts, estates, wills, church records, immigration records, criminal records, military records, census records and community records about birth, marriage and death. In the case of slavery sales record may be of relevance (Duster, 2006). Shriver and Kittles (2004); Larmuseau et al. (2013) also discusses the case of surname matching to introduce previously unconnected relatives.

- **Stable isotopes**: according to Balding et al. (2010) stable isotope analysis has been championed in archeology to distinguish locals from non locals. The method is based on specifying a signature which is defined as the ratio of two isotopes of an element. Animal and human consume food and drinks which leads to a record of signatures in tissue types, such as bone, teeth, hair and nail. Isotopes can be used for different purposes, e.g. element Strontium can be used to date geologies of different ages while other elements, such as oxygen and hydrogen, are derived from water and are utilized as proxy for environmental and climatic effects. According to Cannon et al. (1999) dog bones have been used to extract signatures for carbon and nitrogen which hints at food patterns, i.e. consumption of certain fish types or uptake level and reliance on marine sources of proteins.

  However, there are some limiting factors to this approach. Firstly, there is incomplete sampling of isotope signatures across the world which leads to a lack of comparative values to spot locations which have matching signature values. Related to the last point, lack of fine granularity in sampling global signatures leads to inaccuracy of pointing to a specific location. Furthermore, certain isotope signature exhibit limited variability over large areas while other signatures vary over smaller distances but may show similar values in very remote regions of the world. These signatures have been used by the UK border agency which sampled tissue types, such as hair and nails, of asylum seekers but those tissue types have signatures which only record the immediate past of the last few months. Finally, isotope analysis has been developed for the recognition of outlier detection in the form of non-locals but not for the inference of origin of non-locals.

- **Lineage-based analysis**: lineage-based analysis studies a single maternal or paternal lineage by looking at uniparental inherited haploid markers, i.e. the maternally, mitochondrial DNA (mtDNA), and paternally, Y-chromosome (NRY), inherited DNA (Shriver and Kittles, 2004; Sarata, 2008; Via et al., 2009; Balding et al., 2010; Royal et al., 2010). In other words, due to this focus on a single maternal or paternal lineage, the contributions of a large set of an individual's ancestors to their genome are not accounted for. The mtDNA is passed from the mother to all children but male children do not transmit to their own offspring, whereas the Y-Chromosome is only passed from father to son. However, lineage-based testing only composes a small amount of an individual's total DNA, i.e. less than 1% of the entire human genome, which provides a limited view of one's ancestry (Shriver and Kittles, 2004; Bolnick et al., 2007). In particular, by

examining only the maternal and paternal lineage and assuming g past generations I focus on only $2/2^g$ parts of our biological contribution (assuming distinct ancestors), e.g. 1/2 for grandparents and 1/4 for great-grandparents. Lineage-based data is characterized by a lack of recombination and a slow mutation rate (Shriver and Kittles, 2004; Budowle and Van Daal, 2008).

Lineage-based analysis is used to infer population demographic events, such as migrations (Shriver and Kittles, 2004; Balding et al., 2010). At an individual level, a match of lineage-based data with reference samples in the database shows the existence of a recent common ancestor (Royal et al., 2010). But it is hard to argue which individual migration events gave rise to this match and how many generations ago the shared common ancestor lived. For example, van Oven et al. (2011) used mtDNA to infer the maternal ancestral origin at a continental level. Furthermore, this data type cannot determine kinship or ancestral origin precisely, but only identify a place where people with similar DNA types currently live (Via et al., 2009).

- **Autosomal DNA markers**: multiple autosomal DNA markers yield an average estimate of the ancestry fractions for each of the lineages contributing to the genetic makeup of an individual (Bolnick et al., 2007; Via et al., 2009). In particular, autosomal markers offer much more comprehensive insight into ancestry because cumulatively they act as proxy of a much larger part of the genome history, i.e. many biparentally inherited loci versus a single uniparental marker, such as mtDNA or Y-chromosome (Royal et al., 2010).

  However, information obtained from autosomal markers is also limited because of its finite size. Due to recombination and mutation events the inherited ancestral segment length from any particular individual halves every generation and renders its impact rapidly negligible. In other words, an ancestor does not pass on all of his or her genomic segments which leads to a limited contribution of ancestors to the particular genome of a descendant.

  Autosomal variation can be either measured by short genomic sequences formed by ancestry informative marker (AIM) panels of up to 1,5000 to 5,000 SNPs, or by large marker sets obtained from whole-genome genotyping (Seldin, 2007; Seldin et al., 2011; Churchhouse, 2012). However, whole-genome sequencing still has a considerable cost associated with it which makes genotyping of subsets of markers found in AIM panels appealing.

  Ancestry inference of personalized genetic histories using AIM panels is referred to as biogeographical ancestry (Shriver and Kittles, 2004). For each marker, its binary state on each chromosome is called an allele. The number of markers is a an important factor for ancestry inference (Tsai et al., 2005; Aldrich et al., 2008; Via et al., 2009). AIM panels are composed of population-specific alleles which were selected to yield maximum informativeness with respect to robust, discriminatory power between two chosen populations (Shriver and Kittles, 2004). Rosenberg et al. (2003) reviews different criteria to select ancestry informative markers, such as absolute difference in allele frequency δ (Campbell et al., 2003) or the fixation index F_{st} related to the variance of allele frequencies (Weir and Cockerham, 1984; Excoffier, 2001). Other approaches for marker selection include sparse regression or variable selection techniques (Zhang, 2010; Lee et al., 2012).

  Typically, DNA can be obtained either from cheek or blood samples, but due to technology advances cheek cells are the preferred method for human and canine applications. In particular,

humans donate saliva samples while for canine applications cheek cells are extracted from the mouth using a cheek swab.

- **Microsatellite**: short tandem repeat (STR) or microsatellite loci are abundant molecular markers in the human genome (Budowle and Van Daal, 2008). STRs are formed by tandemly repeated sequences where each sequence is composed of about four DNA bases. An allele corresponds to the number of times a sequence is repeated. For example, in forensics common allele counts are 5 to 20 and STRs are used for kinship detection among other applications.

1.3.2 Local and global ancestry estimation

In this thesis I perform ancestry inference using autosomal markers, which is mainly due to the following two reasons:

- **Ancestry informative**: autosomal markers are inherited from both parents, i.e. the maternal and paternal lineages are equally represented in the genomic sequence while lineage-based haploid groups only represent one particular maternal or paternal lineage.

- **Mars Veterinary**: a very practical reason why I use autosomal markers is that Mars Veterinary provided us with dog genetic sequences from a wide range of breeds.

The focus of ancestry inference can be either local or global level (Alexander et al., 2009; Churchhouse, 2012). In the local mode I imagine that the genome is split into chunks of one or more SNPs of a definite ancestral population. Then, the aim at the local level is to infer the chunk boundaries of these segments and to assign the supposedly ancestral population. At the global level I estimate the overall ancestry contribution of each population as an average of an individual's genome. My emphasis in this study is at the global level because my aim is to infer breed contributions from mixed-breed dogs. For the purpose of identifying distinct clusters and assigning individuals to them, Alexander et al. (2009) further dichotomized the global approach into model-based and algorithmic ancestry estimation. In the model-based approach it is assumed that I jointly estimate parameters, such as population of origin and admixture proportion, and infer cluster membership for each individual (Pritchard et al., 2000). In the algorithmic view multivariate techniques, such as cluster analysis and principal component analysis, are utilized to visualize data in a non-parametric way.

In general, admixture modeling deals with the local level to answers questions related to admixture proportions, ancestry for each marker and number of generations since admixture whereas my focus is on global estimation of population proportions (McKeigue, 2007).

1.3.3 Applications

In Section 1.2 I reviewed motivations for the purchase of DTC kits by actual and potential private users. However, inference of ancestry and genetic tests also form a crucial constituent in academic research, clinical decision-making in a medical context, forensic practice and public management with regard to dog safety.

- **Population genetics**: ancestry estimation is a relevant topic in population genetics which interrogates genetic data to reveal insights about the demographic history, between-population relationships, population structure and migration patterns (Falush et al., 2003; Hellenthal et al., 2008; Vonholdt et al., 2010; Henn et al., 2012). For example, this type of approach enables inference about the contribution of admixed population contributions as well as how far back in time admixture

events took place. Furthermore, ancestry estimation approaches help to compute recombination maps to study the loci of rapid evolutionary change (Price et al., 2009).

- **Health implications**: human health practitioners investigate how racial/ethnic identity contributes to differential health results, i.e. how ancestry is related to environmental and genetic risk factors (The American Society of Human Genetics, 2008). In other words, health-related traits are correlated with specific genetic variations which are more prevalent in a particular ancestral group than another one. While differential health outcomes are explained by a modest impact by genetics those outcomes also tend to be heavily determined by environmental covariates, such as diet, education, access to health care and social economic class.

  Case-control genetic association studies and admixture mapping are two main approaches how ancestry information is utilized to detect genetic risk factors for disease (The American Society of Human Genetics, 2008; Via et al., 2009). Recently, Chang et al. (2009) also looked at how canine case-control GWAS studies can be confounded by population structure giving rise to spurious correlation of phenotype and genotype. However, the focus of admixture mapping is not infer breed contributions as in my project but local estimation to unravel genomic regions containing population-specific risk alleles for disease.

- **Forensic applications**: there are a variety of forensic applications, such as detecting identity in criminal cases, mass disasters or from human remains, prediction of general phenotypic appearance of certain very heritable traits (e.g. coloring, height and facial feature), paternity testing, kinship analysis and inheritance related questions (Budowle and Van Daal, 2008).

- **Public safety concerns about dog attacks on humans**: given the event of safety concerns due to dog attacks, ranging from dog bites (Harris et al., 1974) to fatalities, on humans government bodies, housing associations and insurance agencies in certain countries, such as Canada (Dog Owners' Liability Act, 2005) and UK (Dangerous Dogs Act, 1991), enacted breed specific legislation (BSL) which place restrictions, such as public muzzling, registration with local authorities, or even banning (Ledger et al., 2005). Owners who do not comply with this legislation are penalized with fines and jail sentences. Affected breeds often include those which are aggressive and suitable for dog fighting, e.g. American pit bull terrier and Argentinian Dogo Mastiff.

  Secondly, there are no clear standards, as to which breeds are considered to be dangerous. Legislator use metrics, such as bite incidence, to measure breed danger. Thirdly, the ability of victims, shelter employees and representatives of other organizations to accurately identify canine ancestry composition is of utmost importance. Wrong prediction of breeds may lead to inappropriate dealings for certain breeds which are mistaken for restricted breeds. On the other hand, banned breed dogs may not be recognized and seized by local authorities because phenotypic variation within breeds is not well understood, and some breed owners circumvent breed banning by creating mixed breed dogs with a high proportion of the actually desired breed. For optimal BSL enforcement an objective method for determining the breed of a dog is required, such as DNA testing, because breeds have to be identified with high certainty.

1.3.4 Limitations

Although human and canine DTC-based test kits are offered by a variety of commercial providers these tests have a range of limitations which will be discussed in this section. DTC test user should not expect

a definite answers because the genome only carries partial information on an individual's whole ancestry. Furthermore, estimation of these ancestral proportions is often performed using statistical methods which are uncertain in nature and may lead to wrong inference (Bolnick et al., 2007).

- **Definition of ancestry**:

  1. **Dogs**: Kennel Clubs specify breed membership based on parental breed registration certificates. This recursive definition goes back to its founders leading to closed breed populations (Crowley and Adelman, 1998).
  2. **Humans**: concept of ancestry is much harder to define for humans and can be accomplished using different criteria assessing ancestry at various levels (The American Society of Human Genetics, 2008; Via et al., 2009; Lee et al., 2009).

     Firstly, with respect to time, I could either refer to immediate ancestors, such as parents or grandparents, or going back to earliest hominids or first homo sapiens. However, going back in time each individual has hundreds of ancestors for just a few centuries and due to the fixed genome size only a fraction of those ancestral contributions are recorded. Although going back many generations may have limited interest for private DTC kit users, it is very important to study past migrations and genetic variation from a population genetics perspective. Secondly, ancestry may be interpreted as matches against the reference populations in the database by DTC companies but a private user may equate ancestry with place of birth. Some private user may even interpret ancestry as location where their last name, identity, family narrative, language, religion or culture originated although reference populations are only a proxy of current but not ancestral locations. Finally, ancestry is also relative to size of area, such as specific to villages, countries, continents.

- **Data quality**: as discussed in Section 1.3.1 the quality of the ancestral proportion estimates is strongly influenced by the number of markers and their informativeness.

- **Data source**: most ancestral test either used lineage-based or autosomal markers (Bolnick et al., 2007). In Section 1.3.1 I explained that autosomal markers yield an average view on an individual's genome for ancestral inference while lineage-based markers only represent a minor part of an individual's genome contributing only one maternal and paternal ancestor per generation. Wagner et al. (2012) showed that for human DTC test kit providers that each mtDNA, NRY and autosomal test have similar median price while a combination of mtDNA and NRY is slightly more expensive while combining all three information sources is most expensive.

- **Incomplete sampling**: to represent a population well I need a large number of samples to represent its genetic diversity (The American Society of Human Genetics, 2008; Sarata, 2008). However, I encounter sampling bias where many of these samples are not random samples from the population of interest but convenience samples which were derived from published research or sampled only in certain geographical regions (Sarata, 2008; Balding et al., 2010). For example, if a population is not well represented by its constituting samples, a statistical methods may wrongly increase proportions from the wrong population to the admixture estimate (Royal et al., 2010). A related factor is the amount of within- and between variability among the reference populations which puts a limit on how well those populations can be possibly discriminated. Furthermore, the number of populations is also of marked relevance because it may represent a purebred dog

as admixture of its genetically closest populations. Finally, the investigator needs to judge how to deal with samples which are recent migrants to the target population or are admixed individuals themselves, i.e. whether to exclude those samples from the analysis (Balding et al., 2010). Within the context of forensic applications, Balding and Nichols (1994) showed that an innocent suspect may appear more likely to be the criminal due to difference in the suspect and reference populations in the database, i.e. the innocent suspect's alleles may overlap with the criminal due to kinship, resemblance of physical appearance and genetic composition of people living in his or her area.

- **Proprietary confidentiality in DTC testing**: DTC test providers are very interested to protect their corporate intellectual property and operations, such as statistical techniques used with their assumptions, population sampling technique and size and composition of their reference populations although 'The American Society of Human Genetics' asked for more transparency (The American Society of Human Genetics, 2008; Duster, 2009). Therefore, way of communication of test result to customers is of paramount relevance because due to lack of details test details results may not be fully understood or open to independent replication. Furthermore, due to different implementation of those details by companies I do not always expect consistent results across corporate DTC providers for the same individual (Sarata, 2008; Duster, 2009). With respect to samples, there is little shared use of genotype individuals across different DTC providers. However, some special databases for specific populations, such as West Africans and Native Americans exist (Shriver and Kittles, 2004), are available.

  With respect to canine testing, typically no further details on the computation on breed proportions are provided except that a statistical algorithm is used to compare test dogs with a reference database composed of purebred ones. If information on SNPs is given their number is typically in the hundreds (Mars Veterinary Wisdom Panel, 2013; DNA MY DOG, 2013; Dog DNA, 2013).

- **Identity conflict**: unexpected or undesired results of ancestry testing could lead to psychological responses corresponding to emotional distress (Nordgren and Juengst, 2009). Furthermore, canine testing result may change how owners view their dog. With respect to human testing, customers may question their ancestral history and change how they report ethnicity in the future (Bolnick et al., 2007; The American Society of Human Genetics, 2008; Wagner and Weiss, 2012). Furthermore, identity is often connected with race which as political sensitive term may invoke particular positive or negative associations.

- **Legal implications**: group membership has a variety of legal implications: preferential admission to universities, acceptance of job applications, access to social welfare and health-care, and utility for immigration purposes, such as seeking dual citizenship (Bolnick et al., 2007; The American Society of Human Genetics, 2008). For example, Native American tribal affiliation leads to benefits, such as financial help, housing support and health-care. In the context of canines samples dog owners may dispute whether a dog is purebred, or appeal to the claim that a designer dog is composed of equal breed proportions of the purebred parents. Finally, with respect to breed specific legislation owner may either prove or disprove that their dog is part of the list enumerating either banned or restricted breeds.

1.4 Review canine research

In this section I give some background about work related to canine data in statistical genetics. Firstly, I start with a historical account of the development of dogs. Then, I summarize a few more details about autosomal SNP markers and how they are related to a haplotype representation. After that I look at between-population variability which is related to the theoretical limits of how well an algorithm possibly could discriminate the reference populations. This same section also looks at a measure of correlation between autosomal markers which indicates high canine within breed genetic similarity. Finally, I show how breeds cluster using genetic distance.

1.4.1 Canine history

The dog family, Canidae, has about 34 species, such as different types of foxes and jackals, which diverged about 10 million years ago (Wayne et al., 1997).

Domestic dogs can be traced back to the original domestication from the grey wolves in East Asia around 15K to 100K years ago (Vilà et al., 1997; Savolainen et al., 2002; Ardalan et al., 2011; Ding et al., 2011; Wayne and vonHoldt, 2012). Fossil discoveries show that humans and dogs share a long history, such as sharing home and food sources, going back to caves in Belgium 31K years before present (Germonpré et al., 2009). Pang et al. (2009) use mtDNA to date domestication in Southern China between 5.4K to 16.3K years ago from at least 51 female wolf founders. The transformation from wolf to domestic dog were likely to be accompanied by changes in the dog's environment, food source and human adapted behavioural patterns. Until recently, given the high phenotypic breed diversity, dogs were assumed to be an admixture of several canid species, such as jackals, coyotes and wolves. However, recent analysis using mtDNA data provided evidence that gray wolves are the sole ancestor of modern dogs (Wayne, 1993; Vilà et al., 1997).

More recently, many of the domestic dog breeds acknowledged by national fancier clubs, such as Kennel clubs, are based on human breeding efforts in the Victorian era (1837-1901). In this Victorian period many new modern breeds were defined through backcrossing and inbreeding to yield specific phenotypic and behavioral traits which leads to a loss of genetic variability (Parker et al., 2004; Hedrick and Andersson, 2010). This genetic variability reduction was reduced by dog fanciers' restricted breeding preferences and rules which specified to register a dog as a member of a breed it is required that its parents also have been registered as members of the same breed (Parker, 2012). Although these breeds still show some within-breed variability there are some chromosomal regions related to certain traits which are fixed or near to fixation (Lindblad-Toh et al., 2005; Akey et al., 2010; Vaysse et al., 2011).

On the other hand, some breeds, such as greyhound and pharaoh hound, may be very old breeds or re-creations thereof. According to the Kennel Clubs breeds form a closed population (Crowley and Adelman, 1998). Consequently, those purebred dogs have lower levels of genetic heterogeneity than mixed bred dogs. Furthermore, this selective breeding of purebred leads to huge differences in morphology, e.g. a Great Dane is 40 times the size of a Chihuahua.

In the 1990s, the genetic study of disease, morphology and behavior were the central motivation to establish the 'dog genome project' and to detect genes which may be leading to disease. Geneticists aimed to sequence a dog's 38 autosomal pairs as well as the sex chromosome. Genomes are commonly genotyped in many long overlapping segments, and to increase the probability that the whole genome is genotyped at least once, estimations suggest that there should be about 7 to 8 genome reads across

the entire genome. In 2003, a standard poodle was genotyped to yield a genome coverage of about 80 percent for its about 2.8 billion base pairs (Kirkness et al., 2003). Subsequently, the first fully genotyped dog was a member of the boxer breed whose genome has been genotyped to cover about 99 percent of the genome (Lindblad-Toh et al., 2005). The number of genes is estimated at about 19,000 genes, and for 75% of these genes the similarity due to shared ancestry between dog, human and mouse is very high (Ostrander, 2007). Furthermore, genome wide association studies suggest that breed standards caused a stronger differentiation at loci which are due to morphology than other breed-specific measures which are more difficult to quantify, such as susceptibility to disease, longevity and certain behaviors (Boyko, 2011).

1.4.2 SNP marker data

In this document I model variation in the canine genome using autosomal single nucleotide polymorphisms (SNPs) markers along a chromosome which represents the most commonly studied class of sequence variation data (Suh and Vijg, 2005).

With respect to humans, large amount of data has recently been made available in the HapMap and 1000 Genomes projects (Gibbs et al., 2003; Altshuler et al., 2010). SNP data has been adopted by evolutionary biologists who are interested in inference of population history (Brumfield et al., 2003). Furthermore, SNPs proved to be a powerful data source to reveal markers and genetic processes related to trait inheritance and study of the genetic bases of disease (Gray et al., 2000; Marchini et al., 2004; Altshuler et al., 2008).

A SNP describes genetic variability at an individual marker position (locus) where a mutation has occurred in the evolutionary past of a species. In particular, constant loci which only assume one value do not provide information on modeling variability. A locus can attain four different values which are known as alleles taken from the quaternary vector (G, A, T, C). Markers which take three or more values are rare. Most SNPs are biallelic, i.e. only take two different different values, and form the focus in our work. Then, these two possible nucleotide base values are encoded binary as $\{0, 1\}$ and each variant is referred to as an allele. Typically, allele 0 refers to the major allele wile 1 corresponds to the minor allele. In other words, allele 0 has a higher frequency of occurrence than allele 1. Based on this encoding a haplotype is formed by a contiguous sequence of s biallelic SNP markers on each chromosome. The number of common canine SNPs has been estimated at around 2 million in total across chromosomes, of which current chips are capable of measuring 127K SNPs (Ostrander, 2007).

In diploid species, such as humans and dogs, and individuals have two copies of each chromosome, one for each parent. In this case, maternal and paternal haplotypes $H_M, H_P \in \{0, 1\}^s$ are combined to form a sequence of unordered allele pairs at each locus which is known as genotype (Hodge et al., 1999). In other words, I am uncertain about the loci phases, i.e. based on the genotype I do not know which allele at each locus is derived from which parent. In particular, the sequence of s SNPs $X = [X_1^j, \ldots, X_C^j] \in \{0, 1, 2, ?\}^s$ is referred to as genotype where $X_k^j \in \{0, 1, 2, ?\}^{s_k}, k \in \{1, \ldots, C\}$ is the genotype on chromosome k of length s_k such that $\sum_{k=1}^{C} s_k = s$, C is the number of chromosomes and ? represents a SNP marker which is missing. Due to recombination events these genotypes tend to be composed as a 'mosaic' of longer sub-sequences of the genotype of other individuals in the population (Daly et al., 2001). Furthermore, the genotype is defined based on the parental haplotypes which have to sum up to the respective allele value at each locus t of the genotype, i.e formally it is expressed as $H_M[t] + H_P[t] = X[t] \; \forall t \in \{1, \ldots, s\}$. Genotype alleles $0, 2$ are known as homozygous because it im-

Locus 1/Locus 2	Locus 2 with allele B	Locus 2 with allele b	Marginal counts Locus 1
Locus 1 with allele A	$n_{11} = N(p_A p_B + D)$	$n_{12} = N(p_A p_b - D)$	$n_{1\cdot}$
Locus 1 with allele a	$n_{21} = N(p_a p_B - D)$	$n_{22} = N(p_a p_b + D)$	$n_{2\cdot}$
Marginal counts Locus 2	$n_{\cdot 1}$	$n_{\cdot 2}$	$N = 2n$

Table 1.1: Observed allele distributions under LD.

plies either both 0s or 1s for the respective haplotype values. SNP allele 1 is referred to as heterozygous where one parent assumes value 0 and the other one a 1. The heterozygous case deals with an ambiguous phase because it is unknown which parent attains value 0 and which parent takes value 1. In statistical genetics, phasing (also known as haplotype inference) deals with the detection of the phase at each locus from the genotype. Typically, biological methods to obtain genotypes do not offer phase information and I need to resolve the phase at several heterozygous loci algorithmically. Phase reconstruction can also be achieved using molecular haplotyping (Patil et al., 2001) but it is more expensive than genotyping.

1.4.3 Linkage disequilibrium

The creation of breeds had an impact on the genome structure of the dog affecting linkage disequilibrium (LD). LD is a measures of allelic association between two sites on the genome for n dog individuals. To illustrate this concept, let us assume that allele at locus 1 may take values $\{A, a\}$ and allele at locus 2 may assume values $\{B, b\}$, which have population frequencies p_a, p_A, p_b, p_B (Foulkes, 2009). Given that each dog carries two homologous chromosomes there are a total of $N = 2n$ homologs where n is the number of individuals. The expected distribution of allele pairs for the 2 loci is shown in Table 1.1, and under the assumption of linkage equilibrium I expected the counts shown in this table with constant D assumed to be zero.

This contingency table assumes independence, e.g. joint probability p_{AB} is decomposed as $p_A \cdot p_B$. The degree of independence in this table can be assessed by a Pearson's χ^2 test which compares expected counts with actually observed counts. However, having $D \neq 0$ in this table, I no longer assume independence. I can view D as covariance, or normalized as correlation, given its definition as $D = p_{AB} - p_A \cdot p_B$. Small values of D indicate observed counts close to expected counts and small deviations from linkage equilibrium. Large values for D leads to n_{11}, n_{22} counts greater than expected under independence while n_{12}, n_{21} will take smaller values, indicating larger departure from LD. A commonly used measure for LD is quantity r^2 which is a function of D. This quantity is based on Pearson's χ^2 test for no association of rows and columns and can be expressed as

$$r^2 = \frac{\chi_1^2}{N} = \frac{D^2}{p_A p_B p_a p_b}.$$

Absent of equilibrium in gene linkage, I have LD which means markers are correlated leading to segments of the chromosome which are similar and I denote those ones as haplotype blocks.

According to Falush et al. (2003); Kaeuffer et al. (2007) there are three types of LD types:

- **Mixture LD**: individuals from the same population have correlated markers across the genome due to higher prevalence of certain alleles.
- **Admixture LD**: contiguous segments of markers on the chromosome are correlated where each segment corresponds to a particular ancestral population.

- **Background LD**: correlation among physically close markers due to genetic drift, i.e. random sampling of individuals used for reproduction.

Lindblad-Toh et al. (2005) found that LD within any breeds mostly depends on two bottlenecks. The first bottleneck is an ancient one at the time of domestication (7K-50K generations ago) shared by all breeds while the second one occurred during breed formation in the Victorian era (50-100 generations ago). These bottlenecks had influence on LD and haplotype patterns. The domestication bottleneck can also be seen from the SNP rate: current dog breeds have 1 SNP in about 0.8-0.9 kbp while grey wolves have 1 SNP in 0.58 kbp and coyotes 1 SNP in 0.42 kbp which shows the latter ones exhibit more variability. These two bottlenecks lead to high LD and limited haplotype diversity. Comparably short-range LD extends in humans (4k generations ago) to ancestral haplotype blocks of 20s kbp while LD extends 10-100 fold longer within modern dog breeds; albeit shorter between breeds. On the other hand, ancestral breed founded around 9K generations ago have shorter haplotype blocks of size 10 kbp than humans. This decline in genome variability during the bottlenecks events is due to a selection of subset of dogs for mating leading to certain long-range patterns becoming more common for a given breed and also creating long-range LD. And these long-range patterns, especially since the second bottleneck, have not been shortened much yet due to recombination events.

Lindblad-Toh et al. (2005) measured long-range correlations of SNP markers by evaluating quantity r^2. The authors found that average LD in humans and the overall dog population quickly declines to background level at about 200 kbp. The common decline in LD across breeds shows the still underlying short-range haplotypes of the ancestral dogs. However, LD for individual breeds also decreases although less quickly and remains at a moderately high level for several megabases, i.e. a sharp decline at around 90 kbp and saturates at background level after 5-15 mbp for most breeds which corresponds to long-range breed specific haplotypes.

However, dogs do not just have longer regions of LD than humans but also fewer allelic variants per locus (Sutter et al., 2004; Ostrander and Wayne, 2005). Given this genomic redundancy, genome-wide studies in dogs require less SNPs than in humans. The authors also found there is a high negative correlation between LD and population size of a particular breed. The actual level of the LD curves appears to be related to breed history (Sutter and Ostrander, 2004), e.g. LD extends less than 1 million bp in the popular US breed Labrador Retriever (mixed origin and large population size) which received little genetic pressure from those historical bottlenecks while Pekingese has LD regions of about 3.2 mbp which is representative of tight population bottlenecks in the last 100 years.

Intra-breed diversity has also been measured by studying haplotype structure, i.e. the combination of alleles on a given chromosome. Sutter et al. (2004); Lindblad-Toh et al. (2005) found that in 10 regions across the genome covering each a length of about 100 kbp there are often less or equal to 3 different haplotypes on a chromosome. According to Sutter and Ostrander (2004), on average, 4.5 haplotypes explain 80% of the total chromosome variability at at locus across five dog breeds. Within a breed, 80% of the time, each locus has 2.1-3.4 haplotypes. This shows a low level of diversity and given that any two of these breeds share more than half of their haplotypes these observations reflect common joint history. Typically, 5 haplotypes are observed across each 10 to 500 kilo base pair window where 1 to 2 of them are frequent and the others are rare. Sampling the same haplotype on a chromosome increases by a factor two if the same breeds were selected compared to different ones.

1.4.4 Breed clustering and diversity

Given that many breeds represent closed gene pools, it is reasonable to study between-breed similarity and differentiation of dog breeds. Evolutionary history suggests that related breeds may group into clusters of joint common ancestry (Ostrander and Wayne, 2005). Parker et al. (2004) applied the STRUCTURE clustering algorithm (Pritchard et al., 2000) with a pre-specified number of K clusters. STRUCTURE aims to detect genetically distinct breeds based on patterns of the allele frequencies of the individuals. Investigating the role of K, Parker et al. (2004) found that only $K = 4$ led to consistent clustering. Setting K less than 4 led to partial and inconsistent clustering of some breeds while setting $K > 4$ did not infer further consistent subpopulations.

Before I investigate breed diversity in dogs I take a brief look at the modest genetic differentiation in human populations (Romualdi et al., 2002; Conrad et al., 2006; Li et al., 2008; Jakobsson et al., 2008; Via et al., 2009). Most of the total human genetic variation is focused within members of the same population, i.e. 85% within variability using autosomal SNP markers (Romualdi et al., 2002; Via et al., 2009). Furthermore, there is around 5-10% variability between populations of the same continent, and only about 10% of the total genetic variability is due to continental differences. Finally, less than 2% of the total variability is available in only one continent.

Diversity in dog breeds has been studied using different genetic data sources (Wade, 2011). Dogs have evolved an enormous level of between-breed variation combined with rather small intra breed variability (Ostrander, 2007). The following account discusses breed similarity in terms of data source:

- **mtDNA (maternal lineage)**: Pang et al. (2009); Webb and Allard (2010) looked at mtDNA and found 217 haplotypes which were clustered into ten haplogroups based on a set of 1543 dogs from Asian and Europe, 33 dogs from Artic America and 40 Eurasian wolves. The results showed more variability in Central China (7 haplogroups) compared to Europe (4 haplogroups). Wade (2011) noted that dogs which all carry some common mtDNA haplotype may have very different phenotypes. Furthermore, Sutter and Ostrander (2004); Sundqvist et al. (2006); Parra et al. (2008) showed the mtDNA is not sufficient to reliably discriminate among modern dog breeds due to its slow evolution.

- **Y-Chromosome (paternal lineage)**: to study paternal inheritance Bannasch et al. (2005); Sundqvist et al. (2006) studied Y-chromosome microsatellite data based on 824 male dogs from 50 domestic dog breeds and discovered a total of 67 haplotypes. Among those breeds there are 26 breeds with unique haplotypes while 15 breeds only had a single haplotype present. Furthermore, a large number of haplotypes was shared across breeds indicating a common origin. The authors also found that the total canine variability can be decomposed in 36% within-breed variability an 67% between-breed variability. On the other, based on 20 human populations Kayser et al. (2001) found 77% within-population variability and 23% within-population variance. In summary, there are fewer paternal than maternal haplotypes due to the focus on popular sires for mating, and a large number of breeds either has limited or no Y-chromosome variability (Wade, 2011).

- **Autosomal microsatellite**: early studies of canine genomic patterns for a small number of breeds were based on microsatellite markers and showed the existence of allele frequency difference between breeds which is essential for their discrimination (Parker, 2012). Later, Parker et al. (2004) utilized 96 microsatellite loci from 414 purebreed dogs taken from 85 breeds to compute a between-breed variability of 27% while the corresponding value shown above is 5-10% for humans. Parker et al. (2004) concludes that dog between-breed variability is estimated at five times

Reference	F_{st} (Marker Type)	Dataset description
Kim et al. (2001)	0.154 (8 STRs)	213 dogs from 11 Asian native dog populations
Parker et al. (2004)	0.27 (96 STRs), 0.33 (75 SNPs)	414 purebred dogs from 85 breeds
Parra et al. (2008)	0.11 (236 STR), 0.14 (4 NRY STR) 0.13 (18 mtDNA STR)	173 dogs from five pointing breeds
Boyko et al. (2010)	0.25-0.3 (61K SNPs)	915 dogs from 80 domestic breeds, 83 wild canids, 10 outbred African shelter dogs
Mellanby et al. (2012)	0.11 - 0.26 (15 STR)	285 dogs from 13 popular UK breeds

Table 1.2: This table shows values for the fixation index for a variety of marker types and dog datasets.

the between-ethnic group variability in humans. However, since the first canine whole-genome sequence (Lindblad-Toh et al., 2005) many investigators preferred SNPs over microsatellite markers due to mainly two reasons, easy of genotyping and a large number of markers can be retrieved within one reaction (Parker, 2012).

- **Autosomal SNP markers**: Lindblad-Toh et al. (2005); Karlsson et al. (2007); Vonholdt et al. (2010) measure genetic breed variability using heterozygosity along chromosomes. Between-breed variability estimates ranges from 0.29-0.32 and within-breed variability at around 0.26, e.g. Akita (0.21) varies less than Labrador Retriever (0.30) (Wade, 2011). Therefore, breed formation led to a loss of around 13% (proportion 0.26 from 0.3) of the genomic variability. With other words, dog breeds retained 87% of their original diversity level. Lindblad-Toh et al. (2005) looked at 7.8x boxer sequence and found that long stretches of homozygous regions (62% of genome) are combined with heterozygous regions (38% of genome) across all chromosomes. Homozygous regions were on average 6 times longer than heterozygous regions. In total, there were 770K SNPs, with homozygous regions having 1 SNP in 20kb and with heterozygous regions having 1 SNP in 1kbp.

Furthermore, limited intra-breed genetic diversity is due to strong inbreeding of a small number of purebred founder dogs. However, there may exist stratification, systematic allele differences among geographically disparate members of a given breed, such that some of them appear to be divergent as distinct breeds, while others exist as single breed across continents (Quignon et al., 2007). A particular loss of diversity is associated with breed splits, e.g. the existence of different Poodle sizes (Björnerfeldt et al., 2008; Wade, 2011). Given limited intra-breed genetic diversity and the close link of domestic dogs with humans, it is not very surprising that some of these diseases also affect humans (Patterson et al., 1988). Due to very dedicated and intense inbreeding efforts to yield certain morphologies many diseases are restricted to particular breeds (Patterson, 2000; Calboli et al., 2008).

Another important measure of breed differentiation is given by the fixation index F_{st} which measures how randomly selected alleles differ within a given breed compared to the entire dog population, and in Table 1.2 I show how those values depend on breed selection and data types.

These fixation indices can be compared to human datasets, such as the Perlegen SNP dataset (Hinds et al., 2005) which is based on about 1 million SNPs and the HapMap Project (Gibbs et al., 2003) which uses about 0.6 million SNPs. Weir et al. (2005) studied the fixation index for these two datasets: the HapMap populations have an average $F_{st} = 0.13$ whose four individual populations have $F_{st} = 0.1$ for (CEU) Caucasians of European descent, $F_{st} = 0.12$ for (YRI) Yoruba from Ibadan, $F_{st} = 0.15$ for

Chr ID	#SNPs	Chr ID	#SNPs	Chr ID	#SNPs
1	14	**11**	13	**27**	13
2	13	**12**	13	**29**	14
4	15	**13**	15	**31**	7
5	17	**14**	16	**32**	12
6	12	**15**	9	**34**	8
7	12	**17**	13	**35**	17
8	12	**20**	16	**38**	13
9	13	**21**	10		
10	10	**25**	13		

Table 1.3: The distribution of the 320 SNPs across 25 of the 38 chromosomes of the dog genome.

(HCB) Han Chinese from Beijing and $F_{st} = 0.15$ for (JPT) Japanese from Tokyo. On the other hand, the Perlegen has a mean fixation index of 0.1 whose composing populations have $F_{st} = 0.08$ for (EA) European American, $F_{st} = 0.12$ for (HC) Han Chinese and $F_{st} = 0.1$ for (AA) African American.

1.5 Datasets and marker profile

In this section I discuss the marker profile and the different synthetic variants of the Mars dataset I utilize for data analysis in this thesis.

1.5.1 Marker profile

According to Mars Veterinary Wisdom Panel (2013) Mars tested over 4600 autosomal SNP markers from loci spread across the whole dog genome based on an analysis of 3200 dogs. Then, within an intermediate step 1536 SNPs were selected using another set of 4400 dogs from different breeds. This intermediate marker profiles was reduced in a final selection step to 320 markers on $C = 25$ of the dogs' 38 autosomal chromosomes according to the following quality criteria (Martin et al., 2010)

- **Easy of sequencing**: does the retrieval of a SNP often lead to missing of low quality calls? If the allele has not been reliably identified then preference is given to marker who have repeat SNPs nearby which have the same value due to linkage.

- **Variability across breeds** With breed composition prediction in mind a given marker need to vary at least for a subset of the breeds.

In table 1.3 I see that for each of the chromosomes I have about 10-20 SNP markers on each of the C chromosomes. In Figure 1.1 I illustrate SNP marker density where each point corresponds to a chromosome and is represented by the median and standard deviation of the distance between consecutive SNP marker positions as measured in base pairs. Given that I have only 320 SNP marker it is to be expected that I have a high median SNP marker distance of 968K taken over all chromosomes. However, the high standard deviation of 880K suggests that the SNP markers are not very equally spaced. An inspection of individual SNP marker distances shows a range from a few 100s to several million base pairs. But this unequal spacing of markers is expected because Mars selected markers by an initial investigation of discriminatory power as described above Martin et al. (2010). Given this high median SNP distance, I expect that correlation information can only be utilized to a limited extent in the Mars dataset.

1.5.2 Datasets

In this section I discuss which datasets are used for estimation of model parameters (training dataset) and which datasets are utilized for prediction (test dataset).

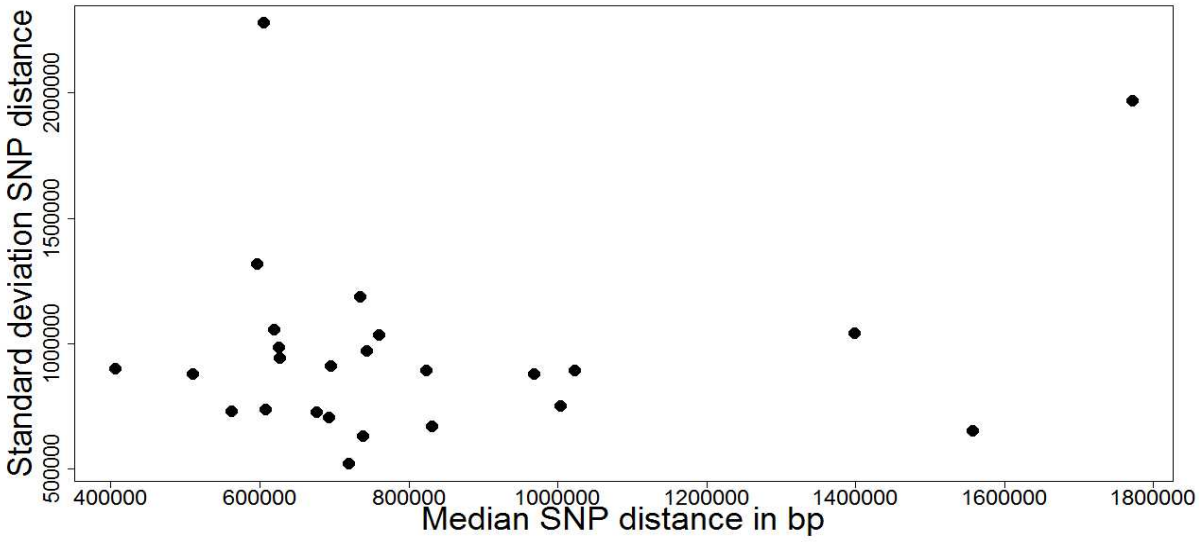


Figure 1.1: This figure shows the median distance and standard deviation in base pairs between consecutive SNP markers. Each of the points represents one chromosome. The data suggests that a few SNP have strong linkage while other can be considered unlinked given their distance of more than a million base pairs.

1.5.3 Purebred reference dataset

Mars provides a purebred dataset based on a total 8552 dog samples (**BigPure**) selected from 149 Kennel Club recognized breeds which are listed in Table A.1. Some of these breeds have subpopulations, i.e. with respect to country (UK vs. USA) or utility (show, working, field) and in the pure breed identification algorithm I tend to count confusion of these subpopulations as wrong prediction while for mixed breed identification I joined many breeds with sufficient similarity. Collaboratively with Mars' staff I decided with sub-breeds to join based on visual proximity inspection of two dimensional breed reconstructions using multidimensional scaling (see Section 2.4). In Tables A.2, A.3, A.4 the reader sees which sub-breeds where merged.

Mars segments these 149 breeds into three classes of non-overlapping meta dog groups: ancient (mostly Asian breeds), Guard (mostly mastiff type dogs) and European Hedge (most of those breeds were created in Victorian era). Most of the breeds (about 125 out of 149) were assigned to Hedge. For computational reasons, initially I also looked at a small dataset denoted as **SmallPure** which is composed of the first 34 alphabetical breeds from BigPure, i.e. breeds with initial starting with either character A or B ranging from Afghan Hound to Bullmastiff.

Within the pure breed identification I split the pure breed dataset of $n_b = train_b + test_b$ dogs for breed b in two datasets where the training data has $train_b$ training dogs and $test_b$ test dogs for each breed. Then, for SmallPure and BigPure I performed the following data splits:

- For the SmallPure I used $n_b - 3$ dogs for training and the remaining 3 dogs for testing. In Table A.1, I list $train_b$ under the column labelled *Small*. I see in this table that I have $\sum_b^{B'} n_b = 1646$ dogs where 1544 dogs are used for training and the remainder $3 \cdot B' = 102$ dogs for testing.
- For BigPure I have a total of $\sum_b^B n_b = 8552$ dogs with a total of 7250 training dogs and 1302 test dogs. If a breed has less than $n_b = 20$ dogs I use $n_b - 3$ for the training set and three dogs for

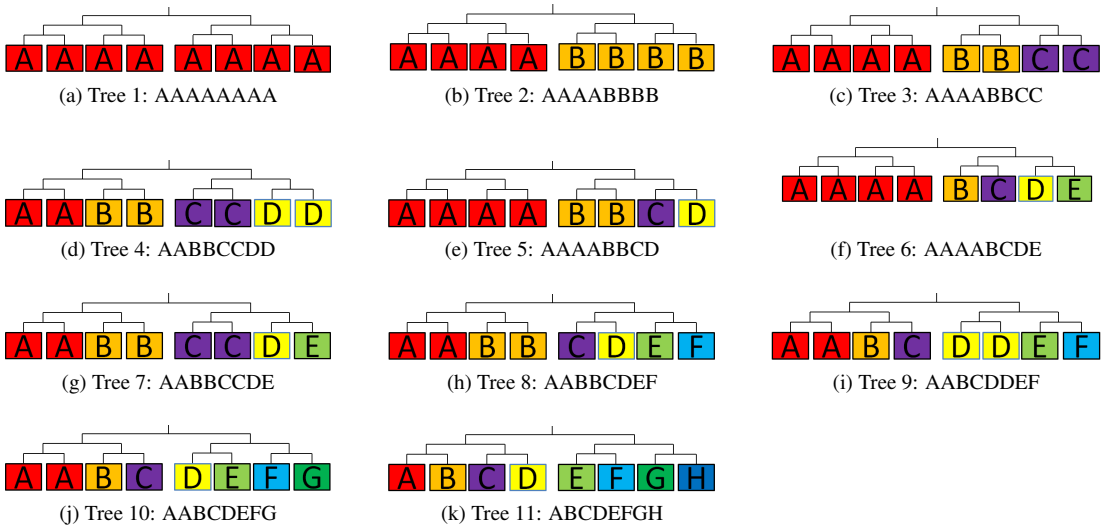
Possible TAP	Lineage trees covering TAP
100%	1
50%	2-3,5-6
25%	3-5,7-10
12.5	5-11
0%	1-11

Table 1.4: For each true ancestral proportion level I list which of the lineage trees have corresponding levels. Lineage tree IDs correspond to those defined in Figure 1.2.

the testing set. However, if $n_b > 20$, I put $n_b - \lfloor 0.15 \cdot n_b + 0.5 \rfloor$ dogs in the training set and the remaining $\lfloor 0.15 \cdot n_b + 0.5 \rfloor$ in the testing set. The training set $train_B$, I obtain for each breed is shown in Table A.1 under column labelled Big.

For BigPure, breeds Schipperke (UKX), Miniature Pinscher (UKX), Weimaraner (US2X), Saint Bernard (UKX) and Soft Coated Wheaten Terrier (UKX) have the five least number of training dogs with a count of $n_b = 4, 5, 5, 7$ and and $n_b = 10$, respectively. On the other hand, I also note that Shih Tzu, Labrador Retriever (US Field), Labrador Retriever (US Show), Yorkshire Terrier (US) and Poodle (Miniature) have the five highest numbers of training dogs with a count of $n_b = 332, 276, 236, 235$ and $n_b = 183$, respectively.

1.5.4 Test Sample of synthetic dogs



(a) Tree 1: AAAAAAAA (b) Tree 2: AAAABBBB (c) Tree 3: AAAABBCC

(d) Tree 4: AABBCCDD (e) Tree 5: AAAABBCD (f) Tree 6: AAAABCDE

(g) Tree 7: AABBCCDE (h) Tree 8: AABBCDEF (i) Tree 9: AABCDDEF

(j) Tree 10: AABCDEFG (k) Tree 11: ABCDEFGH

Figure 1.2: This figure shows the way Mars arranged lineage trees Φ by complexity. Each ggp leaf is encoded by colour according to the breed it represents. The most complex case is shown in tree 11 which assumes eight distinct pure breed ggps. Although these ggps may not pure breed ancestors I make this assumption because it is unlikely to detect ancestry on a finer scale than this. All other trees form special cases, e.g. tree 1 represents purebred dogs assuming that all eight ggps are from the same breed while tree 2 shows designer dogs which are a cross of two purebred dogs. The trees are roughly sorted in order of complexity although tree 6 (1 parent, 4 ggps) is more complex than tree 7 (3 gp, 1 ggp). These figures represent a classification perspective where time flows upwards (current individual at top) which is contrary to coalescent genealogical trees (Felsenstein, 2013) where the most recent individual is placed at the bottom of the figure.

To quantify breed contributions in mixed bred dogs I used dataset BigPure for training and a synthetic dataset for testing. To evaluate algorithmic approaches for breed proportion estimation Mars

Lineage Tree	Purebred	Parent	Grandparent	Great-grandparent
1	1	0	0	0
2	0	2	0	0
3	0	1	2	0
4	0	0	4	0
5	0	1	1	2
6	0	1	0	4
7	0	0	3	1
8	0	0	2	4
9	0	0	2	4
10	0	0	1	6
11	0	0	0	8

Table 1.5: For each lineage tree shown in Figure 1.2 I list the counts for the different true ancestry proportion levels. For example, lineage tree 3 is composed of one parent and two grandparents while tree 10 has 1 grandparent and 6 ggps. This table also implies that all trees have a unique combination of purebred, parents, gps and ggps except for lineage trees 8 and 9 which both have 2 gps and 4 ggps.

ID	Tree	Count	ID	Tree	Count	ID	Tree	Count
1	AAAAAAAA	953	5	AAAABBCD	798	9	AABCDDEF	641
2	AAAABBBB	906	6	AAAABCDE	748	10	AABCDEFG	615
3	AAAABBCC	846	7	AABBCCDE	735	11	ABCDEFGH	529
4	AABBCCDD	791	8	AABBCDEF	662			

Table 1.6: For OrgSyntheticRed I list number of test samples in each of 11 lineage trees.

decided to specify different scenarios of different complexity ranging from a lineage tree with eight different great-grandparent breeds to the special case where all great-grandparents are from the same breed to form a purebred synthetic dog. Figure 1.2 shows all lineage trees composed of different quantities for the true ancestry proportion **(TAP)**: purebreds (100%), parents (p,50%), grandparents (gp,25%), great-grandparents (ggps,12.5%) or breeds not part of the ancestry (0%). In Table 1.4 I outlined which TAP levels are present in which lineage trees. Furthermore, in Table 1.5 I show in more detail how many leaves for each TAP level are present in each of the 11 lineage trees.

Based on this lineage structure shown in Figure 1.2 Mars generated 1,000 synthetic test dogs for each of the 11 lineage trees to yield a total of 11,000 synthetic test dog genotypes (**OrgSynthetic**). Synthetic genotypes were created by independently combining the information on different chromosomes according to their respective haplotype frequencies derived from PHASE (Stephens et al., 2001; Li and Stephens, 2003). The 175 breeds in the lineage tree were selected according to the breed distribution of records stored in Banfield The Pet Hospital's central database which approximately maps breed proportions in the dog population (Martin et al., 2010).

However, there are two constraints why I do not use OrgSynthetic as test dataset for mixed breed. The first constraint is due to data inconsistency. OrgSynthetic (175 breeds) and BigPure (149 breeds) neither have a subset nor a superset relationship with respect to the breeds. Therefore, I only consider breeds which are in the union of OrgSynthetic and BigPure breed set, reducing the number of consistent breeds to $B_M = 125$ breeds. In Tables A.2, A.3, A.4 I show how I arrived at a consistent breed set. In particular, I first deleted all breeds in OrgSynthetic but not in BigPure, and vice versa. Finally, a few breeds were aggregated which contain subpopulation structure. Secondly, I restrict my attention to singleton breed assignments in the lineage tree, i.e. each breed can be assigned only to those ggp

slots which have the same colour encoding. This second constraint is used to simplify possible breed proportions which in this case are in the set $\{0, 0.125, 0.25, 0.5, 1\}$. So, in summary I have 8448 samples in a reduced purebred training dataset **BigPureRed** with a median number of 55 dogs per breed, and another 8224 synthetic test samples in **OrgSyntheticRed** distributed in lineage trees according Table 1.6.

Based on the way test dataset OrgSyntheticRed was simulated I will comment on the inherited breed proportions. Firstly, there is sampling error due to incomplete sampling described in Section 1.3.4 although according to Table A.1 the number of purebred dogs per breed varies widely in my study. Secondly, there is biological variation within the selected parents from a given breed and some parents have genomes closer to the mean dog of a given breed than others. Thirdly, according to Section 1.3.1 the number of markers and their informativeness for ancestry inferences technically limits how well breed proportions can be estimated. Furthermore, there is measurement error when sequencing the genetic data of training or test dogs (Visscher et al., 2006, 2008).

There is also variation in genome sharing because the choice of which parental genome chromosome segment is inherited is probabilistic, i.e. the child inherits a segment from the maternal or paternal breed with equal probability. Let us illustrate this point with an example where I assume lineage tree AAAABBBB in Table 1.2 based on four parents from breed A and B each and each of the eight great-grandparents is observed on $C = 25$ chromosomes. In diploid species there are $2 \cdot C$ Bernoulli trials to select either parents' chromosomal segment with probability $p = 0.5$, i.e. a Binomial distribution with $N = 2 \cdot C$, $p = 0.5$. To construct a $100 \cdot (1 - \alpha) = 95\%$ confidence interval for $\alpha = 5$, I use the Wilson interval (Wilson, 1927) in expression 1.1 based on critical value $z_{\alpha/2}$ (i.e. $(1 - \frac{\alpha}{2})-$percentile of standard normal distribution) to yield interval [0.37, 0.63]. According to Brown et al. (2001) due to the central limit theorem a Binomial confidence interval can be approximated with a normal distribution for $np > 5$ and $n \cdot (1 - p) > 5$ which is valid in my case. Therefore, an approximate normal confidence interval in expression 1.2 yields a similar interval [0.36,0.64]. Therefore, in 95% of the time simulated test samples from lineage tree AAAABBBB will have breed proportion which deviate by $\pm 13\%$ from the truth of 50%.

$$\frac{N}{N + z^2_{\alpha/2}} \left[p + \frac{z^2_{\alpha/2}}{2 \cdot N} \pm z_{\alpha/2} \sqrt{\frac{p \cdot (1 - p)}{N} + \frac{z^2_{\alpha/2}}{4N^2}} \right] \tag{1.1}$$

$$p \pm z_{\alpha/2} \sqrt{\frac{p \cdot (1 - p)}{N}} \tag{1.2}$$

Furthermore, the novel simulation-based algorithm which will be presented in Chapter 4 underestimates breed proportions due its exploration of the breed space. Therefore, in some cases my novel algorithm yields a lower breed proportion estimate than the true ancestral proportion, e.g. an estimate may suggest ggp but it is actually a grandparent breed or a breed may be judged not part of the ancestry but it is actually a ggp breed.

1.5.5 Recombination test sample of synthetic dogs

Mars' synthetic test dataset OrgSyntheticRed only models recombination up to the chromosome level, i.e. in the simulated data there is exactly one breed assigned to the maternal **(M)** and exactly one breed

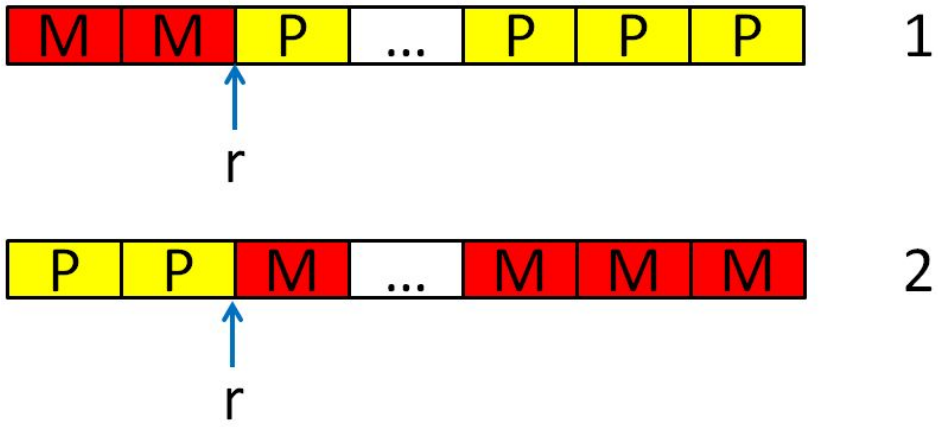


Figure 1.3: The two sequences correspond to two examples for haplotypes on chromosome c where either the maternal or paternal part of the parental haplotype is copied first, respectively. Each haplotype is a combination of maternal **(M)** breed and paternal **(P)** breed haplotype subsequence sampled according to haplotype frequency estimates, such that with 50 percent chance either parent forms the first part of the sequence. Without loss of generality, I assume that the first sequence is related to the maternal haplotype: then, the first r alleles are copied from the maternal haplotype H^c_M, i.e. $H^c_M[1:r]$ while the remaining alleles from $r+1$ to n_c (chromosome length) are copied from the paternal haplotype H^c_P, i.e. $H^c_P[(r+1):n_c]$.

to the paternal **(P)** haplotype. To investigate the effect of more complicated ancestries I derive 100 synthetic canine test samples for the case of recombination. For both cases I create a common lineage painting of singletons according to lineage tree ABCDEFGH for each of the test samples uniformly sampling breeds from the set $B_M = 125$ consistent breeds.

1. **OrgSynRecC1**: this case resembles data generation for lineage tree ABCDEFGH in test dataset OrgSyntheticRed. Firstly, for each test sample I created a genome painting by uniformly assigning breeds to the $2 \cdot C$ chromosomes. Then, the genotype X for each test sample is created, such that for each maternal and paternal breed each on chromosome c I sample a haplotype from the maternal and paternal breed according to their PHASE haplotype frequency estimates. Then, I add maternal (haplotype 1) H^c_M and paternal (haplotype 2) H^c_P haplotypes to yield the genotype on chromosome c. I repeat this procedure for all chromosomes and combine results to yield the full genotype X.

2. **OrgSynRecC2**: in this case I create a dataset which helps to study the impact of recombination. According to Wong et al. (2010) the canine sex-averaged, chromosome-averaged recombination rate is 0.92 cM/Mb where 1 cM/Mb refers to about 1% probability of crossover between two markers a mega base apart (Hellenthal, 2006). Furthermore, the purebred dataset described in Section 1.5.3 has chromosome length ranging from 9-15Mb. Therefore, the chromosome recombination probability α is between 8.3% and 13.8% and for simplicity I assumed $\alpha = 10\%$ in this simulation study.

   For each test sample I assume lineage tree ABCDEFGH shown in Figure 1.2 and sample 1 haplotype for each of the eight great-grandparents according to the haplotype frequencies in the purebred dataset BigPure in Section 1.5.3. In this setup for each chromosome c recombination could occur at the following six ancestral positions:

- Four pairs of great-grandparents (PARENTS) form the four grandparent breeds (CHILD).
- Two pairs of grandparent breeds (PARENTS) form the parents (CHILD).

At each of these ancestral positions recombinations occur with probability α, or with probability $1-\alpha$ one of the two PARENTS haplotypes H_M, H_P is chosen with equal probability. In the case of recombination PARENTS haplotypes H_M^c, H_P^c are combined to form CHILD haplotype H_C as follows

$$H_C^c = [H_M^c[1:r], H_P^c[(r+1):n_c]] \quad \text{or}$$
$$H_C^c = [H_P^c[1:r], H_M^c[(r+1):n_c]]$$

where the recombination position $r \sim U[1, n_c - 1]$ is chosen uniformly over all markers (except for the chromosome boundaries) where n_c is the number of loci on chromosome c. This process is repeated for all chromosomes of a test dog. Once the two parental haplotypes from the sire and dam have been selected I combine them to form the genotype for the test dog.

1.6 Thesis contributions and further chapters' outline

According to the previous sections, there are a wide range of reasons and motivations to provide commercially available DTC canine ancestry testing kits. Having this in mind, the objective of this thesis is the development of a novel computational approach which estimates the proportions of breed contributions in mixed breed dogs using SNPs and compare its results with those obtained from a technique which represents the current state of the art for human ancestry inference in the literature. In particular, I am interested in the investigation of complex ancestries composed of multiple breeds, such that breeds contribute only 1/8th of the ancestry of a dog. This prediction is performed based on a comparison with the specific genetic sequence patterns of the purebred dogs in the ancestral dataset composed of purebred training data.

The **main novel contributions in this thesis** are as follows:

1. A detailed study of breed similarity in dogs to further exploratory investigations of how well breeds can be visually discriminated as well as a mechanism to bias breed update proposals according to a breed proximity matrix within a Markov Chain Monte Carlo (MCMC) algorithm.
2. Research, implementation and testing of novel MCMC-based inference technique DBAncestry which is based on maximization of haplotype frequencies. DBAncestry is used to infer complex ancestry compositions of up to eight great-grandparents and is targeted at datasets composed of short genetic sequences.
3. A comprehensive review of techniques applicable for human and canine ancestry inference due to a lack thereof in the published literature but also to select an advanced technique used as competitor for DBAncestry.
4. As comparison to my novel ancestry inference approach I also applied ChromoPainter to the same test set. ChromoPainter is a recent advanced techniques based on a hidden markov model.

The following part outlines an overview of the following chapters:

- **Chapter 2**: this chapter can be viewed as a continuation of Section 1.4.4 which discussed breed similarity as it applies to clustering of breeds from the existing literature. However, this chapter focuses on an investigation of breed proximity based on dataset BigPure introduced in Section 1.5.3.

  This chapter is motivated by two main factors: firstly, I will explore a variety of breed similarity measures visually to assess their potential for breed discrimination. Then, secondly, given the different explored proximity measures I seek to select an appropriate one which is promoted to form the basis of a breed-biased update proposal mechanism within a MCMC algorithm in Chapter 4.

  I will look at a range of distance and correlation measures to calculate breed proximity based on different purebreed data representations which are either based on the original genotype data, or utilizes haplotype frequencies in two different ways. Furthermore, I visually assess the proximity measures using different visual representations, such as heatmaps, dendrograms and low-dimensional breed similarity reconstructions, to determine breeds which either form clusters or are spread out. These plots show that some breeds are very separated out while other breeds even show some overlap.

  Results will be compared with published work by Vonholdt et al. (2010). For example, I also found that leaf branches in the dendrogram tend to be long suggesting that breeds are distinct, and the clustering structure is flat which is related to little population substructure. I also discovered by visual inspection that a genotype representation is more suitable to separate out clusters of breeds which makes it a suitable choice for the MCMC algorithm.

- **Chapter 3**: this chapter develops the novel methodology DBAncestry for the special case of pure breed identification which will be extended to the crossbred case in subsequent chapters. Furthermore, I review existing work by Mars which does not account for LD in a principled way. In the case of DBAncestry I obtain a characteristic breed representation using genetic information. While the Mars approach will look at SNPs individually the DBAncestry will look at the SNPs jointly through haplotype frequency estimates. Haplotype frequency estimates are computed separately for each breed to avoid confounding of population structure or jointly across breeds to exploit sample size.

  Firstly, the DBAncestry algorithm works by enumerating consistent haplotype pairs. Then, to compute the breed pair assignment on a chromosome DBAncestry averages over the product of maternal and paternal frequency estimates according to the set of consistent haplotypes. However, in some cases not all pairs of consistent haplotype pairs have non-zero frequency estimates. Therefore, I will develop three options to deal with the case where one or more of the haplotype frequencies are zero. Finally, within the experiments on the training/test split discussed in Section 1.5.3 I will look at those options and will report which haplotype frequency imputation strategy works best.

- **Chapter 4**: this chapter extends DBAncestry from the special case for pure breed identification presented in Chapter 3 to inference for mixed breed dogs composed of up to eight great-grandparents.

However, the search space defined by all possible breed compositions is very large. Therefore, to yield estimates for the different breed proportions I sample the space of possible breed compositions using a Metropolis-Hastings algorithm. I will experiment with two different updates rule in the MCMC algorithm: either I uniformly propose a new breed, i.e. all breeds have the same chance to be drawn within the proposal mechanism. Alternatively, I bias the Markov Chain according to the Manhattan rank breed distance defined in chapter 2, such that new breeds are likely to be similar to the current breed assignment. Furthermore, I experiment with different run-length of the MCMC algorithm which differs by a factor of 10, i.e. I either use 700K (short run) or 7 million (long run) main phase iterations.

I found that breeds proportions are well estimated across all ancestral levels from pure breeds down to great-grandparents. Furthermore, possibly due to lack of deep hierarchical structure of the breed space a uniform breed update proposal in the MCMC algorithm is sufficient to explore the breed space.

I also investigated whether knowledge of the lineage tree for the synthetic test has to be assumed or whether it can be derived from breed contribution estimates. Indeed, I find that via the proxy of the ancestral level (purebred, great-grandparent, etc.) the lineage tree can be inferred from the estimated breed contributions. Finally, I found that DBAncestry models recombination well and predicted breed contribution estimates only decrease by a very small amount.

- **Chapter 5**: the final chapter has two parts

  - **Review ancestry inference techniques**: there is a a lack of a comprehensive review of currently available techniques for estimation of local and global ancestry proportions. Due to this shortcoming I extensively reviewed existing approaches at the conceptual level starting from the earlier techniques until the most recent state of the art techniques. These ancestry inference approaches span a wide range of statistical techniques, such as multivariate linear regression with multiple responses and machine learning based techniques, which led us to classify them according to the modelling framework. Furthermore, if possible, I describe the historical development of these techniques based on their identified limitations.

  - **ChromoPainter data analysis**: after this initial review of ancestry inference techniques I defined a list of criteria to select an algorithm as competing technique for DBAncestry. Considered criteria are whether the algorithm is suitable for sparse/dense marker sets, how it scales with the number of breeds, whether there is a publically available implementation and if so whether access to expert users of the software exists. The ChromoPainter techniques satisfied this list of criteria best and I compared it in more detail with DBAncestry.

    After that I discussed how I computed a haplotype representation of the test dogs dataset OrgSyntheticRed (Section 1.5.4) and inferred the recombination and mutation rate from the BigPure dataset (Section 1.5.3) which are used as input to the ChromoPainter software. ChromoPainter reasonably well predicts pure breed estimates. However, for more complex lineages performance sharply drops, such that breed contributions are much more underestimated compared to DBAncestry.

1.7 Conclusions

This chapter provided an introduction to DTC genetic ancestry testing kits along with the motivation of private, public and academic users. While private reasons are mainly due to curiosity about human or canine ancestry while academics use it in the context of population genetics and GWAS studies. Furthermore, public users of human DTC test kits are concerned with forensic applications while canine ancestry prediction is relevant to deal with public safety concerns in human communities with respect to dog attacks.

In this chapter I also discussed the value of five data sources which can be used to assess methods that are predictive of ancestry. In particular, I analyzed the merits and limitations of phenotypic data, language input, genealogical records, stable isotopes as well as genetic data to infer ancestry. With respect to genetic data I reviewed lineage-based markers and autosomal SNP markers whereas the latter one will be used for breed composition estimation in the subsequent chapters. After that I explained the concept of local and global ancestry inference: local ancestry inference is concerned with the ancestry inference at the individual SNP marker while for global ancestry estimation used in this thesis I would like to estimate ancestry proportions averaged over all SNP markers, i.e. ancestry across the whole chromosome.

As next topic I reviewed some background on the historical development of dogs, followed by a discussion of marker correlation leading to similar haplotype segments along the chromosome. After that compared population variability in humans and dogs. Finally I show a few figures which illustrates how breeds cluster according to genetic distance.

The final section in this chapter dealt with the thesis motivation, surveyed its novel contributions along with a more detailed description of the subsequent further chapters.

Chapter 2

Breed similarity in dogs

Many problems in science, such as geometry, probability, statistics, pattern recognition, information retrieval and molecular biology are concerned with classifying objects according to perceived similarities. For examples, ecologists collect quantitative measurements to judge the resemblance of species. In biology the process of classifying species is known as taxonomy (taxa = groups). Taxonomy goes back to Aristotle who classified animal species into vertebrates and invertebrates. Among others, Theophrastus and Linnaeus accounted for the structure and classification of plants. In 1737 Linnaeus wrote in his book 'Genera Plantarum' (Linnaeus, 1737) that 'All the real knowledge which we possess, depends on methods by which we distinguish the similar from the dissimilar. The greater the number of natural distinctions this method comprehends the clearer becomes our idea of things.'

Similarity is also a basic mode of human thinking where I judge how close two physical objects are with respect to length, time period, coolness, etc. Furthermore, distinguishing patters into groups, such as which items are edible or poisonous, and to recognize and classify words in a language, appears to a primitive human characteristic.

In this chapter breed proximity is motivated by two main factors: firstly, I study breed proximity to visually inspect how well groups of breeds are separated out in low-dimensional space. Secondly, I am interested in computing a breed similarity measure which is utilized as information geometry for the discrete breed space in the breed proposal step of a Markov Chain Monte Carlo algorithm. As first step I review different proximity measures and explain how I use techniques, such as multidimensional scaling (MDS), dendrograms and heatmaps, to visualize these measures. After that I discuss how I either employ a data representation based on genotypes or haplotype data to compute these proximity measures. Finally, I present results in form of visualizations and offer a comparison with previous canine hierarchical clustering results evaluated by Vonholdt et al. (2010).

2.1 Motivation to infer interbreed relationships

Most species do not reproduce according to the assumptions of a panmictic population which leads to genetically stratified subpopulations which may be due to different reasons, such as geographic barriers (Rosenberg et al., 2005). The study of breed similarity is motivated by two main reasons. Firstly, I perform an exploratory data analysis to get insights into between-breed variability and to see how well groups of breeds can be visually discriminated. And, secondly, I use breed similarity to investigate whether it may yield improvements in predictive modeling, i.e. to more accurately estimate breed ancestry composition. Although I derive the similarity measure for the second case in this chapter its impact on prediction results will be discussed in chapter 4.

1. **Study of breed variability for visual breed discrimination:** starting with Charles Darwin bi-

ologists targeted the reconstruction of evolutionary history (Nei and Kumar, 2000; Wiley and Lieberman, 2011). Classical evolutionists compared the morphology and physiology of species to arrive at conclusions about the shared evolutionary history of species. Due to the subjective nature of this taxonomic categorization many conclusions have been controversial. Advances in molecular biology provided DNA data of the species genetics. Firstly, DNA data allows for an objective comparison of different types of organisms, such as bacteria, plants and animals. Secondly, evolutionary change of DNA pattern is much more regular than morphological changes, such that change and comparison of DNA between distantly related organisms can be modeled mathematically.

Then, in our study of genetic canine data I get insight into breed structure via an exploratory data analysis where I either cluster breeds or plot breeds as point in two-dimensional space according to an optimal reconstruction of high-dimensional proximities.

2. **Impact of update proposal in MCMC algorithm on prediction quality:** the second objective to study breed similarity is derived from the interest to improve breed ancestry prediction. Therefore, I intend to alter the uniform lineage painting proposal, such that new breed proposals in the lineage painting are more likely to be replaced by similar breeds. Breed similarity may yield a better exploration of the lineage painting search space. However, I may be stuck in local maxima of the search space. In particular, I look at two sub-objectives: firstly, I intend to either increase performance metrics, such as the deviation between true and predicted breed contribution. Secondly, improved exploration of the search may improve convergence speed of the Markov Chain, such that fewer iterations are required in the Metropolis-Hastings algorithm to yield a sufficiently good breed ancestry prediction.

2.2 Introduction to similarity measures

I am given n objects measured each with d features forming a $n \times d$ data matrix $X \in \mathbb{R}^{n \times d}$. In the case of the dog dataset I observe n purebred dogs on d features which are given by the SNP markers.

Cluster analysis is the explanatory study which analyzes algorithms for finding natural grouping and classifying of objects through unsupervised learning (Jain and Dubes, 1988; Hair et al., 2009; Deza and Deza, 2009; Everitt and Hothorn, 2011; Legendre and Legendre, 2012; Härdle and Simar, 2012). In other words, I intend to find a valid organization, reduce the data, understand the structure of the data to generate hypotheses given that objects are not tagged with a class label as response. A cluster groups a number of objects together. This cluster can be visualized as a collection of objects which are close together according to the proximity measure.

A proximity measure establishes the alikeness and affinity between pairs of objects. In particular, in population genetics I summarize the relationships between individuals based on patterns in molecular data. The proximity matrix can be derived from the data matrix. The element in the ith row and jth column of the similarity matrix describes the proximity between the ith and jth object. Furthermore, I can ignore the diagonal element of the proximity matrix because I assume objects have the same measure of proximity to themselves. A proximity measures is either a numerical similarity S or dissimilarity (distance) matrix D. According to Cattell (1952) a proximity matrix can be discussed from two main viewpoints. One may either view this matrix to study the variability of objects (Q analysis) given all features or the variance of features (R analysis) given all objects. The proximity matrix has dimension

$n \times n$ for Q analysis and $d \times d$ for R analysis. For nonmetric data, similarities are referred to as association measures.

- Correlation: a measure of correlation does not look at the magnitude difference between two objects but instead looks at pattern of the features between any two objects.

- Distance: a proximity measure based on magnitude between pairs of objects. Distance measures appear to be preferred over correlation measures. Many distance measures are metrics which satisfy the four following conditions: non-negativity ($D_{ij} \geq 0$), identity of indiscernibles ($D_{ij} = 0$ iff objects are identical, i.e $i = j$), symmetry ($D_{ij} = D_{ji}$) and triangle inequality ($D_{ik} \leq D_{ij} + D_{jk}$).

- Association: similarity measures for nonmetric features, such as nominal or ordinal valued features. For example, in the case of binary responses to a survey an association measure judges the amount of agreement or matching between any two pairs of responses.

For example, the Euclidean distance is a dissimilarity measure while a correlation coefficient corresponds to a similarity measure. If required, similarities and distances can be transformed into each other. If $S_{ij} \in [0, 1]$, then I may compute the corresponding distance as $D = 1 - S$. On the other hand, in the case of similarity defined by correlation $S_{ij} \in [-1, 1]$ and I convert to a distance measure by taking $D = (1 - S)/2$. Distances may not have pre-specified lower and upper values so let us define $D_{\text{norm}} = (D - D_{\text{min}})/(D_{\text{max}} - D_{\text{min}})$ where $D_{\text{min}} = min(D)$ and $D_{\text{max}} = max(D)$, Then, I convert normalized distances into similarities by $S = 1 - D_{\text{norm}}$.

Deriving a suitable proximity measure between objects for a scientific problem has turned into a common task. Given the limited information exchange between specialized scientific disciplines the same proximity measures were introduced with different names in various scientific areas (Crochemore et al., 2007; Deza and Deza, 2009; Hadjieleftheriou and Srivastava, 2011). For example, a distance measure which counts the number of different symbols of two objects of same lengths is known as edit distance in natural language processing, Levenshtein distance in coding theory, Hamming distance in information theory and evolutionary distance in computational biology.

2.3 Measures of proximity

Choice of proximity measure is of central importance to judge the closeness of objects. Firstly, I need to examine the data type which is used within the proximity measure. For example, I may consider dichotomizing a continuous variables. In the case of rivers I may be interested on whether a fish is edible based on a threshold of toxicity (Everitt and Hothorn, 2011). Secondly, the analyst has to decide whether shape and pattern of features between objects is more relevant than differences in magnitude. Thirdly, some proximity measures are monotonically related. In this case I might want to consider clustering algorithms which does not consider absolute values but ranks. For example, in a clustering algorithm referred to as hierarchical clustering the distance between two groups of objects might be computed based on single linkage (minimum distance of one object taken from each of the two clusters) or complete linkage (maximum distance of one object taken from each of the two clusters).

A large number of similarity coefficients were first derived for the case where all variables are binary (Gower and Legendre, 1986; Everitt and Hothorn, 2011). In this report I focus on proximity measures for continuous data and will offer more in-depths information for this case.

2.3.1 Proximity measures for continuous data

Proximity measures will be separated in general similarity and dissimilarity measures for ratio scale data, distances in probability theory and ChromoPainter measure.

2.3.1.1 Distance and correlation measures in statistics

At first I define allele sharing and Minkowski distance:

- **Allele sharing** is also known as evolutionary distance and is defined as one minus the proportion of alleles shared, i.e. $d_{ij} = 1 - H_p$ where H_p is the Hamming distance which is defined as $H_p = \frac{1}{p}\sum_{k=1}^{p} \mathbb{1}\left[x_{ik} \neq x_{jk}\right]$ where x_i, x_j are the genotype vectors for individuals i and j, respectively. The Hamming distance is a popular measure in information theory which describes the number of loci at which alleles are different, i.e. the minimum number of substitutions required to change genotype x_i into x_j.
- **Minkowski distance** which is defined as

$$d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r\right)^{1/r}. \tag{2.1}$$

  Popular special cases are given by the Euclidean distance $(r = 2)$ $d_{ij} = \sqrt{\sum_{k=1}^{p} |x_{ik} - x_{jk}|^2}$ and the Manhattan distance $(r = 1)$ $d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$. Given that SNP data is defined over a ternary alphabet Manhattan distance captures more information than the evolutionary distance. Furthermore, given that SNPs with intermediate frequencies change more than rare alleles, some distance measures are normalized by the allele frequency to adjust for better discrimination of individuals (Patterson et al., 2006).

Then, I look at different forms of correlations coefficients $\phi_{ij} \in [-1, 1]$ between objects i, j which can be converted into distance measures using $d_{ij} = \frac{1-\phi_{ij}}{2}$. Let us define the mean value on object i as $x_{i\cdot} = \frac{1}{p}\sum_{k=1}^{p} x_{ik}$ be the mean for object i. Then, the Pearson correlation, cosine angular similarity and Kendall τ correlation coefficient are defined as follows:

- **Pearson correlation** is defined as

$$\phi_{ij} = \frac{\sum_{k=1}^{p}(x_{ik} - x_{i\cdot})(x_{ik} - x_{j\cdot})}{\sqrt{\sum_{k=1}^{p}(x_{ik} - x_{i\cdot})^2 \sum_{k=1}^{p}(x_{ik} - x_{j\cdot})^2}} \tag{2.2}$$

- **Cosine angular similarity** between objects i, j is defined as

$$\phi_{ij} = \frac{\sum_{k=1}^{p} x_{ik}x_{ik}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{ik}^2}}. \tag{2.3}$$

- **Kendall τ correlation coefficient** measures similarity of ordered data by their ranks and are computed as

$$\phi_{ij} = \frac{2}{n \cdot (n-1)} \sum_{1 \leq k < l \leq p} \text{sign}(x_{ik} - x_{jl})\text{sign}(x_{jk} - x_{il}). \tag{2.4}$$

2.3.1.2 Distances in probability theory

I am interested to measure how much information is shared between two probability densities μ and π (Ali and Silvey, 1966; Csiszar, 1967; Csiszár and Shields, 2004; Reid and Williamson, 2011). In other words, I judge how close both probability densities are to each other. Csiszar (1967) introduced the class of *f-divergences* which includes many probabilistic distance measures, such as the Kullback-Leibler divergence. Let $f(u)$ be a continuous, convex function defined for $u > 0$ and $f(1) = 0$. Then, the f-divergence is defined as

$$D_f(\mu||\pi) = \mathbb{E}_\mu \left[f\left(\frac{\mu(x)}{\pi(x)}\right)\right] = \sum_{x \in X} \pi(x) \cdot f\left(\frac{\mu(x)}{\pi(x)}\right) \tag{2.5}$$

Then, I can specify f in different ways to generate a large number of probabilistic distance measures.

- Variational distance (L_1 norm): function $f(u) = |u - 1|$ leads to

$$D_f(\mu||\pi) = \sum_{x \in X} |\mu(x) - \pi(x)|$$

- Hellinger distance squared: function $f(u) = 0.5 \cdot (\sqrt{u} - 1)^2$ leads to

$$D_f(\mu||\pi) = 0.5 \sum_{x \in X} (\sqrt{\mu(x)} - \sqrt{\pi(x)})^2 = H^2(\mu, \pi)$$

  which can be viewed half the L_2 norm squared for the square root of the variables μ, π where the Hellinger distance is defined as

$$H(\mu, \pi) = \frac{1}{\sqrt{2}} ||\sqrt{\mu(x)} - \sqrt{\pi(x)}||_2$$

- Bhattacharyya distance: function $f(u) = -\sqrt{u}$ leads to

$$D_f(\mu||\pi) = -\sum_{x \in X} \sqrt{\mu(x) \cdot \pi(x)}$$

- Harmonic mean: function $f(u) = \frac{-2u}{u+1}$ leads to

$$D_f(\mu||\pi) = -2 \sum_{x \in X} \frac{\mu(x) \cdot \pi(x)}{\mu(x) + \pi(x)}$$

- Kullback-Leibler divergence $KL(\mu||\pi)$: function $f(u) = u \cdot \log(u)$ leads to

$$D_f(\mu||\pi) = KL(\mu||\pi) = \sum_{x \in X} \mu(x) \cdot \log\left(\frac{\mu(x)}{\pi(x)}\right).$$

2.3.1.3 ChromoPainter similarity measure

General continuous and probability proximity measures treat the SNP markers as independent. However, due to strong LD (see Section 1.4.3) I would like to apply the proximity measure ChromoPainter (Lawson et al., 2012; Lawson and Falush, 2012) which accounts for the correlation structure in the genetic data. The recombination map has been inferred from PHASE (Stephens et al., 2001). However, due to

limited SNP density in the Mars dataset correlation information can only be exploited to a modest extent.

The ChromoPainter proximity measure takes as input the best haplotype pair for each training dog. This input data is utilized to compute the square coancestry matrix which estimate the proportion an individual coalesces (copies DNA) from another individual, i.e. how much two genetic marker profiles overlap. This coancestry matrix is aggregated by breed to form a $b_m \times b_m$ matrix where b_m is the number of breeds which indicates how much each breeds copies from all other breeds.

To compute this coancestry matrix I use the expected total genetic length of DNA that a given training individual copies from all other training individuals. In the extreme event of an infinite recombination rate the unlinked case of PCA with genetic drift correction is recovered where the analysis is performed at the SNP level (Price et al., 2006; Patterson et al., 2006; Lawson et al., 2012). In this limiting case all training individuals which carry the same allele are equally likely to be copied from while we do not inherit from any other of the individuals. Further technical details of this ChromoPainter proximity measure will be discussed in Chapter 5.

2.4 Graphical methods for proximity matrices

I use three main visualization approaches to illustrate the different proximity matrices:

- **Heatmap**: heatmaps visualize the breed proximity matrix by encoding each matrix entry by a colour chosen directly proportional to its magnitude. In this document proximity matrix entries are encoded by a colour continuum which runs from red (corresponding to the lowest matrix values), orange, yellow to white (highest values). Looking at a corresponding matrix entry, it is easy to judge how related a breed pair is.

- **Dendrogram**: I apply hierarchical clustering to the breed proximity matrix (Izenman, 2008; Everitt and Hothorn, 2011). Initially each individual forms its own cluster and then the algorithm continues iteratively to join the most similar clusters at each stage until there is only a single cluster. For cluster joining, I use complete linkage which defines the cluster distance as the maximum distance between any two individuals, one from each cluster. Then, the clusters can be drawn as a hierarchical tree diagram, referred to as dendrogram. The height at which two clusters are combined to form a new, larger cluster provides information about the similarity of those two clusters, i.e. similar clusters are joined at low heights (highly correlated clusters; close clusters) while more distant clusters are connected higher up in the dendrogram (large branch lengths before clusters are joined). Therefore, larger differences in height where clusters are merged, implies more substantial structure in the data.

  Similarly to heatmaps, I will also encode each breed branch of the dendrogram according to a pre-specified breed group membership. However, color-encoded dendrogram branches can only be easily distinguished for a small number of groups.

- **MDS**: for the purpose of direct inspection it is desirable to visualize data sets into 2 or 3 dimensions. Ordination (from the Latin ordinatio and German Ordnung) is the arrangement and positioning of experimental objects as points along one or several axes of references in a space that contains less dimensions than the original space (Krzanowski, 2000; Quinn and Keough, 2002; Legendre and Legendre, 2012). Ordination is a term which was coined in the ecology community

and is better known as scaling to statisticians.

I focus on the methods multidimensional scaling (MDS) and principal component analysis (PCA) which are the most popular multivariate data analysis tools to perform dimension reduction for ordination (Gower, 1966; Mardia et al., 1980; Legendre and Legendre, 2012). Both techniques use as input a $n \times d$ data matrix. Then, in MDS I compute the distance between the n objects to obtain a $d \times d$ distance matrix. This distance matrix is used to compute a low-dimensional representation which best retains the distances therein ((Q technique)). PCA can also be viewed as Q-technique if we were to compute it as special case for MDS applying Euclidean distance as proximity measure. In particular, PCA has least-square optimality, such that $V = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(d_{ij}^2 - \hat{d}_{ij}^2 \right)$ is at a minimum where d is the Euclidean norm between objects i, j and $\hat{d}$ in the reduced space of principal component scores (Krzanowski, 2000; Izenman, 2008). However, typically PCA is presented as variance maximization technique which directly uses the input data matrix (R-technique).

McVean (2009) show that PCA can be used to uncover biological knowledge: firstly, location of individuals in the projected space formed by the principal component scores can be related to the coalescent time of two samples. Furthermore, closeness on the principal component axes of an individual to other populations is utilized to judge its admixture contributions. In this document I will proceed using MDS to study proximities in two dimensional coordinates. because of the wide range of distance and similarity measures beyond Euclidean distance I will consider.

2.5 Breed similarity results using genotypes and haplotypes

In this section I will study experimental results in a two-stage process by first evaluating pairwise proximities between dog breeds and then using this information to visualize relationships between breeds. Then, by visual inspection I will judge how well a given approach discriminates among groups of breeds.

Furthermore, I use four different canine data representations whereas one representation is based on the original genotype data while the others use different forms of a haplotype representation.

- **Genotypes**: original genotype data is used.
- **SmallHap**: feature vectors are defined by haplotypes which have non-zero frequencies in at least one of the two breeds whose proximity I evaluated.
- **BigHap**: feature vectors are defined by haplotypes which have non-zero frequencies in any of the breeds.

2.5.1 Selection of subset of breeds for literature comparison

The last topic I discuss relates to how visualize breeds in a 2D MDS plot. There are $b_M = 125$ breeds in our datasets which can be projected and visualized as b_M points in a 2D Cartesian coordinate system. However, this approach makes it challenging to judge the separation between breeds. Therefore, I follow Wilcox and Walkowicz (2010); Rimbault and Ostrander (2012) who suggest a rough categorization of breeds into groups according to their appearance and behavior. This categorization enables the data analyst to use visual displays to judge whether groups of breeds rather than individual breeds are well separated. For example, Vonholdt et al. (2010) performs a hierarchical clustering analysis of a dataset of at least five (and less than 12) dogs across 79 breeds using 48K SNPs. On the other hand, the Mars

dataset has only 0.32K markers but a median of 55 purebred dogs per breed. To compare results with Vonholdt et al. (2010), I first reduce the number of breeds to $b_H = 64$ which is the intersection of breeds present in both datasets. The reduction of breeds leads to those 11 groups of functional/phenotypic with their b_H member breeds (breed IDs are listed in brackets):

- **Ancient, Spitz dog**: Afghan Hound (1), Akita (3), Alaskan Malamute (4), American Eskimo Dog (5), Basenji (9), Chow (40), Saluki (104), Samoyed (105)
- **Toy breeds**: Brussels Griffon (29), Chihuahua (37), Papillon (90), Pekingese (92), Pug (100), Shih Tzu (112)
- **Spaniels**: Cavalier King Charles Spaniel (35), English Cocker Spaniel (51), English Springer Spaniel (53), German Shorthaired Pointer (60)
- **Scent hounds**: Basset Griffon Vendeen Petit (10), Basset Hound UK (11), Basset Hound US (12), Beagle UK (13), Beagle US (14), Beagle US2 (15), Bloodhound (21) Dachshund LH (44), Dachshund MLH (45), Dachshund MWH (46), Dachshund (47)
- **Working dogs**: Dobermann Pinscher (50), German Shepherd Dog (59), Poodle (96), Portuguese Water Dog (99), Schnauzer Giant (107)
- **Mastiff-like breeds**: Boston Terrier (25), Boxer (27), Bulldog (31), Bullmastiff (32), French Bulldog (58), Mastiff (83), Staffordshire Bull Terrier (115)
- **Small terriers**: Briard (28), Norwich Terrier (88), Scottish Terrier (109), West Highland White Terrier (122), Yorkshire Terrier UKKC (124), Yorkshire Terrier US (125)
- **Mastiff-like breeds**: Great Dane (65), Rottweiler (102), Saint Bernard (103)
- **Retrievers**: Golden Retriever UK (62), Golden Retriever US (63), Labrador Retriever UK (78), Labrador Retriever US (79), Labrador Retriever (80)
- **Herding dogs**: Australian Shepherd (8), Border Collie (22), Collie (42), Pembroke Welsh Corgi (93), Shetland Sheepdog UK (110), Shetland Sheepdog US (111)
- **Sight hound**: Borzoi (24), Greyhound (68), Whippet (123)

2.5.2 Discarding correlation structure in genotypes

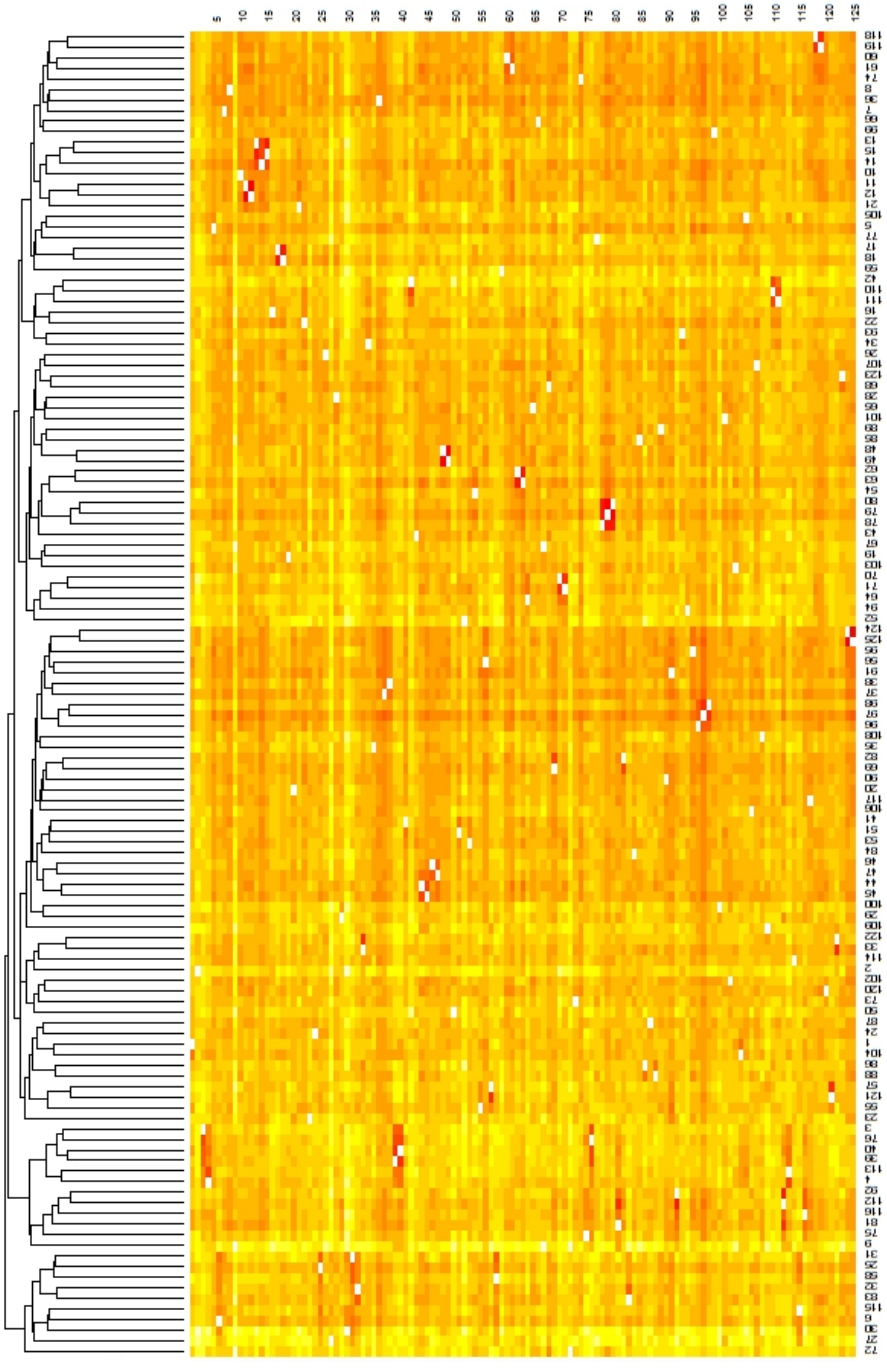


Figure 2.1: This figure shows a heatmap visualization of the Manhattan breed distance matrix using genotype data of 125 breeds. The columns are re-ordered according to the dendrogram of the hierarchical clustering of the distance matrix with complete linkage. Breed pairs which tend towards red have small distance (high similarity) while breed pairs going to the yellow-white spectrum are more distant breeds. Breeds have considerable distance among each other which can be seen from mostly yellow-colored matrix entries and mostly long dendrogram leaf branches. Furthermore, the dendrogram shows a flat cluster structure which implies limit subpopulation structure.

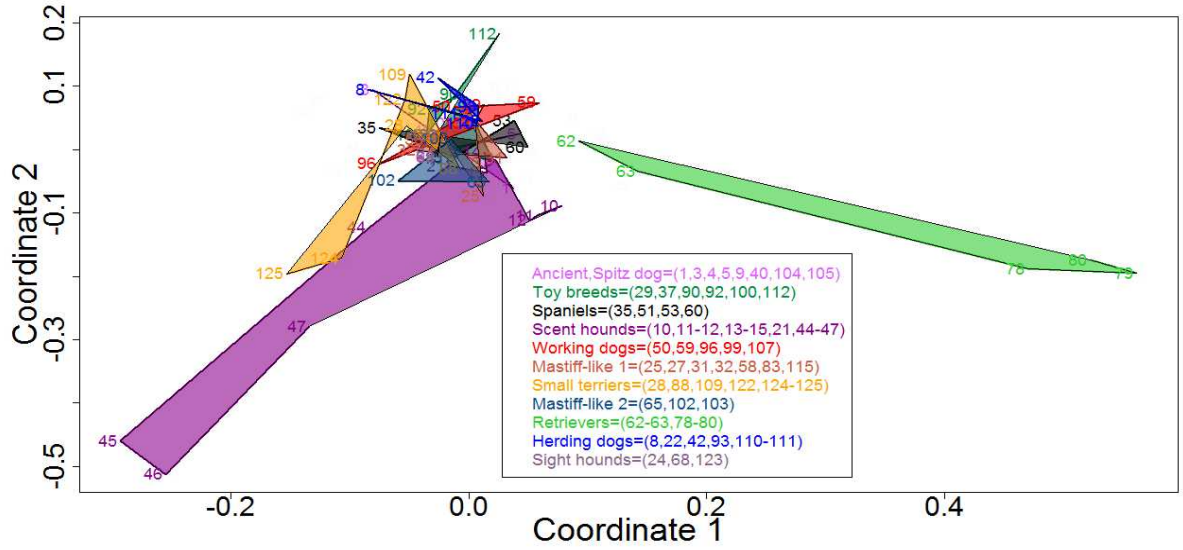
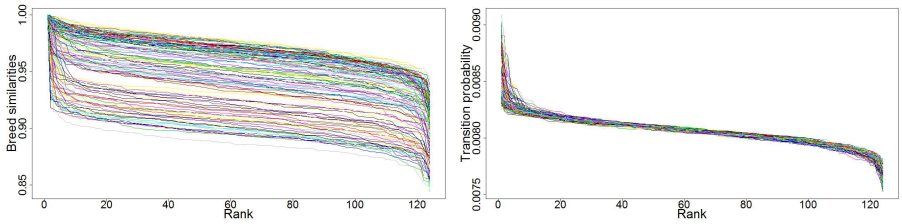


Figure 2.2: This figure shows the MDS plot for the **ChromoPainter** coancestry similarity matrix which has been converted to a distance matrix. This measure visually discriminates Retrievers and Scent hounds while Small Terriers are partially separated out.

This section discusses the use of genotype data, i.e. sequences over the ternary alphabet $\{0, 1, 2\}$, in similarity computations which do not account for LD. Typically, studies utilizing SNP data are based on allele sharing, i.e. Hamming distance (Vonholdt et al., 2010). However, the Hamming distance equally weighs mismatches although allele distance 0-2 is evolutionary more different than mismatches 0-1 and 1-2. Therefore, I focus on Minkowski distances as proximity measure. Furthermore, I would equally weigh all three difference computations which leads to the Manhattan distance.

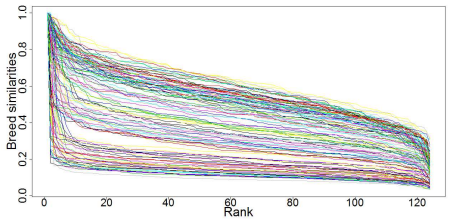
In this analysis I am not interested in the pairwise computation of the proximity measure between any two dogs in the training dataset but rather between any two breeds. However, the naive computation of the Manhattan breed distance matrix is very computationally expensive. There are a median of 55 training dogs per breed. So, to compute the pairwise distance between two breeds on average I need $\frac{55^2}{2} \approx 1500$ symmetry-adjusted computations composed of evaluating the Manhattan distance between the SNP markers of two given training dogs. This previous step needs to be computed pairwise for all breeds (symmetry-adjusted) for $\frac{1}{2}\binom{125}{2} = 3875$ times. Therefore, in total there are $\frac{55^2}{2} \cdot \frac{1}{2}\binom{125}{2} \approx 5.8$ million evaluations of the Manhattan distance. A more efficient way to compute this Manhattan breed distance matrix is by computing the allele frequencies for all SNPs breedwise. And then I can compare these breed frequencies (symmetry-adjusted) for $\frac{1}{2}\binom{125}{2} = 3875$ between any two breeds. The breed frequency comparison between two breeds takes 6 addition and 18 multiplication operations.

I visualize the $b_m \times b_m$ Manhattan breed distance matrix using a heatmap. In Figure 2.1 I see the heatmap for the Manhattan distance matrix based on genotype data and in Figure 2.3 (i) I show the colour continuum with an integrated density plot showing the distribution of the different distance values. I see the same heatmap again in Figure 2.4 (a) while in Figure 2.4 (b) I see that heatmap for the same data but the distances have been converted to similarities. Therefore, more closely related breeds have smaller values using distances and higher values when applied to similarities. In this heatmap in Figure 2.1 I also see the result of a hierarchical clustering algorithm applied to the columns of the breed distance matrix.

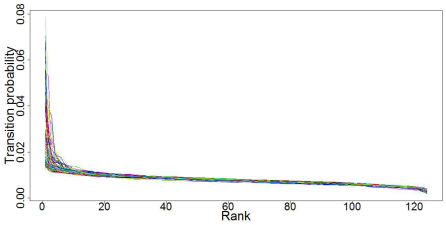


(a) Breed similarity for $f = \exp(0)$

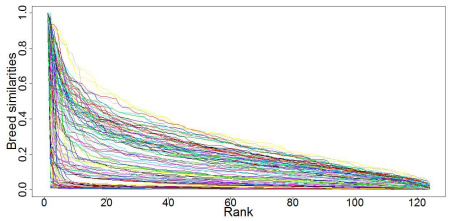
(b) Transition probability for $f = \exp(0)$



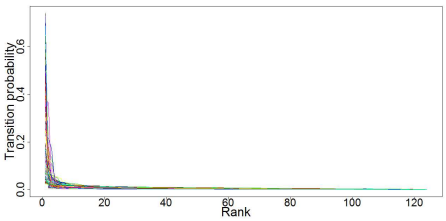
(c) Breed similarity for $f = \exp(3)$



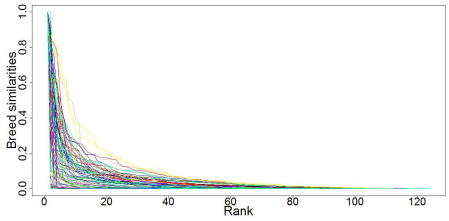
(d) Transition probability for $f = \exp(3)$



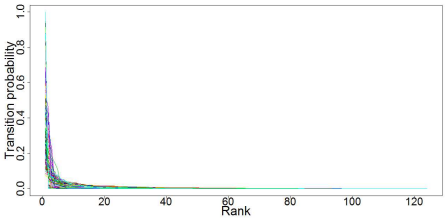
(e) Breed similarity for $f = \exp(4)$



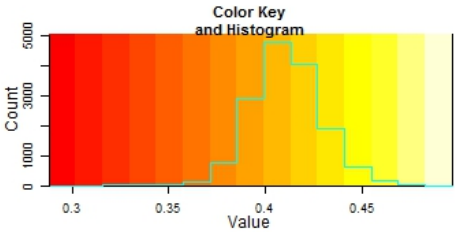
(f) Transition probability for $f = \exp(4)$



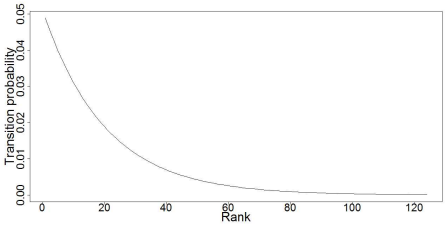
(g) Breed similarity for $f = \exp(5)$



(h) Transition probability for $f = \exp(5)$



(i) Distribution of genotype distances



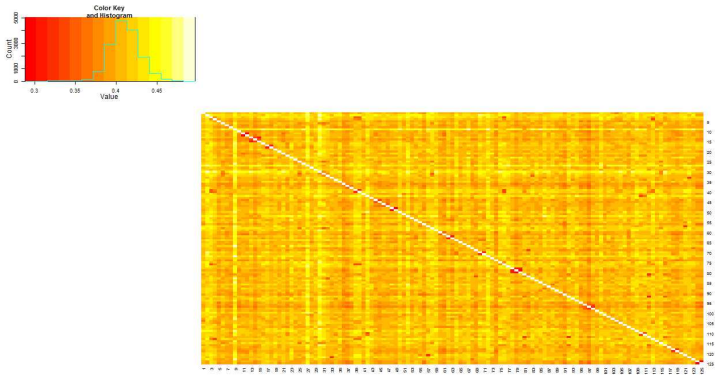
(j) Rank exponential decay function g for $s = 0.05$

Figure 2.3: In Figures (a,c,e,g) each line corresponds to a breed. For each of these breeds I show the decreasingly ordered breed genotype similarities which have been exponentially transformed. In Figures (b,d,f,h) I show the corresponding breed transition probabilities from closest (most similar, left) to furthest breed (right). Figure (i) shows the distribution of distance values according to heatmap in Figure 2.1. Finally, Figure (j) shows the proposed transition probability based on rank in distance sorted breeds.

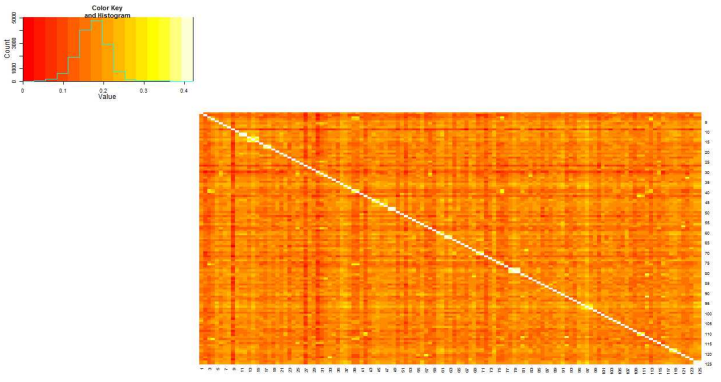
A visual inspection of the integrated density plot in Figure 2.3 (i) and the heatmap in Figure 2.1 itself suggests that the distances are almost normally distributed with a positive skew towards higher distance values. Furthermore, the dendrogram shows that the tree wide and not very deep. These plots show a global picture of the distance distribution. However, first I need to examine how steep the decay of the ordered breed distances is. Given that the decay of the original ordered distances (rank) for each breed is very flat I apply an exponential transformation which contains a scalar scaling factor f which tunes the steepness of the decay. For that purpose for each breed b I define the list of ordered distances to the $i-$th next breed $o' = \left[d_{\sigma(1)} = 0, d_{\sigma(2)}, \ldots, d_{\sigma(b_M)}\right] \in \mathbb{R}^{b_M}$ where $\sigma(1) = b$ corresponds the breed itself, $\sigma(2)$ refers to the closest non-zero distance breed and in general $\sigma(i)$ denotes the $i-$th closest breed in distance. Then, I remove the first element from the list $d_{\sigma(1)}$ and centre by $d_{\sigma(2)}$ to obtain list $o = \left[0, d_{\sigma(3)} - d_{\sigma(2)}, d_{\sigma(4)} - d_{\sigma(2)}, \ldots, d_{\sigma(125)} - d_{\sigma(2)}\right] \in \mathbb{R}^{b_M - 1}$. Then the centered distances in list o are converted to exponentially decaying breed similarities which are given by $s = \exp(-f \cdot o) \in \mathbb{R}^{b_M - 1}$. In Figure 2.3 I see the ordered exponential decay transformed functions for the four scaling factors $f = \exp(0) = 1$ (a), $f = \exp(3) \approx 20.1$ (c), $f = \exp(4) \approx 54.6$ (e) and $f = \exp(5) \approx 148.4$ (g), one line for each of the b_M breeds. These plots show that the breed decay is not very homogeneous across breeds, and for smaller values of f the decay is still very flat. Finally, these exponential decay transformed breed functions can be converted into breedwise transition probabilities using $t = \frac{s}{\sum s} \in \mathbb{R}^{b_M - 1}$ which are shown in Figures 2.3 (b,d,f,h) for the same four scaling factors. These figures show that there is very quick decay for the closest about first five breeds to breed b, then a very flat decay for most breeds except for the last 10 breeds which show another higher negative slope.

We would like to use the transition probabilities to propose breeds within a simulation-based framework discussed in Chapter 4. To ensure the same amount of exploration across breeds I will form a breed-independent decay function for the transition probabilities. However, note that for a fixed amount of exploration the breed proposed depends on the current breed. Furthermore, the transition probabilities should be decaying quickly until about half of the breeds and then saturates at a low level. The exponential decay transition probability function I have in mind with these two characteristics is shown in Figure 2.3 (j) and is defined as function $g'(r) = \exp(-f \cdot r) \in \mathbb{R}$ where the scaling factor is set to $f = 0.05$ and rank is defined for $r = 1, \ldots, b_M$. Then, the closest breed has transition probability of 5 percent, the first six breeds have a cumulative transition probability of 25 percent, the first 14 closest breeds have a cumulative transition probability of 50 percent and the first 62 breeds cover cumulatively 96 percent of the transition probability. Then, to apply function g to our original Manhattan distance matrix I form the ordered list of distances for each breed and replace these values by function g. In other words, for each breed I replace lists $\left[d_{\sigma(1)} = 0, d_{\sigma(2)}, \ldots, d_{\sigma(b_M)}\right] \in \mathbb{R}^{b_M}$ by $[0, g(1), \ldots, g(b_M - 1)] \in \mathbb{R}^{b_M}$. The heatmap for the rank genotype matrix using function g can be seen in Figure 2.4 (c) while the corresponding rank similarity matrix is shown in Figure 2.4 (d). From these figures it can be easily seen that across breeds the majority of breeds have high distance, i.e. small similarity.

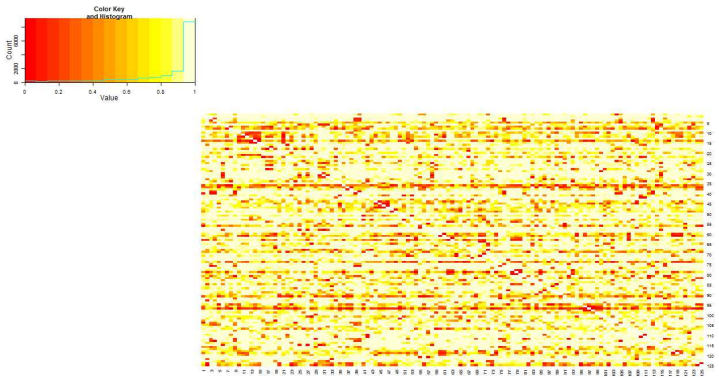
In Figure 2.7 (a) the MDS plot for the original genotype distance is shown. Four breed groups, such as Ancient, Spitz dog, Toy breeds, Mastiff-like and Retrievers are well separated while other breed groups show more overlap. In Figure 2.7 (b) we see the MDS plot for the rank adjusted genotype distance shown in Figure 2.4 (c) which pulls the previously well separated four groups further to the centre of gravity which leads to a lower signal-to-noise ratio.



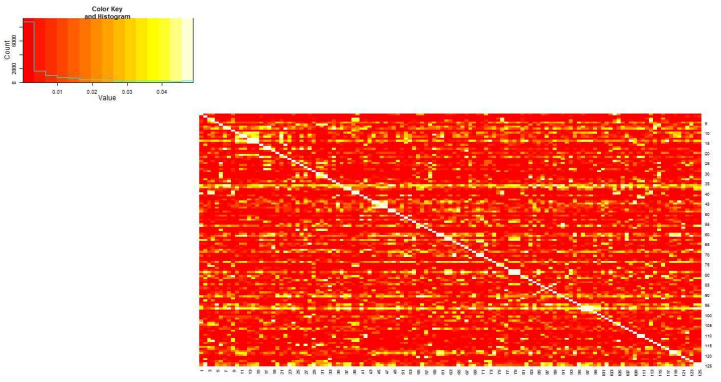
(a) Original genotype distance



(b) Original genotype similarity

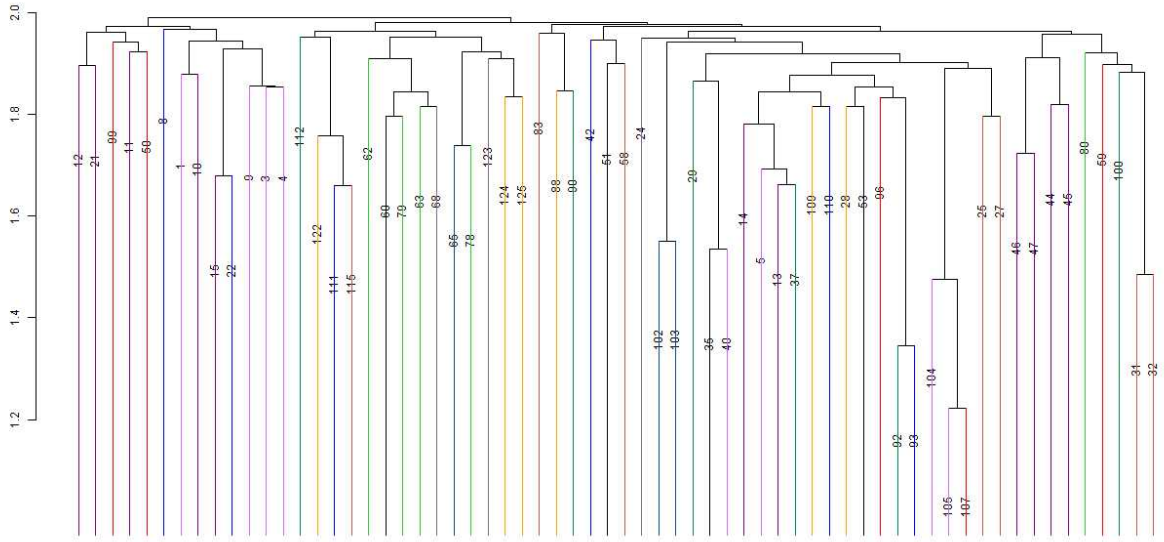


(c) Rank genotype distance

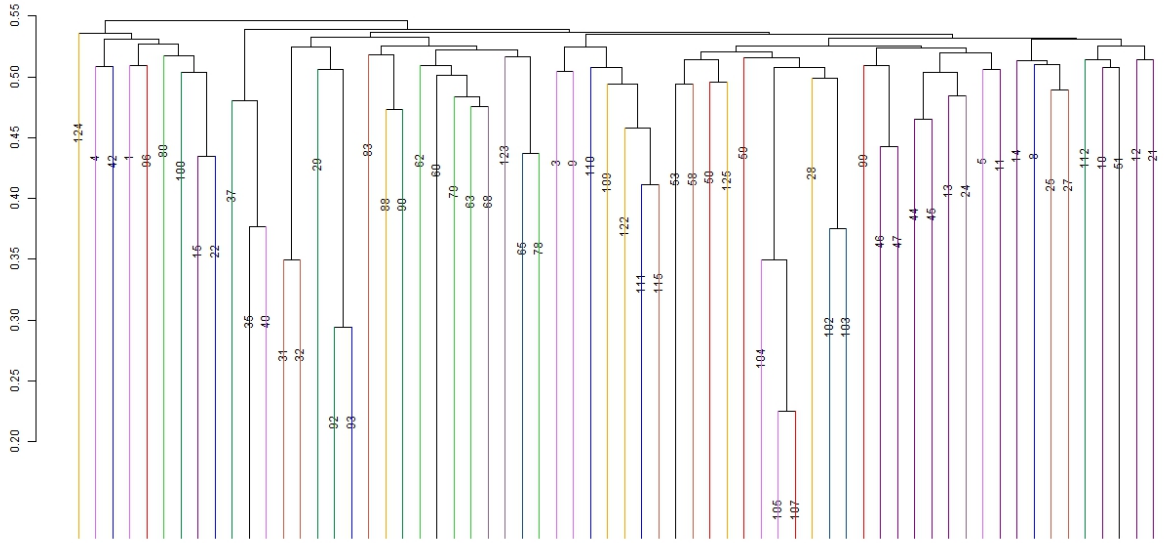


(d) Rank genotype similarity

Figure 2.4: These Figures show a heatmap either based on the distance (left column) or similarity matrix (right column). In the left column red denotes low distance (high similarity) and white refers to high distance (small similarity) while in the right column it is the reverse case: red denotes little similarity and white high similarity. The first row shows the heatmap based on the original distance matrix while in the second row the genotype distances have been transformed by function f from Figure 2.3 (j), such that most breeds are very far away from a given breed (whitish colour) and only few breeds very close (red).



(a) Manhattan distance



(b) Pearson correlation

Figure 2.5: These figures show the hierarchical clustering results using complete linkage for Manhattan distance and Pearson correlation based on the **SmallHap haplotype data**. Breeds from the same group have their branches shown in the same colour. Although breeds from the same breed group tend to be adjacent I notice that breeds from a given group are not distributed homogeneously, i.e. not all breeds from the same group are in the same cluster. As before there is a flat cluster structure confirming limited population substructure. Furthermore, most breeds are associated with long leaf branches suggesting strong differences in breed. Compared with the genotype data dendrogram there are fewer short branches which suggests less strong breed discrimination.

2.5.3 Haplotype analysis discarding LD

In this section I use breed-wise haplotype frequencies to perform the similarity computations. In Section 3.2.2.1 I discuss details on how I derive haplotype frequencies from the purebred training dataset. These haplotype-based proximity measure computations will be distinguished in two cases which I refer to as **SmallHap** and **BigHap**.

SmallHap: in the SmallHap case proximity between two breeds b_1, b_2 on a given chromosome c is defined by comparing haplotype frequency vector from each breed based on the set of haplotypes that have a non-zero frequency in at least one of the breeds. To illustrate the breed frequency vectors v_{b_1} for breed b_1 and v_{b_2} for breed b_2, I give an example over three markers. For each of the two breeds PHASE returns non-zero frequencies f for four haplotypes each:

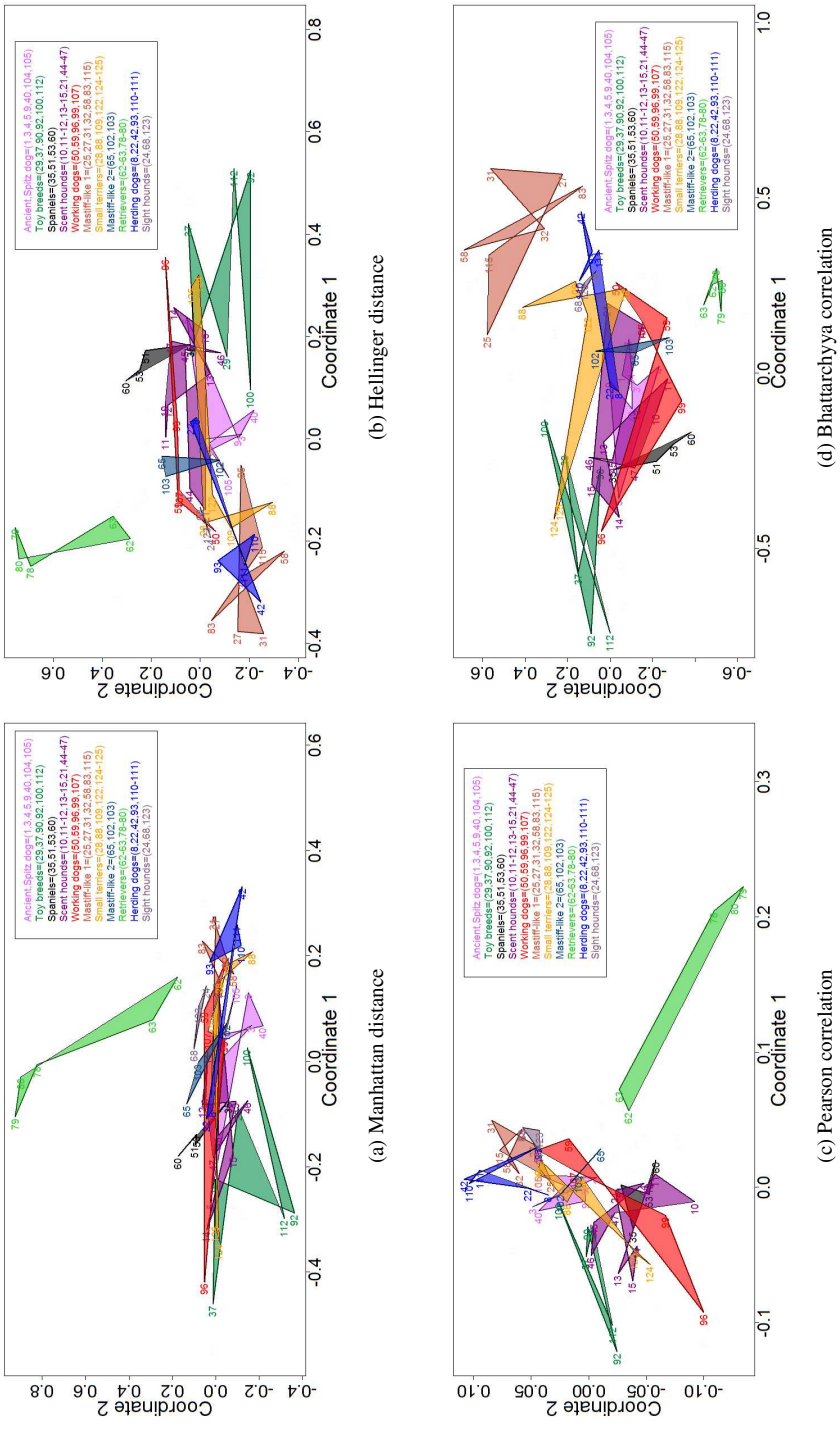
$$f_{b_1}\begin{pmatrix}001\\010\\100\\111\end{pmatrix}=\begin{pmatrix}0.1\\0.4\\0.45\\0.05\end{pmatrix},\quad f_{b_2}\begin{pmatrix}001\\010\\110\\111\end{pmatrix}=\begin{pmatrix}0.2\\0.3\\0.4\\0.1\end{pmatrix} \tag{2.6}$$

I see in equation 2.6 that both breeds share 3 haplotypes and the union contains 5 haplotypes with the following frequencies:

$$f_{b_1}\begin{pmatrix}001\\010\\100\\110\\111\end{pmatrix}=\begin{pmatrix}0.1\\0.5\\0.35\\0\\0.05\end{pmatrix}=v_{b_1},\quad f_{b_2}\begin{pmatrix}001\\010\\100\\110\\111\end{pmatrix}=\begin{pmatrix}0.2\\0.3\\0\\0.4\\0.1\end{pmatrix}=v_{b_2} \tag{2.7}$$

Then, to compute the breed proximity $d^c_{b_1,b2}$ between breeds $b_1, b2$ on chromosome c I use vectors v_{b_1} and v_{b_2} according to equation 2.7 as haplotype frequency representation. To obtain the final proximity $d_{b_1,b2}$ for breeds b_1, b_2 I average $d^c_{b_1,b2}$ over all chromosomes. I will study two distance measures, Manhattan distance and Hellinger distance, and two correlation measures, Pearson correlation and Bhattarchyya correlation, depending on whether the focus is the distance in haplotype frequencies or similarity in shape thereof. None of these proximity measures separates four breed groups as well as the original genotype distance shown in Figure 2.4. All four measures show the retriever group as distinctive cluster, and the Bhattarchyya correlation additionally spreads out the Mastiff-like group. It appears none of these measures distinctly separates the breed better than others. However, depending on the measure certain groups are more spread out, such as (Toy breed) for the Manhattan distance, (Toy breed, Spaniels, Mastiff-like, Herding) for the Hellinger distance, (Toy breeds, Herding) for the Pearson correlation and (Toy breeds, Spaniels, Mastiff-like) for the Bhattarchyya distance.

In Figure 2.6 I see the dendrogram corresponding to hierarchical clustering with complete linkage using each of the four proximity measures. I label branches in the dendrogram by breed ID and colour by breed group. However, I find that this illustration technique is more challenging to interpret by visual inspection because the colors cannot be chosen very distinctly due to the large number of breed groups. The dengrogram seems to confirm the results I discussed for the MDS plots.



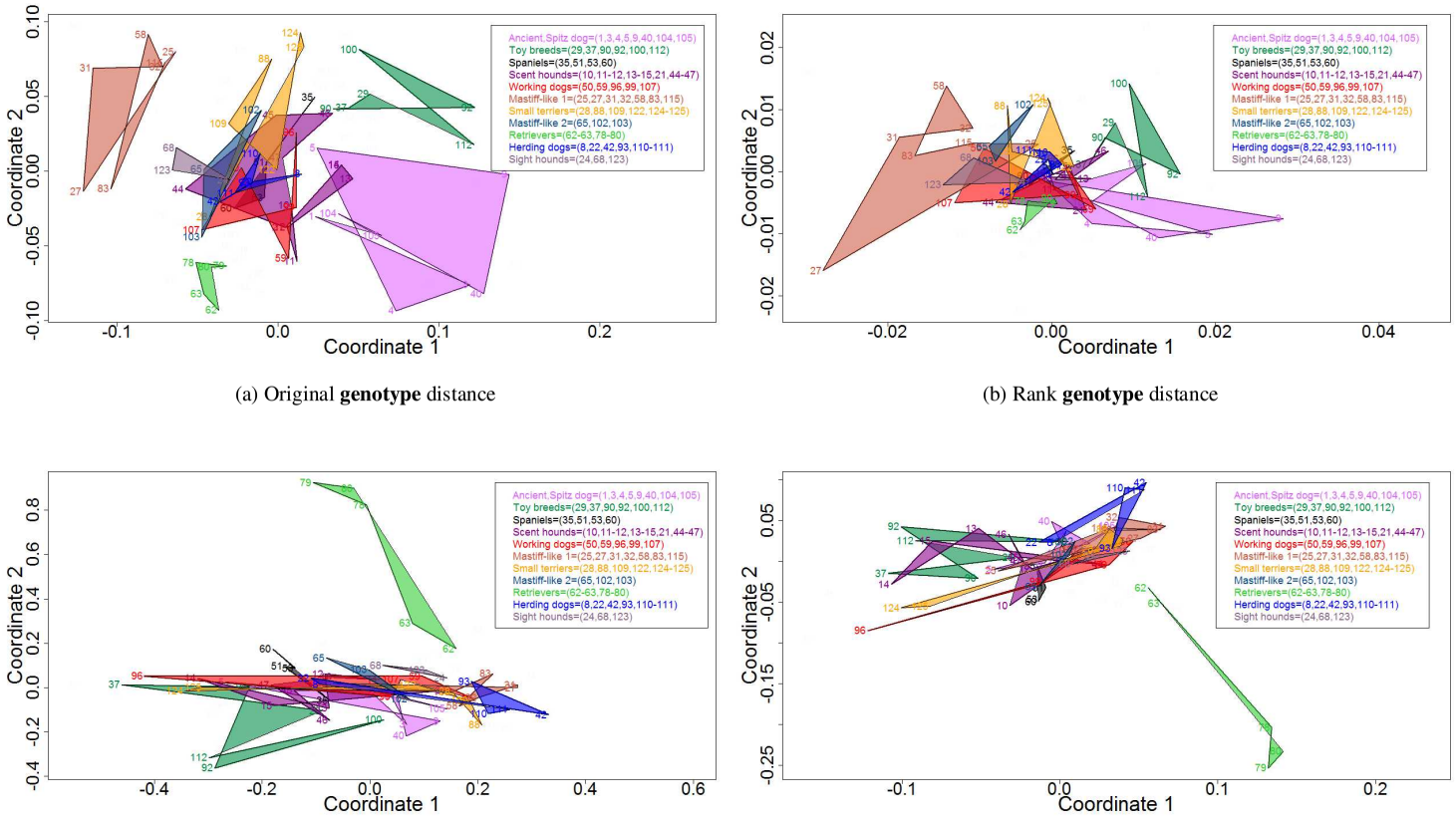
(a) Manhattan distance

(b) Hellinger distance

(c) Pearson correlation

(d) Bhattarchyya correlation

Figure 2.6: These figures shows the MDS plot for two distance, Manhattan distance (ManSmallHap) and Hellinger (HellSmallHap), and two correlation measures, Pearson (PearSmallHap) and Bhattacharyya (BhattSmallHap) based on the **SmallHap haplotype data**. Breeds from the same group are highlighted by transparent polygons. I see that all four proximity measures separate out the Retriever group and BhattSmallHap also separates out Mastiff-like breeds. Furthermore, some breeds are partially visually discriminated, such as (Herding dogs, Toy breeds) for ManSmallHap, (Herding dogs, Mastiff-like breeds, Toy breeds) for HellSmallHap, (Herding dogs, Toy breeds, Working dogs) for PearSmallHap and (Small terriers, Spaniels, Toy breeds) for BhattSmallHap.



(a) Original **genotype** distance

(b) Rank **genotype** distance

(c) Manhattan **BigHap haplotype** distance

(d) Pearson **BigHap haplotype** correlation

Figure 2.7: This figure shows MDS plots where breeds from the same group are highlighted by transparent polygons. The first row shows the original Manhattan distance (MDG) and the rank Manhattan distance based on **genotype data** while the second row illustrates the Manhattan distance (ManBigHap) and Pearson correlation (PearBigHap) based on the **BigHap haplotype data**. The MDG measure separates out Ancient/Spitz dog, Mastiff-like breeds, Retrievers, Toy breeds and partially discriminates Small terriers. Both, ManBigHap and PearBigHap visually discriminate Retrievers and partially out either Herding dogs, Toy breeds or Herding dogs, Scent hounds, Toy breeds, respectively.

#Chr ID	#SNPs	haps	#Chr ID	#SNPs	haps	#Chr ID	#SNPs	haps
1	14	9.5	**11**	13	7.2	**27**	13	7.6
2	13	5.2	**12**	13	7.7	**29**	14	14.5
4	15	21.1	**13**	15	19.6	**31**	7	0.1
5	17	51.7	**14**	16	10.1	**32**	12	4
6	12	4	**15**	9	0.4	**34**	8	0.3
7	12	3.9	**17**	13	5.7	**35**	17	40.4
8	12	4	**20**	16	33.8	**38**	13	7.8
9	13	6.3	**21**	10	1			
10	10	0.9	**25**	13	3.8			

Table 2.1: The number of unique haplotypes in 1000s on a chromosome with non-zero population sample frequency each for at least one of the breeds.

BigHap: this approach also uses haplotype frequencies as SmallHap. However, in this approach in vector v_b I include frequencies for all haplotypes which have non-zero frequencies in at least one of the breeds. The number of these unique haplotypes is shown for each chromosome in Figure 2.8. Then, for breed b I center vector v_b elementwise by the average frequency of the haplotype corresponding to this position. This implies that vector v_b has small entries in most positions. The motivation for these big haplotype vectors is to investigate whether absence of a haplotype for a given breed contains information useful for breed separation. Given this motivation correlation is more suitable because it takes all vectorial entries into account, such as same number or zero entries, while a distance measure discards these entries because they cancel out to zero. In Figure 2.7 (c,d) I show the Manhattan distance and Pearson correlation for the BigHap measure respectively. Note that due to this cancellation behaviour for distances Figures 2.7 (c) and 2.6 (a) look the same. Again both measures separate out the retriever group well. Figure 2.7 (d) suggests that the Pearson correlation also reasonably well distinguishes the groups of Toy breeds and herding dogs.

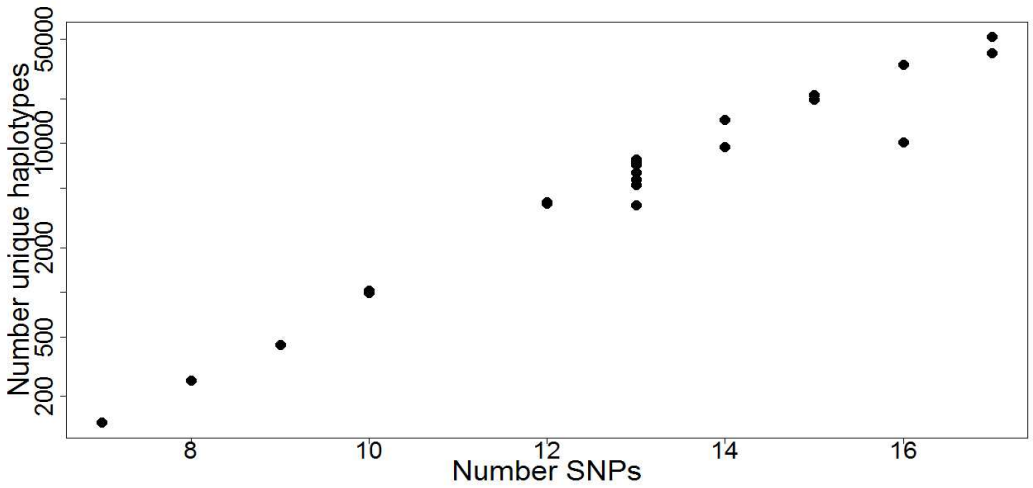


Figure 2.8: This figure shows that the number of haplotypes with frequency > 0.01 for any of the breeds grows exponentially with the number of SNPs on the chromosome. Each of the points represents one chromosome. Note the y-axis is shown in $\log_{10}-$scale.

2.5.4 Haplotype analysis

This section discusses the effect of LD, i.e. correlation of SNPs, on the computation of proximity measures. Not accounting for LD leads to an overestimation bias of the information available in the data. Furthermore, a lack of LD consideration leads to an elevated emphasis and weighting on low SNP density region. However, even if most SNPs are widely spaced out I cannot view them as unlinked with an infinite recombination rate because of the long range LD spanning a large fraction of each chromosome argued in Section 1.4.3. Lawson and Falush (2012) show that LD-enhanced proximity measures improve the signal-to-noise ratio of population separation which is computed as ratio of between-population distance to within-population distance. The authors also give experimental evidence that with more genetic data it is easier to separate populations.

2.5.5 Summary of results

In this section I look how different combinations of proximimity measures and data representations impact visual discrimination of breeds.

Firstly, I note that the genotype data representation provides a slightly better visual MDS separation of breeds groups than either the SmallHap, BigHap or ChromoPainter variant shown in Figures 2.7, 2.6 and 2.2. The genotype representation separates the breed groups Ancient / Spitz dog, Mastiff-like breeds, Retrievers and Toy breeds and partially separates Small terries. The haplotype representations only separate out the Retriever groups and enables partial visual discrimination of the Toy breeds and Herding dogs. Finally, the ChromoPainter measure does not visually separate out more breed groups than other proximity measures, i.e. Retrievers and Scent hounds visually separated out while the Small Terriers group is partially discriminated.

In general, the dendrogram clusterings shown in Figure 2.1 for genotype data and in Figure 2.5 for SmallHap are flat with few strong clusterings which suggests limited population substructure, i.e. there is not much nesting and overlap among breeds. However, there is more cluster structure in the haplotype-based dendrograms than in the the genotypes representations ones which again reflects the fact that the genotype representation offers a stronger separation of breed groups.

Furthermore, most leaf branches in the dengrograms are long and breeds are joined at large height differences. This insight suggests there are large breed dissimilarities showing sizable structure through substantial distance among breed clusters. Figure 2.1 which shows a heatmap of the Manhattan distance using genotype data confirms that most breeds are considerably different because most pairwise breed proximities in the breed distance matrix are encoded as yellow to yellow-orange. However, based on haplotype-based dendrograms in Figure 2.5 some breeds from groups, such as Mastiff-like breeds *(Bulldog (31), Bullmastiff (32), Great Dane (65), Staffordshire Bull Terrier (115))* and Herding dogs *(Border Collie (22), Pembroke Welsh Corgi(93), Shetland Sheepdog US (111))*, have shorter leaf branch length which suggests these breeds are more similar to each other.

Compared to the clustering by Vonholdt et al. (2010) the dendrograms in Figures 2.1, 2.5 show that breeds are less homogeneously clustered in groups. In other words, in our analysis breed groups are less well recovered by the cluster analysis where fewer breeds from the same breed group are members of the same cluster. This may be due to the smaller number of SNPs in our study, i.e. I use 0.32K SNPs compared to 48K SNPs by Vonholdt et al. (2010). Although their paper only has about five dogs per breed

Vonholdt et al. (2010) deal with this limitations by boostrapping 2K replications. Furthermore, similar to our findings Vonholdt et al. (2010) also discovered long branch lengths and flat cluster structure in their dendrograms. Moreover, I did not replicate their finding that the Ancient / Spitz dog form a set of highly divergent breed, such that the top of the dendrogram those breeds are all in one subtree while all other breed groups are in the other subtree. Although this ancient group cannot be separated out in the haplotype-based MDS plots I can see the ancient group as separate polygon in the genotype-based view.

2.6 Conclusions

This work on breed proximity measures was motivated from two directions. Firstly, I explored how breeds are genetically related and cluster in breed space to unravel their similarity relationships. And, secondly, I developed a similarity matrix which can be utilized within a breed-directed update proposal mechanism with a Markov Chain Monte Carlo (MCMC) approach.

As proximity measure I reviewed different distance and similarity measures. Then, based on these measures I used three different ways to visualize breed proximity: heatmaps, dendrograms and MDS plots. Furthermore, I use four different canine data representations whereas one representation is based on the original genotype data while the others use different forms of a haplotype representation.

I also segmented the set of breeds into 11 groups according to Vonholdt et al. (2010). To some extent I replicated some of their findings. For example, I found that the leaf branches in the dendrogram tend to be long suggesting that breeds are distinct, and the clustering structure is flat which is related to little population substructure. However, the detected clusters match the pre-specified breed groups less well than in the work by Vonholdt et al. (2010).

Furthermore, I found that the genotype representation can visually separate out more breed groups in the MDS plots than a haplotype representation, which makes it an attractive choice to be used within the update proposal for the MCMC algorithm.

Chapter 3

Methodology for pure breed identification

In Section 3.1 I will discuss an ancestry inference framework based on maximization of haplotype frequencies which are the foundation for the Mars approach and my novel methodology **DBAncestry**. This chapter deals with the special case of purebred dog prediction while in Chapter 4 I show how I adapt the inference approach to crossbred dogs within the Mars and DBAncestry model.

At first in Section 3.1 I will specify the DBAncestry haplotype frequency model and define the known and unknown parameter for this model followed by a discussion in Section 3.1.1 on how inference is done for the DBAncestry model. Then, in Section 3.2 I will show how chromosomal likelihood $p\,(X_c|\theta)$ in the haplotype frequency model is computed in the Mars (Section 3.2.1) and DBAncestry model (Section 3.2.2). Finally, I will discuss the performance measure for the data analysis in Section 3.3 and the results for purebred prediction in Section 3.4.

3.1 Haplotype frequency model

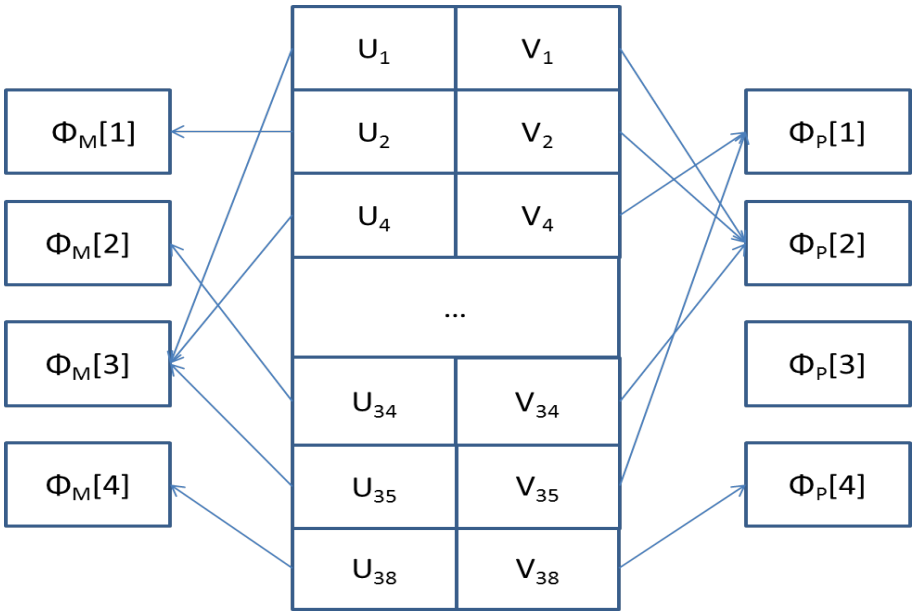


Figure 3.1: Breed data structure Φ : The left column shows the maternal lineage painting and the right column the paternal lineage painting while the centre block refers to the genome painting.

Variables representing the genetic data and the lineage tree type are *known* in the haplotype frequency model:

- **Genotype:** in the haplotype frequency model the genotype data X of a test dog is given where X_c

Chromosome	u_c	breed u_c	v_c	breed v_c
1	$\phi_M[3]$	German Shepherd Dog	$\phi_P[2]$	Pekingese
2	$\phi_M[1]$	Australian Shepherd	$\phi_P[2]$	Pekingese
4	$\phi_M[3]$	German Shepherd Dog	$\phi_P[1]$	German Shepherd Dog
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
34	$\phi_M[2]$	Australian Shepherd	$\phi_P[2]$	Pekingese
35	$\phi_M[3]$	German Shepherd Dog	$\phi_P[1]$	German Shepherd Dog
38	$\phi_M[4]$	Rottweiler	$\phi_P[3]$	Weimaraner

Table 3.1: For a subset of the chromosomes chromosome painting with associated breeds is shown.

refers to the genetic data on chromosome $c \in \{1, \ldots, C\}$.

- **Lineage tree:**
  - **Mars approach:** Φ refers to one of the 11 lineage trees shown in Figure 1.2.
  - **DBAncestry:** I assume only one lineage tree similar to lineage tree ABCDEFGH in the Mars approach. In particular, I assume $\Phi = (\phi_M, \phi_P)$ is composed of two ordered breed lists of length four (with repeats allowed) for the maternal ggps ϕ_M and the paternal ggps ϕ_P, respectively.

The *unknown* variables related to the lineage and genome painting in the haplotype frequency model are estimated:

- **Lineage painting** (specific assignment of breeds to the ggp leaves):
  - **Mars approach:** in this approach variable Φ is reserved for the tree type: therefore, I use an extra variable L which denotes the lineage painting composed of up to eight different ggps, depending on the lineage Φ.
  - **DBAncestry:** the lineage painting is given by the 8-dimensional vector $\Phi = (\phi_M, \phi_P)$. Figure 3.1 illustrates Φ where ϕ_M = [Australian Shepherd, Australian Shepherd, German Shepherd Dog, Rottweiler] and ϕ_P = [German Shepherd Dog, Pekingese, Weimaraner, Weimaraner]. Furthermore, in table 3.1 I show a partial G as textual information. In the special case of pure breed identification in this chapter all eight breeds of Φ are the same, i.e. $\Phi = [\underbrace{b, b, b, b}_{\phi_M}, \underbrace{b, b, b, b}_{\phi_P}]$ where b denotes the purebred breed.
- **Genome painting** G : Given Φ (and L) I generate a painting of each chromosome by choosing one leaf from the maternal leaves and one leaf from the paternal leaves. Let U be a vector length C which lists for each entry $u_c, c \in \{1, \ldots, C\}$ which ggp leaf from the maternal breeds the genotype on chromosome c is inherited from. Vector V is defined for U, but, applied to the paternal breeds. Then, a chromosome painting for chromosome c is defined as a mapping $c \mapsto (u_c, v_c)$. I would like to emphasize that a chromosome painting is defined by a mapping of a leaf pair to a particular chromosome. As the breed assignments to the leaves change, then the breed assignment to the chromosomes changes accordingly. Therefore, any given chromosome painting update cannot assign any breed but only those ones which currently form part of L. The chromosome painting for all chromosomes is referred to as a genome painting $G = (U, V)$. An illustration of the genome painting G for a subset of chromosomes is shown in Figure 3.1.

I integrate these parameters as $\theta = (\Phi, L, G)$ and define the haplotype frequency model in Equation 3.1

Breed	Chromosome	SNP name	0-0	0-1	1-0	1-1
Siberian Husky	1	SNP 1	0.43	0.57	-	-
Siberian Husky	1	SNP 2	0.52	0.48	0.28	0.72
Siberian Husky	1	SNP 3	0.69	0.31	0.46	0.54
Siberian Husky	1	SNP 4	0.06	0.94	0.08	0.92
Siberian Husky	1	SNP 5	0.47	0.53	0.51	0.49
Siberian Husky	1	SNP 6	0.85	0.15	0.66	0.34
Siberian Husky	1	SNP 7	0.93	0.07	0.76	0.24
Siberian Husky	1	SNP 8	0.37	0.63	0.65	0.35
Siberian Husky	1	SNP 9	0.84	0.16	0.95	0.05
Siberian Husky	1	SNP 10	0.03	0.97	0.25	0.75
Siberian Husky	1	SNP 11	0.56	0.44	0.39	0.61
Siberian Husky	1	SNP 12	0.94	0.06	0.90	0.10
Siberian Husky	1	SNP 13	0.76	0.24	0.29	0.71
Siberian Husky	1	SNP 14	0.08	0.92	0.08	0.92

Table 3.2: Breed-specific SNP transition probabilities on chromosome 1 for breed Siberian Husky. There are 14 SNPs sequenced for this first chromosome. In each row those transition probabilities sharing the same old state sum to 1, i.e. $p(0 \to 0) + p(0 \to 1) = 1$, and $p(1 \to 0) + p(1 \to 1) = 1$. Table courtesy of Davison and Fretwell (2012).

$$p(\theta|X) \propto p(X|\theta)p(\theta) \propto p(X|\theta) = \prod_{c=1}^{C} p\left(X_c|\theta\right). \tag{3.1}$$

where I want to maximize posterior probability $p(\theta|X)$ for parameter θ given data X. In Section 3.2 I will show how to compute quantity $p\left(X_c|\theta\right)$. Equation 3.1 is based on the following three assumptions:

- **MAP:** maximum a posteriori estimation,
- **Prior:** uniform prior over parameter vector θ and
- **Independence:** independence of genetic data across chromosomes X_c.

3.1.1 Inference in purebred haplotype frequency model

Let B be the number of breeds in the dataset. Inference of the optimal θ which maximizes $p(\theta|X)$ is found by evaluating $p(\theta|X)\ \forall b \in \{1, \ldots, B\}$. Then, because Φ is given as $\Phi = [\underbrace{b, b, b, b}_{\phi_M}, \underbrace{b, b, b, b}_{\phi_P}]$ I infer breed b as ancestry which corresponds to the maximum of those B evaluations of posterior probability $p(\theta|X)$. Although, in practise I work with the log-likelihood $\log[p(\theta|X)] = \sum_{c=1}^{C} \log[p\left(X_c|\theta\right)]$.

3.2 Computation of chromosomal likelihood

In Section 3.1 I described the haplotype frequency model along with a description on how to perform purebred dog inference in this model. However, I have not outlined yet how to compute chromosomal likelihood $p\left(X_c|\theta\right)$ which is discussed for the Mars approach in Section 3.2.1 and the novel DBAncestry algorithm in Section 3.2.2.

3.2.1 Mars approach:

In Section 3.2.1, I describe the Mars approach which looks at SNP-wise transition frequencies and uses the hidden markov model forward algorithm to combine all SNPs forming X_c to compute $p\left(X_c|\theta\right)$.

3.2.1.1 SNP-wise frequencies

As first step in the breed composition inference I need to characterize each breed given the pure breed genetic data BigPure. The Mars approach is based on the computation of SNP-wise transition frequencies for each pair of breed and chromosome. For example, in Table 3.2 I see the transition probabilities for breed Siberian Husky on chromosome 1. Initial allele states are listed in the row for SNP 1.

3.2.1.2 Evaluating chromosomal likelihood using HMM forward algorithm

I model probability $p_{cuv} = p(X_c|u_c, v_c)$ where pair (u_c, v_c) is the realization of breeds, i.e. a chromosome painting with maternal leaf u_c and paternal leaf v_c. Probability p_{cuv} is computed from a HMM forward algorithm using SNP transition probabilities which are evaluated as a product along the haplotype together with a marginal probability - obtained from a simple frequency estimate - for the initial allele. Therefore, moving along the haplotype there are transitions from 0 to 0, 0 to 1, 1 to 0 and 1 to 1. Commonly, transition probabilities are inferred by the Baum-Welch algorithm for Hidden Markov Models (HMM). However, the Mars approach seems to use a MCMC implementation to infer SNP-wise transition probabilities which I did not analyze.

The HMM used is non-standard because there are two hidden states (the alleles on each chromosome, coded as 0 or 1) underlying each emitted value. Moreover, given the two hidden states the emitted state is determined as their sum, and is therefore not stochastic. Once the transition probabilities are derived in this non-standard HMM, they are subsequently run through an 'LD capping' algorithm to account for linkage disequilibrium. However, I have not investigated how this works.

In the appendix A.4 I discuss the HMM forward algorithm in more detail. At first, I will introduce notation for the HMM. The probability of a chromosome painting for the observed genotype could be naively computed by summing the probabilities of all possible haplotype pairs, given the observed genotype data. However, this is not computationally attractive since there 2^{S-1} possible states for the hidden haplotype pair, where S denotes the number of heterozygous SNP markers on the chromosome. It is more efficient to introduce a HMM forward algorithm to compute this probability. It is based on the idea of separating the observed genotype into two parts and proceeds by recursively accumulating partial results for the hidden allele state sequence. Therefore, I utilize a forward variable to efficiently evaluate this probability. Finally, I provide a toy example for the HMM forward algorithm. For computational reasons, initially, breed probabilities p_{cuv} for all chromosomes and breed pair combinations are precomputed and held in memory.

3.2.2 DBAncestry

In Section 3.2.2 I show the DBAncestry technique which improves upon the Mars approach because it accounts for LD among markers on X_c. As first step as described in Section 3.2.2.1 I phase all purebred training dogs, either separately by breed or jointly across breeds, to obtain estimates $p(H|b, c)$ of haplotype H conditional on a given breed and chromosome. After that in Section 3.2.2.2 I enumerate all haplotype pairs which are consistent with chromosomal segment X_c. A haplotype pair is consistent with a given genotype if at each SNP locus its allele is the sum of the alleles at the same haplotype locus. Then, as final step of DBAncestry in Section 3.2.2.3 I describe three ways to compute $p\left(X_c|\theta\right)$ based on the frequencies of the enumerated haplotype pairs.

3.2.2.1 PHASING

This section discusses how I derive estimates for haplotype frequencies which can be either interpreted estimates of actual sample haplotype which are unknown or estimates of the population from which the

sample was drawn. I look at the whole haplotype rather than at SNP-wise transition frequencies in the Mars approach because I account for the strong canine correlation structure as discussed in Section 1.4.3. Therefore, I performed a literature review which identified four algorithms for phasing genotype data: PHASE (Stephens et al., 2001), fastPHASE (Scheet and Stephens, 2006), BEAGLE (Browning, 2006; Browning and Weir, 2010) and long-range phasing (LRP) (Hickey et al., 2011; Kong et al., 2008). Then, the selected phasing algorithm should at least partially satisfies the following four criteria:

- **R1:** a phasing algorithm which can compute haplotype frequencies.
- **R2:** a phasing algorithm which does not only yield best haplotype pair but also provides as list of high-probability haplotype pairs. I will use this requirement later when I am interested in breed-specific frequencies although I supplied the phasing algorithm with a pooled version of all breeds rather than all breeds separately: breed-specific haplotype frequencies can be inferred in the pooled case by summing probabilities from this list over all dogs in the breed because I know the breed membership in the training data.
- **R3:** phasing algorithm should be applicable for small marker panels of about 10-20 SNPs per chromosome. The criterion is relevant because some of the phasing algorithms are tailored for 10-100K SNPs per chromosome, which is typical for human GWAS.
- **R4:** finally, I would like to choose a phasing algorithm that can take advantage of breed labels in the training data because I may obtain better estimates by accounting for all dogs across breeds in the training but also utilize breed-specific population structure.

An investigation which phasing algorithm requires which of the criteria led to the following list:

✓ **R1:** BEAGLE, PHASE

✓ **R2:** PHASE

✓ **R3:** BEAGLE, fastPHASE, PHASE

✗ **R4:** no algorithm

I decided to use PHASE because it fulfills all the criteria except **R4**, and this motivates a simulation study to investigate whether different ways of segregating dog breeds has an impact on dog classification. Furtermore, with respect to its computational complexity: PHASE scales linearly in the number of markers and quadratically in the number of samples in reference database (Li and Stephens, 2003; Scheet, 2013).

Due to lack of algorithms accounting for R4 I decided to investigate the effect of computing haplotype frequencies for each breed separately or jointly for all training dogs across breeds.

Most phasing algorithms seek haplotypes consistent with genotypes under a panmictic population in Hardy-Weinberg equilibrium (HWE, the independence of an individual's two homologous alleles at a single locus). Phasing separately within each breed allows these assumptions to be satisfied, but accuracy is limited by low numbers of genotyped dogs in many breeds. An alternative is pooled phasing where numbers are increased by jointly phasing the training-set dogs from different breeds. Compared to separate PHASING in pooled phasing (total of 8k training samples in BigPure) I have a factor of 25 or more training samples which is an important computational constraint because of PHASE's computational complexity. The population structure generated by the distinct breeds means that HWE and

long-range linkage equilibrium (LE) are not valid, but there is some evidence from the literature that deviations from these assumptions are not very important in practice: phasing algorithms seek to minimize deviations from these equilibria and this remains a valid strategy even for structured populations which remain far from these equilibria.

Consequently, I conducted simulation experiments to compare the breed classification accuracy, each based on the SNP genotype data from the training dogs as introduced in section 2, for the following two settings:

- **PSEPARATE**: all dogs from the same breed are phased together.,
- **PPOOLED**: dogs across all breeds are phased together.

Intuitively, I expect many haplotypes to be shared across breeds, and pooling dogs over similar breeds will help identify a haplotype in breeds in which that haplotype is rare. Conversely, similar but different haplotypes that segregate in different breeds may be confused by pooling. For PSEPARATE, phasing results may be too strongly biased towards individual breed estimates, which may exaggerate breed differences.

In this part I describe how I compute haplotype frequencies using PHASE for PSEPARATE and PPOOLED and I show which input files PHASE requires and some of the output files it generates. For PSEPARATE I will use Siberian Husky as example based on 66 purebred training dogs where the first chromosome has 14 SNPs (cp. Tables 1.3, A.1). For PPOOLED I use a short artificial example to demonstrate the relevant calculations.

An input file for Siberian Husky on chromosome 1 is specified as follows in PHASE: first I supply the number of individuals followed by information about the number of loci and their physical positions. After that, as input file for PHASE I write the genotype X_1 on chromosome 1 resolving ambiguous phase of the haplotypes arbitrarily. For example, the first purebred Siberian Husky training on chromosome 1 is given by genotype $X_1 = [11022000020002]$ and is written as haplotype pair $H_1 = [11011000010001]$ and $H_2 = [00011000010001]$ where the ambiguous phases correspond to the first two marker positions.

```
66
14
P 1 437564 1681156 2333423 ... 11331464 13204654
SSSSSSSSSSSSS
#1
11011000010001
00011000010001
#2
11111001011001
01110001011001
...
#65
11?111?1011??1
00?100?1010??1
#66
1111?101011001
1101?001011001
```

- **PSEPARATE**:

  For PSEPARATE, I obtain a 'freqs' file with all those haplotypes which have non-zero frequency estimates for each breed b on chromosome c from PHASE. In both cases of PSEPARATE and PPOOLED I refer to this list of n haplotypes as

$$\{H_1, H_2, \ldots, H_n\} \tag{3.2}$$

  which have the following non-zero haplotype frequency estimates

$$\{\hat{p}\left(H_1|b,c\right), \hat{p}\left(H_2|b,c\right), \ldots, \hat{p}\left(H_n|b,c\right)\}. \tag{3.3}$$

  In the following example variable b corresponds to breed Siberian Husky and I use the first chromosome $c = 1$. Each row corresponds to one of the indices $1, \ldots, n$. In this case I obtained $n = 444$ haplotype frequency estimates shown in the third column along with its haplotype representation in column 2. Finally, the last column corresponds to the standard errors of the frequency estimates.

```
index      haplotype     E(freq)          S.E
1 11011000010001    0.036574    0.017191
2 11011000010000    0.000136    0.001032
3 11011000010011    0.004509    0.005507
4 11011000011001    0.000734    0.002307
5 11011000011011    0.000997    0.002797
 ...
442 00101001011001    0.002220    0.004519
443 00100000010001    0.001525    0.004122
444 00100001011001    0.001008    0.003250
```

- **PPOOLED**:

  I do not directly obtain useful haplotype frequencies for PPOOLED from PHASE because they are confounded by the joint breed structure. However, PHASE does not only return pooled population frequencies as output but also a 'pairs' file with the most likely pairs of haplotypes for each of the individuals' genotypes with its posterior probability. The pairs file has the following format on each chromosome: in each row the first two binary strings represent the haplotype pair and the third column represents the posterior probability for this pair. I denote the *i*th haplotype pair of individual j as $[H_1^i(j), H_2^i(j)]$ on some chromosome c and breed b which are omitted from the notation for brevity. Furthermore, each haplotype has an associated posterior probability $p^i(j)$. A new individual is denoted by 'IND' and its integer ID j.

```
IND: #1
10011100000011 , 10110100010101 , 0.034
10011100010111 , 10110100000001 , 0.062
10011100000011 , 10110100000011 , 0.161
10010100000011 , 10111100010101 , 0.186
```

```
10010100000001 , 10111100010111 , 0.017
10010100010101 , 10111100000011 , 0.533
IND: #2
00011100010101 , 10111100010100 , 0.183
00011100010100 , 10111100010101 , 0.792
00111100010101 , 10011100010100 , 0.013
00111100010100 , 10011100010101 , 0.012
IND: #3
10011100000011 , 10011100010101 , 0.835
00011100010101 , 10011100000001 , 0.165
IND: #4
00011100010100 , 00011100010100 , 1.000
IND: #5
00011100000000 , 00011100010100 , 0.990
```

Given these pairs files, in PPOOLED, I list all haplotypes which are part of at least one haplotype pair across all individuals for a fixed breed. Then, these pair files are used to infer breed-specific haplotype frequencies by summing probabilities over all dogs in the breed. To better illustrate the way I deduce breed-specific haplotype frequencies, let us give an example using the verbatim quoted information above. Firstly, suppose that the first four dog individuals all belong to breed br. The first individual from this breed has 6 likely haplotype pairs, the second one has 4, the third one has 2 and the last individual only has one possible haplotype pair. Now, assume I would like to infer the estimated population frequency for haplotype a='10011100000011'. The first computation step consists of calculating a normalization constant $fSUM = \sum_{i \in br} \sum_{j=1}^{J_i} 2 \cdot p^i(j)$ where J_i denotes the number of likely haplotype pairs for individual i. Then, to infer the breed-wise frequency for $f(a)$ I sum over all pairs which contain haplotype 'a', and normalize by $fSUM$. In this example haplotype 'a' occurs 3 times: as first and third pair of the first individual, and as first pair of the third individual. So for $f(a)$ I get

$$f(a) = \frac{0.034 + 0.161 + 0.835}{2 \cdot (0.034 + 0.062 + 0.161 + 0.186 + 0.017 + 0.533 + 0.183 + 0.792 + 0.013 + 0.012 + 0.835 + 0.165 + 1)} = \frac{1.03}{7.986} \approx 0.129$$

3.2.2.2 Enumeration of consistent haplotype pairs

In this section I exhaustively enumerate consistent haplotype pairs for $X_k^j, k \in \{1, \ldots, C\}$ on k-th chromosome on j-th test dog which is described in Section 3.2.2.2. For each X_k^j I enumerate all its m_{jk} consistent haplotype pairs $[H_1^i(j), H_2^i(j)]$ where $i \in \{1, \ldots, m_{jk}\}$. Now that I introduced the notation above consistency can be expressed more formally as $H_1^i(j)[t] + H_2^i(j)[t] = X_k^j[t]\ \forall t \in \{1, \ldots, s_k\}$. The list of haplotypes consistent with X_k^j is of length $2 \cdot m_{jk}$ and referred to as

$$\text{HAPS} = \left\{ \left[H_1^1(j), H_2^1(j)\right], \ldots, \left[H_1^{m_{jk}}(j), H_2^{m_{jk}}(j)\right] \right\}. \tag{3.4}$$

Then, I would like to investigate the number of haplotype pairs m_{jk} consistent with X_k^j, which I compute as product of the ambiguous markers 1 and ?. For SNP 1 I could have had two haplotype pairs

Enumerated: Pair.HaplD	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
Genotype	0	?	0	1	2	?
1.1	0	0 [a]	0	0 [a]	1	0 [a]
1.2	0	1	0	1	1	1
2.1	0	1 [b]	0	0 [a]	1	0 [a]
2.2	0	0	0	1	1	1
3.1	0	1 [c]	0	0 [a]	1	0 [a]
3.2	0	1	0	1	1	1
4.1	0	0 [d]	0	0 [a]	1	0 [a]
4.2	0	0	0	1	1	1
5.1	0	0 [a]	0	1 [b]	1	0 [a]
5.2	0	1	0	0	1	1
6.1	0	1 [b]	0	1 [b]	1	0 [a]
6.2	0	0	0	0	1	1
7.1	0	1 [c]	0	1 [b]	1	0 [a]
7.2	0	1	0	0	1	1
8.1	0	0 [d]	0	1 [b]	1	0 [a]
8.2	0	0	0	0	1	1
9.1	0	0 [a]	0	0 [a]	1	1 [b]
9.2	0	1	0	1	1	0
10.1	0	1 [b]	0	0 [a]	1	1 [b]
10.2	0	0	0	1	1	0
11.1	0	1 [c]	0	0 [a]	1	1 [b]
11.2	0	1	0	1	1	0
12.1	0	0 [d]	0	0 [a]	1	1 [b]
12.2	0	0	0	1	1	0
13.1	0	0 [a]	0	1 [b]	1	1 [b]
13.2	0	1	0	0	1	0
14.1	0	1 [b]	0	1 [b]	1	1 [b]
14.2	0	0	0	0	1	0
15.1	0	1 [c]	0	1 [b]	1	1 [b]
15.2	0	1	0	0	1	0
16.1	0	0 [d]	0	1 [b]	1	1 [b]
16.2	0	0	0	0	1	0
17.1	0	0 [a]	0	0 [a]	1	1 [c]
17.2	0	1	0	1	1	0
18.1	0	1 [b]	0	0 [a]	1	1 [c]
18.2	0	0	0	1	1	1
19.1	0	1 [c]	0	0 [a]	1	1 [c]
19.2	0	1	0	1	1	1
20.1	0	0 [d]	0	0 [a]	1	1 [c]
20.2	0	0	0	1	1	1
21.1	0	0 [a]	0	1 [b]	1	1 [c]
21.2	0	1	0	0	1	1
22.1	0	1 [b]	0	1 [b]	1	1 [c]
22.2	0	0	0	0	1	1
23.1	0	1 [c]	0	1 [b]	1	1 [c]
23.2	0	1	0	0	1	1
24.1	0	0 [d]	0	1 [b]	1	1 [c]
24.2	0	0	0	0	1	1
25.1	0	0 [a]	0	0 [a]	1	0 [d]
25.2	0	1	0	1	1	0
26.1	0	1 [b]	0	0 [a]	1	0 [d]
26.2	0	0	0	1	1	0
27.1	0	1 [c]	0	0 [a]	1	0 [d]
27.2	0	1	0	1	1	0
28.1	0	0 [d]	0	0 [a]	1	0 [d]
28.2	0	0	0	1	1	0
29.1	0	0 [a]	0	1 [b]	1	0 [d]
29.2	0	1	0	0	1	0
30.1	0	1 [b]	0	1 [b]	1	0 [d]
30.2	0	0	0	0	1	0
31.1	0	1 [c]	0	1 [b]	1	0 [d]
31.2	0	1	0	0	1	0
32.1	0	0 [d]	0	1 [b]	1	0 [d]
32.2	0	0	0	0	1	0

Figure 3.2: This table shows the enumerated haplotype pairs for genotype $X_k^j = [0, ?, 0, 1, 2, ?]$. The first columns list the haplotype pair ID i followed by a dot and either a 1, first part of haplotype pair, or 2, second part of haplotype pair. In particular, in this example there are $m_{jk} = 32$ haplotype pairs. Details for the recursive enumeration are listed in Appendix A.2: the other columns shows for $t \in \{1, \ldots, 6\}$ which pattern has been used for haplotye pair $[H_1^i(j)[t], H_2^i(j)[t]]$.

$\{(0,1),(1,0)\}$ while missing SNP ? may represent SNPs $\{(0,1,2)\}$ leading to four consistent haplotype pairs at this locus $\{(0,0),(0,1),(1,0),(1,1)\}$. Therefore, I have

$$m_{jk} = 4^{\text{number of markers '?'}} \cdot 2^{\text{number of markers '1'}}. \tag{3.5}$$

I enumerate the m_{jk} haplotype pairs recursively, i.e. similar to the way one would enumerate all binary number of fixed length. Further details are presented in Appendix A.2. In Figure 3.2 I illustrate this concept for a short genotype shown in Equation 3.6 leading to $m_{jk} = 32$ haplotype pairs.

$$X_k^j = [0,?,0,1,2,?]. \tag{3.6}$$

3.2.2.3 Evaluating chromosomal likelihood using enumerated haplotype pairs

To evaluate $p(X_c|\theta)$ I discuss three different ways

- **Default probability (DFPB)**: *when I encounter zero haplotype frequency estimate replace its value by a small pre-specified constant.*

  The probability $p\left(X_c|\theta\right)$ is defined as

  $$p\left(X_c|\theta\right) = \frac{1}{m_{jc}} \sum_{H_1,\, H_2 \text{ consistent with } X_c} \hat{p}(H_1|b,c) \cdot \hat{p}(H_2|b,c) \tag{3.7}$$

  but can also be approximated using maximization rule in Equation 3.8

  $$p\left(X_c|\theta\right) = \max_{H_1,\, H_2 \text{ consistent with } X_c} \hat{p}(H_1|b,c) \cdot \hat{p}(H_2|b,c) \tag{3.8}$$

  However, in some cases, there are no consistent haplotype pairs because either $\hat{p}(H_1|b,c) = 0$ and/or $\hat{p}(H_2|b,c) = 0$. In these cases I assign a small default probability d to adjust for the zero haplotype frequency estimate, i.e. $\hat{p}(H|b,c) = d$. This default probability is a trade-off: it should be non-zero because a matching haplotype pair may exist but is not covered by the training set, but it should also be chosen slightly less than the smallest observed value for $p\left(X_c|\theta\right)$.

- **Rare default probability (OneHap)**: *if all the products of haplotype frequency estimate in the definition of $p(X_c|\theta)$ in Equation 3.7 evaluate to zero, only take the haplotype with the highest frequency estimate in account based on all haplotypes in the list of consistent haplotypes.*

  If there is at least one consistent haplotype pair both non-zero haplotype frequency estimates use 3.7 to evaluate $p\left(X_c|\theta\right)$. Otherwise, I look for an approach which minimizes the impact of default variable d by using it only in one term where it is multiplied with the $hMaxFreq = \max$ HAPS where HAPS is defined in Equation 3.4 to include only one haplotype frequency estimate in the definition of $p\left(X_c|\theta\right)$. This definition also balances out small d values with larger

values of hMaxFreq. Therefore, I evaluate

$$p\left(X_c|\theta\right) = \frac{1}{m_{jc}} \cdot d \cdot hMaxFreq.$$

- **Pseudo counts (PSEUDO)**: *due to underestimation of some small haplotype frequency estimates are not retrieved. Therefore, I allow for a further unobserved haplotypes which leads to a re-scaling, such that all frequency estimates for all haplotypes are non-zero.*

The rationale for pseudo counts d is that I observe R non-zero haplotype frequency estimates x returned from PHASE and allow for a further d unobserved haplotypes. To derive the pseudo counts idea I assume R non-zero frequency haplotypes plus d further unobserved haplotypes which are normalized by scalar k. Furthermore, these two terms are assumed to sum to one. Therefore, I write

$$\sum_R \frac{\text{observed haps}}{k} + \frac{d \text{ unobserved haps}}{k} = \sum_R \frac{x+d}{k} + \frac{d}{k} = \frac{1+Rd}{k} + \frac{d}{k} = 1.$$

Therefore, normalization constant k evaluates to $k = 1 + [R+1]d$. Then, I can compute the adjusted haplotype frequencies x' from the original PHASE haplotype frequencies x as

$$x' \mapsto \frac{x+d}{1+[R+1]d}$$

which simplifies to $d/(1+[R+1]d)$ for haplotypes not originally observed by PHASE, i.e. those ones assigned zero frequencies by PHASE. Given these adjusted haplotype frequencies $x' > 0$ I can evaluate $p\left(X_c|\theta\right)$ using Equation 3.7.

3.3 Performance measures for pure breed identification

As performance criterion to judge prediction quality for correct breed b the mean posterior probability in Equation 3.9 taken over the test set query dogs $test_b$ is computed as defined in Section 1.5.3.

$$q_b = \frac{1}{|test_b|} \sum_{j \in test_b} p(\theta|X^j) \tag{3.9}$$

Then, I aggregate these results by averaging q_b across B breeds as shown in Equation 3.10.

$$\bar{q}_{1:B} = \frac{1}{|B|} \sum_{b \in B} q_b, \quad \bar{r}_{1:B} = 1 - \bar{q}_{1:B}. \tag{3.10}$$

To see which tuning parameter d according to Section 3.2.2.3 maximizes $\bar{q}_{1:B}$ I experimented with 15 values of d shown in Equation 3.11.

$$d = \{1,3,5,7,9\} \cdot 10^{\{-3,-4,-5\}} \tag{3.11}$$

SmallPure	**DFPB**	**OneHap**	**PSEUDO**
PSEPARATE Max $\bar{r}_{1:B}$	0.011	0.090	0.463
d_{max}	$d = 9 \cdot 10^{-3}$	$d = 9 \cdot 10^{-3}$	$d = 3 \cdot 10^{-4}$
PSEPARATE Min $\bar{r}_{1:B}$	0	0.028	0.029
d_{min}	$d \in \{10^{-5}, \ldots, 5 \cdot 10^{-5}\}$	$d = 3 \cdot 10^{-3}$	$d = 9 \cdot 10^{-3}$
PPOOLED Max $\bar{r}_{1:B}$	0.050	0.047	0.516
d_{max}	$d = 9 \cdot 10^{-3}$	$d = 9 \cdot 10^{-3}$	$d = 9 \cdot 10^{-3}$
PPOOLED Min $\bar{r}_{1:B}$	0.030	0.029	0.029
d_{min}	$d = 3 \cdot 10^{-5}$	$d \in \{7 \cdot 10^{-4}, \ldots, 10^{-3}\}$	$d = \{5 \cdot 10^{-4}, 7 \cdot 10^{-4}\}$

Table 3.3: Results for pure breed identification on **SmallPure** using performance criterion $\bar{r}_{1:B}$ defined in Equation 3.10. I selected the extreme values for $\bar{r}_{1:B}$ with its corresponding position d from Table A.6 according to different combinations of haplotype frequency estimation and evaluation of $p\left(X_c|\theta\right)$. Detailed results for combination *(PSEPARATE,DFPB)* are in Table A.8, for *(PPOOLED,DFPB)* are in Table A.9, for *(PSEPARATE,OneHap)* are in Table A.10, for *(PPOOLED,OneHap)* are in Table A.11, for *(PSEPARATE,PSEUDO)* are in Table A.12 and for *(PPOOLED,PSEUDO)* are in Table A.13.

However, I prefer to provide small values, i.e. minimize $1 - q_b$ and $\bar{r}_{1:B}$. Therefore, I obtain result tables of $(B + 1) \times 15$ where rows correspond to breeds and columns correspond to values of d. Then, each cell, contains value $1 - q_b$ as a function of d. In the final $B + 1$-th row of this matrix I give the columns means (averaged across breeds) $1 - \bar{q}_{1:B}$. However, given that for some breeds $\bar{r}_{1:B} = 0 \; \forall d$, I mostly omit those breeds in the tables in the Appendix. Furthermore, for simplified visual inspections I leave cells with $1 - q_b = 0$ blank and give a precision of 2 decimal places for $1 - q_b$ and 4 digits for $\bar{r}_{1:B}$.

3.4 Results for DBAncestry

In this section I offer results for pure breed identification for the novel methodology DBAncestry. Firstly, I present results for dataset SmallPure which is split in training and test dataset according to the description in Section 1.5.3. I compare options (PSEPARATE, PPOOLED) for the estimation of haplotype frequencies and conclude which of the options (DFPB, OneHap, PSEUDO) to evaluate $p\left(X_c|\theta\right)$ performs best. After that I perform tests and offer results for BigPure and conclude which haplotype frequency estimation options is chosen.

3.4.1 SmallPure case

In this section I discuss performance criterion $\bar{r}_{1:B}$ for dataset SmallPure. A summary of results is shown in Table 3.3. To evaluate $p\left(X_c|\theta\right)$ I used averaging in Equation 3.7 which outperformed the maximization rule in Equation 3.8 for SmallPure and BigPure.

The minimum for $\bar{r}_{1:B}$ is attained for values of ess or equal than $d = 3 \cdot 10^{-3}$ for all approaches. This suggests that too large values of d are in conflict with actual haplotypes returned by PHASE which have low population frequencies. Therefore, I remove the following larger values of d-values $7 \cdot 10^{-3}, 9 \cdot 10^{-3}$ for the BigPure analysis in subsequent Section 3.4.2.

The minimum for $\bar{r}_{1:B}$ is similar for most approaches at around 0.03 with a absolute minimum of $\bar{r}_{1:B} = 0$ for DFPB PSEPARATE. Furthermore, for BigPure I will not look at the PSEUDO approach because the minimum of $\bar{r}_{1:B}$ is similar to OneHap but the results are much more sensitive to d. In particular, the PSEUDO approach may lead to not very good results in the worst case of up to $\bar{r}_{1:B} \approx 0.5$ while the other combinations have a worst case of less than 0.1. Finally, PSEPARATE and PPOOLED show similar minimum values for $\bar{r}_{1:B}$. Therefore, due to the higher computational costs of PPOOLED I only will investigate PSEPARATE in the Section 3.4.2.

BigPure	**DFPB**	**OneHap**
PSEPARATE Max $\bar{r}_{1:B}$	0.009	0.011
d_{max}	$d = \{5 \cdot 10^{-3}\}$	$d = 10^{-5}$
PSEPARATE Min $\bar{r}_{1:B}$	10^{-4}	0.0
d_{min}	$d = \{3 \cdot 10^{-5}, 5 \cdot 10^{-5}\}$	$d = \{3 \cdot 10^{-4}, \ldots, 10^{-3}\}$

Table 3.4: Results for pure breed identification on dataset **BigPure**. The minima and maxima of performance criterion $\bar{r}_{1:B}$ according to Equation 3.10 for separate phasing with either DFPB or OneHap are shown in Table A.7. Detailed results $1 - q_b$ for each breed b are given in Table A.14 for *(PSEPARATE,DFPB)* and in Table A.15 for *(PSEPARATE,OneHap)*.

3.4.2 BigPure case

In this section I discuss pure breed identification results in Table 3.4 for BigPure as defined in Section 1.5.3 for PSEPARATE phasing and I used options DFPB and OneHap to compute $p\left(X_c|\theta\right)$

With respect to the misclassifications for DFPB the DBAncestry algorithm wrongly predicted between 1 and 28 test dogs according to the choice of tuning parameter d. In the case where $\bar{r}_{1:B}$ is at its minimum, DBAncestry confuses two instances forming different subpopulations of Labrador Retriever, i.e. 1 Labrador Retriever (US Field) was confused as Labrador Retriever (US Show). For the best d, I predict all test dogs correctly using OneHap.

Furthermore, a close look at the wrongly predicted dogs shows that most of those dogs were confused with the same breed but from a different region. For example, in those cases where at most 3 query dogs are wrongly predicted, they come from the following 3 breeds: 1 Australian Shepherd (UKX), 1 Labrador Retriever and 1 Weimaraner (US2X). According to breed Table A.1 it is known that there is a a total of 18 test query dogs from Australian Shepherd (UKX+USX) and 91 test query dogs for Labrador Retriever (US Field + US Show). Although there are many test dogs the same ones are misclassified using either DFPB or OneHap which suggests these test instances were on the regional boundary of these breeds, or mislabelled in the database.

Results in Table 3.4 are all very good with OneHap slightly better than DFPB. Also note that the best values of d, such that $\bar{r}_{1:B}$ is smallest, it can be seen that DFPB takes smaller values of d than OneHap. This observation suggests that when I frequently impute a missing haplotype frequency, such as in DFPB, the algorithm preferrably picks smaller values of d to avoid that d too strongly dominates other haplotype frequency estimates. However, in case of OneHap if value d is rarely used DBAncestry assigns it a larger number to have sufficient influence. From Tables A.14, A.15 I also notice that for less optimal choices of d the number of misclassifications is much bigger for less suitable values of d in DFPB than OneHap.

Finally, due to smaller sensitivity to the values of tuning parameter d and the small number of misclassifications for the best choice of d I will proceed to use OneHap for estimation of mixed breed composition.

3.5 Conclusions

In this chapter I dealt with pure breed identification based on the previous Mars approach and the novel DBAncestry methodology. At first I described the data structure where I assume up to eight pure breed different great-grandparents which in turn are assigned to chromosomes to form a genome painting. Then, I describe how I encode a breed representation using genetic information: Mars uses SNP-wise

transition frequencies which I extend to haplotype frequency estimates to account for the correlation structure among SNPs. Furthermore, I discussed whether it is beneficial to derive all frequency estimates separately for each breed or jointly across breeds. After that I describe how I use the breed representation to evaluate $p\left(X_c|\theta\right)$, i.e. the likelihood of a genotype on chromosome c using a breed pair θ. Within the Mars approach I use the HMM forward algorithm to calculate $p\left(X_c|\theta\right)$ while for DBAncestry for average over the product of maternal and paternal haplotype frequency estimates with haplotype pairs chosen which are consistent with X_C. However, some of the estimates for the haplotype frequencies are zero which led me to design three imputation options to deal with this case: DFPB, OneHap and PSEUDO. After that I discussed a haplotype-based approach to maximizing likelihoods to infer pure breeds in test dogs along with performance measures based on the mean posterior breed probabilities for the correct breed. Finally, I presented results for DBAncestry for datasets SmallPure and BigPure: I concluded that phasing breeds jointly does not yield any performance improvements over separate phasing but incurs a higher computational cost. Furthermore, I found that imputing of missing haplotype frequency estimates works best for OneHap, i.e. where imputation is only applied in case none of the terms in the product $p\left(X_c|\theta\right)$ are non-zero.

Chapter 4

MCMC methodology for mixed breed identification

This chapter deals with the question on how to infer ancestry for crossbred dogs: an intuitive approach to find the optimal variable assignments in a discrete space, such as inference of the most likely lineage painting, is based on a group of techniques known as combinatorial optimization algorithms, such as genetic algorithms (Mitchell, 1998). However, typically these models yield a point estimate, i.e. the algorithm returns a binary prediction of which breeds are present/absent in the estimated ancestry, and do not address uncertainty. In particular, due to the probabilistic behavior of genetic inheritance a single run of the optimization technique may not be very useful and the optimum may not represent the true lineage.

Recently, Cussens et al. (2013) developed a combinatorial optimization approach which retrieves the top k most likely pedigree reconstructions which can be averaged to yield proportion estimates. Specifically, the authors added constraints to prevent previously found maxima from being selected again. However, at the moment their approach assumes independent SNPs and their proposed extension to account for LD is likely to be only computationally feasible for a small number of populations.

Therefore, in this chapter I focus on how the Mars algorithm and DBAncestry deals with crossbred dogs based on the haplotype frequency model described in Section 3.1. However, the purebred inference technique provided in Section 3.1.1 which searches over all possible θ is no longer computionally feasible to infer canine breed composition with complex ancestries involving multiple breeds. The search space for Φ is huge and an enumeration of all paintings for the lineage tree and genome is prohibitive. Therefore, I extend DBAncestry to sample this space using a Metropolis-Hastings (MH) algorithm on a discrete space over Φ.

In Section 4.1 I give a brief introduction to Markov Chain Monte Carlo (MCMC) including the MH algorithm. Furthermore, I will outline how the random walk idea for continuous random variables is adapted to deal with nominal breed random variables. After that in Section 4.2 I will specify the details for inference of θ in the crossbred haplotype frequency model. In particular, I will describe the update proposals of the Mars approach in Section 4.2.1 and DBAncestry in Section 4.2. Then, I will continue to describe how I derive breed proptions from the haplotype frequency model in Section 4.3 and I will suggest a few further MCMC extensions for mixed breed inference in Section 4.4. As next step I will explain how I measure breed prediction performance in Section 4.5 and in Section 4.6 I will specify the number of MCMC iterations used to obtain the breed proportion estimates. Finally, in Section 4.7 experimental results for the simulated dataset will be

4.1 Ancestry inference for crossbred dogs using MCMC

I compute the posterior breed probability $p(\theta|X) = \prod_{c=1}^{C} p\left(X_c|\theta\right)$ as outlined in Section 3.1. Based on the results on dataset BigPure shown in Section 3.4.2 I decided to apply the OneHap approach for $p\left(X_c|\theta\right)$ using $d = 5.0 \cdot 10^{-4}$ as parameter value. To infer the ancestry of a mixed breed test datum observed as X I study the posterior probability $p(\theta|X)$.

I use Markov Chain Monte Carlo (MCMC) to approximate the posterior distribution $p(\theta|X)$ (Robert and Casella, 2004, 2010) as target distribution. MCMC is a class of methods which draws a sequence of $M + N$ dependent samples $\theta^{(1)}, \ldots, \theta^{(M)}, \theta^{(M+1)} \ldots, \theta^{(M+N)}$, such that the probability distribution of $\theta^{(t)}$ only depends on $\theta^{(t-1)}$. This conditional probability distribution is referred to as Markov or transition kernel $K, i.e.$

$$\theta^{(t+1)}|\theta^{(1)}, \ldots, \theta^{(t)} \sim K\left(\theta^{(t)}, \theta^{(t+1)}\right).$$

After a sufficiently large burn-in time of M iterations corresponding to samples $\theta^{(1)}, \ldots, \theta^{(M)}$ the markov chain will converge to the stationary distribution, i.e. the next N samples drawn after the burn-in phase are $\theta^{(M+1)}, \ldots, \theta^{(M+N)}$ will approximately correspond to a random sample of $p(\theta|X)$. Those later iterations N are referred to as the main phase of the MCMC approach. In other words, a stationary distribution exists $p(\theta|X)$ exists, such that if $\theta^{(t)} \sim p(\theta|X)$, then $\theta^{(t+1)} \sim p(\theta|X)$. Assuming $K(\theta^{(t)}, \cdot) > 0$, existence of a stationary distribution implies a kernel K which allows for free moves across the state space with positive probability of reaching any region in the state-space. These ideas can be traced back to Metropolis et al. (1953) who present the Metropolis-Hastings algorithm which constructs a Markov chain with stationary distribution $p(\theta|X)$. At each iteration t, I perform the following two steps in this algorithm:

Step 1: Sample $\theta^{'} \sim q\left(\theta^{'}|\theta^{(t)}\right)$ where q is the proposal density and is used as transition kernel K.

Step 2: Draw $r \sim U[0,1]$. Let $a = a_1 \cdot a_2$ where $a_1 = \frac{p(\theta^{'}|X)}{p(\theta^{(t)}|X)}$ and $a_2 = \frac{q\left(\theta^{(t)}|\theta^{'}\right)}{q\left(\theta^{'}|\theta^{(t)}\right)}$ Then, if $r < a$, then accept new proposal $\theta^{(t+1)} = \theta^{'}$ otherwise reject $\theta^{(t+1)} = \theta^{(t)}$. Note that the special case of a symmetric propsal density in step 1, such that $q\left(\theta|\theta^{'}\right) = q\left(\theta^{'}|\theta\right)$ leads to $a_2 = 1$ in this step 2.

Given sufficient dispersion of proposal density q, every q can be used in the Metropolis-Hastings (MH) algorithm to construct a kernel, such that $p(\theta|X)$ is a stationary distribution.

A popular choice for a symmetric proposal density is $q\left(\theta^{'}|\theta\right) \propto \exp[-(\theta^{'} - \theta)^2\sigma^{-2}]$, such that new proposals are chosen as $\theta^{'} \sim N(\theta, \sigma^2)$ which is a random walk algorithm of the form $\theta^{'} = \theta + \sigma \cdot Z$ with $Z \sim N(0,1)$. Large values of variance σ^2 encourage quick exploration of the state space of θ but proposals will frequently be rejected while small variances lead to small proposal update steps with slow mixing, long burn-in times and poor state space exploration.

However, this popular symmetric proposal density is only applicable for continuous random variables while $\theta = (\Phi, L, G)$ is composed of categorical (nominal) variables. Therefore, to define the proposal densities, in our study I use q_L propose new breeds to create lineage painting L and then use q_G select slots from Φ to generate genome painting G.

4.2 Inference in crossbred haplotype frequency model

In Section 4.2.1 I show how the Mars algorithm updates q_L and q_G in each MCMC iteration and which ideas are utilized to improve mixing. Then, in Section 4.2.2 I will discuss how I use a uniform q_L and a breed similarity lineage painting (BSPL) proposal density. With regard to the BSPL option I will explain which breed proximity measures is selected and how it is used to define an update mechanism.

4.2.1 Mars approach

This section is a summary of the Mars algorithm comprehensively described in Section A.5: firstly, the Mars approach does not only look at the most complex lineage tree ABCDEFGH but at several different representations of the state space given by the 11 lineage trees shown in Figure 1.2. Therefore, the inference of θ is performed for each of the trees and the best tree is chosen using the deviance information criterion (DIC) (Spiegelhalter et al., 2002). DIC is a hierarchical Bayesian model comparison measure which are computed from quantities in an MCMC run and is defined as model fit which is penalized by the model complexity through the effective number of parameters in the model.

For a given lineage tree at each MCMC iteration the algorithm:

- **Update lineage painting L:** all slots in the lineage painting are updated uniformly and then this new state is used as proposal q_L in the MH algorithm. The update of all lineage tree slots in a single MCMC iterations leads to more exploration and less acceptance than in DBAncestry.
- **Genome painting G:** in 10% of the iterations a new genome painting q_G is proposed which updates all C chromosome paintings.

In the Mars approach any breed from the lineage painting can be assigned to either the maternal or paternal side of the chromosome painting which is an invalid biological assumption. The Mars approach utilizes a few ideas to improve mixing during the MCMC algorithm:

- **Jumbling**: in each MCMC iteration breeds are jumbled in the lineage painting, i.e. breeds currently forming the lineage painting are randomly swapped with each other. However, I recommend to only swap breeds within the maternal and paternal lineage painting. Furthermore, regarding the number of previous generations, I am more certain with respect to more recent ancestors which should be swapped less often.
- **Averaging Model**: an average model ignores the actual realization of the chromosome painting on chromosome c but instead it is defined as an average likelihood over all possible lineage paintings where each maternal breed is joined with each paternal breed: $A = \frac{1}{16}\sum_{d=1}^{4}\sum_{e=1}^{4} p(X_C|u_c = \phi_M[d], v_c = \phi_P[e])$. At first in the burn-in phase of the algorithm the averaging model is exclusively used and phased out at a linear rate and replaced by the actual chromosome painting model M which is exclusively used at the start of the MCMC main phase, i.e. a linear combination $p'_{cuv} = p(X_c|u_c, v_c)' = r \cdot A + s \cdot M$ is computed with $r = 1, s = 0$ at burn-in phase iteration 1, and $r = 0, s = 1$ at main phase iteration 1. An average model has similarities with simulated annealing (Robert and Casella, 2004, 2010) where the temperature is decreased with the number of iterations: a high temperature corresponds to more randomization while a lower temperature corresponds to a focus on optimizing the fitness function.

4.2.2 DBAncestry

At each MCMC iteration t I either update the lineage painting $\theta^{'} = (\Phi, L', G)$ via q_L with probability $\rho = 0.01$ or I update the genome painting $\theta^{'} = (\Phi, L, G')$ using q_G with probability $1 - \rho = 0.99$.

Therefore, typically for each fixed lineage painting I try about 100 genome painting configurations out of the possible $8^{2\times 25}$ assignment possibilities.

Compared to the Mars approach in Section 4.2.1 which proposes the whole genome painting as update while DBAncestry proposes only painting updates of a single chromosome. Firstly, on average for each lineage the DBAncestry algorithm considers a factor of 1000 more updates of G than the Mars approach. Therefore, the DBAncestry enables a much better exploratory granularity and insight how good the current lineage painting is. Furthermore, more incremental changes of a single chromosome rather than updates of the whole genome painting is much more likely to be accepted due to its smaller proposed step size from the current G, i.e. in the Mars approach new genome update proposals are rarely accepted relative to the DBAncestry algorithm. However, given that Mars runs the MCMC algorithm for each of the lineage trees it is reasonable to focus on finding an optimal lineage painting although the genome painting is less likely to be well explored.

As proposal density I either uniformly sample new breeds for the lineage or I bias the Markov Chain, such that breeds in the lineage are more likely to be replaced by similar breeds. Therefore, the proposal step to yield θ' can be distinguished into the following two cases: a uniform update proposal (see the Mars approach in Section 4.2.1) and a breed-biased update proposal to improve mixing and to offer a better exploration of the lineage painting search space.

Update L: With regard to the **lineage painting** update I uniformly either update a paternal slot $\phi_s = \phi_p$ or a maternal slot $\phi_s = \phi_m$. Then, I uniformly select $t \in \{1, 2, 3, 4\}$ to select a great-grandparent slot. Then, I propose A new breed

$$q\left(\theta^{'}|\theta\right) = q_L(\phi_s[t]' = b'|\phi_s[t] = b). \tag{4.1}$$

I look at two different ways to define q_L :

1. *uniform lineage painting* **(ULP)** Each of the breed proposals b' for slot $\phi_s[t]$ is distributed according to a discrete uniform density $U[1, b_M]$ where b_M is the number of breeds in the dataset, such that the probability mass function for breed b' is given as

$$q_L(\phi_s[t]' = b'|\phi_s[t] = b) = \frac{1}{b_M} \tag{4.2}$$

Furthermore, in this case the transition kernel q_L is symmetric. Given $r \sim U[0, 1]$, I accept breed proposal b' if $r < \frac{p(\theta^{'}|X)}{p(\theta^{(t)}|X)}$.

The ULP update proposal corresponds to a random walk with a large σ and a proposal which is independent of the previous state θ. In other words, each update proposal is possibly associated with a large transition step which explores wide areas of the search space. However, most update proposals will not be accepted.

2. *breed similarity lineage painting* **(BSLP)**

According to Section 2.1 there are two main objectives to investigate an alternative proposal mechanism based on breed similarity:

- **Mixing:** improve mixing time, such that the stationary distribution is reached more quickly which leads to less iterations, shorter run-time and lower investment of compu-

tational resources. In other words, I suggest to propose large transition steps with are accepted with high probability.

– **Exploration granularity:** assumed that I am in the correct search space close to the true breed composition I would like to further explore this region of the space instead of proposing breeds possibly entirely different to the current breed configuration. Although the mechanism may cover the whole breed composition space less densely I may achieve higher granularity of the assumed correct subspace, such that I can successfully discriminate among very similar breeds or those breeds which have subpopulations.

The random walk algorithm is not necessarily the best option to explore the search space because it requires many iterations to move between low-probability regions between modal regions of the target distribution, and its symmetric nature implies that half the time previous regions are explored again (Robert and Casella, 2010). Recently, geometric concepts have been incorporated in the proposal density to exploit the structure in the data. Roberts and Rosenthal (1998) suggest the Langevin algorithm which includes the gradient of the target distribution in the proposal density and Roberts and Stramer (2002) propose the Metropolis Adjusted Langevin Algorithm (MALA). Girolami and Calderhead (2011) look at more advanced sampling schemes incorporating information geometry which applies geometric concepts on manifolds which follow a direct path on the surface, such as geodesics (shortest paths between two points on manifold). Looking at more sophisticated sampling schemes is not sufficient because I deal with Markov Chains over a discrete domain instead of the standard continuous assumption.

This insight poses the question on whether I can think of a mechanism which allows us to control the size of the transition step. The idea I will implement for dogs will be based on a distance measure between breeds. In particular, I will bias the proposal mechanism, such that the breed currently assigned to $\phi_s[t]$ is more likely to be replaced by a breed similar to it.

To implement this idea of breed similarity proposals I need a measure of breed similarity and based on that how I propose a new breed given the current breed assignment.

– **Breed similarity measure:** I compute the breed distance using Manhattan distance on genotypes as shown in Figure 2.4 (a). Then, this Manhattan distance metric is converted into a measure for breed similarity given in Figure 2.4 (b). To account for the varying levels of interbreed similarity shown in Figures 2.3 (a,c,e,g) I replace raw breed similarities by their ranks of how close a new breed is given the current one. Then, I use this rank breed similarity measure as transition kernel q_L whose heatmap is given in Figure 2.4 (d). Furthermore, given that the breed rank similarity matrix is not symmetric the proposal density ratio $a_2 = \frac{q\left(\theta^{(t)}|\theta'\right)}{q\left(\theta'|\theta^{(t)}\right)}$ does not cancel out.
– **Proposal mechanism based on breed similarity:** this topic is concerned with the question on how much exploration is possible at each MCMC iteration. As information geometry I use rank breed similarity. However, given the lack of a well-developed theory for information geometry defined on discrete spaces I use an ad-hoc approach to propose new breeds given the current breed assignment. In particular, as proposed transition probability I use an exponential decay function which gives most weight to similar breeds and very little weight to breeds in the lower half of breed similarity. This

exponential decay transition kernel is shown in Figure 2.3 (j).

Update G: With regard to the **genome painting** update at first I select a chromosome c. Then, uniformly I either propose a breed v_c for a single slot within the paternal chromosome painting $\phi_s = v_c$ or a breed u_c within the maternal chromosome painting $\phi_s = u_c$ to be updated. Finally, I propose the new chromosome painting G_c with probability

$$q\left(\theta^{'}|\theta\right) = q_G\left(G_c \in \phi_s[t]\right)$$

where $t \in \{1, 2, 3, 4\}$ and $s \in \{p, m\}$. The chromosome painting update $q_G(G_c)$ is performed in such a way that I uniformly select among the 4 options from t.

4.3 Breed proportion estimates using MCMC

Typically, MCMC is applied in Bayesian inference to compute expectations of the form $\mathbb{E}\left(f[\theta]|x\right)$, e.g. in the case of ratio scale data f may correspond to the posterior mean $f[\theta] = \theta$, or the posterior variance $\text{Var}[\theta]$ using the expansion $\text{Var}[\theta|x] = \mathbb{E}\left(\theta^2|x\right) - \mathbb{E}\left(\theta|x\right)^2$ (Gilks et al., 1996). I compute the expectation

$$\mathbb{E}\left(f[\theta]|x\right) = \int f[\theta]p(\theta|X)d\theta = \frac{\int f[\theta]p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}.$$

However, these integrals are often high-dimensional and complex which lack a closed form solution. For M large enough, I can use samples $\theta^{(M+1)}, \theta^{(M+2)}, \ldots, \theta^{(M+N)} \sim p(\theta|x)$ to approximate $\mathbb{E}\left(f[\theta]|x\right)$ using Monte Carlo integration as standard average over the main phase MCMC iterations $\frac{1}{N}\sum_{t=1}^{N} f\left[\theta^{(t)}\right]$. For $(M+N) \to \infty$ this standard average converges to $\mathbb{E}\left(f[\theta]|x\right)$ based on the Law of Large Numbers.

However, as before these type of expectations are more targeted at ratio scale data. Breeds represent a categorical level of measurement. One way to compute an expectations on the breeds, referred to as breed proportion. The *proportion of breed* b is computed over the main phase MCMC iterations as

$$\frac{1}{N} \cdot \frac{1}{8} \sum_{t=1}^{N} \mathbb{1}\left[\text{number of occurrences for breed } b \text{ in } L \text{ at iteration } t\right]. \tag{4.3}$$

Then, I fix a breed cut-off, and each breed which has a predicted breed proportion exceeding this specified breed cut-off will form part of the ancestry composition. There are two main reasons for this bias in breed proportion estimation:

1. New breeds are sequentially explored in the proposal step of the MCMC algorithm and therefore breeds are swapped in and out of the lineage painting. Therefore, breeds which form part of the ancestry are underestimated while true negatives are overestimated. In other words, I face an exploration bias where many breeds not part of the ancestry are proposed during the MCMC which consequently will capture some minor breed fractions which again leads to an underestimations of breeds forming the true ancestry.

2. Breeds have limited training sample size and may not have certain haplotypes represented in the training data which leads to low estimated haplotype frequencies and, therefore, to underestimated breed proportions.

In other words, there are true underlying breed proportions assumed to have at most 8 non-zero values. DBAncestry starts off with a uniform prior where each entry equal to one over the number of breeds. Data moves the algorithm from this maximum entropy prior distribution towards the low-entropy true distribution, but because the data is not perfectly informative the results will have higher entropy than the truth which implies under-estimating the large fractions and over-estimating the zero values.

I either look at raw breed fraction estimates UCBC or an amended version thereof referred to as CCBC. These two breed proportion estimates are defined as follows:

- *uncorrected computed breed proportion* **(UCBC)**: this estimate is defined in Equation 4.3.
- *corrected computed breed proportion* **(CCBC)**: define set $CCBC_{\text{breeds}}$ as the union of those breeds with the top 8 highest breed proportion and all breeds forming the true breed ancestry. Then, set all breed proportion estimates to zero of breeds $\notin CCBC_{\text{breeds}}$. Finally, rescale all breed proportions of those breeds $\in CCBC_{\text{breeds}}$ to unit measure, such that breed proportions sum to one.

Based on these UCBC and CCBC values I compute the divergence between the true breed ancestry proportion and the predicted breed proportions. True ancestry proportions for each lineage tree were defined in Table 1.4 depending on whether they form a parental, grandparental, great-grandparental relationship in the lineage trees shown in Figure 1.2 or are absent from the breed composition. Therefore, within each lineage tree I estimate the breed divergence by taking the median for all TAP values over all samples from a specific lineage tree. In the second step these median TAP values are aggregated across all lineage trees which have an ancestor at the specified TAP value according to Table 1.4.

4.4 Extensions

There are a few MCMC alternatives and extensions to the DBAncestry and Mars approach presented above. There are two mechanisms that utilize information on genetic interbreed distance to bias the direction of the Markov Chain to improve mixing:

- **Discrete geometric surface information (Brooks et al., 2011):** I use breed similarity to direct the proposals of the Markov Chain. I can do that either in the ad-hoc way described in Section 4.2, or in a principled way by extending previous Hamiltonian MCMC approaches (Girolami and Calderhead, 2011) to discrete surfaces.
- **Adaptive Independent Metropolis-Hastings Algorithms (AIMHA) (Holden et al., 2009; Robert and Casella, 2010; Liang et al., 2011; Givens and Hoeting, 2012):** this type of algorithm would propose updates which are independent of the current state of the Markov Chain, such as in the ULP proposal. Over time, AIMHA keeps a history vector of update proposals to adapt its update rule. However, to ensure AIMHA is Markovian and the Markov Chain will converge to its stationary distribution I need to demand that adaptation either stops at some point or is decreased, i.e. changing update rule by smaller increments or changing the update rule not at every iteration. Otherwise the algorithm depends too strongly on previous iterations and does not fully explore the breed space. Therefore, in our case I could learn the proportion for the breed acceptance rate for each breed in the burn-in phase. Then, in the main phase I propose new breeds according to these learnt proportions.

Another way to tune the amount of exploration is to introduce a temperature parameter into the transition kernel. A stochastic optimization technique known as simulated annealing (Robert and Casella, 2004, 2010) decreases the temperature logarithmically over time to limit the step size in the proposal.

This setup leads to more state space exploration and avoids the trapping of local maxima in the beginning when the temperature is higher and when it decreases to 0, accepted proposals will concentrate in the neighbourhood of the maximum of θ. However, Robert and Casella (2010) state that the logarithmic temperature decrease leads to very slow convergence.

As a final technique I describe population-based MCMC (popMCMC) methods which are reviewed in Liang et al. (2011). In popMCMC I run several Markov Chains in parallel (Rosenthal, 2000) which exchange information to learn from past samples. An example of popMCMC is given by evolutionary Monte Carlo (EMC) (Liang and Wong, 2001) which encodes a population of individuals as binary strings, or in our case as multinomial strings $\{1, \ldots, b_m\}^8$. Then, to improve mixing EMC employs three genetic operators:

- **Mutation:** select uniformly a character in one of the individuals in the population and change it ULP to a new one.
- **Cross-over:** select to individuals in the population to form parents which are replaced by their two children if state proposal is accepted. Each children is formed by random draws at each position which indicates from which parent the child inherits.
- **Exchange:** assume indexed individuals in the population. Then, swap neighboring indices, i.e. I swap individual j with either individual $\{j-1, j+1\}$.

4.5 Performance measures

Performance of the ancestry algorithm will be measured in two ways: at first I will describe the currently implemented commercial approach by Mars which views the breed prediction problem as a multi-class classification problem (ROC curves). After that I will argue why I prefer to frame this setup as estimation problem in DBAncestry where predicted breed proportions are visualized using boxplots.

- **Classification view (ROC curves):** Mars framed the problem of measuring performance as a multiclass extension of binary ROC curves (Hand and Till, 2001; Fawcett, 2004; Krzanowski and Hand, 2009) which represent a yes (positive class)/no (negative class) criterion of success. A positive breed is defined to be part of the true ancestry while a negative breed is absent from the true ancestry. Standard ROC curves plot sensitivity (true positive rate is defined as proportion of predicted true positives to all positives) of recognizing breeds versus false positive rate (i.e incorrectly predicting that a breed is part of the ancestry although it is not).

  Firstly, there is a large class imbalance, i.e. there are only up to eight different positive breeds r in the true ancestry while there are $125 - r$ negative breeds. Therefore, independently of the quality of estimating breed composition there is only a small false positive rate is expected. A more suitable measure than false negative is the positive predictive value (PPV) which is the proportion of true positives among breeds predicted as positives. Therefore, in the standard ROC curve I replace the false positive rate by PPV. To study these performance measures I either plot true positive rate (TPR) versus PPV or the F_1 score (harmonic mean of TPR and PPV) as a function of breed proportion. A more detailed description about definition and computations for multiclass ROC curves is given in Section A.6.

  A limitation of the ROC curves approach is that it is binary measure for success, i.e. a breed is predicted to form part of the ancestry if it exceeds a minimum pre-specified breed cut-off for

the predicted breed proportion. However, this ROC measure does not account for bias in the estimation, such that although I may correctly recognize a breed forming part of the genuine ancestry this estimate may not form a close approximation of the true ancestral breed fraction level.

- **Accuracy of breed fractions using boxplots:** this type of measurements views breed composition as estimation problem of ancestral proportions at the parental, grandparental (gp), and great-grandparental (ggp) level. In other words, it is not sufficient to merely recognize a breed as part of the ancestry but I also would like to infer its proportional contribution to the ancestry. In particular, for each possible ancestral level within each lineage tree I draw a boxplot to study accuracy of breed estimation by looking at the difference between true breed proportion and median predicted breed proportion over all test samples from a given lineage tree. Additionally, I also study divergences between true and estimated breed proportions for each ancestral levels taken over all lineage trees containing the corresponding levels according to Table 1.4.

  If a true breed fraction of 0.5 is estimated to be 0.2, this may be regarded as a true positive (for example if a positive is defined as a fraction > 0.05). However there is a substantial error in the estimate (difference of 0.3) A further limiting factor of ROC curves is that all correct predictions are weighted equally by their ancestral contribution. Therefore, although breeds at the ggp level (50%) are harder to predict I weigh them equally with predictions at other levels, such as the parental level (50%). A possible extension is to weight each breed by the reciprocal of their true ancestry proportion in the loss function of the ROC curves, similarly to AdaBoost (Freund and Schapire, 1997) where training cases which are harder to predict are given a higher weight in the loss function.

4.6 MCMC run length

I consider two different setups for the number of iterations of the Metropolis-Hastings algorithm. Under the assumption of the uniform lineage painting I discuss how well breed composition search space is explored.

1. Short MCMC Run for DBAncestry **(Short)**: I assume $M = 1.25 \times 10^5$ burn-in iterations and $N = 7 \cdot 10^5$ main phase iterations. Assuming ULP, on average each breed is proposed $7 \cdot 10^5 \cdot 0.01/125 = 56$ times in the lineage painting update of the MCMC main phase run.

2. Long MCMC Run for DBAncestry **(Long)**: in the long MCMC run I assume 4 times more burn-in phase iterations and 10 times more main phase iterations. In particular, I have $M = 5 \cdot 10^5$ burn-in iterations and $N = 7 \cdot 10^6$ main phase iterations. Assuming ULP, on average each breed is proposed $7 \cdot 10^6 \cdot 0.01/125 = 560$ times in the lineage painting update of the MCMC main phase run. A Java implementation which is run on the UCL cluster infrastructure LEGION requires about 30 minutes for each synthetic test dog until completion.

Finally, during a fixed lineage painting I attempt $1/\rho = 1/0.01 = 100$ chromosome painting updates q_G and propose each of the 8 $\phi_s[t]$ slots to a given chromosome G_c as a new assignment for $100/8 = 12.5$ times.

4.7 Experimental results

In this section I will discuss experimental results for the synthetic test dataset OrgSyntheticRed outlined in Section 1.5.4. I will apply the DBAncestry algorithm using either the uniform (ULP) or the breed-similarity biased (BSLP) update proposal according to Section 4.2.2 while varying the run length (short

True breed ancestry Median [1st,3rd] Quartiles	UCBC ULP	UCBC BSLP	CCBC ULP	CCBC BSLP
100%	**0.70** [0.58,0.79]	**0.64** [0.52,0.74]	**0.85** [0.74,0.92]	**0.81** [0.70,0.88]
50%	**0.30** [0.24,0.36]	**0.27** [0.21,0.33]	**0.38** [0.32,0.44]	**0.36** [0.29,0.42]
25%	**0.14** [0.12,0.19]	**0.13** [0.11,0.17]	**0.19** [0.15,0.24]	**0.18** [0.14,0.23]
12.5%	**0.08** [0.01,0.13]	**0.07** [0.01,0.13]	**0.11** [0.02,0.16]	**0.09** [0.013,0.16]
0%	**0.00** [0.00,0.00]	**0.00** [0.00,0.00]	**0.00** [0.00,0.00]	**0.00** [0.00,0.00]

Table 4.1: **Inference for short MCMC run:** a list of the median and lower/upper quartiles for the breed proportion estimates for all true underlying ancestral levels. The results show that the ULP proposal mechanism leads to less underestimation than for BSLP updates. Furthermore, re-scaling estimates from UCBC to CCBC shows there is a considerable exploration bias for the ancestral breeds, i.e. [3,15] percent for ULP and up to [2,17] percent for BSLP.

run, long run) discussed in Section 4.6. I will also discuss the effect of re-scaling breed proportions from UCBC to CCBC introduced in Section 4.3. Breed proportion estimates were averaged over all 11 lineage trees. Based on the results on dataset BigPure shown in Section 3.4.2 I decided to apply the OneHap approach for $p\left(X_c|\theta\right)$ using $d = 5.0 \cdot 10^{-4}$ as parameter value. To infer the ancestry of a mixed breed test datum observed as X I study the posterior probability $p(\theta|X)$.

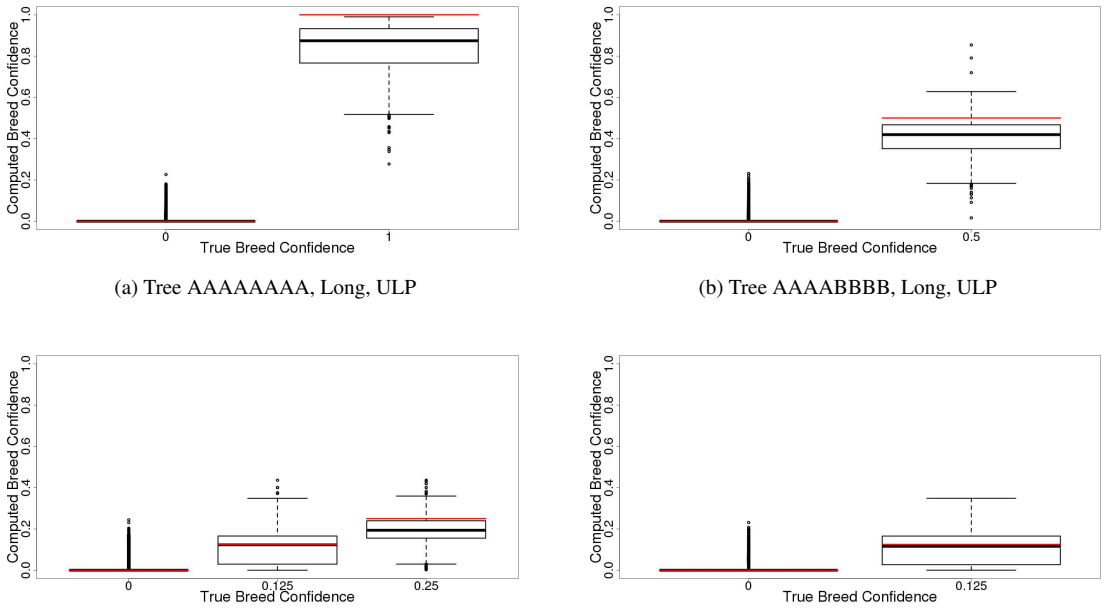
4.7.1 Short Run

In Table 4.1 I list results of the median and lower, upper quartiles for the short MCMC run. This table shows that ULP performs slightly better than BSLP, i.e. by about 1 to 6 percent better depending on the underlying true ancestral level. This may be due to BSLP's in-depth exploration of some part of the breed space while other parts have been only not been densely investigated. A comparison of UCBC and CCBC breed fractions suggests that breeds below the top eight most likely have very small breed proportion estimates. However, their sum is not negligible and the deviation between UCBC and CCBC from purebred down to ggp level ranges from 3 to 15 percent for ULP and from 2 to 17 percent for BSLP.

Furthermore, there is more underestimation for the most recent TAPs, e.g. for a TAP of 50 percent there is a higher impact of exploration because four out of eight leaves have to be updated in the lineage painting while a ggp corresponds to randomization of only one leaf. For example for re-scaled uniform update proposal results (CCBC, ULP), there is an underestimation of 15 percent for pure breed test dogs, 12 percent for parents, 6 percent for gps and 1.5 percent for ggps. Finally, I notice that breeds not forming the ancestry have median predicted breed fractions of 0 looking at the first two leading digits.

4.7.2 Long Run

I provided the results for the long run as boxplots in Figure 4.1. In these figures I show how far the median (bold black line) deviates from the true underlying ancestral level (bold red line) for the two least complex lineages (i.e. AAAAAAAA and AAAABBBB) and for the two most complex lineages (i.e. AABCDEFG, and ABCDEFGH). These results indicate that more main-phase iterations indeed improve estimation results although by a small margin. It can be seen from Table 4.2 that for the different TAP level the underestimation is reduced by 1 percent for the gp and ggp level while at the parent and purebreed level there is a further improvement in estimation accuracy by 2-3 percent. For the long run the BSLP also catches up with ULP proposal mechanism to yield exactly the same performance, even the lower and upper quartile estimates are almost identical. This insight suggests that even the uniform proposals allows for a thorough exploration of the breed space and no further benefit towards fine granularity are obtained using BSLP. Although initially I expected an improvement in estimation



(a) Tree AAAAAAAA, Long, ULP

(b) Tree AAAABBBB, Long, ULP

(c) Tree AABCDEFG, Long, ULP

(d) Tree ABCDEFGH, Long, ULP

Figure 4.1: **CCBC breed proportion estimates for the ULP update proposals within the long MCMC run:** these figures show true ancestral proportion as bold red line while the boxplots offers information on the breed proportion estimates, i.e. the median, lower and upper quartiles, whiskers which extend to the most extreme data point less than 1.5 IQR from the box, and outliers which are drawn as small circles. These plots represent a subset of the 11 lineage trees: in particular the top two rows correspond to the simplest trees while the bottom two columns match the most complex lineage trees.

True breed ancestry Median [1st,3rd] Quartiles	CCBC Long,ULP	CCBC Long,BSLP	CCBC Short,ULP
100%	**0.87** [0.77,0.93]	**0.87** [0.77,0.94]	**0.85** [0.74,0.92]
50%	**0.41** [0.35,0.46]	**0.41** [0.35,0.46]	**0.38** [0.32,0.44]
25%	**0.20** [0.17,0.25]	**0.20** [0.17,0.25]	**0.19** [0.15,0.24]
12.5%	**0.12** [0.03,0.16]	**0.12** [0.03,0.16]	**0.11** [0.02,0.16]
0%	**0.00** [0.00,0.00]	**0.00** [0.00,0.00]	**0.00** [0.00,0.00]

Table 4.2: **Inference for short/long MCMC run:** a list of the median and lower/upper quartiles for the breed proportion estimates for all true underlying ancestral levels. The results show that a longer MCMC run further reduces underestimation by about to 1 to 3 percent. Furthermore, for the long MCMC run the BSLP update proposal yields almost identical estimates to the ULP update which suggests that the breed-biased update mechanism directed proposals over larger parts of the breed space in the additional iterations. However, due to limited hierarchical population structure the uniform updates are sufficient to explore the breed space. Full results for all combinations of MCMC run length, update proposals and breed proportion computations are shown in Table A.19.

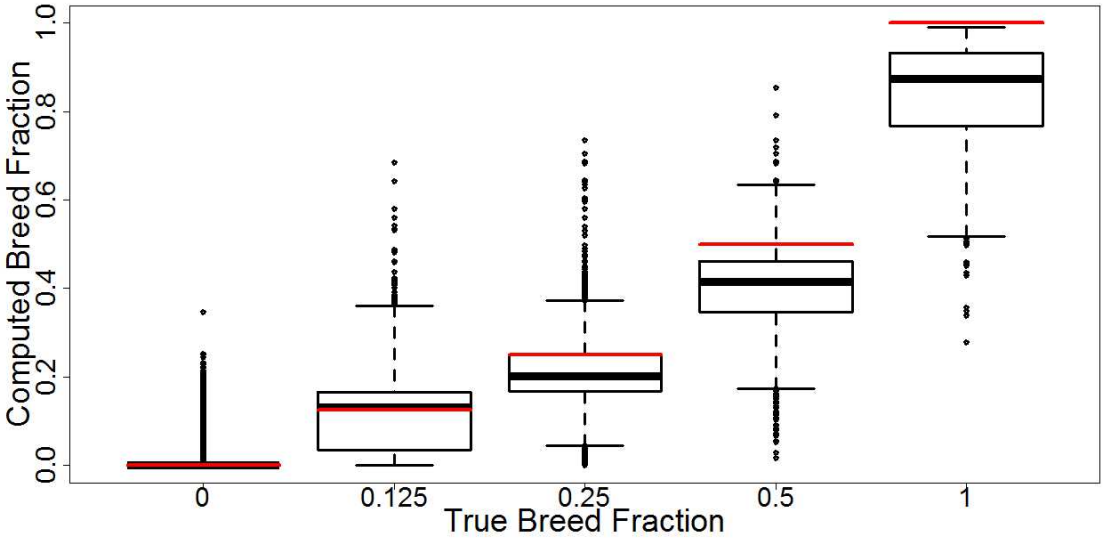


Figure 4.2: This figure shows breed proportion estimates averaged across lineage trees. More in detail, CCBC breed proportion estimates for the ULP update proposal within long MCMC run are visualized as boxplot showing median, lower and upper quartiles of all ancestral levels. Numerical values for the median and quartiles are also given in Table 4.2.

accuracy of BSLP over ULP due to better exploration of subpopulations. However, I did not expect huge improvements because according to the breed similarity results explored in Section 2.5.5 there are only a few strong clusterings in the breed space and limited deep population substructure. A comparison with Mars' classification view in Section A.7.2 confirms the estimation view findings from above.

So far I focussed on predicting breed contributions separately for each lineage as shown in Figure 4.1. However, I would like to investigate how much lineage specific TAP median levels deviate from averaged median breed contributions across all trees as given in Figure 4.2. Indeed, the deviation between lineage-specific and global, i.e. averaged median, breed contribution is less or equal than 0.01 except for three cases: for the parental level in lineage tree AAAABCDE (lineage tree 5) the deviation is 0.016 and at the ggp-level for tree ABCDEFGH (lineage tree 11) the deviation is 0.014. This behavior suggests for both exceptions that when all other TAP levels in the lineage tree are at the ggp level there will be more deviation for the specific lineage tree. But in general it has to be concluded that difference between the global and lineage specific medians are minor.

4.7.2.1 Lineage tree inference

Although in simulations I know which lineage tree a sample comes from in setups with real dogs this information is not available and I need to rely on predicted breed contributions only. Therefore, I would like to explore the possibility whether the original lineage tree can be inferred from the TAP counts only, i.e. the number of breeds predicted for each ancestral level ranging from ggp to purebred. Indeed, looking at Table 1.5 this is possible for all trees except lineage trees for AABBCDEF (lineage tree 8) and AABCDDEF (lineage tree 9) which have the same combination of ancestral TAP levels (2 gps and 4 ggps each). For example, one predicted parent and two grandparents implies AAAABBCC (lineage tree 3) while one estimated parent, 1 grandparent and 2 great-grandparents suggests AAAABBCD (lineage tree 5).

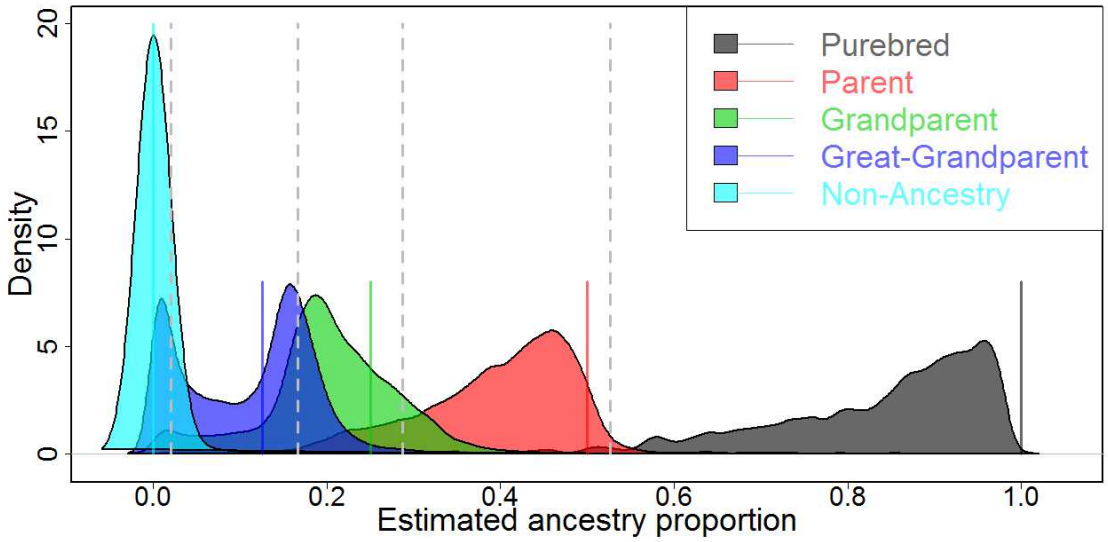


Figure 4.3: This figure shows a density plot (gaussian kernel smoother with bandwidths 0.01-0.02) for global estimated breed contributions (CCBC, ULP, Long). Each colored transparent polygon corresponds to one of the ancestral levels, i.e. purebred, parent, gp, ggp and non-ancestry. Furthermore, vertical lines colored according to the TAP level illustrate the true breed contributions. Vertical grey dashed lines show the inferred breed contribution cut-offs between consecutive TAP levels, i.e. the cut-off between parental and purebred level is about 0.5258.

- View	Max Density	Max Quantiles
Non-Ancestry	**0.02**	**0.00**
Quantiles	-	$\text{NonAnc}_{\alpha=0.96}, \text{GGP}_{1-\alpha=0.04}$
Great-Grandparent	**0.17**	**0.17**
Quantiles	-	$\text{GGP}_{\alpha=0.76}, \text{GP}_{1-\alpha=0.24}$
Grandparent	**0.29**	**0.29**
Quantiles	-	$\text{GP}_{\alpha=0.86}, \text{P}_{1-\alpha=0.14}$
Parent	**0.53**	**0.53**
Quantiles	-	$\text{P}_{\alpha=0.98}, \text{Purebred}_{1-\alpha=0.02}$
Purebred	**1**	**1**
Quantiles	-	-

Table 4.3: The breed intervals for each TAP level for the quantile and density view are shown. For the quantile view I also provide the cumulative density threshold. For example, if a dog has a predicted breed contribution in the interval [0.17,0.29] I assign it to the TAP class grandparent. Furthermore, the upper bound 0.29 for the grandparent level is associated with the 86-th quantile and 0.14-th quantile for the cumulative density of grandparent and parent, respectively. Note the results for the quantile and density views are slightly different because for the quantile I chose a 200-quantile as approximation while for the density view I selected the default option of machine precision.

True breed ancestry	**OrgSynRecC1**	**OrgSynRecC2**
12.5%	**0.13** [0.08,0.16]	**0.14** [0.06,0.16]
0%	**0** [0,0]	**0** [0,0]

Table 4.4: **Inference for ULP update proposal within long MCMC run:** simulation dataset OrgSynRecC1 yields similar median ggp estimates when compared to Mars' dataset OrgSyntheticRed. Furthermore, DBAncestry is robust towards recombination with roughly equal medians but slightly smaller lower quartile estimates.

However, TAP levels are not directly available either. Therefore, estimated breed contributions need to be utilized to infer the mapping from predicted breed contribution intervals to TAP level. In particular, breed contribution cut-offs need to be found which distinguish between consecutive TAP levels, i.e. between purebred-parent, parent-gp, gp-ggp and between ggp versus non-ancestry. I present two views based on cumulative distribution of the predicted breed contributions:

1. **Density view:** in Figure 4.3 I show a density plot $p(\text{TAP}, x)$ of the cumulative distribution function for the estimated breed contributions jointly for each TAP class and estimated breed proportion x. Then, to find the optimal breed contribution cut-off t between the BC densities f, g of consecutive TAP levels I minimize the difference $|\int_0^t f(t)dt - \int_t^1 g(t)dt|$. The integral difference is minimized numerically using golden section search for unimodal functions (Brent, 1973). Alternatively, one could define the decision boundary where the densities of consecutive ancestral levels intersect. Results for BC cut-offs are again summarized in in Table 4.3. According to Webb (2003) in elementary decision theory the Bayes' decision rule minimizes error of TAP assignment where the ancestral level TAP is selected which has the highest a posteriori probability $p(\text{TAP}|x)$ due to the conditional probability relationship $p(\text{TAP}, x) = p(\text{TAP}|x) \times p(x)$.

2. **Quantile view:** alternatively, inference of the optimal separation boundary between consecutive ancestral levels can be based on the estimated breed contributions across all ancestral levels which is described in Section A.7.3.

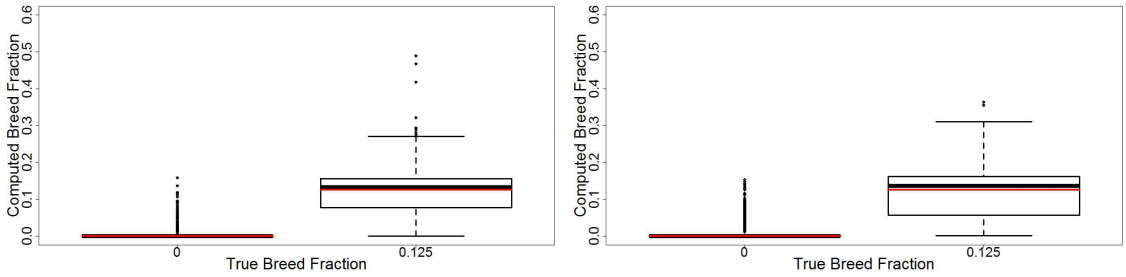
Finally, based on the results for the BC cut-offs shown in Table 4.3 I would like to explore the question which proportion of TAPs is recovered:

- **Purebred:** 85% TAP predictions are correctly recognized.
- **Parental:** 64% TAP predictions are correctly recognized.
- **Grandparent:** 37% TAP predictions are correctly recognized.
- **Great-Grandparent:** 74% TAP predictions are correctly recognized.
- **Non-Ancestry:** 97% TAP predictions are correctly recognized.

Most TAP levels have reasonable recovery levels given the overlapping nature of the TAP classes. As expected from Figure 4.3 the gp level shows a strong overlap with the ggp and parental which suggest more uncertain assignments. In particular, it suggests that breeds at the gp level are often confused as ggps.

4.7.2.2 Impact of recombination

In this section I investigate how much recombination influences the breed contribution estimates at the great-grandparent level within the ABCDEFGH lineage tree. In particular, I apply the MCMC algorithm using the uniform breed update proposals for the long MCMC run. Firstly, to confirm the results



(a) **OrgSynRecC1**: without recombination

(b) **OrgSynRecC2**: with recombination.

Figure 4.4: Breed proportion proportion estimates of lineage tree ABCDEFGH for ULP update proposal within long MCMC run based on own simulation data discussed in Section 1.5.5. The red vertical line represents the true non-ancestry and great-grandparent levels while the bold black line corresponds to the median estimate in the boxplot. Numerical values for the median estimates and quartiles are also given in Table 4.4.

for ABCDEFGH lineage tree in Mars's dataset OrgSyntheticRed I replicate these results using dataset OrgSynRecC1: from tables 4.2, 4.4 it can be seen that the median breed contribution estimates are very similar with 0.12 and 0.13 for datasets OrgSyntheticRed (CCBC, Long, ULP) and OrgSynRecC1, respectively.

In the next setup I simulate data test data OrgSynRecC2 with recombination according to Section 1.5.5. The median ggp breed contribution estimates for OrgSynRecC1 are very similar to OrgSynRecC2 with 0.133 and 0.137, respectively. The upper quartiles are almost identical (i.e. 0.16) while the recombination dataset has a slightly smaller lower quartile value (0.08 vs. 0.06). On the hand, I did not expect very different results because the impact of recombination rate is limited due to a low recombination probability of $\alpha = 10\%$. However, the two following factors should be taken into account

1. Number of loci across chromosomes is very small, i.e. in the range [7,17] (cp. Table 1.3),
2. DBAncestry does not explicitly account for recombination

to appreciate the high prediction quality in the case of recombination events.

4.8 Conclusions

In this chapter I reviewed Mars' MCMC algorithm and developed the ideas leading to my novel MCMC approach DBAncestry which either updates breeds uniformly across breeds or biases the proposal mechanism, such that similar breeds are more likely to be explored. I found that breed proportions are well estimated, and even are improved by up to three percent if I increase the number of main phase MCMC iterations by a factor of 10 from 700K for the short run to 7 million for the long MCMC run. Although the breed proportion estimates are only improved by a small margin for the long run there is a large improvement to classify breeds correctly, i.e. the area under the ROC curve is increased by up to ten percent for the most complex lineage tree.

I also found that that global median breed contributions for the different TAP levels only have minor differences with lineage-specific results. After that I presented an approach which shows how the lineage tree can be inferred from the counts for the different TAP levels (i.e. number of breeds at each ancestral level) which again can be derived from the predicted breed contributions.

Furthermore, I also found that a uniform breed update proposal in the MCMC algorithm is sufficient to explore the breed space. Furthermore, due to a lack of a deep clustering structure in the breed space, a breed-biased update proposal does not lead to further improvements in breed proportion estimates but approximately yield equal breed proportion estimates to the uniform update proposal.

Moreover, I showed how the type of lineage tree can be inferred from the true ancestral proportions, such as purebred and parents, which are again based on the predicted breed contributions. I also showed using simulation data that DBAncestry is robust against recombination events typical for canine data.

Chapter 5

Alternative ancestry inference models

This chapter deals with alternative ancestry estimation techniques and contains two parts:

1. **Review alternative ancestry techniques**: there is a lack of a comprehensive review of alternative approaches applicable for local and global ancestry estimation. Therefore, I look at a variety of approaches, group them by topic and describe how certain model limitations led to more involved approaches. An overview of the techniques discussed is shown in Table 5.1.

2. **ChromoPainter data analysis**: furthermore, a detailed literature review is required to select a very competitive algorithm in Section 5.2.1 which I utilize as comparison technique for the DBAncestry technique from Chapter 4. In particular, I will argue why I choose the ChromoPainter model as comparison technique and how I prepare the data and set tuning parameters for the analysis. After that I will offer results for the data analysis using the same performance measures as for the DBAncestry technique. Finally, I will discuss why the MCMC approach from Chapter 4 is tailored towards small marker panels while ChromoPainter is more specific to dense datasets covering a larger number of SNPs.

5.1 Review ancestry inference techniques

This Section reviews different approaches to estimate local or global genetic ancestry composition in humans and animals. The current literature lacks an integrated, comprehensive review of the different approaches at a conceptual level. There are partial reviews (Churchhouse, 2012; Churchhouse and Marchini, 2013; Liu et al., 2013) for a subset of ancestry inference techniques I cover in this chapter which includes global model-based clustering in Section 5.1.2, local window-based approaches in Section 5.1.3, local hidden Markov models in Sections 5.1.4, 5.1.4.8, PCA-based approaches in Section 5.1.6. But none of these reviews discussed how these approaches are grouped into different categories and motivated their development. A further limitation of these reviews is that they do not include techniques based on either multivariate, multiple regression in Section 5.1.1, machine learning-based approaches in Section 5.1.7 and recent nonparametric Bayesian methods in Section 5.1.5.

5.1.1 Regression

Global ancestry inference can be framed as **multivariate linear regression with multiple responses (MMLR)** (Izenman, 2008) which is visualized in Equation 5.1. For the ancestral training dataset I am given multivariate covariate design matrix $X \in \{0,1,2\}^{m\times p}$ where each row corresponds to the p SNP markers of a purebred dog. Furthermore, I have a multiple response matrix $Y \in [0,1]^{m\times b}$ where each row corresponds to the breed contributions of a given training dog, such that breed proportions normalize to 1 for each dog $\cdot$ and I have $\sum_{j=1}^{b} Y_{\cdot j} = 1$. Then, response matrix Y is regressed on predictor matrix X to estimate regression coefficient matrix $\hat{\beta} \in \mathbb{R}^{p\times b}$. Finally, for a test dataset of n admixed dogs I

Method	>2 pops	Correlation	Type
MMLR (Izenman, 2008)	yes	no	Regression
PCR (Massy, 1965)	yes	predictors	Regression
Ridge Regression (Hoerl and Kennard, 1970)	yes	no	Regression
LASSO (Tibshirani, 1996)	yes	no	Regression
MTRL (Argyriou et al., 2008)	yes	predictors	Regression
Curds & Whey (Breiman and Friedman, 1997)	yes	responses	Regression
RRR (Izenman, 1975)	yes	predictors	Regression
PLS2 (Wold, 1975)	yes	predictors, responses	Regression
MRCE (Rothman et al., 2010)	yes	responses	Regression
MMLRC (Sohn and Kim, 2012)	yes	predictors, responses	Regression
Structure (Pritchard et al., 2000)	yes	no	Clustering
Frappe (Tang et al., 2005)	yes	no	Clustering
Admixture (Alexander et al., 2009)	yes	no	Clustering
Lamp (Sankararaman et al., 2008b)	yes	no	Window
WinPop (Paşaniuc et al., 2009b)	yes	no	Window
AncestryMap (Patterson et al., 2004)	no	no	HMM
ADMIXMAP (Hoggart et al., 2004)	no	no	HMM
Structure 2 (Falush et al., 2003)	yes	no	HMM
Saber (Tang et al., 2006)	yes	yes	HMM
Switch (Sankararaman et al., 2008a)	yes	yes	HMM
Hapmix (Price et al., 2009)	no	yes	HMM
Hapaa (Sundquist et al., 2008)	yes	yes	HMM
GEDI-ADMX (Paşaniuc et al., 2009a)	yes	yes	HMM
LAMP-LD (Baran et al., 2012)	yes	yes	HMM
Alloy (Rodriguez et al., 2013)	yes	yes	HMM
Multimix (Churchhouse and Marchini, 2013)	yes	yes	HMM
ChromoPainter (Lawson et al., 2012)	yes	yes	HMM
iHMM (Sohn et al., 2012)	yes	yes	Non-parametric
FragCoag (Teh et al., 2011)	yes	yes	Non-parametric
SmartPCA (Patterson et al., 2006)	yes	yes	PCA
ipPCA (Intarapanich et al., 2009)	yes	yes	PCA
PCAdmix 2 (Brisbin et al., 2012)	yes	yes	PCA, HMM
GEM (Lee et al., 2010a)	yes	yes	PCA
ETHNOPRED (Hajiloo et al., 2013)	yes	no	ML
SVM with RBF (Haasl et al., 2012)	yes	yes	ML, PCA
SVM with strings (Do et al., 2012)	yes	yes	ML

Table 5.1: A summary of the reviewed ancestry inference techniques is shown. In particular, listed types are based on regression in Section 5.1.1, global model-based clustering in Section 5.1.2, local window-based approaches in Section 5.1.3, hidden markov models in Section 5.1.4, non-Parametric Bayesian approaches in Section 5.1.5, PCA-based approaches in Section 5.1.6 and Machine Learning (ML) based approaches in Section 5.1.7. For each of these methods I show whether this technique is applicable for more than two populations. I also list whether a technique deals with correlation. In particular, for regression techniques I list whether correlation in predictor, response space or both is modelled.

predict breed proportions $Y' \in [0,1]^{n \times b}$ using the test dog genotype data $X' \in \{0,1,2\}^{n \times p}$ and the inferred regression coefficient estimate $\hat{\beta} \in \mathbb{R}^{p \times b}$ from the ancestral data.

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & & & \\ x_{m1} & x_{m2} & \dots & x_{mp} \end{pmatrix} \cdot \begin{pmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1b} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2b} \\ \dots & & & \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pb} \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1b} \\ y_{21} & y_{22} & \dots & y_{2b} \\ \dots & & & \\ y_{m1} & y_{m2} & \dots & y_{mb} \end{pmatrix} \tag{5.1}$$

Standard multivariate regression, i.e. a special case of MMLR with a single response, has been used to predict crossbred bull breed proportions based on 17 bull breeds. Kuehn et al. (2011) specify a design matrix X which is formed by estimated frequencies of a given allele for all loci in the ancestral bull population as previously applied to human populations in Chiang et al. (2010). Then, the alleles of the crossbred bull is regressed on this design matrix and the estimated regression coefficient corresponds to the breed contributions. The measured squared correlation R^2 between estimated regression coefficients and the truth of pedigree based analysis was at 89 percent based on 52K SNP markers while a chip with 3K markers led to an average $R^2 = 0.83$ which indicates strong LD in the 52K chip.

Returning to the general MMLR setup I would like to adjust for the case that the response depends on the covariates nonlinearly, and to account for correlation in the data. There are two types of correlation I would like to account for:

- **Correlated covariates**: this setup where two or more independent predictors are highly correlated is also known as multicollinearity (Freund and Wilson, 1998; Yan and Su, 2009). Firstly, correlated markers leads to a loss of information due to redundancy, such that a given predictor can be represented as a linear combination of the others. Secondly, predictor correlation affects computation of a stable least square estimate which contains the term $X^T X$ which in this case is nearly singular with very small eigenvalues and its inverse $\left(X^T X\right)^{-1}$ may have very large eigenvalues. In particular, $X^T X$ is ill-conditioned because this term may undergo a large change for a small change in the covariate matrix X. One way to check collinearity is by computing the variance inflation factor $VIF_i = 1/(1 - R_i^2)$ where R_i^2 is the coefficient estimate by regressing the $i-$th predictor against all others. VIF attains its minimum at 1 for $R_i^2 = 0$ where the $i-$th predictor is independent from the other covariates. Then, as rule of thumb, a VIF value exceeding 10 may indicate potential collinearity problems (Yan and Su, 2009).

- **Correlated responses**: with respect to the response space I model correlation among the different breeds in the training data. In general, each response could be modeled using separate linear regressions although I may lose information on breed relationships.

Firstly, I investigate remedial techniques which can be utilized to deal with multicollinearity:

- **Redefine variables**: if I knew a subset of predictors which are highly correlated I may remove all those variables and replace it a by a single variable which is a linear transformation of those predictors, such as their sum or difference.

- **Create new variables**: I perform principal component regression (PCR) (Massy, 1965) because it decorrelates the data in projected space: I compute principal component scores for data matrix X using the leading eigenvectors capturing most variability in the data. Then, each response is regressed on these principal component scores which form the new set of predictors.

- **Penalized regression**: in penalized regression problems estimation of coefficients is bounded to improve stability and select relevant variables.
  - In **ridge regression** (Hoerl and Kennard, 1970) I penalize the least square fit by using the L_2 norm, i.e. I adjust the least square estimator by adding a small multiple of the identity matrix to the covariance matrix, i.e. $\left(X^T X + k \cdot I_{p\times p}\right)^{-1}$, to improve numerical stability of the matrix inversion. Ridge regression is different to PCR by keeping all variables and not discarding principal component directions which explain little variability. However, weights of these low-variability directions are shrunk more.
  - **LASSO** (Tibshirani, 1996) is another example of penalized regression where we penalize the least square fit by using the L_1 norm. LASSO selects a subset of variables which contribute most in predicting the response. HYPERLASSO (Hoggart et al., 2008) is a Bayesian version for penalized regression.
  - **Multi-task relationship learning (MTRL)** (Argyriou et al., 2008; Zhang and Yeung, 2010): in multitask learning predicting each response is a task, and the correlation structure of the predictors is explicitly modeled. In particular, this MTRL is a least square problems which is penalized by coupling the univariate response regression coefficient estimate β with the inverse covariance matrix Σ of the predictors, i.e $tr(\beta\Sigma^{-1}\beta^T)$.

Secondly, I deal with correlated responses. In general, for independent outputs I could run a separate ridge regression for each response variable. However, if the responses are correlated I can do better by accounting for this joint relationship. I, continue with reviewing four approaches related to either covariance or correlation information between predictors and responses.

- **Curds & Whey (CW)** procedure (Breiman and Friedman, 1997) and **reduced rank regression (RRR)** (Izenman, 1975, 2008) are two examples of least square regression approaches built on the foundation of canonical correlation analysis (CCA). In particular, the responses are projected onto the canonical coordinate systems using the eigenvectors of the CCA decomposition. Then, a separate OLS regression fit is computed for each response in this new response space. After that, this estimated regression coefficient is used to predict each of the responses in canonical coordinate system. As next step, these predicted responses are shrunk by a diagonal matrix based on the eigenvalues of the CCA decomposition. Finally, the shrunk predicted responses are transformed back from the canonical coordinate system to the original space using the inverse of the CCA decomposition eigenvector matrix.
  - CW is motivated by performing a separate least square regression for each response. Then, to compute the $j-$th response, I form a linear combination of the b separate response estimates. Then, the coefficients of this linear combination are found through performing a CCA onto the covariates and responses.
  - RRR is formed by a MMLR which is weighted by the inverse of the covariate covariance matrix Σ^{-1} and features a rank constraint on the matrix of estimated regression coefficients. Breiman and Friedman (1997) showed RRR is also related to performing least square regression in the canonical coordinate system with a slightly different shrinkage matrix than CW.
- **Partial Least Squares (PLS)** is a class of methods for modeling relations between sets of observed predictors by using orthogonal latent score vectors (factors) (Wold, 1975; Rosipal and Krämer, 2006; Boulesteix and Strimmer, 2007; Postma et al., 2011). PLS is also well-known

to model strongly collinear data due to its projection on a smaller number of latent factors (Yeniay and Goktas, 2002). **Two-block PLS (PLS2)** models multivariate responses and the latent factors are found by maximising the covariance between the two blocks which are formed by predictors and responses whereas **PLS1** only models univariate responses. In other words, I look for linear combinations of covariates and responses, such that their covariance is maximized. Then, in PLS2, a separate regression can be performed for each response which is regressed on the canonical predictor covariance linear combination, i.e. on the transformed original covariates.

- **Multivariate regression with covariance estimation (MRCE)** (Rothman et al., 2010): MRCE is a least square regression problems which treats the predictor space as IID but weights the least square problem with the inverse of the response covariance matrix Ω, i.e. $\mathrm{Tr}\left[(Y - X\beta)\Omega^{-1}(Y - X\beta)\right]$ is minimized.

Frkonja et al. (2011, 2012) applied PLS1 and LASSO to model a univariate response which is used to predict breed composition in an crossbred Swiss Fleckvieh (SF) cattle population (305 admixed bulls) which has Simmental (SI, 90 pure bulls) and Red Holstein Friesian (RHF, 100 pure bulls) as founder populations. The goal in this study is to estimate two-way SF admixture proportions where 0 corresponds to zero percent RHF (100 percent SI) and 1 corresponds to 100 percent RHF (0 percent SI) using 40.4K SNPs. Performance has been measured using Pearson correlation between pedigree based analysis and either PLS1 or LASSO prediction. Performance for PLS1 has been 0.976 and LASSO 0.934. This excellent performance is expected because the average fixation index $F_{st} = 0.11$ is quite high and corresponds to a human population distance of Caucasian vs. Chinese/Japanese. The authors also studied the impact of reducing the number of markers to smaller panels based on either fixation index or evenly spaced subsets. Selecting 594 SNPs by $F_{st} > 0.5$ reduces PLS1 performance to 0.957 and keeps LASSO performance at 0.934. Finally, choosing 48 SNPs with the highest F_{st} further reduces PLS1 and LASSO performance to 0.903. Furthermore, PLS1's performance is less sensitive than LASSO to the way SNPs are selected, i.e. either selecting SNPs at even spacings or by the highest F_{st}. In particular, choosing 1 percent of the SNPs (404 SNPs) evenly spaced slightly reduced PLS1 performance to 0.955 but led to a strong decrease to 0.847 for LASSO. This shortcoming of LASSO may be due to its limitations to deal with correlated predictors (Zou and Hastie, 2005). In general, the performance of PLS1 and LASSO is even very strong for small panels which suggests strong LD patterns in the chip.

Returning to modelling correlation in response space, all of these approaches are restrictive to some extent. MRCE only models the output covariance but does not account for collinearity in the predictor space. CW, RR and PLS project the original data onto a different space which may discard relevant information if there are more responses than covariates (Rai et al., 2012). Although, in my context this is less relevant because there are more SNPs than breeds. Therefore, the current state of the art techniques aim to model covariate and response structures jointly using the inverse covariance information of predictors and responses (Lee and Liu, 2012; Sohn and Kim, 2012; Rai et al., 2012). I refer to these models as 'multivariate linear regression with multiple responses correlated' (MMLRC). For example, Rai et al. (2012) choose a framework which combines MTRL with MRCE, i.e. the following expression

$$\mathrm{Tr}\left[(Y - X\beta)\Omega^{-1}(Y - X\beta)\right] + \mathrm{Tr}(\beta\Sigma^{-1}\beta^T) + \mathrm{Tr}(\beta\beta^T) + m\log|\Omega| + p\log|\Sigma|$$

is minimized alternating between the inverse of the response covariance matrix Ω^{-1}, inverse of the predictor covariance matrix Σ^{-1} and regression coefficient β.

Finally, once I simultaneously modeled the covariate and response space, there may still be some structure left which has not been considered. In particular, the chosen hypothesis class of responses which depend linearly on the covariates may have limited expressive power given that the data may exhibit nonlinear behavior. There are two main ways to handle nonlinearity in the data according to Rosipal and Krämer (2006):

- **Basis function expansion (Ruppert et al., 2003)**: the original data is fit to various basis functions expansions, such as polynomials, splines for non-period data and a Fourier basis expansion for periodic data. Then, the fitted coefficients are used as new predictors to account for the nonlinearity. Alternatively, a nonlinear function, such as sigmoid, is applied to composite linear functions, such as artificial neural networks (ANN). Then, to perform nonlinear regression or classification, this model is fit by a technique using gradient descent and the chain rule for the multiple layers of linear functions which is known as backpropagation. The ANN with multiple outputs could be applied to our problem. However, the basis function approach is known to only work well with a small number of covariates which is not satisfied in our case.

- **Kernel approach (Shawe-Taylor and Cristianini, 2004)**: another approach is based on kernelizing the PLS2 approach, i.e. by replacing inner products of the covariates in the optimization algorithm by a kernel function which maps the original data to a new space of higher dimensionality. The kernel approach will be described more in detail in Section 5.1.7 where I discuss this topic with respect to the support vector machine (SVM) approach. Bennett and Embrechts (2003) showed that for their datasets kernel PLS achieves the least average misclassification rate compared to different SVM versions and a kernelized version of ridge regression.

- **Generalized linear model (GLM) (MacCullagh and Nelder, 1989):** GLM is a generalization of OLS where the response variable is linearly related to the covariates using a link function.

5.1.2 Global model-based clustering

5.1.2.1 Structure

Global model-based clustering approaches attempt to infer the ancestry proportions of individuals from K ancestral populations based on genotype data. In one of the earliest global model-based ancestry inference approaches, Pritchard et al. (2000) introduce the Bayesian clustering approach **STRUCTURE** that assigns admixed individuals to one or more populations. Given genotype data X the algorithm jointly estimates the following vectorial parameters: allele frequencies S for each population which follow a Dirichlet distribution (i.e. markers are considered unlinked), the Dirichlet-distributed admixture proportion Q of each individual that it was derived from a particular population, and the population of origin Z of each locus for each individual. STRUCTURE relies on Gibbs sampling to sample from the posterior distributions of all those parameters (Gilks et al., 1996). In other words, given genotype data, allele frequencies and admixture proportions, I predict the population of origin for each locus. And given the genotypes with the population of origin, I update allele frequencies and individual ancestry proportions. STRUCTURE has been successfully used in clustering of two-way cattle admixtures (Frkonja et al., 2011, 2012) and the prediction of different types of dairy cattle in Kenya (Gorbach et al., 2010).

Later, Tang et al. (2005) developed a frequentist approach of STRUCTURE, known as **FRAPPE**, using maximum likelihood estimation. Firstly, Tang et al. (2005) are concerned about a reliable assessment of convergence of MCMC approaches and their sensitivity to priors. Secondly, their frequentist approach is less computationally intensive although the root mean squared error between true and estimated admixture proportion is marginally bigger than that of STRUCTURE. Finally, FRAPPE shows

less bias of admixture proportion estimates in information-poor setups which have less than 100 markers and individuals per ancestral population. Frappe has also been implemented in two software packages: in the ethnic admixture option of Mendel (Lange et al., 2013) and in PSMIX (Wu et al., 2006). FRAPPE has successfully been applied for prediction of ancestry proportions of cross bull breeds (Kuehn et al., 2011) as well as human admixture (Wu et al., 2006).

5.1.2.2 Structure 2

Falush et al. (2003) developed an extension of the original STRUCTURE model which is referred to as **STRUCTURE 2** and accounts for correlations between linked markers in admixed populations. Although STRUCTURE 2 belongs to the local HMM category in Section 5.1.4 it is described here because it is designed based on the limitations of STRUCTURE. The STRUCTURE model assumes that the allele population origin Z within each individual are independent although inter-marker correlations are expected along the chromosome. Therefore, STRUCTURE can only model one type of linkage disequilibrium known as mixture LD which looks at within-breed variation according to breed-specific allele frequency patterns (Alexander et al., 2009). The STRUCTURE 2 model also accounts for admixture LD which takes correlation in ancestry along the chromosome into account. In particular, each chromosome is viewed as mosaic of chunks (contiguous SNP sequences) which are inherited as unbroken units from the ancestral population (Daly et al., 2001; Li and Stephens, 2003; Scheet and Stephens, 2006; Lawson et al., 2012). In Falush et al. (2003), chunks of chromosomes are inherited as units from one or more of the K populations and all markers composing a particular chunk derive from the same population. In the STRUCTURE 2 model breakpoints, i.e recombination events which change ancestry, between consecutive chunks are distributed according to a Poisson process with rate r per unit of genetic distance. Random variable r is estimated as part of the MCMC algorithm and the limiting case of an infinite recombination rate corresponds to the admixture case of STRUCTURE with independent loci. Therefore, STRUCTURE 2 is a hidden markov model where the hidden state Z corresponds to the unobservable ancestral population proportions Z which generate the observed genotypes X.

5.1.2.3 Admixture

Alexander et al. (2009) noted that prior model-based clustering approaches focused on the use of AIM markers which were known to have different frequency patterns in different populations. As noted by Tang et al. (2005) this is due to the slow convergence of STRUCTURE in the case of many markers. Therefore, Alexander et al. (2009) developed **ADMIXTURE**, which is a faster frequentist version of FRAPPE, which uses an extension of Newton's method for constrained optimization and maximizes a second-order Taylor expansion of the log-likelihood by alternating between maximizing allele frequencies and ancestry proportions. The speed of ADMIXTURE is similar to singular value decompositions approaches, such as PCA, which makes it possible to use large number of SNPs for analysis, or select AIMs for different analyzes based on inferred allele frequencies.

5.1.3 Local window-based approaches

5.1.3.1 Lamp

In window-based approaches locus ancestry inferences are carried out separately for each window, such that in each window there is either none or at most one recombination event. These window-based approaches are motivated by the fact that other models, such as HMMs, estimate a large number of parameters including the exact position of recombinations which requires a search over a prohibitively large search space. Sankararaman et al. (2008b) developed an approach called **LAMP** which looks at recently admixed populations where the number of generations g since the start of mixing is small. An

extension of their model known as **LAMP-ANC** uses knowledge of ancestral genotype information. Resulting chromosomes are viewed as mosaic of K populations. LAMP-ANC assumes certain parameters to be known in the admixed population, such as the Poisson-distributed recombination rate r, individual admixture proportions α and generations g. Furthermore, allele frequencies f for each population are estimated from the ancestral populations. LAMP moves a sliding window of length l over the genotype. For each window, LAMP-ANC infers the most likely population pair assignment given f and the genotype of the current individual. Locus-specific ancestry prediction errors are minimized by taking a majority vote of all windows which include a given marker. The window size l is computed using g, α, r and is chosen to be short enough to have mostly zero recombination events but also long enough to yield sufficient information for population discrimination. Sankararaman et al. (2008b) compare LAMP-ANC based on 40K SNPs with STRUCTURE and SABER (Tang et al., 2006) (to be introduced later) using 4K SNPs due to computational complexity. They perform experiments on two-way admixtures (Nigeria vs. European descent, Japanese vs. European descent, Japanese vs. Chinese). They find that based on allele frequency distance of considered ancestral populations LAMP-ANC outperforms STRUCTURE by 11 to 46 percent and SABER by 4 to 11 percent in accuracy of prediction the loci ancestries for each population.

5.1.3.2 WinPop

LAMP-ANC performs very well for admixtures from distant populations, such as Nigerian vs. European descent, but works less well for closely related populations, such as Japanese vs. Chinese, because the algorithm does not consider population differentiation as modeled by allele frequency differences. This leads to a setup where the window size is chosen without respect to the genetic population affinity. Furthermore, the assumption of no recombination events in a given window leads to shorter window with higher rate of errors. Therefore, Paşaniuc et al. (2009b) proposed an extension of LAMP with adaptive window size known as **WinPop**, such that each locus the local genetic distance between two ancestral populations is taken into account. WinPop estimates locus ancestries by using dynamic programming which iterates over all markers in the window, and for each position computes the probability of having one ancestry upstream while the other is located downstream. For distant populations the accuracy gain of Win Pop is 4 percent over LAMP and HAPAA (Sundquist et al., 2008) (introduced below) and 8 percent over SABER. Furthermore, for closely related populations WinPops yields an improvement in accuracy of 17 percent over LAMP, 10 percent over HAPAA and 14 percent over SABER.

5.1.4 Local HMMs

5.1.4.1 Early HMM approaches

Due to very high sequencing cost the earliest local ancestry inference algorithms using hidden markov models relied on small marker panels which were composed of 1,500 to 5,000 markers (Smith and O'Brien, 2005; Seldin et al., 2011). In these early HMM implementations the ancestry states consitute the hidden chain and must be inferred from the observed genotypes. Furthermore, transition probabilities model recombination events while emission probabilities follow ancestral marker-wise and population-specific allele frequencies. These HMMs combine information on genotypes at adjacent loci which leads to higher utility of models focussing on separate analysis of each marker. Two of the most well-known approaches are known as **ADMIXMAP** and **AncestryMap**, respectively (Hoggart et al., 2004; Patterson et al., 2004). In these two approaches certain parameters, such as recombination rate and allele frequencies, are assumed to be known.

5.1.4.2 Saber and Switch

Tang et al. (2006) developed a second-order HMM model known as **SABER** for genome wide data which explicitly accounts for LD, i.e. correlation of markers in dense genetic data. Motivated by computational tractability early HMMs assumed that observed SNPs are independent from each other conditional on the hidden ancestry state. However, SABER uses a Markov-HMM (MHMM) where the genotype evolves in a non-Markovian along the chromosome, such that the current allele at time t depends on the hidden state at time t according to the HMM emission probabilities but also on the SNP at time $t-1$. The transition probabilities in SABER closely follow STRUCTURE 2 where hidden state transitions are Poison distributed based on marker distance and time since admixing. The MHMM is estimated via an extension from the standard forward and backward algorithm but due to the larger parameter space compared to HMMs requires more marker information. SABER as first approach to model admixture LD in local ancestry inference motivated further investigation of linkage patterns and development of more sophisticated LD models.

Sankararaman et al. (2008a) developed the **Switch** model which is closely related to SABER due to its implementation as a MHMM model for two-way admixture. The Switch model augments SABER by including explicit indicators for recombination events, such that in the case of no recombination events ancestries remain unchanged while for a recombination event the authors select SNP ancestry independently. In other words, SABER conditions on ancestry states and does not account for recombination events which do not change ancestries. On the other hand, in Switch emission probabilities are additionally conditioned on recombination events directly which is beneficial if the algorithm encounters new haplotypes unseen in the ancestral reference populations that were formed due to recombination events without ancestral change. Imagine that the first ancestral population has two haplotypes 00 and 11 and the admixed population has haplotype 01 created by a recombination event of the two haplotypes from the first population: in this case SABER would assign a negligible probability that this haplotype comes from the first population only. On the other hand, Switch assigns a probability equal to the product of the allele frequencies for these two markers.

5.1.4.3 Hapmix and HAPAA

The next step in the development of HMM models leads to an even more accurate modeling of LD although their model complexity and run-time, which scales quadratically with the size of the ancestral population, are limiting factors (Seldin et al., 2011). Early HMM ancestry inference models and second-order HMM approaches, such as SABER and Switch, model genome-wide data which do not fully account for long-range dependencies. This limitation could be partly overcome by switching to higher order models at the expense of exponential model growth and decreased estimation tractability. However, these higher fixed-order models still do not fully examine information contained in the full haplotypes. This incomplete modeling of LD between non-adjacent markers leads to systematic biases causing an increased number of false positives in ancestry inference (Price et al., 2008). Furthermore, an accurate modeling of long-range LD improves the capture of the ancestry signal of recombination events further back in time which are characterized by shorter and more remotely inherited blocks. Therefore, the ancestry inference community developed the models 'HMM-based analysis of polymorphisms in admixed ancestries' (**HAPAA**) and **HAPMIX** which model entire haplotypes using nested HMMs (Sundquist et al., 2008; Price et al., 2009). HAPAA and HAPMIX both rely on a hierarchical hidden layers in the HMM which models transitions at different levels. HAPAA is a model which has transitions at the level of population, individual and haplotype, such that a compound hidden state is formed by one of the two haplotypes of a particular individual in one of the ancestral populations.

HAPMIX also implements a nested HMM which involves transitions at two layers, i.e. small-scale transitions between haplotypes of ancestral panel and a large-scale transition switching between populations. Both, HAPAA and HAPMIX, assume that test genotypes are formed as mosaic of the ancestral haplotypes (Li and Stephens, 2003; Falush et al., 2003) while allowing for phase switch errors, and miscopying from the 'wrong' population in the case of HAPMIX or providing a small allowance for unseen haplotypes in the ancestral population, mutations and genotype errors for HAPAA. Due to finite sample size certain haplotypes may not be seen in the ancestral population. Therefore, to smooth out short inferred ancestry chunks HAPAA applies a post-processing filtering step which discards inferred ancestry chunks below a threshold of particular genetic length. On the other, HAPMIX explicitly specifies a miscopying probability within the model because a test individual may coalesce first with an ancestor from the 'wrong population' while further back in time admixed with the true population. As final limitation I note that the model size grows linearly and time complexity scales quadratically with the number of parental individuals which makes inference for prohibitively costly for large datasets.

5.1.4.4 GEDI-ADMX

The **GEDI-ADMX** as developed by Paşaniuc et al. (2009a); Kennedy (2009) implements a factorial hidden markov model (FHMM) (Ghahramani and Jordan, 1997) which assumes a distributed state representation, such that there are several hidden chains interacting which jointly emit a symbol at each position of the observed time series. The hidden chains of this first order HMM are formed by the maternal and paternal ancestral populations, respectively, which emit haplotype values at the corresponding loci. Then, these two HMMs are combined by emitting the sum of both haplotypes values at a particular locus with probability 1 if the sum adds up to the corresponding allele, and 0 otherwise. GEDI-ADMX runs along the genotype assuming each marker as missing (given all other SNPs) and imputes alleles under all possible local ancestries. Finally, the ancestry with the highest imputation accuracy is chosen using a weighted-voting scheme of window lengths' between 100 and 1500 loci. For distant two-way admixtures GEDI-ADMIX performs better than HMM-based models (Saber, Switch, HAPAA) and is competitive with window-based models (LAMP, WinPop). Furthermore, for closely related populations, such as Japanese vs. Chinese, GEDI-ADMIX significantly improves these HMM- and window-based models by 11 to 26 percent. Given that haplotype value imputation scales cubic with the number of ancestral population, GEDI-ADMIX is rendered impractical for large number of populations.

5.1.4.5 LAMP-LD

Baran et al. (2012) invented the **LAMP-LD** model as accelerated extension of HAPMIX by implementing a variant of the standard model by Li and Stephens (2003) and incorporating ideas from LAMP. LAMP-LD was motivated by three key limitations: high time complexity and number of populations limited to two-way admixtures, and a bias in the Li and Stephens model which overpredicts ancestry changes because recombination events constitute supersets thereof. Firstly, the authors noted that according to Price et al. (2009) the Li and Stephens model introduces too many ancestry switches although some recombination events, i.e. between the same ancestral population, leaves the ancestry switched. Therefore, LAMP-LD borrows the windowing idea from LAMP, such that the fully haplotypes are segmented into haplotype blocks formed by non-overlapping windows of length L will have constant ancestry, to avoid false, short ancestral segments. Furthermore, to be able to deal with large ancestral reference panels in a computationally efficient way the number of hidden states in each windows needs to be limited. In particular, HAPMIX scales linearly in model complexity and quadratically in time complexity as a function of the number of ancestral individuals n. Therefore, LAMP-LD restricts the

number hidden window states to $S << n$. Experiments showed that S in the range of 10-15 is sufficient to capture genetic variation and yield low ancestry estimation error. Similarly, window sizes of 50-100 were found reasonable to capture enough LD information. As last limitation, Seldin et al. (2011); Baran et al. (2012) noted that many approaches, such as SABER, HAPAA, HAPMIX, LAMP and WinPop, are targeted at two-way admixtures, such as African-American admixtures, but are either not applicable or insufficiently accurate for admixtures of three populations, such as Latinos formed by European, African and Native American influence, or, generally, for local ancestry inference of multi-way admixed populations. Although LAMP-LD scales linearly in the number of windows its run time is proportional to $\binom{K}{2}^2$ which is prohibitively large for high numbers of ancestral populations K. Based on these shortcomings LAMP-LD applies a hierarchical HMM where the top level HMM transitions between windows representing different ancestral populations while the nested window HMM models the probability of a haplotype segment formed by L markers over S hidden states. Given the constant ancestry within the nested window HMMs any ancestry switches are only allowed at the boundary between consecutive windows. Finally, as a post-processing step their algorithm searches for ancestry breakpoints close to the endpoints of windows where an ancestry change has been inferred because the actual algorithm only allows ancestry switches at the boundary of the nested HMM windows.

5.1.4.6 Alloy

With the prospect to further improve background LD modeling in ancestral populations Rodriguez et al. (2013) developed **ALLOY** which implements an inhomogeneous variable-length Markov chain model (VLMC) which can be represented as FHMM for efficient computations (Ron et al., 1995; Browning and Browning, 2007). Previous simpler models AncestryMap and SABER implemented first order and second order Markov models, respectively, which only capture limited marker correlation and over-simplified background LD in admixture cases. Later, HAPAA and HAPMIX, developed the other extreme case of over-specification where full ancestral haplotypes are explicitly used. Furthermore, Halperin and Eskin (2004); Rodriguez et al. (2013) notice that fixed haplotype blocks do not adequately model inter-marker correlation structure because LD may either extend beyond short ranges or exhibit block patterns. Therefore, ALLOY uses haplotype clusters as hidden layer states for all loci. The number of hidden states are adapted according to the amount of locally encountered LD separately for each hidden chain corresponding to an ancestral population in the FHMM. These localized haplotype clusters at each locus are formed by a group of locally similar haplotypes which share an edge transitioning between consecutive SNP markers with the current locus as parent node, such that adjacent loci are also likely to be in the same cluster. Then, each of the haplotype clusters emits one and only one haplotype value with certainty. Rodriguez et al. (2013) found that ALLOY which represents a flexible LD model of intermediate complexity improves ancestry estimation accuracy over either fixed-order HMMs or explicit use of haplotypes.

5.1.4.7 Multimix

Motivated by HAPMIX, Churchhouse and Marchini (2013) developed **MULTIMIX** which is an extension of two-way to multiway admixtures using multiple ancestral source populations. MULTIMIX segments chromosomes into windows of equal length of constant ancestry. For each pair of window and ancestral population the algorithm estimates SNP-wise mean allele frequencies and the covariance matrix over SNPs in this particular window to model pairwise LD. Then, the probability of an admixed haplotype segment for a given pair of window and ancestral population is evaluated assuming a multivariate normal density with parameters corresponding to the estimated mean and covariance of the

allele frequencies of the given source population. Furthermore, similarly to the miscopying process in HAPMIX, the authors assume a misfitting property where an ancestral population is copied for a window which is different to the estimated ancestral population from the normal density. This misfitting property is assumed to decrease spurious ancestry switches at window boundaries. Finally, a first-order markov process is assumed over the ancestries in consecutive windows which is formulated as a function of the genetic distance between the center marker of consecutive windows. As before first-order models may be limited when LD between SNPs extends over several windows. Inference for the most likely hidden ancestry state sequence is computed using an EM algorithm and MCMC approach. Although Churchhouse and Marchini (2013) did not apply MULTIMIX to more than five source populations, the number of populations K does not seem to be a limitation given that MULTIMIX linearly scales in the number of windows and ancestral populations.

5.1.4.8 ChromoPainter

The final HMM approach which is suitable for multiway admixture is **ChromoPainter** (Lawson et al., 2012; Hellenthal, 2012; Hellenthal et al., 2014). ChromoPainter models shared ancestry through haplotype similarity patterns in genome-wide studies. In particular, the algorithm paints individuals ('recipients') as linear combination of segments ('chunks') taken from individuals ('donors') in the ancestral panel. In other words, this painting process at each locus can be interpreted as finding the closest related haplotypes of length L SNPs for in the recipient individual. The similarity of the ancestral dataset formed of two haplotypes for each individual is given by a coancestry matrix X. Element x_{ij} in this **copying matrix** shows the predicted number of chunks recipient individual i copies from donor individual j and takes larger values if individuals i and j are sampled from genetically close populations.

ChromoPainter is based on a haploid copying model which was developed to relate genetic variation to two underlying scaling parameter, i.e. crossover recombination rate ρ per unit physical distance and per locus mutation rate θ (Li and Stephens, 2003; Hellenthal et al., 2008). The Li and Stephens (2003) approach features a computationally tractable likelihood based model which covers key properties of a genealogical process. Their model is based on the conditional likelihood decomposition p of the joint distribution over all n haplotypes $h_1, \ldots, h_n$ where the conditional densities are approximated using $\hat{\pi}$ which is referred to as 'product of approximate conditionals' (PAC)

$$p(h_1, \ldots, h_n, \rho) \approx \hat{\pi}(h_1|\rho) \cdot \hat{\pi}(h_2|h_1, \rho) \cdot \ldots \cdot \hat{\pi}(h_n|h_1, \ldots, h_{n-1}, \rho) = L_{PAC}(\rho). \tag{5.2}$$

However, ChromoPainter does not order the haplotypes in the same order as the Li and Stephens (2003) PAC model but each next haplotype is constructed using haplotypes from all other individuals as potential donors, i.e. the $n-$th haplotype is modeled using other training haplotypes $1, \ldots, n-1$. This process is repeated for all haplotypes, such that all haplotypes are painted in terms of all other haplotypes.

The approximations for those conditionals $\hat{\pi}(h_{n+1}|h_1, \ldots, h_n)$ should capture certain properties of an evolving population: the next haplotype is more likely to be similar to haplotypes observed frequently and often only deviates by a small number of mutations from a previous one. Clearly, I would also expect that the probability of a novel haplotype is a function as a Markov process of n previously encountered haplotypes and mutation rate θ where k decreases the chances and θ makes new haplotypes more likely. Finally, due to recombination, I expect the next haplotypes to be slightly similar over contiguous SNP markers from several previous haplotypes where the length of those shared segments depends on the

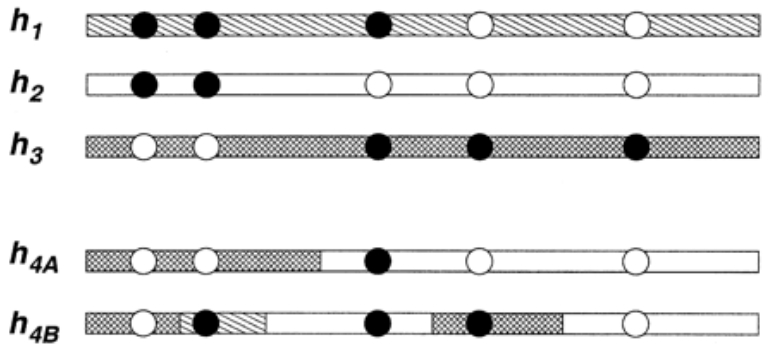


Figure 5.1: This Figure is due to Li and Stephens (2003) and shows the fourth haplotypes are coloured according to an imperfect mosaic of the previous ones, with a probability I denote as $\hat{\pi}(h_{4\{A,B\}}|h_1, h_2, h_3)$. This example captures the copying process for two examples of the 4th haplotype where h_{4A} has a lower switch rate than h_{4B}. The shading for those next haplotypes illustrates which previous haplotypes I copy from. And jumps in the shading indicates switches in the ancestral copying. Given that shadings are unavailable in actual datasets I need to sum over all possible ones. Each circle represents a biallelic SNP marker along the chromosome with white/black color coding according to allelic variation. The third SNP marker illustrates the effect of mutations in the imperfect copying process where both $h_{4\{A,B\}}$ have the black allele but still copy from the second haplotype which has the white allele.

local recombination rate in those genomic regions, i.e. a low recombination rate leads to longer segments shared. In Figure 5.1 I show how the next haplotype is painted as linear combination of the previous ones.

I assume the copying process evolves Markovian along the chromosome which leads to a hidden markov model, see Section 5.1.4 for further HMM approaches. This HMM is parametrized by a vector of genetic distances $\rho = \{\rho_1, \ldots, \rho_{L-1}\}$ where ρ_l refers to genetic distance between sites l and $l+1$, by a per site 'imperfect copying' mutation parameter θ and by parameter vector $f = \{1, \ldots, k\}$ where f_k is a vector of copying probabilities which indicates the probability of copying from haplotype h_j at any locus. Then, ChromoPainter structures the conditional probability $\hat{\pi}\left(h_{n+1}|h_1, \ldots, h_n, \rho\right)$ as hidden markov model. The hidden state sequence in the HMM is given by $Y = \{Y_1, \ldots, Y_L\}$ where Y_l indicates which of the k ancestral haplotypes I copy from at each locus. The switches are distributed according to a Poisson process with rate ρ_l and the default for the prior haplotype copying probabilities is uniform, i.e. $f_k = \frac{1}{n}$. Then, the transition probabilities for the hidden chain are given in Equation 5.3.

$$p(Y_{l+1} = y_{l+1}|Y_l = y_l) = \begin{cases} \exp(-\rho_l) + (1 - \exp(-\rho_l))\, f_{y_{l+1}} & \text{if } y_{l+1} = y_l; \\ (1 - \exp(-\rho_l))\, f_{y_{l+1}} & \text{otherwise.} \end{cases} \tag{5.3}$$

Therefore, the observed sequence in the HMM most of the time, i.e. with probability $1 - \theta$, emits the allele according to the copied haplotype. However, to account for the effects of mutation the mosaic of copied donor individuals may be imperfect, and with probability θ, the allele not found in the donor haplotype is copied. Therefore, the emission probabilities are defined as in Equation 5.4 where $h_{k,l}$ refers to the allelic type of the $k-$th ancestral haploid k at SNP l.

$$\gamma_{n+1,l}(y_l) = p(h_{n+1,l} = a|Y_l = y_l, h_1, \ldots, h_n) = \begin{cases} 1 - \theta & \text{if } h_{y_l,l} = a; \\ \theta & h_{y_l,l} \neq a. \end{cases} \tag{5.4}$$

where $p(h_{n+1,l})$ refers to the allelic type of next haploid $n+1$ at locus l. Based on these transition and emission probabilities ChromoPainter computes the expected number of chunks copied from each donor individual. ChromoPainter scales linearly, i.e. its computational complexity has order $o(LDR)$ where L is the number of SNPs, D the number of donor individuals and R the number of recipient individuals.

5.1.5 Non-Parametric Bayesian approaches

Nonparametric methods have long been popular in classical statistics (Wasserman, 2006) and attracted increased interest in the machine learning community recently (Teh and Jordan, 2010; Orbanz and Teh, 2010; Ghahramani, 2013). There are two main benefits of nonparametric modelling over a parametric approach which has driven its development: firstly, non-parametric models show flexibility by growing adaptively with the amount of data. Secondly, non-parametric do not depend on model selection, e.g. there is no need to specify the number of hidden states in a HMM.

In Bayesian noparametric modelling an infinite-dimensional parameter space is utilized; but on a given dataset D only a finite number of parameters is used which grows with the dataset. The infinite hidden markov model (Beal et al., 2002) is an example for a methodology which has been extended nonparametrically. This infinite HMM can deal with an infinite number of latent states to circumvent the model selection issue of choosing among varying number of hidden states.

In this Section I describe two nonparametric Bayesian approach for modelling of genetic data: the first approach performs local genetic ancestry inference using **iHMM** in an admixed population (Sohn, 2011; Sohn et al., 2012) while the second approach **FragCoag** clusters past genetic data as mosaic using fragmentation-coagulation processes (Teh et al., 2011; Elliott and Teh, 2012; Teh et al., 2013).

5.1.5.1 iHMM and FragCoag

The iHMM is motivated by two main factors: extension of HAPMIX to multiway admixture and by modeling of genetic relatedness among ancestral populations. In particular, most ancestry inference approaches did not account for genetic relatedness in ancestral populations but assumed them to be independent which led to high sensitivity with respect to size and choice of individuals in the ancestral populations. To incorporate this sharing of genetic information across populations iHMM models ancestry inference as a composite process where at the top level admixed individuals inherit from ancestral populations which on a finer level depend on a mosaic of founder haplotypes. In other words, based on these jumps between these two genetic resolutions the founders create the ancestral populations and the admixed individuals. In this iHMM each ancestral population is modeled using a separate infinite HMM where the hidden states correspond to a joint space of hypothetical founder haplotypes and ancestral populations, and the observed sequence refers to the observed haplotypes of the admixed individuals. Genetic relatedness is modeled by sharing a set of founders across multiple populations, such that each infinite HMM selects its own unique subset of the shared founder haplotypes with their own recombination and mutation patterns. Then, an admixed individual is formed by a mosaic of founders with transitions according to the recombination process, and emissions are based on mutations of founder individuals, i.e. segments in the admixed individual are due to haplotypes in the ancestral populations which are based on the founder haplotypes.

FragCoag is another approach which clusters a set of genetic sequences to form a mosaic structure, such that at each locus sequences are unlabeled partitioned into groups of variable size according to

similarity. The group membership of this partition evolves along the chromosome by splitting ('fragmenting') and merging ('coagulating'), such that two clusters are merged or one cluster is split into two.

5.1.6 PCA-based approaches

Principal component analysis (PCA) is an established multivariate technique which belongs to the algorithmic view (see Section 1.3.2) (Krzanowski, 2000; Izenman, 2008). PCA is utilized for dimensionality reduction and can be derived by two main ways: either as a projection of the data in the direction of maximum variance (Hotelling, 1933), or, by using the least-square optimality criterion to construct a low-rank matrix approximation of the original data (Izenman, 1975, 2008). More, than three decades ago, Menozzi et al. (1978) was the first to apply PCA to genetic data. Menozzi et al. (1978); Cavalli-Sforza et al. (1993) collected allele counts for a few loci from population samples at several European, Asian and African locations. Then, these allele counts were converted into frequencies and interpolated between sample locations. As low-dimensional summary statistic, for each of the first few principal components a heatmap ('synthetic map') according to the corresponding PC score is plotted which is overlaid with the matching spatial location. Novembre and Stephens (2008) noticed that an application of PCA to spatially varying allele frequencies results in sinusoidal PC score patterns due the decay of allele frequencies similarities with geographic distance.

5.1.6.1 SmartPCA

In general, allele frequencies differences are due to evolutionary factors, such as migration, selection and drift, and lead to population structure. Then, PCA can be used as a fast non-parametric approach to recover this population structure. However, PCA is an unsupervised technique which projects individuals along principal component axes of major variation. But PCA does not classify individuals according to pre-specified population labels although a test individual of unknown mixture can be assigned to its closest population by clustering ancestral panel individuals. A popular implementation of PCA is **SmartPCA** (Patterson et al., 2006) from the EIGENSOFT softare package which also contains formal tests for the existence of population structure. In particular, a dataset exhibits population substructure if the eigenvector corresponding to the largest eigenvalue is significantly greater than the remaining eigenvalues. For this purpose, Patterson et al. (2006) developed a test statistic based on Tracey-Widom (TW) distribution while Limpiti et al. (2011) consider the EigenDev heuristics. Both of these test statistics describe the distribution of the largest eigenvalue whereas the EigenDev test statistic is more advanced by also taking into account the remaining eigenvalues.

5.1.6.2 ipPCA

Intarapanich et al. (2009) developed ipPCA which aims to detect population substructure, find the optimal number of populations and to assign individuals accordingly. Previous approaches which used clustering of individuals in multi-dimensional PC score space were found to be insufficient for closely related subpopulations. Firstly, this is because it is not clear how many leading PC directions should be utilized, and secondly, clusters may or may not be merged depending on the axis considered. Therefore, **ipPCA**, which is a top-down algorithm, iteratively builds a population tree by splitting individuals into two clusters if population substructure was found according to the TW test statistic. Then, the optimal number of sub-populations can be found by counting terminal nodes in the final population tree. Limpiti et al. (2011) extended the ipPCA model by employing the EigenDev test statistic which decreased the number of type I errors in large datasets over the TW test statistic, i.e. fewer spurious patterns are promoted as predicted population substructure.

5.1.6.3 PCAdmix 1 and 2

So far all these methods have been focussed on population structure inference. However, I am more concerned with admixture estimation. To this end, Bryc et al. (2010) implemented **PCAdmix 1** which combines PCA with a first-order HMM to estimate two-way admixture proportions. Firstly, the authors compute the first principal component score for all individual genotypes of the two ancestral populations. Then, for a test datum I compute its principal component score by projecting its genotype on the leading principal component of the ancestral populations. Then, I define as a and b the chord distances between the centroids of population 1 and 2, respectively, and I compute ratio $P = a/(a+b)$. This PC 1 distance analysis is performed for all consecutive non-overlapping 15 SNP windows. Then, the observed sequence for the HMM is formed by the window ratios P while the hidden states represent the ancestry states and are formed over the tenary alphabet $\{0, 1, 2, \}$ to express how many alleles are copied from population 2.

Brisbin et al. (2012); Henn et al. (2012) developed **PCAdmix 2** which extends PCAdmix 1 to multiway admixtures using a haploid version of the HMM presented in Bryc et al. (2010) for haplotype data. Furthermore, in multiway-admixture of K populations the algorithm projects an admixed individual onto the first $K-1$ principal components of the K ancestral populations. Then, similarly to the P ratio in PCAdmix 1, I define a ratio $Q_{i,j} = D_{i,j}/\sum_{k=1}^{K} D_{i,k}$ in PCAdmix 2 which describes the average ancestry of admixed haplotype i in population $j \in \{1, \ldots, K\}$. In particular, $D_{i,j}$ is defined as the Euclidean distance of the admixed individual in PC space to the hyperplane consisting of the mean scores of all ancestral populations not including population j.

I discussed in chapter 2 on breed similarity that PCA and MDS compute different projections unless I use Euclidean distance as distance measure in MDS. Therefore, to take advantage of a wide range of distance and correlation measures it is beneficial to explore MDS-based approaches. Furthermore, PCA is very sensitive to outliers and may lead to more significant dimensions which are spurious and do not represent population structure, i.e. dimensions which are only required to separate outliers (Luca et al., 2008). In other words, clusters of projected individuals are biased towards small and tight arrangements. Therefore, Lee et al. (2010a) developed **spectral GEnetic Matching (GEM)** using a spectral decomposition based on spectral clustering (Von Luxburg, 2007) where individuals are represented as nodes in a graph connected by edges which have weights corresponding to the strength of their connections. The weights in this spectral clustering approach are chosen corresponding to the entries of the covariance matrix. Then, this spectral clustering approach finds a partition of the graph such that edges between different clusters have low weights and within cluster weights are high.

5.1.7 Machine Learning based approaches

Recently, machine learning based approaches emerged to tackle the problem of ancestry inference.

5.1.7.1 Ethnopred

Hajiloo et al. (2013) developed the **ETHNOPRED** software which predicts either continental or subcontinental populations, such as European subpopulations (North-Western vs. Southern) and East Asian subpopulations (Chinese vs. Japanese), using data from the HapMap projects. The authors method of choice are an **ensemble of disjoint decision trees** (Breiman, 1993). Decision trees are an established statistical technique which attempt to classify an individual by making consecutive binary choices based on the value of SNP markers. The marker chosen at each step is based on the expected value of the information gain, i.e. how certain I am to classify an individual once I observe the current marker.

Furthermore, (Hajiloo et al., 2013) chose an ensemble of about 30 decision trees, such that with high likelihood at least 9 decision trees are available which do not have missing markers. The authors show that 150 SNPs are sufficient for successful continental classification and less than 500 markers are required for strong sub-continental discrimination.

5.1.7.2 SVM-based approaches

The following approaches I discuss are based on support vector machines (SVMs) (Burges, 1998; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004). In their most basic form, SVMs estimate a separating hyperplane between two classes of individuals so as to maximize the margin, i.e. distance between closest points of either class to hyperplane. An important property of SVMs which I use is that its discriminatory power does not depend on the dimensionality of the space where the hyperplane is learnt, and I can learn in very high dimensions without overfitting. In particular, SVMs use a kernel which implicitly maps the original genotype data into high dimensional feature space, such that the linear separation in feature space may correspond to a non-linear separation in the original space. The mapping from the input to the feature space is accomplished using a kernel function K. Most of the kernels were developed for individuals with continuous feature spaces although they are used for discrete biological sequences (Wessel and Schork, 2006; Kwee et al., 2008; Wu et al., 2010; Lin et al., 2011; Schifano et al., 2012).

A common kernel function for the genotype of two individuals $x, y \in \mathbb{R}^p$ of fixed length p include the inner product kernel $K(x,y) = x^T y$ which measures cosine angle similarity. This linear kernel can also be augmented by a weight vector w to define $K(x,y) = \sum_{i=1}^{p}(w_i \cdot x_i \cdot y_i)/\sum_{i=1}^{p} w_i$ where a w_i is often defined as $w_i = (q_i)^{-0.5}$ where q_i is the minor allele frequency for the $i-th$ locus. This weighting scheme increases the weights for rare variants and downweighs common alleles, i.e. rare alleles carry more information than similarity in rare alleles. Further common choices are polynomial kernels of degree d $K(x,y) = \left(x^T y + c\right)^d, c \in \mathbb{R}$, $K(x,y) = \left(I(x=y) + c\right)^d, c \in \mathbb{R}$ and $K(x,y) = \left(|x-y| + c\right)^d, c \in \mathbb{R}$ based on inner products, Hamming distance and L_1 norm, respectively. Another popular choice is given by the radial basis function (RBF) kernel $K(x,y) = \exp\left(-\frac{1}{2\sigma^2}||x-y||^2\right)$. In statistical genetics, the identical by state (IBS) kernel is a popular kernel function because it incorporates SNP-SNP interactions. This IBS kernel measures the number of alleles shared at each haplotype locus which can be either 0,1 or 2 and is defined as $K(x,y) = \frac{1}{2p}\sum_{i=1}^{p} IBS(x_i, y_i)$. Similarly to the weighted linear kernel the IBS kernel can be integrated with weights, such that rare alleles are more emphasized.

Haasl et al. (2012) use a **SVM using a RBF kernel** for prediction of eight unadmixed European regional classes, such as British isles (Ireland, Great Britain, Scotland) and Central (Austria, Germany, Netherlands). Firstly, they apply SmartPCA using five iterations of outlier removal and then PCA to the genotype data matrix G of the eight ancestral population classes. In particular, they compute matrix V spanned by the first two leading eigenvectors of the covariance eigendecomposition of G. Then, to compute 2D feature representations of the test individuals, their genotype is projected onto matrix V to compute the principal component scores. Given that projected test individuals are often biased towards zero compared to the ancestral individuals used to compute V, I reduce this projection bias by post-multiplying test individual principal component scores by the reciprocal of a shrinkage factor computed in Lee et al. (2010b). Although I could compute higher order PCs beyond the second principal component score but Haasl et al. (2012) noted that this leads to little increase in prediction accuracy. But this is expected given that the first two PCs reflect geographical distribution of European populations

(Heath et al., 2008; Novembre et al., 2008).

5.1.7.3 String kernel

In the last decade there has been an increasing interest in the design of kernels for discrete symbol sequences over an alphabet, such as ternary SNP strings $\{0, 1, 2\}$ (Shawe-Taylor and Cristianini, 2004). Do et al. (2012) developed a window-based **SVM approach using string kernels for multi-class classification** for admixed canine ancestry. Firstly, they segment the genotype in windows of size 100 and learn a SVM classifier with multiple classes for each window where a class corresponds to one of the ancestral populations. Then, for prediction of admixed individuals the algorithm runs along the chromosome and records the most likely ancestry for each window. Finally, the ancestry contributions are given by the proportion of windows each ancestral population was observed.

However, Do et al. (2012) do not specify which string kernel is used. There are a wide variety of options:

- **Pairwise comparison kernel** (Liao and Noble, 2003): compute a pairwise similarity score between query sequence and reference sequence using similarity metrics, such as edit distance, L_1 norm, BLAST local alignment via dynamic programming using the Smith-Waterman algorithm (Smith and Waterman, 1981; Vert et al., 2004).

- **Composition kernel** (Ding and Dubchak, 2001): count of properties, such as numbers of 0s, 1s in sequence, structural and physio-chemical properties, etc.

- **Spectrum kernel** (Leslie et al., 2002): count common subsequences of length k in both sequences known as $k-$mers

- **Mismatch kernel** (Leslie et al., 2004): same as spectrum kernel but allows for l mismatches in $k-$mers which may be due to mutations

- **Marginalized kernels** (Tsuda et al., 2002): all previous kernels do not account for contextual information, e.g. symbols in DNA sequences have different meanings depending on whether they are part of the coding or non-coding region. Therefore, a HMM kernel is constructed where the observed symbol sequence is associated with a latent variable related to its context. Tsuda et al. (2002) also present an extension using a second-order HMM where the number of two adjacent symbols is counted to give limited correlation information.

5.2 ChromoPainter data analysis

In this section I will discuss which algorithm I select to compare DBAncestry, i.e. ChromoPainter, based on a set of criteria outlined below. After that I will discuss the ChromoPainter data preparation and analysis results. In particular, I will discuss how I compute the haplotype representation from the genotypes. After that I explain how I obtain our estimates for the recombination and mutation rate, as well as for the copying matrix. Finally, I will discuss how I use either use a HMM and regress on a breed-aggregated version of the copying matrix to compute breed fraction estimates.

5.2.1 Algorithm selection for a comparison with DBAncestry

Beyond presenting an integrated overview of ancestry inference techniques in Section 5.1 I aim to select an approach which represents the current state-of-art for a comparison with the DBAncestry technique. In particular, the comparison technique is sought to satisfy the following characteristics:

5.2.1.1 Criteria to select algorithm

(I) **Number of markers**: the algorithm can deal with small marker panels outlined in Section 1.5.1 representative for my data. Furthermore, the algorithm should scale well for a larger set of markers which may be available for future work.

(II) **Number of populations**: many of the ancestry inference techniques have been developed for the admixture of two source populations. However, due to the large number of breeds according to Section 1.5.3 a technique is required which performs inference for datasets composed of K populations.

(III) **Scalability for number of populations**: although some of the inference techniques can be used for large number of populations a few of them exhibit inapplicable scaling behaviour, i.e. the algorithm may scale exponentially in the number of populations.

(IV) **Correlation structure**: canine markers show strong LD according to Section 1.4.3 which favours model that either captures correlation between two markers or preferably for whole haplotype segments.

(V) **Experienced research collaborator**: to optimally use an ancestry inference implementation a collaboration with either its developer or an experienced user thereof is beneficial.

5.2.1.2 Comparing ancestry groups against criteria

Now, I check these criteria against the most sophisticated approaches from the different ancestry inference groups outlined in Section 5.1 to select a suitable algorithm:

- **Regression**: one of the most involved regression techniques is MMLR with inverse covariance information which satisfies criteria I, II, III and partially IV. However, modeling of correlation information (IV) is limited to pairwise SNPs.

- **Global model-based clustering**: Admixture is a fast version of Structure which satisfies I, II and III. However, it does not account for LD.

- **Local window-based approaches**: Lamp/WinPop satisfies I, II and III. However, LD is not formally modelled. Furthermore, admixture proportions have to be supplied to the algorithm and are not inferred. However, assuming these global admixture proportions the algorithm finds the best population pair assignment for the current individual.

- **Local HMM-based models approaches**: Multimix satisfies I, II, III and IV. However, for criterion IV the algorithm only models covariance information between two SNPs but does not consider long-range LD. ChromoPainter satisfies I, II, III, IV and V (Garrett Hellenthal). In particular with respect to IV, ChromoPainter looks at haplotype segments instead of correlation between pairwise SNPs and represents test dogs as linear combination of haplotype segments in the training data. Furthermore, both MultiMix and ChromoPainter have implementations available which are free for academic use.

- **Non-Parametric Bayesian approaches**: the iHMM satisfies I, II and III and IV. In particular, LD is accounted for modeling ancestry as a composite process of two layers: within the first layer individuals inherit from ancestral populations which again depend on a subset of founder haplotypes. However, at the moment there is no publically released code available for the iHMM technique.

- **PCA-based approaches**: the iPCA technique satisfied I, II and III. However, the technique is less aimed at ancestry inference but rather targeted at unsupervised learning, i.e. if the interest is in finding the optimal number of populations or to detect population substructure. Furthermore, PCA decorrelates SNPs in eigenspace to maximize retained variance rather than to account for correlation structure. On the other hand, the hybrid technique PCAdmix also satisfies criteria I, II and III but also does not account for correlation in the genotype data.

- **Machine Learning based approaches**: SVMs with string kernels satisfies I and II. Criterion III is computationally expensive and can be either implemented as each class '1-against-1' or '1-against-the rest'. Furthermore criterion IV is not explicitly accounted but rather two haplotype segments are compared using a similarity measure, such as an alignment score.

Based on this comparison the strongest competitors are MultiMix, ChromoPainter and iHMM. However, Multimix does not model long-range dependencies while iHMM does not offer a publically available implementation. I selected ChromoPainter because it models LD by considering haplotype segments and is freely available available for academic use. Furthermore, researcher Garrett Hellenthal collaborated with me to smoothly obtain results. Finally, I will conclude this part by comparing the DBAncestry algorithm with ChromoPainter.

5.2.1.3 ChromoPainter vs. DBAncestry assumptions

- **Time complexity for markers and training individuals**: both approaches, ChromoPainter (prediction run-time) and the computation of population frequency estimates of PHASE for DBAncestry scale linearly in the number of SNP markers (Hellenthal, 2012; Scheet, 2013). However, ChromoPainter is also linear in the number of purebred training individuals while PHASE is quadratic in the number of samples in the reference database (Li and Stephens, 2003; Scheet, 2013).

- **Constant ancestry on chromosome**: DBAncestry assumes constant ancestry within a chromosome of two breeds assigned on the maternal and paternal side accordingly. Recombination can only occur between adjacent chromosomes. In the case of a small number of markers on each chromosome this assumption may be appropriate. However, for denser sets of markers I would like to unravel the different breeds which contribute to a given chromosome. In particular, ChromoPainter chooses the length of haplotype segments which form constant ancestry adaptively, i.e. according to the recombination rate and whether a test dog still copies from the same training individual.

- **Marker correlation**: in the age of GWAS data, 320 SNP markers is very sparse and far from a typical dense GWAS dataset. Furthermore, SNPs are not evenly spaced and distance among adjacent SNP may range from a few hundred to million of base pairs as indicated in Figure 1.1. Therefore, it is hard to exploit correlation pattern of markers close in distance. According to Section 1.4.3 LD information given by a recombination map can be utilized to some extent.

- **Amount of information extracted from training data**: to discriminate breeds at the ggp level on average the algorithm uses $\frac{320}{8} = 40$ SNPs to make this prediction. DBAncestry uses more information from the genotype because for each chromosome I compute a large number of haplotype frequencies typical for a given breed while for ChromoPainter I only use the most likely phasing, i.e. haplotype pair, for each training individual.

- **Disk storage and working memory requirements**: the storage requirements do not change much for ChromoPainter because each additional marker only adds another column in the file of the most likely haplotype pair phasings of the training data. With respect to DBAncestry, more markers lead to a combinatorial increase of possible haplotypes which all need to be enumerated, stored and queried with their haplotype frequencies from file and loaded to memory. Although due to strong LD in many dense SNP datasets the number of likely haplotypes is much smaller than all possible enumerations 2^S where S is the number of markers (Gattepaille and Jakobsson, 2012). As outlined in Section 1.4.3 in windows of 10-500 kbp there around 5 haplotypes covering 80% of the observed haplotypes (Lindblad-Toh et al., 2005; Parker, 2012). However, to yield a good coverage of rare haplotypes I expect to retrieve a substantial number of haplotype frequencies leading to very heavy use of disk read operations and large amount of working memory. Alternatively, I could store fewer haplotype frequencies but haplotypes may be more often missed in the test dog which would lead to the demand of a measure of deviation to account for imperfect copying due to mutations.

5.2.2 Computation of haplotype representation

Phasing of the training data BigPure was discussed in Section 3.2.2.1. For the synthetic test dataset I looked at three options:

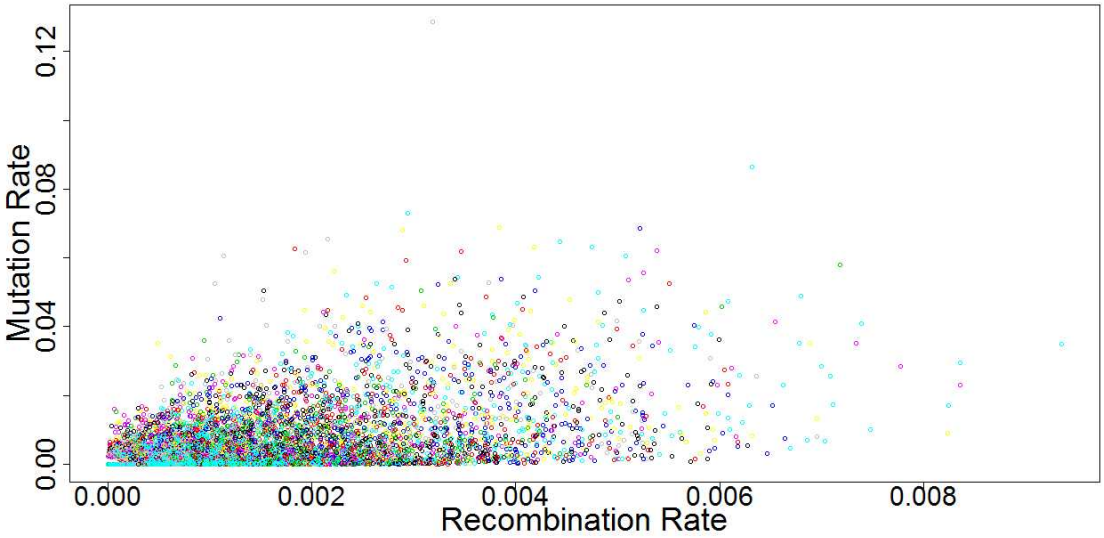
- Utilizing **PHASE**: (Stephens et al., 2001),
- Utilizing **BEAGLE**: (Browning, 2006; Browning and Weir, 2010),
- **Enumeration**: according to Section 3.2.2.2 I enumerate all possible haplotypes for each chromosome for a given test dog. Then, we run ChromoPainter separately on each chromosome. Then, on the first chromosome ChromoPainter returns the expected log-likelihoods for the haplotypes from each enumerated haplotype pair. Then, on the first chromosome I find that enumerated haplotype pair which has the highest value for the sum of the two expected log-likelihood estimates, and set this haplotype pair as the best enumerated haplotype pair. I repeat this same process for all chromosomes. Finally, to find the best phasing of the whole genotype for the test dog I concatenate the best haplotype presentations across all chromosomes.

On a small subset of the test dataset I compared all three options and noticed there is little impact on the predicted breed fractions and classification results. For some chromosomes PHASE can be very slow to compute the phasing and takes several hours for 10-20 SNPs. Similarly, for the enumeration approach on some chromosomes there are thousands of enumerated consistent haplotype pairs which makes the approach very computationally intensive when ChromoPainter is run for so many times. BEAGLE, on the other hand, computes the phasing typically in a few seconds. Given that neither of the methods appears to be more accurate, I decided to use BEAGLE to phase test dog genotypes.

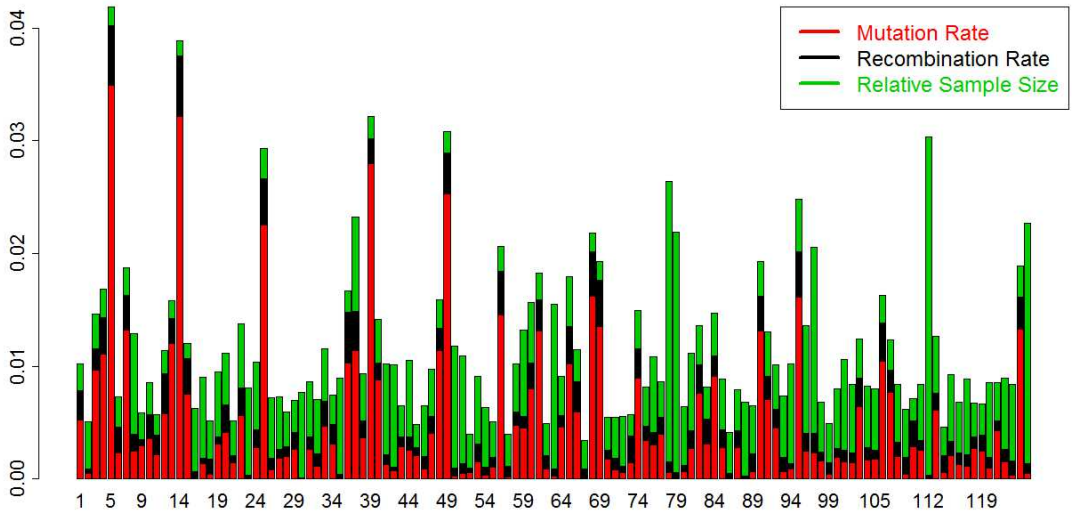
5.2.3 Recombination and mutation rate

I used PHASE to estimate a recombination map (Li and Stephens, 2003; Crawford et al., 2004) as input for ChromoPainter to obtain more accurate breed fraction estimates. For each pair of training breed and chromosome I obtain a file which has as many columns as there are SNPs on the chromosome. The first column corresponds to the background recombination rate while subsequent columns match the recombination rate between adjacent SNPs. The rows in this file are given by samples from the posterior distribution for the recombination parameters. To yield point estimates of these estimates incorporating the background rate, I take the median of each column, and then I define the new SNP-wise recombination rates ρ with units in base pairs as the median background recombination rate $\bar{\rho}$

multiplied the SNP-wise median recombination rate. Based on this recombination map from PHASE I use ChromoPainter to estimate the recombination scaling constant $N_e = 0.0011$ for the recombination map as well as a global mutation rate θ estimated at 0.0018 for the emission probabilities.



(a) Mutation rate against recombination scaling constant. Each dot represents one of the dogs in the training set. Dots are colour-coded by breed although due to the large number of breeds some breeds share the same colour encoding.



(b) As a function of the breed I plot time series for mutation rate, recombination rate and the proportion of training samples with respect to the maximum number in any of the breeds.

Figure 5.2: These figures show the distribution of the ChromoPainter estimates for recombination scaling and mutation rate either separately for each pure breed individual or aggregated by breed.

In Figure 5.2 (a) I plot the recombination scaling constant against the mutation rate for all purebred training dogs. There are five training dogs (3 American Eskimo dogs with rates (0.008, 0.008, 0.009), 1 Beagle US with rate 0.008, 1 Pomeranian with rate 0.008) which have recombination rate greater than 0.008. The median value for the rescaling constant is 0.0011. With respect to the mutation rate there are

two dogs with a rate higher than 0.08 (1 American Eskimo dog with rate 0.08, 1 Weimaraner with rate 0.12). The median mutation rate for this training dataset is 0.0018. Finally, after having examined the recombination and mutation rate I compute the copying matrix and final values for the rescaling constant and mutation rate.

I also aggegrated the estimates for the median recombination and mutation rates by breed. To visualize the impact of the number of purebred samples per training breed I also plot a scaled version of the ratio of number of samples per breed divided by the maximum number of samples in any breed (i.e. 391 samples is maximum in our dataset according to Table A.1). These three sequences for the recombination scaling constant, mutation rate and scaled sample size ratio are shown with respect to their breed ID on the abscissa Figure 5.2 (b). I notice that recombination and mutation rate are very correlated at 0.83 while each recombination rate or mutation rate are slightly negatively correlated at -0.38 and -0.33 with the scaled sample number ratio, respectively.

5.2.4 Computation of breed fraction estimates

Finally, I look at two different ways to compute the copying matrix outlined in Section 5.1.4.8.

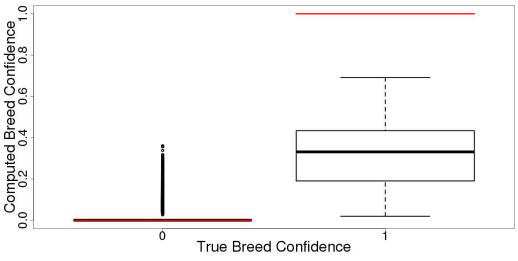
- **Standard**: the **standard** breed fraction estimates are obtained from ChromoPainter upon supplying and running it with the set of enumerated training haplotype pairs, test dog haplotype pairs, a recombination map, estimates for the recombination scaling constant and mutation rate, as well as information on which training samples form a population. Then, ChromoPainter computes a chunk length row vector which has as values the expected total genetic length of DNA the test dog copies from each training dog population across all SNPs under the ChromoPainter model (Hellenthal, 2012). Then, to yield breed fraction estimates this row vector is normalized to total sum of length 1.

- **NNLS**: breed fraction estimates in the standard approach will be biased towards those breeds which have more training samples because a priori I assume each test dog has the same chance to copy from each training dog. To deal with this bias I review a regression approach referred to as **NNLS**. Firstly, I aggregate the square copying of size number total training samples by breed, i.e. I sum column-wise and then average it row-wise. Then, this aggregated matrix is standardized by length 1, i.e. each row sums to 1. Then, I take the ChromoPainter output from a test dog given by the standard chunk length row vector and standardize it. This standardized recipient vector is regressed on the standardized aggregated matrix. However, standard regression might be insufficient because it leads to negative breed fraction estimates which lack interpretation (Kuehn et al., 2011). Therefore, I perform the regression using non-negative least squares (NNLS) (Lawson and Hanson, 1974) which seems to be sparse in our dataset leading to few non-zero breed fraction estimates.

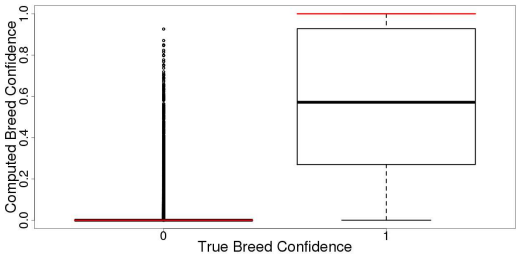
5.2.5 ChromoPainter results

As alternative to the DBAncestry results presented in Section 4.7.2 I also will apply ChromoPainter to dataset OrgSyntheticRed discussed in Section 1.5.4 using either the standard or NNLS variant according to Section 5.2.4.

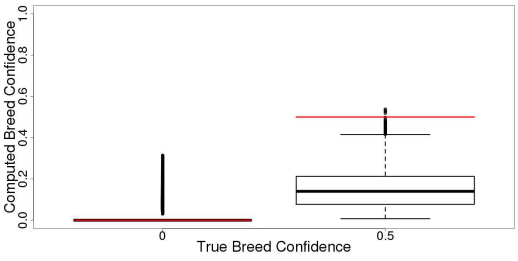
Results for 4 specific lineage trees, i.e. the two simplest and most complex two lineage trees are shown in Figure 5.3 while global breed contribution estimates which were averaged across all lineage trees are shown for all TAP levels in Figures 5.4 (NNLS) and 5.5 (Standard). A comparison between



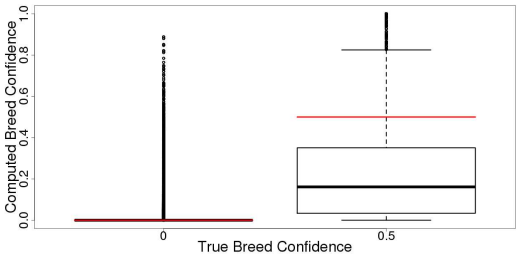
(a) Tree AAAAAAAA, Standard



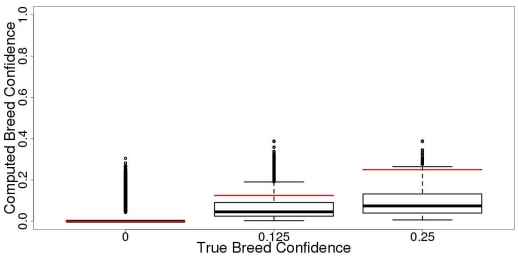
(b) Tree AAAAAAAA, NNLS



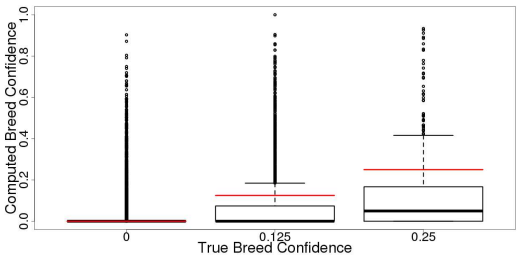
(c) Tree AAAABBBB, Standard



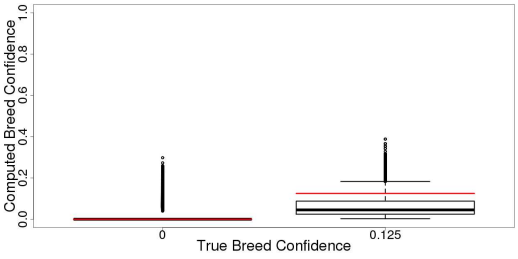
(d) Tree AAAABBBB, NNLS



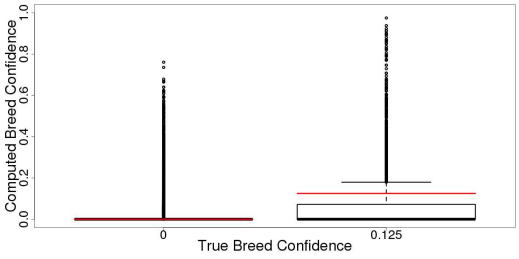
(e) Tree AABCDEFG, Standard



(f) Tree AABCDEFG, NNLS



(g) Tree ABCDEFGH, Standard



(h) Tree ABCDEFGH, NNLS

Figure 5.3: **CCBC breed proportion estimates for the ChromoPainter standard and NNLS variant:** These figures show true ancestral proportion as bold red line while the boxplots offers information on the breed proportion estimates, i.e. the median, lower and upper quartiles, whiskers which extend to the most extreme data point less than 1.5 IQR from the box, and outliers which are drawn as small circles. These plots represent a subset of the 11 lineage trees: In particular the top two rows correspond to the simplest trees while the bottom two rows match the most complex lineage trees. Furthermore, the left column shows the standard variant wile the right column corresponds to the NNLS variant. These figures show that NNLS has a much larger IQR than the standard variant.

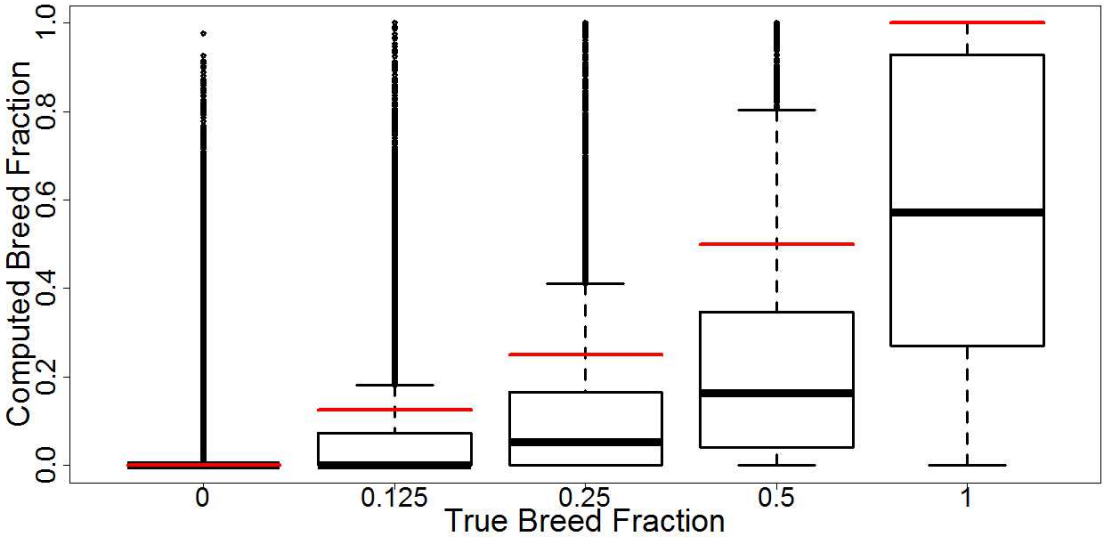


Figure 5.4: This figure shows breed proportion proportion estimates averaged across lineage trees. More in detail, CCBC breed proportion estimates for the **NNLS** variant are visualized as boxplot showing median, lower and upper quartiles of all ancestral levels. Numerical values for the median and quartiles are also given in Table 5.2.

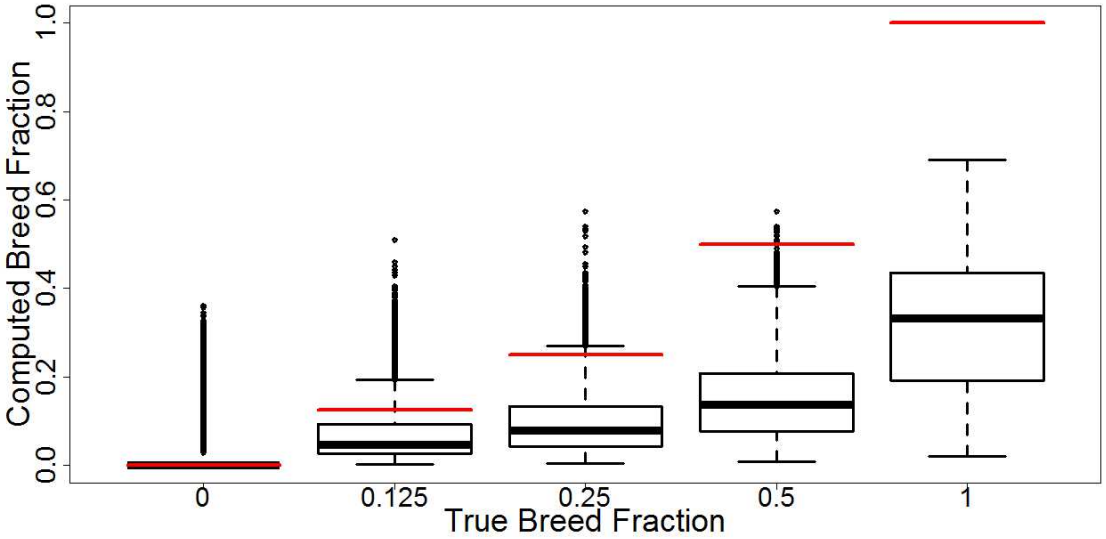


Figure 5.5: This figure shows breed proportion proportion estimates averaged across lineage trees. More in detail, CCBC breed proportion estimates for the **standard** variant are visualized as boxplot showing median, lower and upper quartiles of all ancestral levels. Numerical values for the median and quartiles are also given in Table 5.2.

True breed ancestry Median [1st,3rd] Quartiles	CCBC Standard	CCBC NNLS
100%	**0.33** [0.19,0.43]	**0.57** [0.27,0.93]
50%	**0.14** [0.08,0.21]	**0.16** [0.04,0.35]
25%	**0.08** [0.04,0.13]	**0.05** [0.00,0.16]
12.5%	**0.05** [0.02,0.09]	**0.00** [0.00,0.07]
0%	**0.00** [0.00,0.00]	**0.00** [0.00,0.00]

Table 5.2: **CCBC breed proportion estimates for the ChromoPainter standard and NNLS variant:** The median, lower and upper quartile estimates of the re-scaled (CCBC) breed proportion for all possible true ancestral levels are shown. Breed fraction estimates are computed taking predictions from all lineage trees into account. Further visualization and discussion of these estimates is given in Figure 5.3.

global and lineage specific estimates their deviation is very minor, i.e. less than 0.01 for any tree. A tabular description of the median, lower and upper quartiles for the global estimates is shown in Table 5.2.

Firstly, it can be seen that the NNLS approach has a much larger interquartile range for the predicted breed proportions. In particular, at TAP level 100 percent the IQR is 0.66 for NNLS vs. 0.24 for the standard variant, at the parental level there is an IQR of 0.31 for NNLS vs. 0.13 for the standard variant, at the gp level the IQR for NNLS is 0.16 versus 0.09 for the standard variant while both variants have an identical IQR of 0.07 at the ggp level. According to Table 4.2 the reader sees that DBAncestry CCBC ULP has very similar IQR ranges to the ChromoPainter standard variant although at the ggp level ChromoPainter has an even half as large IQR of 0.07 vs. 0.14 for DBAncestry.

Furthermore, NNLS estimates have much less underestimation at the purebred level with a median breed proportion of 0.57 vs. 0.33 for the standard variant. According to Table 4.2 DBAncestry yields a much higher median breed proportion estimate of 0.87. At the parental level ChromoPainter yields similar median proportion estimates of 0.16 and 0.14 for NNLS and the standard variant, respectively. Therefore, these parental proportion estimates are substantially underestimated and are just about one third of the median estimate 0.41 for DBAncestry. At the TAP level of 25 percent Chromopainter estimates median breed proportions of 0.05 and 0.08 for NNLS and the standard variant, respectively, which is only a quarter to two fifth of the DBAncestry median estimate of 0.2. Furthermore, the NNLS approach does not detect breeds at the ggp level which is reflected by its median breed proportion estimate 0 while the standard approach strongly underestimates these proportions with a median of 0.05.

For DBAncestry ggp breed proportions are very accurately predicted with a median breed proportion of 0.12. To sum up, ChromoPainter works reasonably well to detect pure breed proportions. Furthermore, the NNLS approach has a minor advantage over the standard variant at the parental level but has worse performance at the grandparental and ggp level. Finally, for all TAP levels DBAncestry performs much better than either ChromoPainter variant. Similar conclusion are obtained examining the classification view outlined in Section A.8 which also shows that ChromoPainter only achieves a reasonable classification performance for pure breed synthetic test dogs.

5.3 Summary of ancestry inference techniques and conclusions

In the first part of this chapter I provided a comprehensive review of different techniques for estimation of local and global ancestry proportions. This overview extends previous partial reviews due to Churchhouse (2012); Churchhouse and Marchini (2013); Liu et al. (2013). In this review I looked at different techniques, such as regression, clustering and window-based, non-Parametric Bayesian approaches,

PCA-based approaches and machine learning approaches with a focus on kernel-based methods.

I also discussed the evolution of local HMM-based model approaches: initially researchers applied first order HMMs before they moved to second order HMMs to account for LD between adjacent SNP markers. Due to the insight that none of the fixed order HMM can model long range dependencies researchers focused on modelling entire haplotypes within a hierarchical HMM with layers either corresponding to populations or haplotypes. Then, due to computational constraints HMMs were developed whose number of hidden states is much less than the number of ancestral individuals. Alternatively, to avoid oversimplification by using fixed order HMMs and overspecification by modelling whole haplotypes, researchers studied haplotype clusters as hidden states which are adaptive according the amount of locally observed LD in each ancestral population. All of these HMM variants described before are either not applicable for many populations or have time complexity which makes inference in multiple population setup infeasible. This limitation of multiple populations was successfully implemented by models, such as MULTIMIX and ChromoPainter. MULTIMIX segments chromosomes into windows of constant ancestry and estimates SNP frequencies and covariance over those SNPs in the window assuming normality. Then, for an unknown admixed individual the most likely hidden ancestry states are inferred using a first order HMM. On the other hand, ChromoPainter represents test dog individuals as linear combination of haplotype segments in the training data.

After this review of ancestry inference techniques in Section 5.2.1 I defined a set of characteristics to select one of these approaches as comparison technique with DBAncestry. In particular, the required characteristics include whether the algorithm is suitable for sparse/dense marker sets, how many populations it can process, how it scales with the number of populations, whether it accounts for correlation among SNPs, whether there is a publicly available implementation and feedback opportunities of expert users.

Based on these characteristics I chose the ChromoPainter technique which I compared with DBAncestry in Section 5.2.1 based on factors, such as time complexity as a function of the number of SNPs and training samples, whether ancestry changes are considered within the chromosome, whether it deals with LD, how much information is extracted from the training samples as well as disk storage and working memory requirements.

After that in Section 5.2.2 I discussed how I obtain the haplotye representation for the test dogs, how I estimate recombination and mutation rates in Section 5.2.3 which are used as input to ChromoPainter, how I compute breed proportion estimates using either the standard or NNLS variant according to Section 5.2.4. Finally, I described the results in Section 5.2.5 and compared them with those ones obtained from DBAncestry.

Results showed that ChromoPainter predicts pure breed proportions reasonably well with median estimates of 0.33 and 0.57 for the standard and NNLs variant, respectively. However, performance sharply declines for more complex lineages involving multiple breeds. In particular, parental breeds only have median breed proportions estimates of 0.14 to 0.16 which is around 0.35 less than the truth. Similary, at the gp level median estimates are 0.05 to 0.08 which is much less than the truth of 25 percent. Finally, at the ggp level the NNLs variant almost does not detect breeds at this ancestral level while the standard variant underestimates these breed proportions at 0.05.

A comparison with the DBAncestry technique showed DBAncestry still yields a high AUC score for highly complex lineage trees (i.e. lineages with several gps and ggps) which is only achieved for pure breed trees using ChromoPainter. This difference in performance is most likely due to the fact that ChromoPainter extracts less information from the training dataset than the MCMC approach.

Chapter 6

Conclusions and future work

This chapter briefly summarizes the thesis contributions along with pointers to future work.

6.1 Conclusions

This research on canine ancestry inference was initiated by Mars Veterinary and Prof. David Balding. The underlying motivation for this work was to improve upon currently available ancestry inference technique for the interesting case where canine samples are represented by short genetic sequences from a large number of breeds. In particular, we were interested in the estimation of breed composition of synthetic test dogs at the level of recent ancestry going up to three generations back which reflects great-grandparent kinship.

As first step in this thesis we provided some background about commercial ancestry testing and dog research. In particular, we discussed direct-to-consumer (DTC) tools for ancestry inference and why private users, academic researchers, medical professionals and employees of government bodies, housing associations and insurance agencies are interested in accurate global canine breed composition estimation. There are a variety of data sources which are predictive of ancestry and we review some of them including data based on phenotypes, language, stable isotopes and genetic data. Genetic data can be further subdivided into SNP and microsatellite markers which both contain autosomal components as well as lineage-based parts for the maternal and paternal lineage. However, these ancestry inference tools also have certain limitations, i.e. their results depend on a definition of ancestry which is interpreted differently depending on audience and purpose, the number of markers and their breed discriminating characteristics, the utilized data sources and how complete ancestral populations where sampled with respect to their database use. Furthermore, there are limitations of those DTC tools based on the accuracy and interpretation of genetic testing of disease and other disorders. Ancestry inference can also lead to emotional distress of dog owners who expected a different breed composition or may lead to legal disputes with respect supposedly purebred dogs. Finally, many of these DTC providers enforce proprietary confidentiality rules to protect their business which may lead to the unexpected fact that different, even contradictory, inference results are obtained from different DTC companies for the same test sample. In the next section we explained how dog breeds diverged from wolves based on two bottlenecks related to domestication and Victorian breed formation, respectively. Analysis different types of genetic data shows that these breed creations led to more between-breed variability and a decline of genetic diversity within breeds which can be utilized for ancestry predictions but has consequences on whether certain disorders are inherited.

Firstly, I explored the genetic proximity among breeds for the datasets used in this thesis. Therefore, we look at several multivariate distance and similarity measures which can be computed using either the

original genotype data or the frequencies of the phased haplotypes. We also discussed one similarity measure based on the ChromoPainter software which take the marker correlation into account although we could not exploit this strength much due to limited LD in our sparse SNP marker profile composed of short genetic sequences. Then, we visualized these derived proximity matrices using heatmaps, dendrograms as well as 2D reconstructions using multi-dimensional scaling which retain original, high-dimensional distances approximately optimal. We noticed that some breeds can be identified as very distinct while some breeds have individuals which overlap with members of other breeds.

After that I continued with a review of Mars Wisdom' proprietary canine ancestry inference implementation which I used as foundation for the development of our own algorithm DBAncestry for breed composition inference. At first our work was concerned with the development of an inference algorithm for the special case of purebreed synthetic dogs. To characterize different breeds in our algorithm we enumerated typical haplotypes with their frequencies for each pair of breed and chromosome. These frequencies were computed using the statistical genetics software PHASE. More specifically, we either phased the purebred training dataset by breed which is adversely affected by small samples sizes in some of the breeds, or we phased all training samples together across breed which confounds population structure. Then, based on the test dog genotype for a chromosome we enumerated consistent haplotype pairs and their corresponding frequencies to compute the most likely assignment of a breed pair to this very chromosome. However, there are cases where the frequencies of those consistent haplotypes are zero, and we experimented with three inference options to deal with this case. Although a few breeds, especially subpopulations were confused, results were very encouraging.

In the next chapter of this thesis we extended our novel algorithm DBAncestry from the special case of purebred dogs to mixed breed dogs of varying lineage complexities. In this crossbred case each of 25 chromosomes has breed pairs assigned from a database composed of 125 breeds. Therefore, assigning breed pairs across the genome leads to a prohibitively large number of possible breed compositions. Therefore, we sample the space of possible breed configurations using a Metropolis-Hastings algorithm. As proposal update in this MCMC algorithm we either uniformly sample new breeds or we bias the Metropolis-Hastings algorithm, such that breed proposals is likely to be similar to the current breed assignment. As breed-biased update proposal we use information based on a breed rank Manhattan distance matrix using the original genotype data. We also experimented with the run-time of the MCMC algorithm using either a short run of 700K or a long run of 7 million iterations in the main phase. As performance measures we used an adaptation of classical classification literature, i.e. an extension of binary ROC curves to multiple class, and we calculated how predicted breed proportion estimates deviate from the genuine ones. The results we obtained are very good although the breed contribution of genuine breeds in the ancestry are underestimated due to exploration in the MCMC. More complex ancestries as well as the short MCMC run lead to a slight drop in performance. Furthermore, we expected that breed-biased update rules either improves mixing time or prediction accuracy over a uniform update rule but results do not show any evidence for this hypothesis. Then, I continued to show how lineage tree is derived from the estimated breed contributions via the proxy of true ancestral proportions. Finally, I showed that the DBAncestry algorithm is robust against recombination events typical for canine datasets and only shows a very small decrease in prediction quality.

In the final chapter of this thesis we were seeking for a sophisticated algorithm for the inference of complex ancestries which should be utilized as comparison technique for DBAncestry to predict predict complex canine breed compositions. However, there is a lack of a comprehensive review of ancestry

inference techniques in the literature and how they are related to each other. Therefore, we reviewed a large number of techniques in the literature at a conceptual level and grouped them into different statistical modelling approaches, such as regression, hidden markov models and principal component analysis. Furthermore, if possible, we outlined the evolutionary development of those techniques within a particular group, i.e. we discussed limitations of earlier approaches which led to subsequent work. Finally, we selected a recent a advanced technique, ChromoPainter, for our data analysis comparison. ChromoPainter is based on a hidden markov model which represents the test dog genotype as linear combination of haplotype segments from training dogs in the purebred, ancestral reference dataset. We noticed that the breed estimates of the genuine ancestry are more underestimated than those ones from DBAncestry which is most likely due to lesser use of information from the purebred training data. Especially, DBAncestry uses hundreds of haplotype frequencies for each combination of breed and chromosome while ChromoPainter only uses the most likely phased haplotype pair for each purebred training dog. Furthermore, ChromoPainter was not developed for short genetic sequences but for dense whole-genome datasets which show strong marker correlation.

6.2 Future work

In future research we would like to explore a few different directions.

- **More samples**: as we see in Table A.1 there are a few breeds which have fewer than 40-50 purebred ancestral training samples. However, to minimize biases related to incomplete sampling described in Section 1.3.4 we propose to obtain more samples for those breeds, especially breeds which have subpopulations. For example, the Beagle UK, Beagle US Field and Beagle US Show only have 21, 18 and 18 purebred samples, respectively.

- **More markers**: firstly, we would to incorporate more markers to possible improve performance. Based on previous results for human datasets using dense marker sets (Lawson et al., 2012) ChromoPainter is expected to show a steep performance improvement because it can exploit marker correlation patterns. Furthermore, more markers are required to model complex ancestries which show evidence of recombinations within a chromosome. In particular, in the within-chromosome recombination case the DBAncestry algorithm does not perform well as discussed in Section 4.7.2.2.

  A large number of markers may pose new challenges for DBAncestry. Absent from the small number of common haplotypes (Sutter and Ostrander, 2004; Sutter et al., 2004; Lindblad-Toh et al., 2005) I first need to investigate the number of rare haplotypes. If there is a large number of rare haplotypes I either need to store large PHASE output files with their corresponding haplotype frequencies, or I may have to investigate some form of approximate matching between enumerated haplotypes and those haplotypes listed in the PHASE output files to avoid many zero frequencies. Furthermore, long genotypes on a chromosome requires us to rethink the DBAncestry assumptions of only two breeds on each chromosome and no modelling of recombination, e.g. I could investigate chunking the genotype in windows whose length is adaptive towards the amount of locally encountered LD.

- **Other animals**: I would like to apply DBAncestry to other animal species. Recently, genetic testing already has been extended to cats (Leroy et al., 2013; Lyons' Veterinary Genetics Laboratory, University of California, Davis, 2013).

- **Common genetic proximity measures**: In the chapter on proximity measures I was mostly motivated by methods from the multivariate and information theory literature. However, in the genetics literature people prefer to use slightly different measures or terminology, such as absolute difference in allele frequency δ or the fixation index F_{st} (Rosenberg et al., 2003). According to Section 2.5.2, the δ measure corresponds to the Manhattan distance which either naively can be evaluated by averaging the pairwise computations between any two individuals, one individual from each population, or more efficiently by taking the difference of allele-wise frequencies in both populations. The fixation index also measures population differentiation, and a simple estimator is based on the ratio of between-population and within-population variability. The fixation index has three main differences compared to the approach in Section 2.5.2: firstly, a popular difference measure is based on nucleotide diversity which is related to Hamming distance. Secondly, it accounts for within-population diversity whose idea is related to the Mahalanobis distance (Deza and Deza, 2009), i.e. in case there is only modest differentiation among populations they still can be discriminated if there is little variability within those populations. Thirdly, between-population variability is either defined as pairwise differences between any two individuals, one individual from each population, or as pairwise differences between members of one population versus members of the remaining populations pooled.

  Assumed I obtain a dense marker profile I propose a ChromoPainter fixation index, i.e. I use the fixation index where I use as proximity measure how much each ancestral populations copies from each other. Therefore, this ChromoPainter fixation index also accounts for marker correlation.

- **Alternative sampling schemes**: according to the DBAncestry results in Section 4.7 I noticed that there is no evidence that a breed-biased proposal leads to better breed composition estimation performance. This behavior might be due to the fact that I run the sampling scheme long enough, such that even a uniform breed proposal mechanism leads to satisfactory exploration of the breed space. An alternative explanation might be that due to the sparse SNP marker profile breed similarity may not yield a detailed enough representation. Finally, within DBAncestry I utilize the breed similarity in an ad-hoc way: therefore, I might investigate approaches which uses the breed structure in a sound way, or even learn breed similarity in an adaptive way within the burn-in phase of the MCMC algorithm (cp. Section 4.4).

- **Shared haplotypes across populations in phasing**: when I started this research I was not aware of any approach which accounts for shared haplotypes across multiple populations. Therefore, in Section 3.2.2.1 I developed the PPOOLED option which phases all training dogs across breeds together and computes breed-wise frequencies from the joint haplotype frequencies using the list of most likely haplotype pair phasings from each jointly phased training individual. Due to comparable performance and less computational cost I continued to use the naive approach PSEPARATE which segments the training data by breed and infers estimates haplotype frequencies separately for all breeds. However, certain haplotypes might be rare in most breeds but more common across populations which may lead to biased estimates thereof. For example, Pritchard (2001) suggests that population bottlenecks leads to population-specific genetic diversity, such that certain haplotypes are shared but are present at different frequencies in the different subpopulations.

  To approach this problem Sohn and Xing (2009); Sohn (2011) developed a non-parametric Bayesian clustering algorithm which incorporates several layers. Firstly, at the first level the different populations draw a subset of founders from a common pool of founder haplotypes. Therefore,

the same haplotype from the common pool can be present in multiple populations although at different frequencies in each population. Then, at the second level the individual haplotype pairs are drawn from different populations in a process which allows for mutations of the population-specific haplotypes.

Appendix A

Appendix

A.1 Overview of breeds and analysis pure breed dogs

In table A.1 I show the used breeds for the small and big dataset along with total numbers of dogs per breeds and the number training dogs.

Breed	# Train Small	# Train Big	# Total n_b	Breed	# Train Small	# Train Big	# Total n_b
Afghan Hound	28	26	31	Greater Swiss Mountain Dog		28	33
Airedale Terrier	51	46	54	Greyhound		18	21
Akita	37	34	40	Havanese		19	22
Alaskan Malamute	30	28	33	Irish Setter::UK		32	38
American Eskimo Dog	19	19	22	Irish Setter::US		41	48
American Staffordshire Terrier	32	30	35	Irish Wolfhound		48	57
Australian Cattle Dog	29	27	32	Italian Greyhound		20	24
Australian Shepherd::UKX	29	27	32	Italian Spinone		37	44
Australian Shepherd::USX	81	71	84	Japanese Chin		39	46
Basenji	28	26	31	Japanese Shiba Inu		74	87
Basset Griffon Vendeen (Petit)	34	31	37	Keeshond::UKX		13	16
Basset Hound::UK	21	20	24	Keeshond::USX		21	25
Basset Hound::US	24	23	27	Labrador Retriever::UK		57	67
Beagle::UK	18	18	21	Labrador Retriever::US Field		276	325
Beagle::US Field	15	15	18	Labrador Retriever::US Show		236	278
Beagle::US Show	15	15	18	Lhasa Apso::UKX		32	38
Bearded Collie	70	62	73	Lhasa Apso::USX		43	51
Belgian Sheepdog	91	80	94	Maltese::UKX		19	22
Belgian Tervuren	41	37	44	Maltese::USX		20	23
Bernese Mountain Dog	72	64	75	Mastiff		31	37
Bichon Frise	56	50	59	Miniature Pinscher::UKX		5	8
Bloodhound	37	34	40	Miniature Pinscher::USX		36	42
Border Collie	72	64	75	Newfoundland		49	58
Border Terrier	98	86	101	Norfolk Terrier		40	47
Borzoi	75	66	78	Norwegian Elkhound		40	47
Boston Terrier	32	30	35	Norwich Terrier		72	85
Bouvier Des Flanders	67	59	70	Old English Sheepdog		48	56
Boxer	57	51	60	Papillon		34	40
Briard	37	34	40	Parson Russell Terrier		43	51
Brussels Griffon	35	32	38	Pekingese		44	52
Bull Terrier	64	57	67	Pembroke Welsh Corgi::UKX		25	30
Bull Terrier (Miniature)	28	26	31	Pembroke Welsh Corgi::USX		35	41
Bulldog	61	54	64	Pointer		98	115
Bullmastiff	60	54	63	Pomeranian::UKX		25	30
Cairn Terrier		51	60	Pomeranian::USX		25	30
Cardigan Welsh Corgi		29	34	Poodle		68	80
Cavalier King Charles Spaniel		94	111	Poodle (Miniature)		48	57
Chesapeake Bay Retriever		21	25	Poodle (Miniature)::miscwithtoy		183	215
Chihuahua::UKX		60	71	Poodle (Toy)		38	45
Chihuahua::USX		32	38	Portuguese Water Dog		39	46
Chinese Crested		47	55	Pug		59	69
Chinese Shar-Pei		22	26	Rhodesian Ridgeback		89	105
Chow Chow		43	51	Rottweiler		68	80
Cocker Spaniel		89	105	Saint Bernard::UKX		7	10
Collie::UKX		49	58	Saint Bernard::USX		30	35
Collie::USX		51	60	Saluki		60	71
Curly Coated Retriever		31	36	Samoyed::UKX		36	42
Dachshund (Longhaired)		23	27	Samoyed::USX		25	29
Dachshund (Miniature Longhaired)		50	59	Schipperke::UKX		4	7
Dachshund (Miniature Shorthaired)		88	104	Schipperke::USX		21	25
Dachshund (Miniature Wirehaired)		47	55	Schnauzer (Giant)		30	35
Dachshund (Shorthaired)		31	37	Schnauzer (Miniature)		57	67
Dachshund (Wirehaired)		43	51	Scottish Terrier		48	56
Dalmatian::UK		28	33	Shetland Sheepdog::UK		22	26
Dalmatian::US		21	25	Shetland Sheepdog::US		55	65
Dobermann Pinscher		120	141	Shih Tzu		332	391
English Cocker Spaniel		86	101	Siberian Husky		56	66
English Cocker Spaniel::Field		20	23	Soft Coated Wheaten Terrier::UKX		10	13
English Setter		33	39	Soft Coated Wheaten Terrier::USX		17	20
English Springer Spaniel::UKX		47	55	Staffordshire Bull Terrier		65	77
English Springer Spaniel::USX		20	24	Tibetan Spaniel		49	58
Flat Coated Retriever		59	69	Tibetan Terrier		74	87
Fox Terrier (Smooth)		35	41	Vizsla::UK		34	40
Fox Terrier (Toy)		25	29	Vizsla::US		31	36
Fox Terrier (Wire)		32	38	Weimaraner::UKX		45	53
French Bulldog		48	56	Weimaraner::US2X		5	8
German Shepherd Dog		84	99	Weimaraner::USX		22	26
German Shorthaired Pointer::UKX		37	43	Welsh Terrier::UKX		24	28
German Shorthaired Pointer::USX		23	27	Welsh Terrier::USX		13	16
German Wirehaired Pointer		26	31	West Highland White Terrier		70	82
Golden Retriever::UK		32	38	Whippet::UKX		44	52
Golden Retriever::US		161	190	Whippet::USX		31	36
Gordon Setter		38	45	Yorkshire Terrier::UK		31	36
Great Dane		49	58	Yorkshire Terrier::US		235	277
Great Pyrenees		31	37				

Table A.1: **Breed overview**: for the small and big pure breed dataset I show the used breeds. The big dataset covers all breeds while the small dataset covers those 34 breeds ranging from Afghan Hound to Bullmastiff. The breed identifier is shown in the first column. The fourth column has the total number of dogs n_b per breed. The second column 'Small' shows the number of training dogs for the small dataset, and column 'Big' the number of training dogs for the big dataset.

breedM ID	breedM Name	breedM ID from breedP	breedM Name from breedP	breedP ID	breedP Name
1	Afghan Hound	1	Afghan Hound	1	Afghan Hound
2	Airedale Terrier	2	Airedale Terrier	2	Airedale Terrier
3	Akita	3	Akita	3	Akita
4	Alaskan Malamute	4	Alaskan Malamute	4	Alaskan Malamute
5	American Eskimo Dog	5	American Eskimo Dog	5	American Eskimo Dog
6	American Staffordshire Terrier	6	American Staffordshire Terrier	6	American Staffordshire Terrier
7	Australian Cattle Dog	7	Australian Cattle Dog	7	Australian Cattle Dog
8	Australian Shepherd	8—9	Australian Shepherd::UKX—Australian Shepherd::USX	8	Australian Shepherd::UKX
9	Basenji	10	Basenji	9	Australian Shepherd::USX
10	Basset Griffon Vendeen Petit	11	Basset Griffon Vendeen (Petit)	10	Basenji
11	Basset Hound::UK	12	Basset Hound::UK	11	Basset Griffon Vendeen (Petit)
12	Basset Hound::US	13	Basset Hound::US	12	Basset Hound::UK
13	Beagle::UK	14	Beagle::UK	13	Basset Hound::US
14	Beagle::US	15	Beagle::US Field	14	Beagle::UK
15	Beagle::US::2	16	Beagle::US Show	15	Beagle::US Field
16	Bearded Collie	17	Bearded Collie	16	Beagle::US Show
17	Belgian Sheepdog	18	Belgian Sheepdog	17	Bearded Collie
18	Belgian Tervuren	19	Belgian Tervuren	18	Belgian Sheepdog
19	Bernese Mountain Dog	20	Bernese Mountain Dog	19	Belgian Tervuren
20	Bichon Frise	21	Bichon Frise	20	Bernese Mountain Dog
21	Bloodhound	22	Bloodhound	21	Bichon Frise
22	Border Collie	23	Border Collie	22	Bloodhound
23	Border Terrier	24	Border Terrier	23	Border Collie
24	Borzoi	25	Borzoi	24	Border Terrier
25	Boston Terrier	26	Boston Terrier	25	Borzoi
26	Bouvier Des Flanders	27	Bouvier Des Flanders	26	Boston Terrier
27	Boxer	28	Boxer	27	Bouvier Des Flanders
28	Briard	29	Briard	28	Boxer
29	Brussels Griffon	30	Brussels Griffon	29	Briard
30	Bull Terrier	31—32	Bull Terrier—Bull Terrier (Miniature)	30	Brussels Griffon
31	Bulldog	33	Bulldog	31	Bull Terrier
32	Bullmastiff	34	Bullmastiff	32	Bull Terrier (Miniature)
33	Cairn Terrier	35	Cairn Terrier	33	Bulldog
34	Cardigan Welsh Corgi	36	Cardigan Welsh Corgi	34	Bullmastiff
35	Cavalier King Charles Spaniel	37	Cavalier King Charles Spaniel	35	Cairn Terrier
36	Chesapeake Bay Retriever	38	Chesapeake Bay Retriever	36	Cardigan Welsh Corgi
37	Chihuahua	39—40	Chihuahua::UKX/Chihuahua::USX	37	Cavalier King Charles Spaniel
38	Chinese Crested	41	Chinese Crested	38	Chesapeake Bay Retriever
39	Chinese Shar-Pei	42	Chinese Shar-Pei	39	Chihuahua::UKX
40	Chow Chow	43	Chow Chow	40	Chihuahua::USX
41	Cocker Spaniel	44	Cocker Spaniel	41	Chinese Crested
42	Collie	45—46	Collie::UKX—Collie::USX	42	Chinese Shar-Pei
43	Curly Coated Retriever	47	Curly Coated Retriever	43	Chow Chow
44	Dachshund::LH	48	Dachshund (Longhaired)	44	Cocker Spaniel
45	Dachshund::MLH	49	Dachshund (Miniature Longhaired)	45	Collie::UKX
46	Dachshund::MWH	51	Dachshund (Miniature Wirehaired)	46	Collie::USX
47	Dachshund	52—53	Dachshund (Shorthaired)—Dachshund (Wirehaired)	47	Curly Coated Retriever
48	Dalmatian::UK	54	Dalmatian::UK	48	Dachshund (Longhaired)
49	Dalmatian::US	55	Dalmatian::US	49	Dachshund (Miniature Longhaired)
50	Dobermann Pinscher	56	Dobermann Pinscher	50	Dachshund (Miniature Shorthaired)

Table A.2: Breed Merging Part I: in the first two columns I show the IDs and names for the merged breed. Then, similarly in columns 5 and 6 I show the IDs and names of the pure breeds. In the centre columns 3 and 4 I show which pure breeds correspond to the merged breeds in the respective table row.

breedM ID	breedM Name	breedM ID from breedP	breedM Name from breedP	breedP ID	breedP Name
51	English Cocker Spaniel	57—58	English Cocker Spaniel—English Cocker Spaniel::Field	51	Dachshund (Miniature Wirehaired)
52	English Setter	59	English Setter	52	Dachshund (Shorthaired)
53	English Springer Spaniel	60—61	English Springer Spaniel::UKX—English Springer Spaniel::USX	53	Dachshund (Wirehaired)
54	Flat Coated Retriever	62	Flat Coated Retriever	54	Dalmatian::UK
55	Fox Terrier Smooth	63	Fox Terrier (Smooth)	55	Dalmatian::US
56	Fox Terrier Toy	64	Fox Terrier (Toy)	56	Dobermann Pinscher
57	Fox Terrier Wire	65	Fox Terrier (Wire)	57	English Cocker Spaniel
58	French Bulldog	66	French Bulldog	58	English Cocker Spaniel::Field
59	German Shepherd Dog	67	German Shepherd Dog	59	English Setter
60	German Shorthaired Pointer	68—69	German Shorthaired Pointer::UKX—German Shorthaired Pointer::USX	60	English Springer Spaniel::UKX
61	German Wirehaired Pointer	70	German Wirehaired Pointer	61	English Springer Spaniel::USX
62	Golden Retriever::UK	71	Golden Retriever::UK	62	Flat Coated Retriever
63	Golden Retriever::US	72	Golden Retriever::US	63	Fox Terrier (Smooth)
64	Gordon Setter	73	Gordon Setter	64	Fox Terrier (Toy)
65	Great Dane	74	Great Dane	65	Fox Terrier (Wire)
66	Great Pyrenees	75	Great Pyrenees	66	French Bulldog
67	Greater Swiss Mountain Dog	76	Greater Swiss Mountain Dog	67	German Shepherd Dog
68	Greyhound	77	Greyhound	68	German Shorthaired Pointer::UKX
69	Havanese	78	Havanese	69	German Shorthaired Pointer::USX
70	Irish Setter::UK	79	Irish Setter::UK	70	German Wirehaired Pointer
71	Irish Setter::US	80	Irish Setter::US	71	Golden Retriever::UK
72	Irish Wolfhound	81	Irish Wolfhound	72	Golden Retriever::US
73	Italian Greyhound	82	Italian Greyhound	73	Gordon Setter
74	Italian Spinone	83	Italian Spinone	74	Great Dane
75	Japanese Chin	84	Japanese Chin	75	Great Pyrenees
76	Japanese Shiba Inu	85	Japanese Shiba Inu	76	Greater Swiss Mountain Dog
77	Keeshond	86—87	Keeshond::UKX—Keeshond::USX	77	Greyhound
78	Labrador Retriever::UK	88	Labrador Retriever::UK	78	Havanese
79	Labrador Retriever::2	89	Labrador Retriever::US Field	79	Irish Setter::UK
80	Labrador Retriever::3	90	Labrador Retriever::US Show	80	Irish Setter::US
81	Lhasa Apso	91—92	Lhasa Apso::UKX—Lhasa Apso::USX	81	Irish Wolfhound
82	Maltese	93—94	Maltese::UKX—Maltese::USX	82	Italian Greyhound
83	Mastiff	95	Mastiff	83	Italian Spinone
84	Miniature Pinscher	96—97	Miniature Pinscher::UKX—Miniature Pinscher::USX	84	Japanese Chin
85	Newfoundland	98	Newfoundland	85	Japanese Shiba Inu
86	Norfolk Terrier	99	Norfolk Terrier	86	Keeshond::UKX
87	Norwegian Elkhound	100	Norwegian Elkhound	87	Keeshond::USX
88	Norwich Terrier	101	Norwich Terrier	88	Labrador Retriever::UK
89	Old English Sheepdog	102	Old English Sheepdog	89	Labrador Retriever::US Field
90	Papillon	103	Papillon	90	Labrador Retriever::US Show
91	Parson Russell Terrier	104	Parson Russell Terrier	91	Lhasa Apso::UKX
92	Pekingese	105	Pekingese	92	Lhasa Apso::USX
93	Pembroke Welsh Corgi	106—107	Pembroke Welsh Corgi::UKX—Pembroke Welsh Corgi::USX	93	Maltese::UKX
94	Pointer	108	Pointer	94	Maltese::USX
95	Pomeranian	109—110	Pomeranian::UKX—Pomeranian::USX	95	Mastiff
96	Poodle	111—114	Poodle—Poodle (Toy)	96	Miniature Pinscher::UKX
97	Poodle Miniature ::justtoy	113	Poodle (Miniature)::miscwithtoy	97	Miniature Pinscher::USX
98	Poodle Miniature ::miniature	112	Poodle (Miniature)	98	Newfoundland
99	Portuguese Water Dog	115	Portuguese Water Dog	99	Norfolk Terrier
100	Pug	116	Pug	100	Norwegian Elkhound

Table A.3: Breed Merging Part II: continued, see Table A.2 for a discussion.

breedM ID	breedM Name	breedM ID from breedP	breedM Name from breedP	breedP ID	breedP Name
101	Rhodesian Ridgeback	117	Rhodesian Ridgeback	101	Norwich Terrier
102	Rottweiler	118	Rottweiler	102	Old English Sheepdog
103	Saint Bernard	119—120	Saint Bernard::UKX—Saint Bernard::USX	103	Papillon
104	Saluki	121	Saluki	104	Parson Russell Terrier
105	Samoyed	122—123	Samoyed::UKX—Samoyed::USX	105	Pekingese
106	Schipperke	124—125	Schipperke::UKX—Schipperke::USX	106	Pembroke Welsh Corgi::UKX
107	Schnauzer Giant	126	Schnauzer (Giant)	107	Pembroke Welsh Corgi::USX
108	Schnauzer Miniature	127	Schnauzer (Miniature)	108	Pointer
109	Scottish Terrier	128	Scottish Terrier	109	Pomeranian::UKX
110	Shetland Sheepdog::UK	129	Shetland Sheepdog::UK	110	Pomeranian::USX
111	Shetland Sheepdog::US	130	Shetland Sheepdog::US	111	Poodle
112	Shih Tzu	131	Shih Tzu	112	Poodle (Miniature)
113	Siberian Husky	132	Siberian Husky	113	Poodle (Miniature)::miscwithtoy
114	Soft Coated Wheaten Terrier	133—134	Soft Coated Wheaten Terrier::UKX—Soft Coated Wheaten Terrier::USX	114	Poodle (Toy)
115	Staffordshire Bull Terrier	135	Staffordshire Bull Terrier	115	Portuguese Water Dog
116	Tibetan Spaniel	136	Tibetan Spaniel	116	Pug
117	Tibetan Terrier	137	Tibetan Terrier	117	Rhodesian Ridgeback
118	Vizsla::UK	138	Vizsla::UK	118	Rottweiler
119	Vizsla::US	139	Vizsla::US	119	Saint Bernard::UKX
120	Weimaraner	140—141—142	Weimaraner::UKX—Weimaraner::US2X—Weimaraner::USX	120	Saint Bernard::USX
121	Welsh Terrier	143—144	Welsh Terrier::UKX—Welsh Terrier::USX	121	Saluki
122	West Highland White Terrier	145	West Highland White Terrier	122	Samoyed::UKX
123	Whippet	146—147	Whippet::UKX—Whippet::USX	123	Samoyed::USX
124	Yorkshire Terrier::UKKC	148	Yorkshire Terrier::UK	124	Schipperke::UKX
125	Yorkshire Terrier::US	149	Yorkshire Terrier::US	125	Schipperke::USX
				126	Schnauzer (Giant)
				127	Schnauzer (Miniature)
				128	Scottish Terrier
				129	Shetland Sheepdog::UK
				130	Shetland Sheepdog::US
				131	Shih Tzu
				132	Siberian Husky
				133	Soft Coated Wheaten Terrier::UKX
				134	Soft Coated Wheaten Terrier::USX
				135	Staffordshire Bull Terrier
				136	Tibetan Spaniel
				137	Tibetan Terrier
				138	Vizsla::UK
				139	Vizsla::US
				140	Weimaraner::UKX
				141	Weimaraner::US2X
				142	Weimaraner::USX
				143	Welsh Terrier::UKX
				144	Welsh Terrier::USX
				145	West Highland White Terrier
				146	Whippet::UKX
				147	Whippet::USX
				148	Yorkshire Terrier::UK
				149	Yorkshire Terrier::US

Table A.4: Breed Merging Part III: continued, see Table A.2 for a discussion.

Indicators/Position	$\Psi(1) = 2$	$\Psi(2) = 4$	$\Psi(3) = 6$
SNP allele	$X_k^j\,[\Psi(1] = -1$	$X_k^j\,[\Psi(1] = 1$	$X_k^j\,[\Psi(1] = -1$
Repeats ρ	1	4	8
Cycles π	8	4	1
Type	4	2	4

Table A.5: Values of repeats ρ, cycles π different pattern structures based on the ambiguous marker type based on $X_k^j = [0, ?, 0, 1, 2, ?]$.

A.2 Enumerating haplotype pairs

This section continues from Section 3.2.2.2 and offers more details on how we recursively enumerate haplotype pairs with a given genotype. At first I define four patterns:

- $a = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ corresponds to $H_1^i(j)[t] = 0, H_2^i(j)[t] = 1$,
- $b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ corresponds to $H_1^i(j)[t] = 1, H_2^i(j)[t] = 0$,
- $c = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ corresponds to $H_1^i(j)[t] = 1, H_2^i(j)[t] = 1$,
- $d = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ corresponds to $H_1^i(j)[t] = 0, H_2^i(j)[t] = 0$.

If the SNP at locus t assumes the values $X_k^j[t] \in \{0, 2\}$ there is no ambiguity and lead to pattern d or c, respectively. For $X_k^j[t] = 1$ I apply pattern a and b. Finally for $X_k^j[t] =?$, I apply all four patterns. I express these three cases formally as types

$$\text{type} = \begin{cases} 1, & \text{if marker } X_k^j[t] \in \{0, 2\}, \\ 2, & \text{if marker } X_k^j\,[\Psi(\gamma)] = 1, \gamma \in \{1, \ldots, \Gamma\}, \\ 4, & \text{if marker } X_k^j\,[\Psi(\gamma)] =?, \gamma \in \{1, \ldots, \Gamma\}. \end{cases}$$

where number of ambiguous markers in a genotype is denoted as $\Gamma \leq s$. Returning to the example $X_k^j = [0, ?, 0, 1, 2, ?]$ there are $\Gamma = 3$ ambiguous SNPs at loci 2,4 and 6.

Then, define the map Ψ : ambiguous marker $\mapsto t$ which assigns r-th ambiguous marker ID to its corresponding SNP locus. Furthermore, I define repeats ρ and cycles π at each locus t within four cases:

1. CASE ? (Type = 4): this pattern structure is referred to as $type = 4$ because it includes the four patterns a, b, c and d. If $X_k^j[t] =?$, I define the pattern structure

$$[a\{\rho\}b\{\rho\}c\{\rho\}d\{\rho\}]\{\pi\}$$

At position t pattern a is repeated ρ times for haplotype pairs $i = 1$ to $i = \rho$. From haplotype pair $i = \rho + 1$ to $i = 2 \cdot \rho$ I repeat pattern b, then from haplotype pair $i = 2 \cdot \rho + 1$ to $i = 3 \cdot \rho$ I repeat pattern c, and finally from haplotype pair $i = 3 \cdot \rho + 1$ to $i = 4 \cdot \rho$ I repeat pattern d. Then, from haplotype pair $i = 4 \cdot \rho + 1$, I start with pattern a again, followed by pattern b, c and d. In total, I repeat this process of repeating patterns a, b, c, d for π times.

2. CASE 1 (Type = 2): this pattern structure is referred to as $type = 4$ because it includes the two patterns a and b. If $X_k^j[t] = 1$, I define pattern structure

$$[a\{\rho\}b\{\rho\}]\{\pi\}$$

and this pattern is defined as of $type = 2$. Then, similarly to case ?, for π times I repeat pattern a for ρ, followed by ρ times of pattern b.

3. CASE 0 (Type = 1): $X_k^j[t] = 0$, I define pattern structure $[d\{m_{jk}\}]\{1\}$, i.e. at position t $\forall i$, I have pattern d.

4. CASE 2 (Type = 1): $X_k^j[t] = 2$, I define pattern structure $[c\{m_{jk}\}]\{1\}$, i.e. at position t $\forall i$, I have pattern c.

Returning to Figure 3.2, for example at the second ambigious marker $X_k^j[\Psi(2)] = X_k^j[4] = 1$ I have $[a\{4\}b\{4\}]\{4\}$. This pattern structure implies that in the first cycle the first four haplotype pairs follow pattern a, haplotype pairs 5 to 8 follow pattern b. In the second cycle, haplotype pairs 9 to 13 take pattern a while haplotype pairs 14 to 17 assume pattern b. Similarly, I proceed with cycle 3 and 4. Values for ρ and π for the other ambiguous markers are shown in table A.5.

In general, I obtain the following recursive formulae for ρ. At first, I define ρ recursively with base case $\rho\left[\Psi(1)\right] = 1$:

$$\rho\left[\Psi(\gamma)\right] = \begin{cases} 2 \cdot \rho\left[\Psi((\gamma-1)\right], & \text{if marker } X_k^j\left[\Psi((\gamma-1)\right] = 1, \gamma \in \{1,\ldots,\Gamma\}, \\ 4 \cdot \rho\left[\Psi((\gamma-1)\right], & \text{if marker } X_k^j\left[\Psi((\gamma-1)\right] =?, \gamma \in \{1,\ldots,\Gamma\}. \end{cases}$$

Similarly, I compute π recursively with base case

$$\pi\left[\Psi(1)\right] = \begin{cases} m_{jk}/2, & \text{if marker } X_k^j\left[\Psi((1)\right] = 1 \\ m_{jk}/4, & \text{if marker } X_k^j\left[\Psi((1)\right] =? \end{cases}$$

and general case

$$\pi\left[\Psi((\gamma)\right] = \begin{cases} m_{jk}/\rho\left[\Psi((\gamma)\right]/2, & \text{if marker } X_k^j\left[\Psi((\gamma)\right] = 1, \gamma \in \{1,\ldots,\Gamma\}, \\ m_{jk}/\rho\left[\Psi((\gamma)\right]/4, & \text{if marker } X_k^j\left[\Psi((\gamma)\right] =?, \gamma \in \{1,\ldots,\Gamma\}. \end{cases}$$

A.3 Results for pure breed identification

d	10^{-5}	$3 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$7 \cdot 10^{-5}$	$9 \cdot 10^{-5}$	10^{-4}	$3 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	10^{-3}	$3 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$9 \cdot 10^{-3}$
DFPB PSEPARATE $1 - \bar{q}_{1:B}$	0	0	0	$2 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	0.001	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
DFPB PPOOLED $1 - \bar{q}_{1:B}$	0.032	0.030	0.031	0.032	0.034	0.034	0.036	0.036	0.036	0.036	0.036	0.039	0.043	0.050	0.050
OneHap PSEPARATE $1 - \bar{q}_{1:B}$	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.028	0.036	0.074	0.090
OneHap PPOOLED $1 - \bar{q}_{1:B}$	0.039	0.039	0.039	0.037	0.036	0.035	0.030	0.030	0.029	0.029	0.029	0.030	0.030	0.039	0.047
PSEUDO PSEPARATE $1 - \bar{q}_{1:B}$	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.033	0.041	0.057	0.065	0.198	0.338	0.423	0.463
PSEUDO PPOOLED $1 - \bar{q}_{1:B}$	0.032	0.030	0.031	0.031	0.032	0.032	0.029	0.029	0.029	0.029	0.030	0.117	0.297	0.442	0.516

Table A.6: Each cell in the table refers to $1 - \bar{q}_{1:B}$, i.e. **one minus the breed-averaged** mean posterior probability assigned to the correct breed for the small dataset based on **separate** and **pooled phasing**. I use **DFPB, PSEUDO and OneHap** to define $p\left(X_c|\theta\right)$. I use $B' = 34$ breeds.

d	10^{-5}	$3 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$7 \cdot 10^{-5}$	$9 \cdot 10^{-5}$	10^{-4}	$3 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	10^{-3}	$3 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
DFPB PSEPARATE Misclassifications	1	1	1	1	1	1	6	8	9	11	12	22	28
OneHap PSEPARATE Misclassifications	5	4	3	1	1	1	0	0	0	0	0	0	1
DFPB PSEPARATE $1 - \bar{q}_{1:B}$	0.0023	10^{-4}	10^{-4}	$2 \cdot 10^{-4}$	$4 \cdot 10^{-4}$	$6 \cdot 10^{-4}$	0.002	0.0023	0.0026	0.0029	0.0031	0.006	0.0088
OneHap PSEPARATE $1 - \bar{q}_{1:B}$	0.0112	0.0094	0.0054	0.0026	0.0023	0.0022	0	0	0	0	0	$2 \cdot 10^{-4}$	$9 \cdot 10^{-4}$

Table A.7: Summary of $1 - \bar{q}_{1:B}$, i.e. **one minus the breed-averaged mean posterior probability assigned to the correct breed**, and the number of misclassifications for the **big dataset** based on **separate phasing**. Prediction misclassification is out of the total number of test dogs which is 1302. I use **DFPB and OneHap** to define $p\left(X_c|\theta\right)$. use $B = 149$ breeds.

DFPB	10^{-5}	$3 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$7 \cdot 10^{-5}$	$9 \cdot 10^{-5}$	10^{-4}	$3 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	10^{-3}	$3 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$9 \cdot 10^{-3}$
Australian Shepherd::UKX				0.01	0.02	0.04	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.34	0.34
Australian Shepherd::USX															
Beagle::UK															
Beagle::US Field													0.01	0.02	0.02
Beagle::US Show															
$1 - \bar{q}_{1:B}$				$2 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	0.0013	0.0096	0.0098	0.0098	0.0098	0.0098	0.0099	0.0101	0.0108	0.0108

Table A.8: These are the results for the **small pure breed dataset** based on **separate phasing**. I use **DFPB** to define $p\left(X_c|\theta\right)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d.

DFPB	10^{-5}	$3 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$7 \cdot 10^{-5}$	$9 \cdot 10^{-5}$	10^{-4}	$3 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	10^{-3}	$3 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$9 \cdot 10^{-3}$
Alaskan Malamute														0.01	0.01
American Eskimo Dog													0.03	0.07	0.07
American Staffordshire Terrier														0.01	0.01
Australian Shepherd::UKX	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.34	0.34	0.34	0.42	0.54	0.63	0.63
Australian Shepherd::USX														0.01	0.01
Basset Hound::UK														0.01	0.01
Basset Hound::US														0.01	0.01
Beagle::US Field	0.33	0.32	0.31	0.3	0.3	0.29	0.26	0.26	0.26	0.26	0.26	0.31	0.37	0.44	0.44
Beagle::US Show													0.01	0.03	0.03
Belgian Tervuren		0.02	0.07	0.13	0.18	0.2	0.3	0.31	0.31	0.31	0.31	0.26	0.22	0.19	0.19
Bull Terrier	0.41	0.35	0.34	0.34	0.34	0.34	0.33	0.33	0.33	0.33	0.33	0.31	0.29	0.28	0.28
$1 - \bar{q}_{1:B}$	0.0316	0.0303	0.0312	0.0326	0.0338	0.0343	0.0363	0.0363	0.0364	0.0364	0.0364	0.0386	0.0431	0.0495	0.0495

Table A.9: These are the results for the **small pure breed dataset** based on **pooled phasing**. I use **DFPB** to define $p\left(X_c|\theta\right)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d.

OneHap	10^{-5}	$3 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$7 \cdot 10^{-5}$	$9 \cdot 10^{-5}$	10^{-4}	$3 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	10^{-3}	$3 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$9 \cdot 10^{-3}$
American Eskimo Dog													0.27	0.33	0.33
Australian Cattle Dog														0.08	0.08
Australian Shepherd::UKX	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.4	0.98	0.98
Australian Shepherd::USX															
Beagle::UK															
Beagle::US Field													0.06	0.3	0.3
Beagle::US Show															
Belgian Tervuren														0.14	0.14
Boston Terrier													0.03	0.33	0.33
Bull Terrier	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.63	0.45	0.37	0.37
Bull Terrier (Miniature)															
$1 - \bar{q}_{1:B}$	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294	0.0283	0.0357	0.0744	0.09

Table A.10: These are the results for the **small pure breed dataset** based on **separate phasing**. I use **OneHap** to define $p(X_c|\theta)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. We only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d.

OneHap	10^{-5}	$3 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$7 \cdot 10^{-5}$	$9 \cdot 10^{-5}$	10^{-4}	$3 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	10^{-3}	$3 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$9 \cdot 10^{-3}$
Australian Cattle Dog													0.01	0.31	0.31
Australian Shepherd::UKX	0.34	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
Australian Shepherd::USX												0.01	0.02	0.02	0.02
Beagle::UK															
Beagle::US Field	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.32	0.33	0.33	0.33
Beagle::US Show															
Bull Terrier	0.67	0.66	0.64	0.6	0.56	0.53	0.35	0.34	0.33	0.33	0.33	0.33	0.33	0.33	0.33
Bull Terrier (Miniature)															
$1 - \bar{q}_{1:B}$	0.0394	0.0391	0.0385	0.0374	0.0359	0.0352	0.0299	0.0295	0.0294	0.0294	0.0294	0.0295	0.0303	0.0393	0.0468

Table A.11: These are the results for the **small pure breed dataset** based on **pooled phasing**. I use **OneHap** to define $p(X_c|\theta)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d.

PSEUDO	10^{-5}	$3 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$7 \cdot 10^{-5}$	$9 \cdot 10^{-5}$	10^{-4}	$3 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	10^{-3}	$3 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$9 \cdot 10^{-3}$
Alaskan Malamute												0.66	0.67	1	1
American Eskimo Dog										0.01	0.03	1	1	1	1
American Staffordshire Terrier												0.07	0.73	1	1
Australian Cattle Dog													0.65	1	1
Australian Shepherd::UKX	0.33	0.33	0.33	0.33	0.33	0.33	0.35	0.55	0.66	0.93	1	1	1	1	1
Australian Shepherd::USX												0.49	0.7	1	1
Basset Hound::US												0.44	0.99	1	1
Beagle::US Field							0.01	0.02	0.23	0.48	0.67	1	1	1	1
Beagle::US Show												0.04	0.76	1	1
Belgian Tervuren								0.02	0.05	0.08	0.1	0.28	0.32	0.59	0.59
Bichon Frise												0.56	1	1	1
Border Collie													0.99	1	1
Borzoi													0.31	0.54	0.54
Boston Terrier												0.82	1	1	1
Bouvier Des Flanders													0.01	0.31	0.31
Bull Terrier	0.67	0.67	0.67	0.67	0.67	0.66	0.62	0.54	0.48	0.43	0.42	0.35	0.34	0.33	0.33
Bulldog													0.03	0.59	0.59
Bullmastiff														0.02	0.02
$1 - \bar{q}_{1:B}$	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294	0.0287	0.0334	0.0414	0.057	0.0652	0.1975	0.3379	0.4228	0.4625

Table A.12: These are the results for the **small pure breed dataset** based on **separate phasing**. I use **PSEUDO** to define $p\left(X_c|\theta\right)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d.

PSEUDO	10^{-5}	$3 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$7 \cdot 10^{-5}$	$9 \cdot 10^{-5}$	10^{-4}	$3 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	10^{-3}	$3 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$9 \cdot 10^{-3}$
Alaskan Malamute												0.07	0.66	0.67	0.67
American Eskimo Dog												0.1	0.98	1	1
American Staffordshire Terrier												0.04	0.68	0.99	0.99
Australian Cattle Dog													0.17	0.65	0.65
Australian Shepherd::UKX	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.34	0.36	1	1	1	1
Australian Shepherd::USX									0.01	0.02	0.04	0.64	0.95	1	1
Basset Hound::US												0.01	0.2	0.6	0.6
Beagle::US Field	0.33	0.32	0.3	0.27	0.24	0.23	0.04	0.01	0.01	0.03	0.04	0.75	0.99	1	1
Belgian Sheepdog												0.01	0.07	0.97	0.97
Belgian Tervuren		0.02	0.07	0.12	0.17	0.19	0.29	0.29	0.28	0.26	0.25	0.06	0.47	0.97	0.97
Bernese Mountain Dog													0.01	0.68	0.68
Bichon Frise												0.91	1	1	1
Border Collie													0.95	1	1
Borzoi													0.45	0.71	0.71
Boston Terrier													0.66	0.98	0.98
Bouvier Des Flanders													0.33	0.33	0.33
Bull Terrier	0.41	0.35	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.35	0.42	0.52	0.52
Bulldog													0.09	0.65	0.65
Bullmastiff														0.29	0.29
$1 - \bar{q}_{1:B}$	0.0316	0.0301	0.0306	0.0314	0.0318	0.0319	0.0292	0.0285	0.0285	0.0291	0.0299	0.1166	0.2968	0.4422	0.5159

Table A.13: These are the results for the **small pure breed dataset** based on **pooled phasing**. I use **PSEUDO** to define $p(X_c|\theta)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B' = 34$ breeds (including those ones not shown) for a fixed value of d.

DFPB	10^{-5}	$3 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$7 \cdot 10^{-5}$	$9 \cdot 10^{-5}$	10^{-4}	$3 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	10^{-3}	$3 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
Australian Shepherd::UKX				0.01	0.04	0.06	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Beagle::US Field													0.01
Cairn Terrier												0.01	0.03
Chihuahua::USX												0.02	0.07
Golden Retriever::US										0.01	0.01	0.07	0.09
Labrador Retriever::US Field		0.02	0.02	0.02	0.02	0.02	0.09	0.14	0.18	0.21	0.23	0.32	0.37
Labrador Retriever::US Show												0.02	0.02
Lhasa Apso::USX												0.02	0.02
Maltese::USX												0.05	0.11
Poodle (Miniature)::miscwithtoy										0.01	0.01	0.16	0.29
Yorkshire Terrier::US												0.03	0.06
$1 - \bar{q}_{1:B}$	0.0023	10^{-4}	10^{-4}	$2 \cdot 10^{-4}$	$4 \cdot 10^{-4}$	$6 \cdot 10^{-4}$	0.002	0.0023	0.0026	0.0029	0.0031	0.006	0.0088

Table A.14: These are the results for the **big pure breed dataset** based on **separate phasing**. I use **DFPB** to define $p\left(X_c|\theta\right)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B = 149$ breeds (including those ones not shown) for a fixed value of d.

OneHap	10^{-5}	$3 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$7 \cdot 10^{-5}$	$9 \cdot 10^{-5}$	10^{-4}	$3 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	10^{-3}	$3 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
Australian Shepherd::UKX												0.03	0.14
Australian Shepherd::USX													
Keeshond::UKX	0.67	0.66	0.24	0.02									
Weimaraner::US2X	1	0.74	0.57	0.36	0.34	0.33							
$1 - \bar{q}_{1:B}$	0.0112	0.0094	0.0054	0.0026	0.0023	0.0022	0	0	0	0	0	$2 \cdot 10^{-4}$	$9 \cdot 10^{-4}$

Table A.15: These are the results for the **big pure breed dataset** based on **separate phasing**. I use **OneHap** to define $p\left(X_c|\theta\right)$. Rows correspond to breeds while columns correspond to values of tuning parameter d. Each cell in the table refers to $1 - q_b$, i.e. one minus the mean posterior breed probability. A blank field refers to $1 - q_b = 0$. I only show a subset of breeds, i.e. those rows with at least one element $1 - q_b > 0$. The last row shows the column mean $1 - \bar{q}_{1:B}$ across all $B = 149$ breeds (including those ones not shown) for a fixed value of d.

Tree	BP_1	BP_2	BP_3	BP_4	Tree CCBC	BP_1^C	BP_2^C	BP_3^C	BP_4^C
AAAAAAAA	$Q_1^0 = 0.001$	$Q_1^{1.0} = 0.569$			AAAAAAAA	$Q_1^0 = 0$	$Q_1^{0.5} = 0.739$		
	$Q_2^0 = 0.001$	$Q_2^{1.0} = 0.712$				$Q_2^0 = 0$	$Q_2^{0.5} = 0.866$		
	$Q_3^0 = 0.002$	$Q_3^{1.0} = 0.805$				$Q_3^0 = 0$	$Q_3^{0.5} = 0.931$		
AAAABBBB	$Q_1^0 = 0.001$	$Q_1^{0.5} = 0.230$			AAAABBBB	$Q_1^0 = 0$	$Q_1^{0.5} = 0.321$		
	$Q_2^0 = 0.001$	$Q_2^{0.5} = 0.310$				$Q_2^0 = 0$	$Q_2^{0.5} = 0.412$		
	$Q_3^0 = 0.003$	$Q_3^{0.5} = 0.373$				$Q_3^0 = 0$	$Q_3^{0.5} = 0.466$		
AAAABBCC	$Q_1^0 = 0.001$	$Q_1^{0.25} = 0.129$	$Q_1^{0.5} = 0.236$		AAAABBCC	$Q_1^0 = 0$	$Q_1^{0.25} = 0.166$	$Q_1^{0.5} = 0.331$	
	$Q_2^0 = 0.001$	$Q_2^{0.25} = 0.150$	$Q_2^{0.5} = 0.317$			$Q_2^0 = 0$	$Q_2^{0.25} = 0.201$	$Q_2^{0.5} = 0.413$	
	$Q_3^0 = 0.003$	$Q_3^{0.25} = 0.196$	$Q_3^{0.5} = 0.371$			$Q_3^0 = 0$	$Q_3^{0.25} = 0.255$	$Q_3^{0.5} = 0.466$	
AABBCCDD	$Q_1^0 = 0.001$	$Q_1^{0.25} = 0.121$			AABBCCDD	$Q_1^0 = 0$	$Q_1^{0.25} = 0.160$		
	$Q_2^0 = 0.001$	$Q_2^{0.25} = 0.146$				$Q_2^0 = 0$	$Q_2^{0.25} = 0.199$		
	$Q_3^0 = 0.003$	$Q_3^{0.25} = 0.188$				$Q_3^0 = 0$	$Q_3^{0.25} = 0.252$		
AAAABBCD	$Q_1^0 = 0.001$	$Q_1^{0.125} = 0.266$	$Q_1^{0.25} = 0.125$	$Q_1^{0.5} = 0.228$	AAAABBCD	$Q_1^0 = 0$	$Q_1^{0.125} = 0.037$	$Q_1^{0.25} = 0.156$	$Q_1^{0.5} = 0.325$
	$Q_2^0 = 0.001$	$Q_2^{0.125} = 0.103$	$Q_2^{0.25} = 0.142$	$Q_2^{0.5} = 0.312$		$Q_2^0 = 0$	$Q_2^{0.125} = 0.135$	$Q_2^{0.25} = 0.193$	$Q_2^{0.5} = 0.409$
	$Q_3^0 = 0.002$	$Q_3^{0.125} = 0.129$	$Q_3^{0.25} = 0.180$	$Q_3^{0.5} = 0.367$		$Q_3^0 = 0$	$Q_3^{0.125} = 0.164$	$Q_3^{0.25} = 0.243$	$Q_3^{0.5} = 0.459$
AAAABCDE	$Q_1^0 = 0.001$	$Q_1^{0.125} = 0.020$	$Q_1^{0.5} = 0.222$		AAAABCDE	$Q_1^0 = 0$	$Q_1^{0.125} = 0.028$	$Q_1^{0.5} = 0.312$	
	$Q_2^0 = 0.001$	$Q_2^{0.125} = 0.101$	$Q_2^{0.5} = 0.304$			$Q_2^0 = 0$	$Q_2^{0.125} = 0.132$	$Q_2^{0.5} = 0.396$	
	$Q_3^0 = 0.002$	$Q_3^{0.125} = 0.128$	$Q_3^{0.5} = 0.362$			$Q_3^0 = 0$	$Q_3^{0.125} = 0.162$	$Q_3^{0.5} = 0.453$	
AABBCCDE	$Q_1^0 = 0.001$	$Q_1^{0.125} = 0.018$	$Q_1^{0.25} = 0.121$		AABBCCDE	$Q_1^0 = 0$	$Q_1^{0.125} = 0.026$	$Q_1^{0.25} = 0.159$	
	$Q_2^0 = 0.001$	$Q_2^{0.125} = 0.099$	$Q_2^{0.25} = 0.146$			$Q_2^0 = 0$	$Q_2^{0.125} = 0.133$	$Q_2^{0.25} = 0.198$	
	$Q_3^0 = 0.001$	$Q_3^{0.125} = 0.129$	$Q_3^{0.25} = 0.186$			$Q_3^0 = 0$	$Q_3^{0.125} = 0.166$	$Q_3^{0.25} = 0.264$	
AABBCDEF	$Q_1^0 = 0.001$	$Q_1^{0.125} = 0.018$	$Q_1^{0.25} = 0.112$		AABBCDEF	$Q_1^0 = 0$	$Q_1^{0.125} = 0.026$	$Q_1^{0.25} = 0.152$	
	$Q_2^0 = 0.001$	$Q_2^{0.125} = 0.094$	$Q_2^{0.25} = 0.142$			$Q_2^0 = 0$	$Q_2^{0.125} = 0.128$	$Q_2^{0.25} = 0.195$	
	$Q_3^0 = 0.002$	$Q_3^{0.125} = 0.128$	$Q_3^{0.25} = 0.181$			$Q_3^0 = 0$	$Q_3^{0.125} = 0.167$	$Q_3^{0.25} = 0.243$	
AABCDDEF	$Q_1^0 = 0.001$	$Q_1^{0.125} = 0.018$	$Q_1^{0.25} = 0.117$		AABCDDEF	$Q_1^0 = 0$	$Q_1^{0.125} = 0.025$	$Q_1^{0.25} = 0.153$	
	$Q_2^0 = 0.001$	$Q_2^{0.125} = 0.093$	$Q_2^{0.25} = 0.142$			$Q_2^0 = 0$	$Q_2^{0.125} = 0.126$	$Q_2^{0.25} = 0.200$	
	$Q_3^0 = 0.002$	$Q_3^{0.125} = 0.128$	$Q_3^{0.25} = 0.180$			$Q_3^0 = 0$	$Q_3^{0.125} = 0.167$	$Q_3^{0.25} = 0.241$	
AABCDEFG	$Q_1^0 = 0.001$	$Q_1^{0.125} = 0.015$	$Q_1^{0.25} = 0.091$		AABCDEFG	$Q_1^0 = 0$	$Q_1^{0.125} = 0.022$	$Q_1^{0.25} = 0.135$	
	$Q_2^0 = 0.001$	$Q_2^{0.125} = 0.082$	$Q_2^{0.25} = 0.137$			$Q_2^0 = 0$	$Q_2^{0.125} = 0.117$	$Q_2^{0.25} = 0.187$	
	$Q_3^0 = 0.003$	$Q_3^{0.125} = 0.127$	$Q_3^{0.25} = 0.173$			$Q_3^0 = 0$	$Q_3^{0.125} = 0.168$	$Q_3^{0.25} = 0.240$	
ABCDEFGH	$Q_1^0 = 0$	$Q_1^{0.125} = 0.014$			ABCDEFGH	$Q_1^0 = 0$	$Q_1^{0.125} = 0.020$		
	$Q_2^0 = 0.001$	$Q_2^{0.125} = 0.075$				$Q_2^0 = 0$	$Q_2^{0.125} = 0.107$		
	$Q_3^0 = 0.003$	$Q_3^{0.125} = 0.126$				$Q_3^0 = 0$	$Q_3^{0.125} = 0.166$		

Table A.16: Breed proportions of all three quartiles for all 11 family trees are shown. The first 5 columns correspond to uncorrected while the last columns represent CCBC breed proportions.

A.4 HMM Forward Algorithm

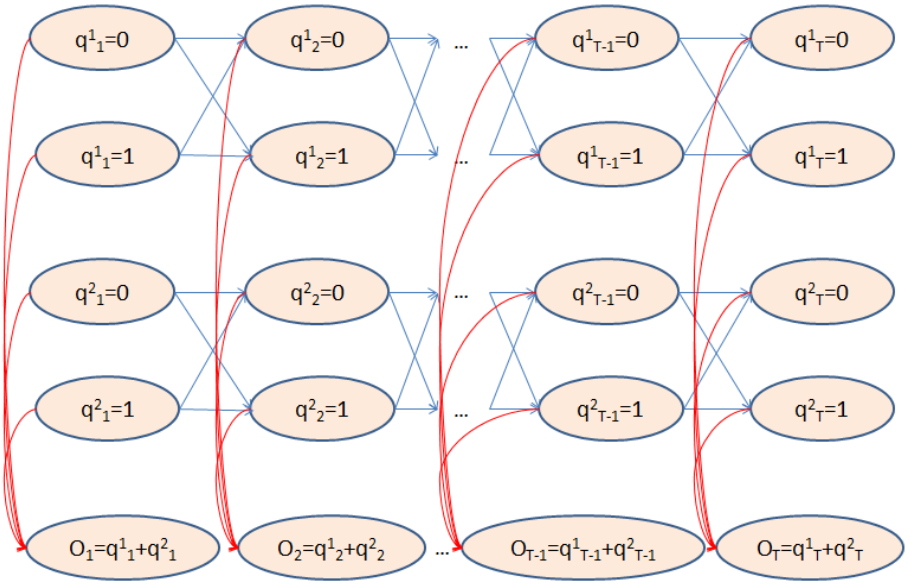


Figure A.1: HMM with two hidden markov chains unfolds in time as a lattice

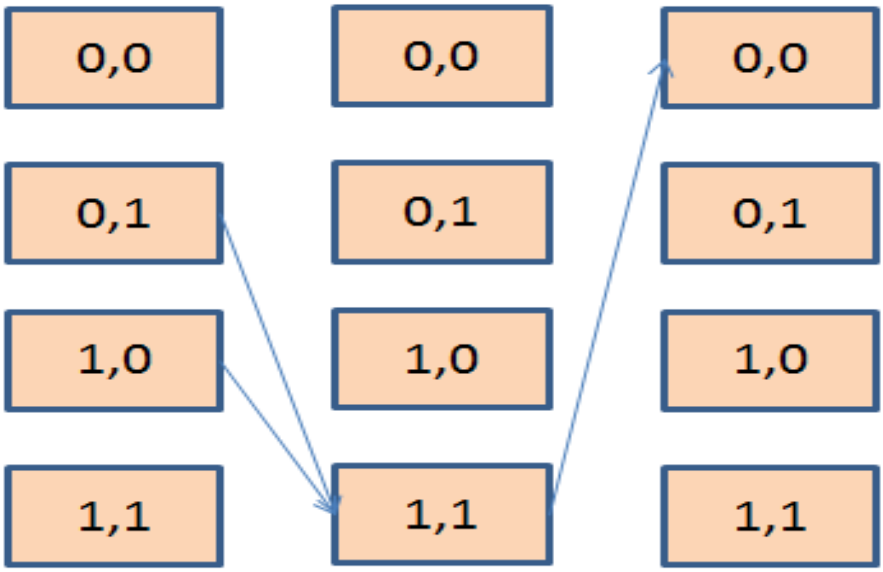


Figure A.2: HMM hidden path illustration for genotype [1,2,0] for three time steps

I frame the computation p_{cuv} as a forward algorithm in a hidden markov model (HMM) (Rabiner and Juang, 1986) whose topology is shown in figure A.1. I observe the genotype X_c on chromosome c which I view as a sequence of T emitted observations in a HMM, bottom row figure A.1, and I write $X_c = [O_1, \ldots, O_T]$ and $O_t \in \{v_1 = 0, v_2 = 1, v_3 = 2, v_4 = -1\}, t \in \{1, \ldots, T\}$. There are two hidden markov chains $q^1_{1:T} \in \{S_1 = 0, S_2 = 1\}^T$ and $q^2_{1:T} \in \{S_1 = 0, S_2 = 1\}^T$, one for each of the haplotypes. In figure A.1 the first two rows correspond to the hidden markov chain for the first haplotype through time, similarly rows 3 and 4 represent the hidden markov chain for the second haplotype. This figure illustrates how each of the hidden markov chain's state trajectory may oscillate between 0 and 1. It is assumed that one of the haplotypes is related to dog breed u_c while the other corresponds to dog

breed v_c. The emitted observation O_t at each time step in the genotype is computed as $O_t = q_t^1 + q_t^2$ as shown in figure A.1. The probability p_{cuv} in HMM terminology refers to probability $p(X_c|\lambda)$ of the observed genotype sequence X_c given the HMM model $\lambda = (\pi^1, A^1, \pi^2, A^2, B)$ which includes the initial state probabilities $\pi_i^1 = p\left(q_1^1 = S_i\right)$, $\pi_j^2 = p\left(q_1^2 = S_j\right)$, the state transition probabilities $a_{ij}^1 = p\left(q_{t+1}^1 = S_j | q_t^1 = S_i\right)$, $a_{kl}^2 = p\left(q_{t+1}^2 = S_l | q_t^2 = S_k\right)$, $S_\bullet \in \{0, 1\}$ for the two haplotypes and observation emission probabilities $b_{j,k}(m) = p\left(O_t = v_m | q_t^1 = S_j, q_t^1 = S_k\right), m \in \{1, 2, 3\}$. Note, that first in contrast to a standard HMM (Rabiner and Juang, 1986) I have two and not just one hidden markov chains associated with the observed markov chain. This can be viewed as a compound hidden state $q_t = (q_t^1, q_t^2)$ which attains the following $4 = |\{0, 1\}^2|$ values $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Secondly, I do not have 'actual' emission probabilities because when the first hidden markov chain at time t has value S_j and the second hidden markov chain attains value S_k I emit with probability 1 the observed state O_t at time t is $v_m = S_j + S_k$. Then, to evaluate p_{cuv} I compute $p(X_c|\lambda)$ which can be computed by marginalizing over all the possible hidden state sequences in $p\left(X_c, q_{1:T}|\lambda\right)$ which can be written as

$$
\begin{aligned}
p\left(X_c, q_{1:T}|\lambda\right) &= p(q_1^1)p(q_1^2)\prod_{t=2}^{T} p(q_t^1|q_{t-1}^1)\prod_{t=2}^{T} p(q_t^2|q_{t-1}^2)\prod_{t=1}^{T} p(O_t|q_t^1, q_t^2) \\
p\left(X_c, q_{1:T}|\lambda\right) &= \pi_{q_1^1}^1 \pi_{q_1^2}^2 b_{q_1^1,q_1^2}(q_1^1 + q_1^2) a_{q_1^1 q_2^1}^1 a_{q_1^2 q_2^2}^2 b_{q_2^1,q_2^2}(q_2^1 + q_2^2) \cdots a_{q_{T-1}^1 q_T^1}^1 a_{q_{T-1}^2 q_T^2}^2 b_{q_T^1,q_T^2}(q_T^1 + q_T^2)
\end{aligned}
$$

As I mentioned above once I know the hidden states at some time step t there is only one possibility for the emitted observation and this symbol is emitted with certainty. Therefore, I can simplify as

$$
\begin{aligned}
p\left(X_c, q_{1:T}|\lambda\right) &= p(q_1^1)p(q_1^2)\prod_{t=2}^{T} p(q_t^1|q_{t-1}^1)\prod_{t=2}^{T} p(q_t^2|q_{t-1}^2) \\
p\left(X_c, q_{1:T}|\lambda\right) &= \pi_{q_1^1}^1 \pi_{q_1^2}^2 a_{q_1^1 q_2^1}^1 a_{q_1^2 q_2^2}^2 \cdots a_{q_{T-1}^1 q_T^1}^1 a_{q_{T-1}^2 q_T^2}^2
\end{aligned}
$$

There are 4^T possible hidden state paths $q_{1:T}$ under the assumption that all probabilities are non-zero. In practise this marginalizing over all the possible hidden states is often computationally expensive and a forward variable is recursively evaluated based on the idea that the observation sequence can be divided into two partial observation sequences: the first starting from time 1 until t and the second one from time $t + 1$ until T. The forward variable $\alpha_t(i, j)$ is defined as

$$
\alpha_t(i, j) = p\left(O_1, \ldots, O_t, q_t^1 = S_i, q_t^2 = S_j|\lambda\right)
$$

and can be calculated recursively. According to Rabiner and Juang (1986) I initialize the forward variable as

$$
\begin{aligned}
\alpha_1(i, j) &= p\left(O_1, q_1^1 = S_i, q_1^2 = S_j|\lambda\right) \\
\alpha_1(i, j) &= \pi_i^1 \pi_j^2 b_{S_i,S_j}(S_i + S_j) \\
\alpha_1(i, j) &= \pi_i^1 \pi_j^2
\end{aligned}
$$

Then, I can write down the recursive equation for the forward variable whose derivation can be found in (Rabiner and Juang, 1986)

Hidden state	SNP 1	SNP 2	SNP 3
(0,0)	0	0	0.17 * 0.46 * 0.46 = 0.04
(0,1)	0.43 * 0.58 = 0.25	0	0
(1,0)	0.58 * 0.43 = 0.25	0	9
(1,1)	0	0.25 * 0.48 * 0.72 + 0.25 * 0.72 * 0.48 = 0.17	0

Table A.17: HMM example forward probability for genotype [1,2,0]

$$\alpha_{t+1}(i,j) = p\left(O_1, \ldots, O_{t+1} q^1_{t+1} = S_i, q^2_{t+1} = S_j | \lambda\right)$$
$$\alpha_{t+1}(i,j) = \left[\sum_{k=1}^{2}\sum_{l=1}^{2} \alpha_t(k,l) a^1_{ki} a^2_{lj}\right] b_{ij}(S_i + S_j)$$
$$\alpha_{t+1}(i,j) = \left[\sum_{k=1}^{2}\sum_{l=1}^{2} \alpha_t(k,l) a^1_{ki} a^2_{lj}\right]$$

The forward variable $\alpha_t(k,l)$ explains the first t observations and ends in state i for the first hidden markov chain and in state j for the second one. I multiply this by transition probabilities a^1_{ki} and a^2_{lj} such that the first hidden markov chain moves to state i and the second markov chain to state j because there are two possible previous states for both hidden markov chains, and I need to sum over all possible such possible previous states. The emission probability $b_{ij}(S_i + S_j)$ is known with certainty again and as such evaluates to one. Finally, when I calculated the forward probabilities up to time T it is easy to evaluate the probability $p\left(X_c|\lambda\right)$ of the observation sequence, i.e. the genotype on chromosome c, given the model λ :

$$p(X_c|\lambda) = \sum_{k=1}^{2}\sum_{l=1}^{2} p(X_c, q^1_T = S_k, q^2_T = S_l|\lambda)$$
$$p(X_c|\lambda) = \sum_{k=1}^{2}\sum_{l=1}^{2} \alpha_T(k,l)$$

where $\alpha_t(k,l)$ is the probability of generating the complete genotype on chromosome c and end in states S_k for the first hidden markov chain and S_l for the second hidden markov chain.

A.4.1 Toy example

I would like to illustrate this forward algorithm with an example on the first chromosome. Assume that both dog breeds u_c and v_c are Siberian Husky as shown in table 3.2 with its transition probabilities. Note that for simplicity I only consider the first three SNP marker and I use only two digits after the comma for transition probabilitis in table 3.2. Let the observed genotype be $X_1 = [1,2,0]$. This example can be found in figure A.2 which shows the possible hidden state sequences for this genotype. In this figure the first number represents the first haplotype and the second number corresponds to the second haplotype. Then, I know at time 1 the allele can be either $0,1$ or vice versa, at time 2 both alleles must be 1 and at time $T = 3$ both alleles must attain value 0. In table A.17 I show the forward probabilities: in column 'SNP 1' I show the values for forward probabilities $\alpha_1(\cdot,\cdot)$, and finally in column 'SNP 3' I have the values $\alpha_3(\cdot,\cdot)$ for time $T = 3$. Then, summing over all hidden states of the forward variable at time $T = 3$ the probability $p\left(X_c|\lambda\right)$ of the observed genotype

$X_1 = [1, 2, 0]$ given model λ evaluates to about 0.04. In some family tree Z this probability corresponds to probability $p_{cuv} = p\left(X_1|Z, u_c = \text{Siberian Husky}, v_c = \text{Siberian Husky}\right)$. These probabilities p_{cuv} are pre-computed for each chromosome c given any two breeds u_c, v_c at the start of the dog classification algorithm.

A.5 Mars algorithm as pseudo code

Now, let us formalize this pipeline for the first run: $\forall$ query test dog t I compute

1. Enable 3 meta breeds $AvgDogs = (Ancient, Guard, EuropeanHedge)$.
2. Load SNP-wise transition probabilities for $B+3$ breeds. Meta breeds have transition probabilities computed by averaging over their composing breed collection.
3. Load genotype X_c $\forall$ chromosomes $c \in \{1, \ldots, C\}$ for query dog t.
4. $\forall$ chromosomes, $\forall$ breeds b_i, b_j with $i, j \in \{1, \ldots, B\}$ of test dog t pre-compute probabilities $p(X_c|B_i, B_j)$ which describes the probability of SNP data X_c given chromosome painting $p(X_c|u_c = B_i, v_c = B_j)$.
5. For all lineage trees Z run the Metropolis-Hastings algorithm and do:
   (a) Initialization 1: create a new object for lineage tree Z.
   (b) Initialization 2: select a lineage tree painting L uniformly for Z such that for each leaf all B breeds have equal choice of selection.
   (c) Initialization 3: select a genome painting G uniformly that for each chromosome painting a leaf u_c is chosen randomly (each leaf has equal probability) from the maternal leaves, and one leaf from the paternal leaves is assigned to v_c.
   (d) Start the MCMC run: burn in phase (BIP) with 10,000 iterations. One BIP iteration works as follows:
      i. Note that in the burn phase I do not use haplotype frequencies $p(X_c|u_c, v_c)$ but $p(X_c|u_c, v_c)'$ which I define subsequently. HoIIver, at first, I introduce an averaging model A which averages over possible genome paintings G given the current lineage tree painting L and I define $A = \sum_{d=1}^{Y} \sum_{e=1}^{Y} \frac{p(X_c|u_c^d, v_c^e)}{Y^2}$ where Y runs over the maximum of paternal, maternal leaves. If lineage tree Z is unbalanced (Y different for maternal and paternal side), then that side of tree side with less leaves has some leaves considered twice in the formula for A to match the corresponding leaves on the other side of the tree. And let us denote $M = p(X_c|u_c, v_c)$ as chromosome painting model.
      At BIP iteration 1 I only use the averaging model A which I linearly phase out to exclusively utilize the chromosome painting model M at the end of BIP. Model M is not utilized all at BIP iteration 1 but the algorithm linearly increase its impact and fully use it at BIP iteration 10,000. In other words I compute a linear combination $p'_{cuv} = p(X_c|u_c, v_c)' = r \cdot A + s \cdot M$ with $r = 1, s = 0$ at BIP iteration 1, and $r = 0, s = 1$ at BIP iteration 10,000.
      ii. Compute likelihood $p = p(X|\theta)$ given current $\theta = (Z, L, G)$.
      iii. Jumble lineage tree painting L to obtain new proposal L' with likelihood p', and compute $acc = p' - p$, and uniform random number $g \sim U[0, 1]$.
      iv. If $(acc >= 0)$ OR $(g < 10^{acc})$ accept proposal L', otherwise keep old lineage tree painting L.

v. $\forall$ leaves $step$ in lineage tree Z do

A. Compute likelihood $p = p(X|\theta)$ given current $\theta = (Z, L, G)$.

B. Update lineage tree painting L to obtain L' such that current leaf $step$ is uniformly assigned a new breed. Compute uniform random numbers $g, h \sim U[0, 1]$.

C. Set scalar $qFactor := 0$.

D. If $h < 0.1$, then update genome painting G. Compute likelihood $q = p(X|Z, L', G)$. And, then, compute likelihood $q' = p(X|Z, L', G')$ after genome update G'. Set $qFactor = q' - q$.

E. If $h < 0.1$, then $p' = p(X|Z, L', G')$ else $p' = p(X|Z, L', G)$.

F. Set $acc = p' - p - qFactor$. Then, if $(acc >= 0)$ OR $(g < 10^{acc})$ accept proposal with new breed assignment to leaf (and potential new genome painting G'). Note, that a Metropolis-Hastings algorithm with uniform proposal density is used such that the proposal density cancels out and I only consider log likelihood ratio $acc = p(\theta') - p(\theta)$ of the new proposed sample θ' compared to the previous sample θ. Definition of acc suggests, the acceptance probability in the Metropolis-Hasting algorithms only depends on the lineage tree painting L likelihoods but not on the genome painting G.

(e) Continue MCMC run: main phase (MP) with 500,000 iterations. One MP iteration works as follows. I proceed the same way as in BIP. However,

i. I only use the chromosome painting model $M = p(X_c|u_c, v_c)$.

ii. I record the likelihood $p = p(X|\theta)$ and the associated lineage tree painting L every 5th MCMC iteration.

(f) Compute DIC score as $DIC = p_D + \bar{D}$ where $p_D = \bar{D} - D(\bar{\theta})$ and $D(\theta) = -2 \cdot \log[p(X|\theta)]$. I evaluate the expected value $E_\theta[D] = \bar{D}$ as average over all samples. And $D(\bar{\theta})$ should be evaluated at the average argument $\bar{\theta}$. However, in the Mars code p_d is computed as $\max_\theta \log[p(X|\theta)]$. Furthermore, in the Mars code there is a further factor 2, and I have $DIC = 2 \cdot \left(\frac{1}{T} \sum_\theta D(\theta) + \max_\theta \log[p(X|\theta)]\right)$ where T is the number of MCMC iterations considered.

(g) Compute breed proportion for each of the B breeds. Breed proportion of a particular breed is defined as the frequency of occurrence as one of the leaves in a lineage tree painting L in the MCMC run.

6. Select lineage tree Z with lowest DIC score.

In the second run, the Metropolis-Hastings algorithm is run only on lineage trees Z which are as or more complex than the best lineage tree in the first run, and the meta breeds are no longer included as possible breed assignments. Moreover, in the second run certain breeds will be locked as leaves in the lineage tree and are not substituted during the MCMC run. The locked-in breeds are those breeds which exceed a certain minimum probability estimated as its relative frequency during the first MCMC run. Finally, once the second run has completed the DIC-optimal lineage tree is chosen. Then, the algorithm scans through the MCMC iteration recordings of the second run which contains the lineage tree painting for every 5th MCMC iteration to compute breed proportions for all breeds. Those breeds which are recognized as high- or medium-proportion according to a breed-specific cut-off table are the final assignments of breeds to the leaves of the optimal lineage tree.

-	True class p	True class n	Row sums
Predicted class p'	True Positives (TP)	False Positives (FP)	P' = TP + FP
Predicted class n'	False Negatives (FN)	True Negatives (TN)	N' = FN + TN
Column sums	P=TP+FN	N=FP+TN	-

Table A.18: Confusion matrix is given as contingency table and shows the relationship between true and predicted class.

A.6 Multiclass ROC Curves

Performance measures in binary classification are widely used (Fawcett, 2004). In these setups each test sample X^i comes from one of two classes. Each X^i either has a positive p or negative n class label as true class. A classifier will take X^i as input and produces class labels $\{p', n'\}$ where Y refers to the positive class and N to the negative class. The Cartesian product of true and predicted class yields a confusion matrix with four options as shown in table A.18:

- True Positive (TP): test sample has positive true class p and is predicted as p'.
- True Negative (TN): test sample has positive true class n and is predicted as n'.
- False Positive (FP): test sample has positive true class n and is predicted as p'.
- False Negative (FN): test sample has positive true class p and is predicted as n'.

In the ancestry prediction problem breeds which are part of the true ancestry are mapped to class p while breeds not part of the ancestry are mapped to n. Similarly, the classifier predicts class p' for breeds to be hypothesized to be part of the ancestry and class n' for breeds which are not prediced to be part of the ancestry.

However, the ancestry prediction problem represents a more advanced setting because for each test sample X^i I can have multiples TPs, FPs, TNs and FNs. Therefore, each performance measures will be defined on a sample-by-sample basis based on TP_i, FP_i, TN_i, FN_i.

Let k corresponds to the number of slots in the family tree, e.g. family tree ABC has slots for up to 3 breeds. Then, I know $TP_i \in \{0, \ldots, k\}$, $FN_i \in \{0, \ldots, k\}$ and $TP_i + FN_i = k$. Furthermore, I know that $TN_i \in \{0, \ldots, b_M - k\}$, $FP_i \in \{0, \ldots, b_M - k\}$ and $TN_i + FP_i = b_M - k$.

Then, given the confusion matrix I can define performance measures. In particular, I are interested in breeds that are part of the breed ancestry, and thus, focus on performance measure related to the positive class. Furthermore, accuracy, which counts the number of correct predictions across $\{p, n\}$ is not a useful measure here because each test sample is expected to have many more breeds in the negative class, i.e. not to be part of the ancestry, than breeds in the positive class. I will look at two performance measures. The first measure is *sensitivity*, which is also known as true positive rate (TPR), hit rate, recall, and the second measure is *positive predictive value (PPV)*, also referred to as precision. The first measure is sensitivity defines how many of positives of all positives samples I detected and is defined as

$$TPR_i = \frac{TP_i}{P_i} = \frac{TP_i}{TP_i + FN_i}.$$

And PPV determines what fraction of positive predicted samples are actually true positives. PPV is computed as

$$PPV_i = \frac{TP_i}{P'_i} = \frac{TP_i}{TP_i + FP_i}.$$

However, I would like to weigh PPV_i and TPR_i. One way to do that is suggested in information retrieval by establishing the F-measure F_β

$$F_i^\beta = \frac{\left(1+\beta^2\right) \cdot PPV_i \cdot TPR_i}{(\beta^2 \cdot PPV_i) + TPR_i}$$

assuming I decide β times as much relevance to TPR_i (recall) then PPV_i (precision). In other words, it is β times as important to obtain more true positives (breeds which are part of ancestry) than it is to avoid introducing false positives (breeds which are not part of the ancestry). I equally weigh both measures setting $\beta = 1$ to yield the F1-score

$$F_i^1 = \frac{2 \cdot PPV_i \cdot TPR_i}{PPV_i + TPR_i}$$

which is the harmonic means of recall and precision. Finally, I compute overall TPR as $TPR = \frac{1}{N_Z}\sum_{i=1}^{N_Z} TPR_i$, overall PPV as $PPV = \frac{1}{N_Z}\sum_{i=1}^{N_Z} PPV_i$ and overall F1-score as $F^1 = \frac{1}{N_Z}\sum_{i=1}^{N_Z} F_i^1$.

ROC curves using sensitivity and positive positive predictive values as axes provide a two-dimensional representation of classifier performance. Based on these ROC curves I compute the following two scalar summary statistics: area under this curve (AUC) (Hanely and McNeil, 1982; Bradley, 1997) provides a measure of quality of prediction. As alternative measure I determine the maximum F_1 value, maxF1, taken over the breed fraction cut-off interval [0,1]. These two measures tend to have correlation of 0.95 or higher.

A.7 DBAncestry Results

A.7.1 Breed estimation view

True breed ancestry Median [1st,3rd] Quartiles	UCBC Long,ULP	UCBC Long,BSLP	UCBC Short,ULP	UCBC Short,BSLP	CCBC Long,ULP	CCBC Long,BSLP	CCBC Short,ULP	CCBC Short,BSLP
100%	**0.73** [0.61,0.81]	**0.73** [0.61,0.81]	**0.70** [0.58,0.79]	**0.64** [0.52,0.74]	**0.87** [0.77,0.93]	**0.87** [0.77,0.94]	**0.85** [0.74,0.92]	**0.81** [0.70,0.88]
50%	**0.32** [0.26,0.38]	**0.32** [0.26,0.37]	**0.30** [0.24,0.36]	**0.27** [0.21,0.33]	**0.41** [0.35,0.46]	**0.41** [0.35,0.46]	**0.38** [0.32,0.44]	**0.36** [0.29,0.42]
25%	**0.15** [0.13,0.19]	**0.15** [0.13,0.19]	**0.14** [0.12,0.19]	**0.13** [0.11,0.17]	**0.20** [0.17,0.25]	**0.20** [0.17,0.25]	**0.19** [0.15,0.24]	**0.18** [0.14,0.23]
12.5%	**0.10** [0.02,0.13]	**0.10** [0.02,0.13]	**0.08** [0.01,0.13]	**0.07** [0.01,0.13]	**0.12** [0.03,0.16]	**0.12** [0.03,0.16]	**0.11** [0.02,0.16]	**0.09** [0.013,0.16]
0%	**0.001** [0.00,0.00]	**0.001** [0.00,0.00]	**0.00** [0.00,0.00]	**0.00** [0.00,0.00]	**0.00** [0.00,0.00]	**0.00** [0.00,0.00]	**0.00** [0.00,0.00]	**0.00** [0.00,0.00]

Table A.19: **Breed proportion estimates:** Median, lower and upper quartile estimates of the raw (UCBC) and re-scaled (CCBC) breed proportion for all possible true ancestral levels are shown. In this experiment I look at combinations of short/long MCMC run-length and uniform (ULP) vs. breed proposal biased breed (BSLP) proposals. Breed fraction estimates are computed taking predictions from all lineage trees into account.

A.7.2 Breed classification view

Results for the classification view are given in Table A.20 and a visual representation for the long run using the ULP update proposal is shown in Figure A.3. It can be seen that the areas under the curves (AUC) are almost identical for the uniform and breed-biased update proposals in the long MCMC run. Although if I were to account for more than the leading two digits AUC according to Table A.21 the reader sees that even for the long run the ULP variant still has a very minor AUC performance advantage. Similar arguments apply for the F_1 score criterion which can be seen in Table A.22. Furthermore, according to Table A.20 there is a much more substantial benefit of the longer run over the shorter one in the classification view, e.g. for the most complex lineage tree ABCDEFGH there is a 10 percent increase in the area under the curve. In other words, although the long MCMC run only offers up to 3 percent improvement in the estimation of breed proportions there is a large improvement in classifying breeds correctly. In general, AUC scores are much better than random guessing as shown in Figure A.4. In particular, I obtain perfect classification for pure breed dogs, and even for the complex lineage ABCDEFGH I get an AUC score of 0.81.

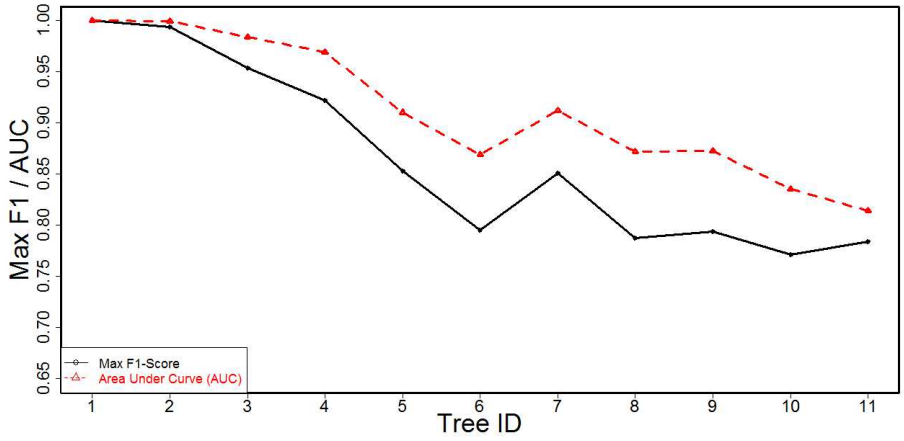
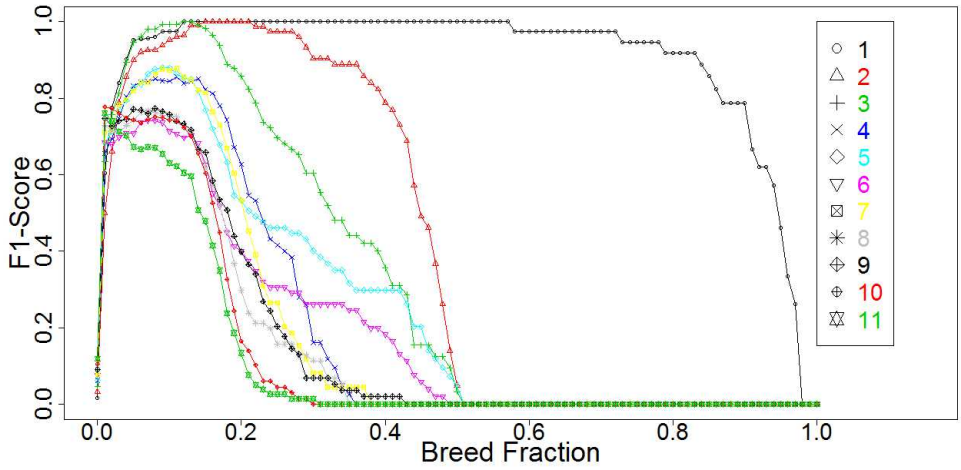
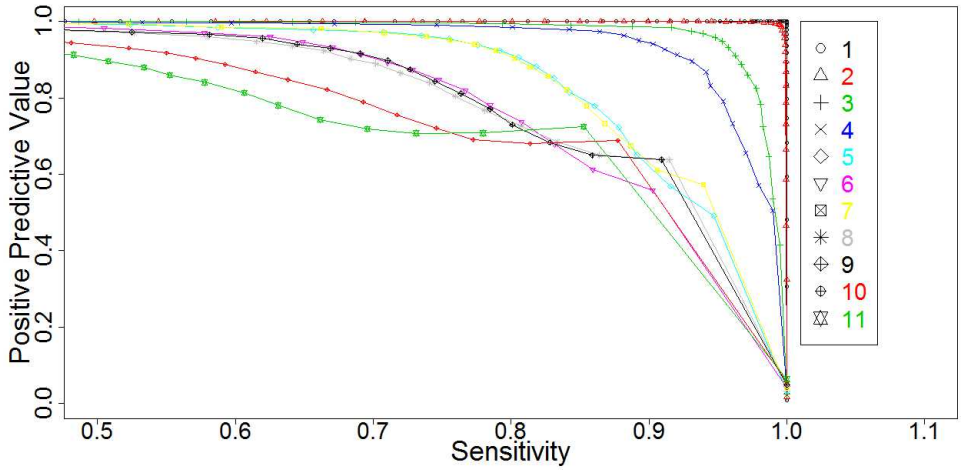


Figure A.3: This figure shows CCBC-based results for the area under the curve and the F_1 criterion using the long MCMC run with ULP update proposal for all 11 lineage trees from Figure 1.2. A visualization of the actual performance plots are shown in Figure A.4. Firstly, it can be seen that both criteria are very correlated, i.e. have a correlation of 0.97. For the simplest tree both performance criteria show almost perfect performance. After that, both criteria decrease quickly until intermediate tree complexity, i.e. around 0.86 for AUC and 0.8 for F_1 score, and then continue to be reduced moderately at a level of 0.81 (AUC) and 0.78 (F_1) for the most complex tree, respectively.



(a) F1-Score, Long, ULP



(b) PPV/SENS-Score, Long, ULP

Figure A.4: **DBAncestry classification performance:** For all 11 lineage trees these two plots show PPV vs. sensitivity and F_1 score as a function of breed proportion, respectively. These results are based on CCBC-based breed proportion estimates from the long MCMC run with a ULP update proposal. According to the AUC-plot it can be seen that the AUC score is near perfect for trees AAAAAAAA (tree 1) and AAAABBBB (tree 2) and very high for AAAABBCC (tree 3) and AABBCCDD (tree 4). Furthermore, tree 5-11 have similar AUC scores at a high level. The F_1 plot suggests that almost independently of the chosen breed proportion cut-off there is a very high F_1 score for tree 1. Even for tree 2 the F_1 score only starts to decrease sharply at around 0.4. All other trees typically have optimal cut-off values of less than 0.2. While the simplest three trees achieve almost perfect F_1 score, with increasing amount of complexity the best F_1 score goes down to 0.78 for the most complex tree ABCDEFGH.

Tree ID -	Tree Label -	CCBC Long,ULP	CCBC Long,BSLP	CCBC Short,ULP
1	AAAAAAAA	**1.00**	**1.00**	**0.99**
2	AAAABBBB	**1.00**	**1.00**	**1.00**
3	AAAABBCC	**0.98**	**0.98**	**0.96**
4	AABBCCDD	**0.97**	**0.97**	**0.94**
5	AAAABBCD	**0.91**	**0.91**	**0.86**
6	AAAABCDE	**0.87**	**0.87**	**0.81**
7	AABBCCDE	**0.91**	**0.91**	**0.86**
8	AABBCDEF	**0.87**	**0.87**	**0.80**
9	AABCDDEF	**0.87**	**0.87**	**0.79**
10	AABCDEFG	**0.84**	**0.82**	**0.74**
11	ABCDEFGH	**0.81**	**0.81**	**0.71**

Table A.20: **Classification view criterion area under the curve (AUC):** TAUC values which were obtained from numerical integration are shown. AUC values evaluated for each lineage tree using UCBC and CCBC breed proportion estimates. Although according to Table 4.2 the long run only yields modest improvements in breed proportion estimates there is a considerable performance enhancement to correctly classify breeds. In particular, for complex lineage trees 5 to 11 there is an improvement from 0.05 to 0.1 for the AUC criterion. Furthermore, the finding from Table 4.2 are confirmed that for the long run performance of the ULP and BSLP updates are almost identical. Full AUC results listing the first four leading digits are shown in Table A.21 for all combinations of MCMC run length, update proposals and breed proportion computations.

Tree ID -	Tree Label -	UCBC Long,ULP	UCBC Long,BSLP	UCBC Short,ULP	UCBC Short,BLP	CCBC Long,ULP	CCBC Long,BSLP	CCBC Short,ULP	CCBC Short,BSLP
1	AAAAAAAA	**1.0000**	**1.0000**	**0.9989**	**0.9996**	**1.000**	**1.000**	**0.9938**	**0.9998**
2	AAAABBBB	**0.9992**	**0.9989**	**0.9977**	**0.9952**	**0.9994**	**0.9991**	**0.9972**	**0.9961**
3	AAAABBCC	**0.9839**	**0.9824**	**0.9638**	**0.9459**	**0.9843**	**0.9824**	**0.9649**	**0.9471**
4	AABBCCDD	**0.9665**	**0.9649**	**0.9341**	**0.9057**	**0.9694**	**0.9685**	**0.9386**	**0.9108**
5	AAAABBCD	**0.9041**	**0.9026**	**0.8564**	**0.8291**	**0.9103**	**0.9100**	**0.8641**	**0.8425**
6	AAAABCDE	**0.8552**	**0.8510**	**0.7944**	**0.7518**	**0.8689**	**0.8655**	**0.8105**	**0.7753**
7	AABBCCDE	**0.9005**	**0.8980**	**0.8373**	**0.7978**	**0.9117**	**0.9089**	**0.8550**	**0.8212**
8	AABBCDEF	**0.8514**	**0.8467**	**0.7686**	**0.7237**	**0.8721**	**0.8688**	**0.7984**	**0.7605**
9	AABCDDEF	**0.8510**	**0.8431**	**0.7612**	**0.7279**	**0.8727**	**0.8651**	**0.7903**	**0.7620**
10	AABCDEFG	**0.8002**	**0.7891**	**0.6964**	**0.6451**	**0.8350**	**0.8291**	**0.7421**	**0.6974**
11	ABCDEFGH	**0.7656**	**0.7568**	**0.6561**	**0.5812**	**0.8142**	**0.8078**	**0.7112**	**0.6452**

Table A.21: **Area under the curve (AUC):** AUC values which were obtained from numerical integration are shown. AUC values evaluated for each lineage tree using raw (UCBC) and re-scaled (CCBC) breed proportion estimates. In this experiment I look at combinations of short/long MCMC run-length and uniform (ULP) vs. breed proposal biased breed (BSLP) proposals.

Tree ID -	Tree Label -	UCBC Long,ULP	UCBC Long,BSLP	UCBC Short,ULP	UCBC Short,BLP	CCBC Long,ULP	CCBC Long,BSLP	CCBC Short,ULP	CCBC Short,BSLP
1	AAAAAAAA	**1.0000**	**0.9997**	**0.9997**	**0.9989**	**1.0000**	**1.0000**	**1.0000**	**0.9989**
2	AAAABBBB	**0.9928**	**0.9937**	**0.9822**	**0.9770**	**0.9942**	**0.9932**	**0.9837**	**0.9848**
3	AAAABBCC	**0.9525**	**0.9522**	**0.9178**	**0.8927**	**0.9540**	**0.9538**	**0.9188**	**0.8975**
4	AABBCCDD	**0.9210**	**0.9195**	**0.8872**	**0.8517**	**0.9220**	**0.9197**	**0.8876**	**0.8517**
5	AAAABBCD	**0.8541**	**0.8523**	**0.8105**	**0.7809**	**0.8527**	**0.8520**	**0.8103**	**0.7799**
6	AAAABCDE	**0.7982**	**0.7896**	**0.7474**	**0.7165**	**0.7948**	**0.7886**	**0.7450**	**0.7147**
7	AABBCCDE	**0.8512**	**0.8468**	**0.7842**	**0.7539**	**0.8511**	**0.8460**	**0.7850**	**0.7570**
8	AABBCDEF	**0.787**	**0.7824**	**0.7220**	**0.6834**	**0.7873**	**0.7841**	**0.7189**	**0.6827**
9	AABCDDEF	**0.7930**	**0.7869**	**0.7167**	**0.6813**	**0.7938**	**0.7875**	**0.7159**	**0.6831**
10	AABCDEFG	**0.7351**	**0.7264**	**0.6622**	**0.6171**	**0.7711**	**0.7675**	**0.6973**	**0.6750**
11	ABCDEFGH	**0.6969**	**0.6890**	**0.6192**	**0.5656**	**0.7837**	**0.7803**	**0.7046**	**0.6644**

Table A.22: **Maximum value for F_1 score (maxF1):** Values for maxF1 values are shown which were obtained from numerical integration. The maxF1 values evaluated for each lineage tree using raw (UCBC) and re-scaled (CCBC) breed proportion estimates. In this experiment I look at combinations of short/long MCMC run-length and uniform (ULP) vs. breed proposal biased breed (BSLP) proposals.

A.7.3 Inferring ancestral boundaries: quantile view

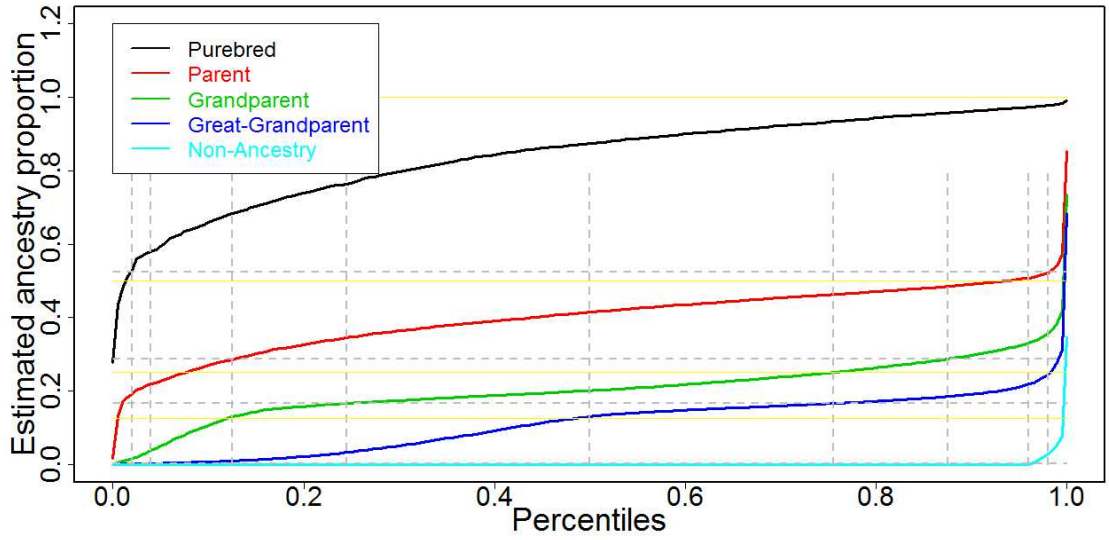


Figure A.5: Estimated global breed contributions (CCBC, ULP, Long) percentiles are shown for each TAP. Each coloured curve corresponds to one of the ancestral levels, i.e. purebred, parent, gp, ggp and non-ancestry. Furthermore, horizontal yellow lines illustrate the true breed contributions for the different TAP levels. Vertical grey dashed lines correspond to the inferred breed contribution cut-offs between consecutive TAP levels. Related to that horizontal grey dashed lines show quantiles associated to those breed cut-offs.

In Figure A.5 I plot the estimated breed contributions as a function of the quantiles (200-quantiles). To find the optimal breed contribution (BC) cut-offs between consecutive TAP levels I minimized the α quantile of the more recent TAP element and the least recent TAP level, i.e. for the purebred and parental level I am required to find that α which minimizes the difference $[BC(\text{Pure}_{\alpha}) - BC(\text{P}_{1-\alpha})]$. For example, according to Table 4.3 the cut-off between purebred-parent is $BC(\text{Pure}_{\alpha=0.02}) = BC(\text{P}_{1-\alpha=0.98}) = 0.52$, i.e. at a breed contribution value of 0.52 which corresponds to 0.02-th quantile for the purebred and the 0.98-th quantile for the parent. Finally, there is a vertical grey dashed line at 0.5 to indicate the median breed contribution for each TAP level.

A.8 ChromoPainter Results: classification view

In Figures A.6, A.7 it can A.3 that the AUC score for tree AAAAAAAA assumes a high value of 0.81 suggesting that the ChromoPainter NNLS variant works reliably at recognizing synthetic purebreed dogs. However, even for tree AAAABBBB the AUC score drops sharply by 0.32 to a low value 0.51. For more complex lineage trees the performance drops to even lower AUC scores of 0.37 to 0.43. Furthermore, recall that according to Table A.20 for the four simplest lineage trees (i.e. AAAAAAAA, AAAABBBB, AAAABBCC, AABBCCDD) DBAncestry achieves a ROC score of above 0.94 and even for the most complex tree ABCDEFGH AUC score takes a value of 0.71. Finally, the AUC score for pure breed tree AAAAAAAA of 0.81 using ChromoPainter NNLS is still achieved for DBAncestry at a high lineage tree complexity, i.e. lineages AABBCDEF (tree 6 with 1 parent, 4 ggps), AABBCDEF (tree 8 with 2 gp, 4 ggps) and AABCDDEF (tree 9 with 2 gp, 4 ggps) have AUC scores of 0.81, 0.8 and 0.79, respectively, which implies that DBAncestry works much better at inferring less recent ancestry contributions.

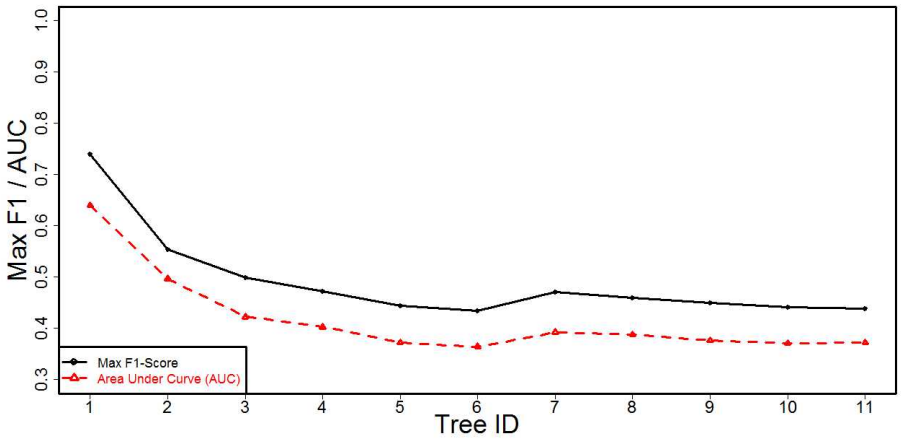
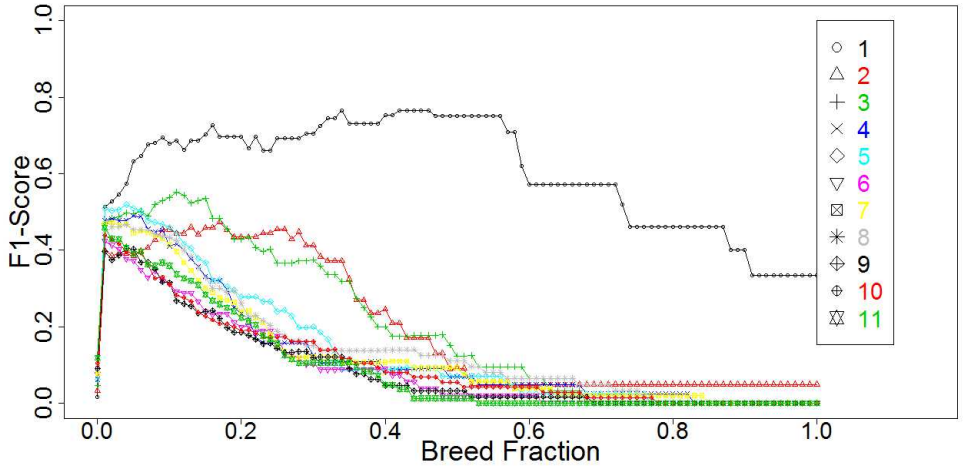
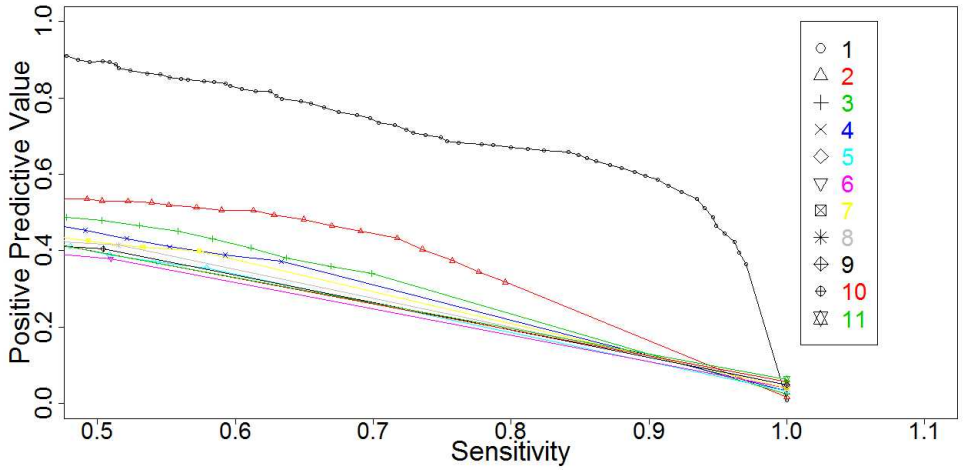


Figure A.6: **ChromoPainter NNLS classification performance:** This figure shows the AUC score and F_1 criterion for all 11 lineage trees. Prediction quality for pure breed tree AAAAAAAA is reasonably good at 0.83 (AUC) and 0.74 (F_1). Then, for more complex lineage trees classification performance drops sharply to [0.37,0,51] (AUC) and [0.44,0.55] (F_1) which suggests ChromoPainter does not deal well with lineages involving multiple breeds.



(a) F1-Score



(b) PPV/SENS-Score

Figure A.7: **ChromoPainter NNLS classification performance:** For all 11 lineage trees these two plots show PPV vs. sensitivity and F_1 score as a function of breed proportion, respectively. These results are based on CCBC-based breed proportion estimates using ChromoPainter with the NNLS variant. According to the AUC-plot it can be seen that the AUC score is high for tree AAAAAAAA (tree 1) with a value of 0.81, i.e. the algorithm performs reasonably well for pure breed synthetic test dogs. For more complex trees there sharp decline in performance to an AUC: while tree AAAABBBB (tree 2) has an AUC value of 0.51 all other lineage trees have an AUC score between 0.37 and 0.43. The same behaviour can be observed for the F_1 score plots: in particular this figure shows that tree AAAAAAAA has a much higher F_1 score value of 0.75 compared to the next best value of 0.45 for AAAABBBB (tree 2).

Bibliography

AKEY, J. M., RUHE, A. L., AKEY, D. T., WONG, A. K., CONNELLY, C. F., MADEOY, J., NICHOLAS, T. J. and NEFF, M. W. (2010). Tracking footprints of artificial selection in the dog genome. *Proceedings of the National Academy of Sciences*, **107** 1160–1165.

ALDRICH, M. C., SELVIN, S., HANSEN, H. M., BARCELLOS, L. F., WRENSCH, M. R., SISON, J. D., QUESENBERRY, C. P., KITTLES, R. A., SILVA, G., BUFFLER, P. A. ET AL. (2008). Comparison of statistical methods for estimating genetic admixture in a lung cancer study of African Americans and Latinos. *American journal of epidemiology*, **168** 1035–1046.

ALEXANDER, D. H., NOVEMBRE, J. and LANGE, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19** 1655–1664. Admixture.

ALI, S. and SILVEY, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)* 131–142.

ALLISON, M. (2012). Direct-to-consumer genomics reinvents itself. *Nature biotechnology*, **30** 1027–1029.

ALTSHULER, D., DALY, M. J. and LANDER, E. S. (2008). Genetic mapping in human disease. *science*, **322** 881–888.

ALTSHULER, D., DURBIN, R., ABECASIS, G., BENTLEY, D., CHAKRAVARTI, A., CLARK, A., COLLINS, F., DE LA VEGA, F., DONNELLY, P., EGHOLM, M. ET AL. (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467** 1061–73.

AMERICAN VETERINARY MEDICAL ASSOCIATION (2002). U.S. pet ownership and demographics sourcebook.

ARDALAN, A., KLUETSCH, C. F., ZHANG, A.-B., ERDOGAN, M., UHLÉN, M., HOUSHMAND, M., TEPELI, C., ASHTIANI, S. R. M. and SAVOLAINEN, P. (2011). Comprehensive study of mtDNA among Southwest Asian dogs contradicts independent domestication of wolf, but implies dog–wolf hybridization. *Ecology and evolution*, **1** 373–385.

ARGYRIOU, A., EVGENIOU, T. and PONTIL, M. (2008). Convex multi-task feature learning. *Machine Learning*, **73** 243–272.

BALDING, D., WEALE, M., RICHARDS, M. and THOMAS, M. (2010). Genetic and isotopic analysis and the UK Border Agency. *Significance*, **7** 58–61.

BALDING, D. J. and NICHOLS, R. A. (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International*, **64** 125–140.

BANNASCH, D. L., BANNASCH, M. J., RYUN, J. R., FAMULA, T. R. and PEDERSEN, N. C. (2005). Y chromosome haplotype analysis in purebred dogs. *Mammalian genome*, **16** 273–280.

BARAN, Y., PASANIUC, B., SANKARARAMAN, S., TORGERSON, D. G., GIGNOUX, C., ENG, C., RODRIGUEZ-CINTRON, W., CHAPELA, R., FORD, J. G., AVILA, P. C. ET AL. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, **28** 1359–1367. Lamp LD.

BATESON, P. and SARGAN, D. R. (2012). Analysis of the canine genome and canine health: A commentary. *The Veterinary Journal*.

BAUMUNG, R. and SOLKNER, J. (2003). Pedigree and marker information requirements to monitor genetic variability. *Genetics Selection Evolution*, **35** 369–384.

BEAL, M. J., GHAHRAMANI, Z. and RASMUSSEN, C. E. (2002). The infinite hidden Markov model. *Advances in neural information processing systems*, **14** 577–584.

BENNETT, K. and EMBRECHTS, M. (2003). An optimization perspective on kernel partial least squares regression. *Nato Science Series sub series III computer and systems sciences*, **190** 227–250.

BJÖRNERFELDT, S., HAILER, F., NORD, M. and VILÀ, C. (2008). Assortative mating and fragmentation within dog breeds. *BMC evolutionary biology*, **8** 28.

BLOSS, C. S., DARST, B. F., TOPOL, E. J. and SCHORK, N. J. (2011). Direct-to-consumer personalized genomic testing. *Human molecular genetics*, **20** R132–R141.

BOLNICK, D. A., FULLWILEY, D., DUSTER, T., COOPER, R. S., FUJIMURA, J. H., KAHN, J., KAUFMAN, J. S., MARKS, J., MORNING, A., NELSON, A. ET AL. (2007). The science and business of genetic ancestry testing. *Science*, **318** 399.

BOULESTEIX, A.-L. and STRIMMER, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, **8** 32–44.

BOYKO, A. (2011). The domestic dog: man's best friend in the genomic era. *Genome Biology*, **12** 216.

BOYKO, A., QUIGNON, P., LI, L., SCHOENEBECK, J., DEGENHARDT, J., LOHMUELLER, K., ZHAO, K., BRISBIN, A., PARKER, H., CARGILL, M. ET AL. (2010). A simple genetic architecture underlies morphological variation in dogs. *PLoS Biology*, **8** e1000451.

BRADLEY, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, **30** 1145–1159.

BREIMAN, L. (1993). *Classification and regression trees*. CRC press.

BREIMAN, L. and FRIEDMAN, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59** 3–54.

BRENT, R. P. (1973). *Algorithms for minimization without derivatives*. Courier Dover Publications.

BRISBIN, A., BRYC, K., BYRNES, J., ZAKHARIA, F., OMBERG, L., DEGENHARDT, J., REYNOLDS, A., OSTRER, H., MEZEY, J. G. and BUSTAMANTE, C. D. (2012). PCAdmix: Principal Components-Based Assignment of Ancestry along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Human biology*, **84** 343–364. PCAdmix.

BROOKS, S., GELMAN, A., JONES, G. and MENG, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. Taylor & Francis US.

BROWN, L. D., CAI, T. T. and DASGUPTA, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, **16** 101–117.

BROWNING, S. (2006). Multilocus association mapping using variable-length Markov chains. *The American Journal of Human Genetics*, **78** 903–913.

BROWNING, S. and WEIR, B. (2010). Population structure with localized haplotype clusters. *Genetics*, **185** 1337.

BROWNING, S. R. and BROWNING, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics*, **81** 1084. VNLC.

BRUMFIELD, R. T., BEERLI, P., NICKERSON, D. A. and EDWARDS, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, **18** 249–256.

BRYC, K., AUTON, A., NELSON, M. R., OKSENBERG, J. R., HAUSER, S. L., WILLIAMS, S., FROMENT, A., BODO, J.-M., WAMBEBE, C., TISHKOFF, S. A. ET AL. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences*, **107** 786–791. RFmix PCA.

BUDOWLE, B. and VAN DAAL, A. (2008). Forensically relevant SNP classes. *BioTechniques: The international journal of life science methods*, **44** 603.

BURGES, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, **2** 121–167.

BURNETT, M. S., STRAIN, K. J., LESNICK, T. G., DE ANDRADE, M., ROCCA, W. A. and MARAGANORE, D. M. (2006). Reliability of self-reported ancestry among siblings: implications for genetic association studies. *American journal of epidemiology*, **163** 486–492.

CALBOLI, F., SAMPSON, J., FRETWELL, N. and BALDING, D. (2008). Population structure and inbreeding from pedigree analysis of purebred dogs. *Genetics*, **179** 593.

CAMPBELL, D., DUCHESNE, P. and BERNATCHEZ, L. (2003). AFLP utility for population assignment studies: analytical investigation and empirical comparison with microsatellites. *Molecular Ecology*, **12** 1979–1991.

CANNON, A., SCHWARCZ, H. P. and KNYF, M. (1999). Marine-based subsistence trends and the stable isotope analysis of dog bones from Namu, British Columbia. *Journal of archaeological science*, **26** 399–407.

CATTELL, R. B. (1952). *Factor Analysis: An Introduction and Manual for the Psychologyst and Social Scientist*. Harper; H. Hamilton.

CAVALLI-SFORZA, L. L., MENOZZI, P. and PIAZZA, A. (1993). Demic expansions and human evolution. *Science*, **259** 639–646.

CHANG, M. L., YOKOYAMA, J. S., BRANSON, N., DYER, D. J., HITTE, C., OVERALL, K. L. and HAMILTON, S. P. (2009). Intrabreed stratification related to divergent selection regimes in purebred dogs may affect the interpretation of genetic association studies. *Journal of Heredity*, **100** S28–S36.

CHIANG, C. W., GAJDOS, Z. K., KORN, J. M., KURUVILLA, F. G., BUTLER, J. L., HACKETT, R., GUIDUCCI, C., NGUYEN, T. T., WILKS, R., FORRESTER, T. ET AL. (2010). Rapid assessment of genetic ancestry in populations of unknown origin by genome-wide genotyping of pooled samples. *PLoS genetics*, **6** e1000866.

CHURCHHOUSE, C. (2012). *Bayesian Methods for Estimating Human Ancestry Using Whole Genome SNP Data*. Ph.D. thesis, University of Oxford.

CHURCHHOUSE, C. and MARCHINI, J. (2013). Multiway Admixture Deconvolution Using Phased or Unphased Ancestral Panels. *Genetic epidemiology*, **37** 1–12. MultiMix.

CONRAD, D. F., JAKOBSSON, M., COOP, G., WEN, X., WALL, J. D., ROSENBERG, N. A. and PRITCHARD, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature genetics*, **38** 1251–1260.

CRAWFORD, D. C., BHANGALE, T., LI, N., HELLENTHAL, G., RIEDER, M. J., NICKERSON, D. A. and STEPHENS, M. (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature genetics*, **36** 700–706.

CROCHEMORE, M., HANCART, C. and LECROQ, T. (2007). *Algorithms on strings*. Cambridge University Press.

CROWLEY, J. and ADELMAN, B. (1998). The Complete Dog Book. *Official Publication of the American Kennel Club*, **19** 27–625.

CSISZAR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Stud. Sci. Math. Hung.*, **2** 299–318.

CSISZÁR, I. and SHIELDS, P. C. (2004). *Information theory and statistics: A tutorial*. Now Publishers Inc.

CUSSENS, J., BARTLETT, M., JONES, E. M. and SHEEHAN, N. A. (2013). Maximum likelihood pedigree reconstruction using integer linear programming. *Genetic Epidemiology*, **37** 69–83.

DALY, M. J., RIOUX, J. D., SCHAFFNER, S. F., HUDSON, T. J. and LANDER, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature genetics*, **29** 229–232.

DAVISON, S. and FRETWELL, N. (2012). Mars Veterinary Project Meeting.

DEZA, M. M. and DEZA, E. (2009). *Encyclopedia of distances*. Springer.

DING, C. H. and DUBCHAK, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17** 349–358.

DING, Z., OSKARSSON, M., ARDALAN, A., ANGLEBY, H., DAHLGREN, L.-G., TEPELI, C., KIRKNESS, E., SAVOLAINEN, P. and ZHANG, Y. (2011). Origins of domestic dog in Southern East Asia is supported by analysis of Y-chromosome DNA. *Heredity*, **108** 507–514.

DNA MY DOG (2013). `http://dnamydog.com`. Accessed: 2013-08-01.

DO, C. B., DURAND, J. M., MACPHERSON, J. M., NAUGHTON, B., MOUNTAIN, J. L. and "23ANDME" (2012). A scalable pipeline for local ancestry inference using thousands of reference individuals.

DOG DNA (2013). `http://www.dog-dna.com`. Accessed: 2013-08-01.

DUSTER, T. (2006). Deep roots and tangled branches.

DUSTER, T. (2009). Ancestry testing and DNA: Uses, limits - and caveat emptor. *Genewatch*, **22** 16–17.

ELLIOTT, L. and TEH, Y. W. (2012). Scalable imputation of genetic data with a discrete fragmentation-coagulation process. In *Advances in Neural Information Processing Systems 25*. 2861–2869.

EVERITT, B. and HOTHORN, T. (2011). *Cluster analysis*. Springer.

EXCOFFIER, L. (2001). Analysis of population subdivision. In *Handbook of statistical genetics* (D. J. Balding, M. Bishop and C. Cannings, eds.), chap. 29. Wiley Online Library, 980–1020.

FALUSH, D., STEPHENS, M. and PRITCHARD, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164** 1567–1587. Structure with LD.

FAWCETT, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, **31** 1–38.

FELSENSTEIN, J. (2013). Theoretical evolutionary genetics. Book.

FOULKES, A. (2009). *Applied Statistical Genetics with R: For Population-based Association Studies*. Springer Verlag.

FREUND, R. J. and WILSON, W. J. (1998). *Regression analysis: Statistical modeling of a response variable*. Academic Press.

FREUND, Y. and SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, **55** 119–139.

FRKONJA, A., GREDLER, B., SCHNYDER, U., CURIK, I. and SOELKNER, J. (2012). Prediction of breed composition in an admixed cattle population. *Animal Genetics*, **43** 696–703.

FRKONJA, A., GREDLER, B., SCHNYDER, U., CURIK, I. and SÖLKNER, J. (2011). How to Use Fewer Markers in Admixture Studies. *Agriculturae Conspectus Scientificus (ACS)*, **76** 187–190.

GATTEPAILLE, L. M. and JAKOBSSON, M. (2012). Combining markers into haplotypes can improve population structure inference. *Genetics*, **190** 159–174.

GERMONPRÉ, M., SABLIN, M., STEVENS, R., HEDGES, R., HOFREITER, M., STILLER, M. and DESPRÉS, V. (2009). Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. *Journal of Archaeological Science*, **36** 473–490.

GHAHRAMANI, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **371**.

GHAHRAMANI, Z. and JORDAN, M. I. (1997). Factorial hidden Markov models. *Machine learning*, **29** 245–273.

GIBBS, R. A., BELMONT, J. W., HARDENBOL, P., WILLIS, T. D., YU, F., YANG, H., CH'ANG, L.-Y., HUANG, W., LIU, B., SHEN, Y. ET AL. (2003). The international HapMap project. *Nature*, **426** 789–796.

GILKS, W., RICHARDSON, S. and SPIEGELHALTER, D. (1996). *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC.

GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73** 123–214.

GIVENS, G. H. and HOETING, J. A. (2012). *Computational statistics*, vol. 708. John Wiley & Sons.

GORBACH, D., MAKGAHLELA, M., REECY, J., KEMP, S., BALTENWECK, I., OUMA, R., MWAI, O., MARSHALL, K., MURDOCH, B., MOORE, S. ET AL. (2010). Use of SNP genotyping to determine pedigree and breed composition of dairy cattle in Kenya. *Journal of animal breeding and genetics*, **127** 348–351.

GOWER, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53** 325–338.

GOWER, J. C. and LEGENDRE, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of classification*, **3** 5–48.

GRAY, I. C., CAMPBELL, D. A. and SPURR, N. K. (2000). Single nucleotide polymorphisms as tools in human genetics. *Human molecular genetics*, **9** 2403–2408.

HAASL, R. J., MCCARTY, C. A. and PAYSEUR, B. A. (2012). Genetic ancestry inference using support vector machines, and the active emergence of a unique American population. *European Journal of Human Genetics*.

HADJIELEFTHERIOU, M. and SRIVASTAVA, D. (2011). *Approximate string processing*. Now Pub.

HAIR, J., ANDERSON, R., BABIN, B. and BLACK, W. (2009). *Multivariate data analysis: A global perspective*. Prentice Hall.

HAJILOO, M., SAPKOTA, Y., MACKEY, J. R., ROBSON, P., GREINER, R., DAMARAJU, S. ET AL. (2013). ETHNOPRED: a novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction. *BMC bioinformatics*, **14** 61.

HALPERIN, E. and ESKIN, E. (2004). Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, **20** 1842–1849.

HAND, D. J. and TILL, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, **45** 171–186.

HANELY, J. and MCNEIL, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143** 29–36.

HÄRDLE, W. K. and SIMAR, L. (2012). *Applied multivariate statistical analysis*. Springer.

HARRIS, D., IMPERATO, P. and OKEN, B. (1974). Dog bites–an unrecognized epidemic. *Bulletin of the New York Academy of Medicine*, **50** 981.

HEATH, S. C., GUT, I. G., BRENNAN, P., MCKAY, J. D., BENCKO, V., FABIANOVA, E., FORETOVA, L., GEORGES, M., JANOUT, V., KABESCH, M. ET AL. (2008). Investigation of the fine structure of European populations with applications to disease association studies. *European Journal of Human Genetics*, **16** 1413–1429.

HEDRICK, P. and ANDERSSON, L. (2010). Are dogs genetically special&quest. *Heredity*, **106** 712–713.

HELLENTHAL, G. (2006). *Exploring rates and patterns of variability in gene conversion and crossover in the human genome*. Ph.D. thesis, University of Washington.

HELLENTHAL, G. (2012). Instruction Manual for 'ChromoPainter: a copying model for exploring admixture in population data'.

HELLENTHAL, G., AUTON, A. and FALUSH, D. (2008). Inferring human colonization history using a copying model. *PLoS Genetics*, **4** e1000078.

HELLENTHAL, G., BUSBY, G. B. J., BAND, G., WILSON, J. F., CAPELLI, C., FALUSH, D. and MYERS, S. (2014). A Genetic Atlas of Human Admixture History. *Science*, **343** 747–751.

HENN, B. M., BOTIGUÉ, L. R., GRAVEL, S., WANG, W., BRISBIN, A., BYRNES, J. K., FADHLAOUI-ZID, K., ZALLOUA, P. A., MORENO-ESTRADA, A., BERTRANPETIT, J. ET AL. (2012). Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS genetics*, **8** e1002397. PCAdmix.

HICKEY, J. M., KINGHORN, B. P., TIER, B., WILSON, J. F., DUNSTAN, N., VAN DER WERF, J. H. ET AL. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet Sel Evol*, **43** 12.

HINDS, D. A., STUVE, L. L., NILSEN, G. B., HALPERIN, E., ESKIN, E., BALLINGER, D. G., FRAZER, K. A. and COX, D. R. (2005). Whole-genome patterns of common DNA variation in three human populations. *Science*, **307** 1072–1079.

HODGE, S. E., BOEHNKE, M. and SPENCE, M. A. (1999). Loss of information due to ambiguous haplotyping of SNPs. *Nature genetics*, **21** 360–361.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12** 55–67.

HOGGART, C. J., SHRIVER, M. D., KITTLES, R. A., CLAYTON, D. G. and MCKEIGUE, P. M. (2004). Design and analysis of admixture mapping studies. *American journal of human genetics*, **74** 965. ADMIXMAP.

HOGGART, C. J., WHITTAKER, J. C., DE IORIO, M. and BALDING, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS genetics*, **4** e1000130.

HOLDEN, L., HAUGE, R. and HOLDEN, M. (2009). Adaptive independent metropolis-hastings. *The Annals of Applied Probability* 395–413.

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *The Journal of educational psychology* 498–520.

INTARAPANICH, A., SHAW, P., ASSAWAMAKIN, A., WANGKUMHANG, P., NGAMPHIW, C., CHAICHOOMPU, K., PIRIYAPONGSA, J. and TONGSIMA, S. (2009). Iterative pruning PCA improves resolution of highly structured populations. *BMC bioinformatics*, **10** 382.

IZENMAN, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, **5** 248–264.

IZENMAN, A. J. (2008). *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Springer.

JAIN, A. K. and DUBES, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

JAKOBSSON, M., SCHOLZ, S. W., SCHEET, P., GIBBS, J. R., VANLIERE, J. M., FUNG, H.-C., SZPIECH, Z. A., DEGNAN, J. H., WANG, K., GUERREIRO, R. ET AL. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451** 998–1003.

KAEUFFER, R., RÉALE, D., COLTMAN, D. and PONTIER, D. (2007). Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity*, **99** 374–380.

KARLSSON, E. K., BARANOWSKA, I., WADE, C. M., HILLBERTZ, N. H. S., ZODY, M. C., ANDERSON, N., BIAGI, T. M., PATTERSON, N., PIELBERG, G. R., KULBOKAS, E. J. ET AL. (2007). Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature genetics*, **39** 1321–1328.

KAYSER, M., KRAWCZAK, M., EXCOFFIER, L., DIELTJES, P., CORACH, D., PASCALI, V., GEHRIG, C., BERNINI, L. F., JESPERSEN, J., BAKKER, E. ET AL. (2001). An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *The American Journal of Human Genetics*, **68** 990–1018.

KENNEDY, J. (2009). *Efficient algorithms for SNP genotype data analysis using hidden Markov models of haplotype diversity*. Ph.D. thesis, University of Connecticut. GEDI-ADMX.

KIM, K., TANABE, Y., PARK, C. and HA, J. (2001). Genetic variability in East Asian dogs using microsatellite loci analysis. *Journal of Heredity*, **92** 398–403.

KIRKNESS, E., BAFNA, V., HALPERN, A., LEVY, S., REMINGTON, K., RUSCH, D., DELCHER, A., POP, M., WANG, W., FRASER, C. ET AL. (2003). The dog genome: survey sequencing and comparative analysis. *Science*, **301** 1898.

KLIMENTIDIS, Y. C., MILLER, G. F. and SHRIVER, M. D. (2009). Genetic admixture, self-reported ethnicity, self-estimated admixture, and skin pigmentation among Hispanics and Native Americans. *American journal of physical anthropology*, **138** 375–383.

KONG, A., MASSON, G., FRIGGE, M., GYLFASON, A., ZUSMANOVICH, P., THORLEIFSSON, G., OLASON, P., INGASON, A., STEINBERG, S., RAFNAR, T. ET AL. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, **40** 1068–1075.

KRZANOWSKI, W. (2000). *Principles of multivariate analysis: a user's perspective*. Oxford University Press, USA.

KRZANOWSKI, W. J. and HAND, D. J. (2009). *ROC curves for continuous data*. CRC Press.

KUEHN, L., KEELE, J., BENNETT, G., MCDANELD, T., SMITH, T., SNELLING, W., SONSTEGARD, T. and THALLMAN, R. (2011). Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 Bull Project. *Journal of animal science*, **89** 1742–1750.

KWEE, L. C., LIU, D., LIN, X., GHOSH, D. and EPSTEIN, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, **82** 386–397.

LANGE, K., PAPP, J. C., SINSHEIMER, J. S., SRIPRACHA, R., ZHOU, H. and SOBEL, E. M. (2013). Mendel: the Swiss army knife of genetic analysis programs. *Bioinformatics*.

LARMUSEAU, M., GEYSTELEN, A. V., OVEN, M. V. and DECORTE, R. (2013). Genetic Genealogy Comes of Age: Perspectives on the use of deep-rooted pedigrees in human population genetics. *American journal of physical anthropology*.

LAWSON, C. L. and HANSON, R. J. (1974). *Solving least squares problems*, vol. 161. SIAM.

LAWSON, D. J. and FALUSH, D. (2012). Population Identification Using Genetic Data. *Annual Review of Genomics and Human Genetics*, **13** 337–361.

LAWSON, D. J., HELLENTHAL, G., MYERS, S. and FALUSH, D. (2012). Inference of population structure using dense haplotype data. *PLoS genetics*, **8** e1002453.

LEDGER, R., ORIHEL, J., CLARKE, N., MURPHY, S. and SEDLBAUER, M. (2005). Breed specific legislation: considerations for evaluating its effectiveness and recommandations for alternatives. *The Canadian Veterinary Journal*, **46** 735.

LEE, A. B., LUCA, D., KLEI, L., DEVLIN, B. and ROEDER, K. (2010a). Discovering genetic ancestry using spectral graph theory. *Genetic epidemiology*, **34** 51–59.

LEE, S., EPSTEIN, M. P., DUNCAN, R. and LIN, X. (2012). Sparse Principal Component Analysis for Identifying Ancestry-Informative Markers in Genome-Wide Association Studies. *Genetic Epidemiology*, **36** 293–302.

LEE, S., ZOU, F. and WRIGHT, F. A. (2010b). Convergence and prediction of principal component scores in high-dimensional settings. *Annals of statistics*, **38** 3605.

LEE, S. S.-J., BOLNICK, D. A., DUSTER, T., OSSORIO, P. and TALLBEAR, K. (2009). The illusive gold standard in genetic ancestry testing. *Science*, **325** 38.

LEE, W. and LIU, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *Journal of Multivariate Analysis*.

LEGENDRE, P. and LEGENDRE, L. (2012). *Numerical ecology*, vol. 20. Elsevier.

LEROY, G. (2011). Genetic diversity, inbreeding and breeding practices in dogs: results from pedigree analyses. *The Veterinary Journal*, **189** 177–182.

LEROY, G. and BAUMUNG, R. (2011). Mating practices and the dissemination of genetic disorders in domestic animals, based on the example of dog breeding. *Animal Genetics*, **42** 66–74.

LEROY, G., VERNET, E., PAUTET, M. and ROGNON, X. (2013). An insight into population structure and gene flow within pure-bred cats. *Journal of Animal Breeding and Genetics*.

LESLIE, C., ESKIN, E. and NOBLE, W. S. (2002). The spectrum kernel: A string kernel for SVM protein classification. In *Proceedings of the Pacific symposium on biocomputing*, vol. 7. Hawaii, USA., 566–575.

LESLIE, C. S., ESKIN, E., COHEN, A., WESTON, J. and NOBLE, W. S. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20** 467–476.

LI, J. Z., ABSHER, D. M., TANG, H., SOUTHWICK, A. M., CASTO, A. M., RAMACHANDRAN, S., CANN, H. M., BARSH, G. S., FELDMAN, M., CAVALLI-SFORZA, L. L. ET AL. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *science*, **319** 1100–1104.

LI, N. and STEPHENS, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165** 2213.

LIANG, F., LIU, C. and CARROLL, R. (2011). *Advanced Markov chain Monte Carlo methods: learning from past samples*, vol. 714. Wiley.

LIANG, F. and WONG, W. H. (2001). Evolutionary Monte Carlo for protein folding simulations. *The Journal of Chemical Physics*, **115** 3374.

LIAO, L. and NOBLE, W. S. (2003). Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *Journal of computational biology*, **10** 857–868.

LIMPITI, T., INTARAPANICH, A., ASSAWAMAKIN, A., SHAW, P., WANGKUMHANG, P., PIRIYAPONGSA, J., NGAMPHIW, C. and TONGSIMA, S. (2011). Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and structure. *BMC bioinformatics*, **12** 255.

LIN, X., CAI, T., WU, M. C., ZHOU, Q., LIU, G., CHRISTIANI, D. C. and LIN, X. (2011). Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genetic epidemiology*, **35** 620–631.

LINDBLAD-TOH, K., WADE, C., MIKKELSEN, T., KARLSSON, E., JAFFE, D., KAMAL, M., CLAMP, M., CHANG, J., KULBOKAS, E., ZODY, M. ET AL. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438** 803–819.

LINNAEUS, C. (1737). *Genera Plantarum*. Lugduni Batavorum.

LIU, Y., NYUNOYA, T., LENG, S., BELINSKY, S. A., TESFAIGZI, Y., BRUSE, S. ET AL. (2013). Softwares and methods for estimating genetic ancestry in human populations. *Human genomics*, **7** 1–7.

LUCA, D., RINGQUIST, S., KLEI, L., LEE, A. B., GIEGER, C., WICHMANN, H., SCHREIBER, S., KRAWCZAK, M., LU, Y., STYCHE, A. ET AL. (2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *The American Journal of Human Genetics*, **82** 453–463.

LYONS' VETERINARY GENETICS LABORATORY, UNIVERSITY OF CALIFORNIA, DAVIS (2013). `http://www.vgl.ucdavis.edu/services/index.php`. Accessed: 2013-08-01.

MACCULLAGH, P. and NELDER, J. A. (1989). *Generalized linear models*, vol. 37. CRC press.

MARCHINI, J., CARDON, L. R., PHILLIPS, M. S. and DONNELLY, P. (2004). The effects of human population structure on large genetic association studies. *Nature genetics*, **36** 512–517.

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1980). *Multivariate analysis*. Academic press.

MARS VETERINARY WISDOM PANEL (2013). `http://www.wisdompanel.co.uk`. Accessed: 2013-08-01.

MARTIN, A., WATSON, A., FRETWELL, N., BUSTAMANTE, C., STARK, R., DAVISON, S., VRATIMOS, A., RAUSCH-DERRA, L., JONES, P. and MARKWELL, P. (2010). Assessing performance of a genetic test for determining the breed composition of mixed breed dogs using in silico generated crosses and F1 hybrids.

MASSY, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, **60** 234–256.

MCKEIGUE, P. (2007). Population admixture and stratification in genetic epidemiology. *Handbook of Statistical Genetics, Third Edition* 1190–1215.

MCVEAN, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, **5** e1000686.

MELLANBY, R. J., OGDEN, R., CLEMENTS, D. N., FRENCH, A. T., GOW, A. G., POWELL, R., CORCORAN, B., SCHOEMAN, J. P. and SUMMERS, K. M. (2012). Population structure and genetic heterogeneity in popular dog breeds in the UK. *The Veterinary Journal*.

MENOZZI, P., PIAZZA, A. and CAVALLI-SFORZA, L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science*, **201** 786–792.

METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, **21** 1087.

MITCHELL, M. (1998). *An introduction to genetic algorithms (complex adaptive systems)*. MIT Press.

NEI, M. and KUMAR, S. (2000). *Molecular evolution and phylogenetics*. Oxford University Press, USA.

NORDGREN, A. and JUENGST, E. T. (2009). Can genomics tell me who I am? Essentialistic rhetoric in direct-to-consumer DNA testing. *New Genetics and Society*, **28** 157–172.

NOVEMBRE, J., JOHNSON, T., BRYC, K., KUTALIK, Z., BOYKO, A. R., AUTON, A., INDAP, A., KING, K. S., BERGMANN, S., NELSON, M. R. ET AL. (2008). Genes mirror geography within Europe. *Nature*, **456** 98–101.

NOVEMBRE, J. and STEPHENS, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, **40** 646–649.

ORBANZ, P. and TEH, Y. W. (2010). Bayesian nonparametric models. In *Encyclopedia of Machine Learning*. Springer, 81–89.

OSTRANDER, E. (2007). Genetics and the Shape of Dogs. *American Scientist*.

OSTRANDER, E. and WAYNE, R. (2005). The canine genome. *Genome research*, **15** 1706–1716.

PANG, J.-F., KLUETSCH, C., ZOU, X.-J., ZHANG, A.-B., LUO, L.-Y., ANGLEBY, H., ARDALAN, A., EKSTRÖM, C., SKÖLLERMO, A., LUNDEBERG, J. ET AL. (2009). mtDNA data indicate a single origin for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. *Molecular biology and evolution*, **26** 2849–2864.

PARKER, H., KIM, L., SUTTER, N., CARLSON, S., LORENTZEN, T., MALEK, T., JOHNSON, G., DEFRANCE, H., OSTRANDER, E. and KRUGLYAK, L. (2004). Genetic structure of the purebred domestic dog. *Science*, **304** 1160.

PARKER, H. G. (2012). Genomic analyses of modern dog breeds. *Mammalian Genome*, **23** 19–27.

PARRA, D., MENDEZ, S., CANON, J. and DUNNER, S. (2008). Genetic differentiation in pointing dog breeds inferred from microsatellites and mitochondrial DNA sequence. *Animal genetics*, **39** 1–7.

PAŞANIUC, B., KENNEDY, J. and MĂNDOIU, I. (2009a). Imputation-based local ancestry inference in admixed populations. In *Bioinformatics Research and Applications*. Springer, 221–233. GEDI-ADMX.

PAŞANIUC, B., SANKARARAMAN, S., KIMMEL, G. and HALPERIN, E. (2009b). Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, **25** i213–i221. WinPop.

PATIL, N., BERNO, A. J., HINDS, D. A., BARRETT, W. A., DOSHI, J. M., HACKER, C. R., KAUTZER, C. R., LEE, D. H., MARJORIBANKS, C., MCDONOUGH, D. P. ET AL. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294** 1719–1723.

PATTERSON, D., HASKINS, M., JEZYK, P., GIGER, U., MEYERS-WALLEN, V., AGUIRRE, G., FYFE, J., WOLFE, J. ET AL. (1988). Research on genetic diseases: reciprocal benefits to animals and man. *Journal of the American Veterinary Medical Association*, **193** 1131.

PATTERSON, D. F. (2000). Companion animal medicine in the age of medical genetics. *Journal of Veterinary Internal Medicine*, **14** 1–9.

PATTERSON, N., HATTANGADI, N., LANE, B., LOHMUELLER, K. E., HAFLER, D. A., OKSENBERG, J. R., HAUSER, S. L., SMITH, M. W., OBRIEN, S. J., ALTSHULER, D. ET AL. (2004). Methods for high-density admixture mapping of disease genes. *American journal of human genetics*, **74** 979. MALDsoft.

PATTERSON, N., PRICE, A. and REICH, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, **2** e190.

POSTMA, G., KROOSHOF, P. and BUYDENS, L. (2011). Opening the kernel of kernel partial least squares and support vector machines. *Analytica chimica acta*, **705** 123–134.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38** 904–909.

PRICE, A. L., TANDON, A., PATTERSON, N., BARNES, K. C., RAFAELS, N., RUCZINSKI, I., BEATY, T. H., MATHIAS, R., REICH, D. and MYERS, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics*, **5** e1000519. Hapmix.

PRICE, A. L., WEALE, M. E., PATTERSON, N., MYERS, S. R., NEED, A. C., SHIANNA, K. V., GE, D., ROTTER, J. I., TORRES, E., TAYLOR, K. D. ET AL. (2008). Long-range LD can confound genome scans in admixed populations. *American journal of human genetics*, **83** 132.

PRITCHARD, J., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155** 945. Structure.

PRITCHARD, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, **69** 124–137.

QUIGNON, P., HERBIN, L., CADIEU, E., KIRKNESS, E., HÉDAN, B., MOSHER, D., GALIBERT, F., ANDRÉ, C., OSTRANDER, E. and HITTE, C. (2007). Canine population structure: assessment and impact of intra-breed stratification on SNP-based association studies. *PloS one*, **2** 1324.

QUINN, G. P. and KEOUGH, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge University Press.

RABINER, L. and JUANG, B. (1986). An introduction to hidden Markov models. *IEEE ASSp Magazine*, **3** 4–16.

RAI, P., KUMAR, A. and DAUME III, H. (2012). Simultaneously Leveraging Output and Task Structures for Multiple-Output Regression. In *Advances in Neural Information Processing Systems 25*. 3194–3202.

REID, M. and WILLIAMSON, B. (2011). Information, Divergence and Risk for Binary Experiments. *Journal of Machine Learning Research*, **12** 731–817.

RIMBAULT, M. and OSTRANDER, E. A. (2012). So many doggone traits: mapping genetics of multiple phenotypes in the domestic dog. *Human molecular genetics*, **21** R52–R57.

ROBERT, C. and CASELLA, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.

ROBERT, C. and CASELLA, G. (2010). *Introducing Monte Carlo Methods with R*. Springer Verlag.

ROBERTS, G. and ROSENTHAL, J. (1998). Markov-chain monte carlo: Some practical implications of theoretical results. *Canadian Journal of Statistics*, **26** 5–20.

ROBERTS, G. and STRAMER, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, **4** 337–357.

RODRIGUEZ, J. M., BERCOVICI, S., ELMORE, M. and BATZOGLOU, S. (2013). Ancestry Inference in Complex Admixtures via Variable-length Markov Chain Linkage Models. *Journal of Computational Biology*. Variable length MCMC.

ROMUALDI, C., BALDING, D., NASIDZE, I. S., RISCH, G., ROBICHAUX, M., SHERRY, S. T., STONEKING, M., BATZER, M. A. and BARBUJANI, G. (2002). Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Research*, **12** 602–612.

RON, D., SINGER, Y. and TISHBY, N. (1995). On the learnability and usage of acyclic probabilistic finite automata. In *Proceedings of the eighth annual conference on Computational learning theory*. ACM, 31–40.

ROSENBERG, N. A., LI, L. M., WARD, R. and PRITCHARD, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics*, **73** 1402–1422.

ROSENBERG, N. A., MAHAJAN, S., RAMACHANDRAN, S., ZHAO, C., PRITCHARD, J. K. and FELDMAN, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS genetics*, **1** e70.

ROSENTHAL, J. S. (2000). Parallel computing and Monte Carlo algorithms. *Far east journal of theoretical statistics*, **4** 207–236.

ROSIPAL, R. and KRÄMER, N. (2006). Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*. Springer, 34–51.

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, **19** 947–962.

ROYAL, C. D., NOVEMBRE, J., FULLERTON, S. M., GOLDSTEIN, D. B., LONG, J. C., BAMSHAD, M. J. and CLARK, A. G. (2010). Inferring genetic ancestry: opportunities, challenges, and implications. *The American Journal of Human Genetics*, **86** 661–673.

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric regression*, vol. 12. Cambridge University Press.

SACKS, J., SINCLAIR, L., GILCHRIST, J., GOLAB, G. and LOCKWOOD, R. (2000). Breeds of dogs involved in fatal human attacks in the United States between 1979 and 1998. *Journal of the American Veterinary Medical Association*, **217** 836–840.

SANKARARAMAN, S., KIMMEL, G., HALPERIN, E. and JORDAN, M. I. (2008a). On the inference of ancestries in admixed populations. *Genome research*, **18** 668–675. Switch.

SANKARARAMAN, S., SRIDHAR, S., KIMMEL, G. and HALPERIN, E. (2008b). Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, **82** 290–303. Lamp.

SARATA, A. K. (2008). Genetic Ancestry Testing.

SAVOLAINEN, P., ZHANG, Y.-P., LUO, J., LUNDEBERG, J. and LEITNER, T. (2002). Genetic evidence for an East Asian origin of domestic dogs. *Science*, **298** 1610–1613.

SCHEET, P. (2013). personal communication.

SCHEET, P. and STEPHENS, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, **78** 629–644.

SCHIFANO, E. D., EPSTEIN, M. P., BIELAK, L. F., JHUN, M. A., KARDIA, S. L., PEYSER, P. A. and LIN, X. (2012). SNP set association analysis for familial data. *Genetic Epidemiology*, **36** 797–810.

SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization and beyond*. the MIT Press.

SELDIN, M. F. (2007). Admixture mapping as a tool in gene discovery. *Current opinion in genetics & development*, **17** 177–181.

SELDIN, M. F., PASANIUC, B. and PRICE, A. L. (2011). New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, **12** 523–528.

SHARIFLOU, M. R., JAMES, J. W., NICHOLAS, F. W. and WADE, C. M. (2011). A genealogical survey of Australian registered dog breeds. *The Veterinary Journal*, **189** 203–210.

SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

SHRIVER, M. D. and KITTLES, R. A. (2004). Genetic ancestry and the search for personalized genetic histories. *Nature Reviews Genetics*, **5** 611–618.

SMITH, M. W. and O'BRIEN, S. J. (2005). Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nature Reviews Genetics*, **6** 623–632.

SMITH, T. F. and WATERMAN, M. S. (1981). Comparison of biosequences. *Advances in Applied Mathematics*, **2** 482–489.

SOHN, K.-A. (2011). *Learning Ancestral Genetic Processes using Nonparametric Bayesian Models*. Ph.D. thesis, Carnegie Mellon University.

SOHN, K.-A., GHAHRAMANI, Z. and XING, E. P. (2012). Robust estimation of local genetic ancestry in admixed populations using a nonparametric Bayesian approach. *Genetics*, **191** 1295–1308.

SOHN, K.-A. and KIM, S. (2012). Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *International Conference on Artificial Intelligence and Statistics*. 1081–1089.

SOHN, K.-A. and XING, E. P. (2009). A hierarchical Dirichlet process mixture model for haplotype reconstruction from multi-population data. *The Annals of Applied Statistics* 791–821.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64** 583–639.

STEPHENS, M., SMITH, N. and DONNELLY, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, **68** 978–989.

SUH, Y. and VIJG, J. (2005). SNP discovery in associating genetic variation with human disease phenotypes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **573** 41–53.

SUNDQUIST, A., FRATKIN, E., DO, C. B. and BATZOGLOU, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome research*, **18** 676–682. HAPAA.

SUNDQVIST, A., BJORNERFELDT, S., LEONARD, J., HAILER, F., HEDHAMMAR, A., ELLEGREN, H. and VILA, C. (2006). Unequal contribution of sexes in the origin of dog breeds. *Genetics*, **172** 1121.

SUTTER, N., EBERLE, M., PARKER, H., PULLAR, B., KIRKNESS, E., KRUGLYAK, L. and OSTRANDER, E. (2004). Extensive and breed-specific linkage disequilibrium in Canis familiaris. *Genome Research*, **14** 2388–2396.

SUTTER, N. B. and OSTRANDER, E. A. (2004). Dog star rising: the canine genetic system. *Nature Reviews Genetics*, **5** 900–910.

TANG, H., CORAM, M., WANG, P., ZHU, X. and RISCH, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, **79** 1–12. SABER.

TANG, H., PENG, J., WANG, P. and RISCH, N. J. (2005). Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology*, **28** 289–301. Frappe.

TEH, Y. W., BLUNDELL, C. and ELLIOTT, L. T. (2011). Modelling genetic variations with fragmentation-coagulation processes. *Advances in Neural Information Processing Systems*, **23** 819–827.

TEH, Y. W., BLUNDELL, C. and ELLIOTT, L. T. (2013). Bayesian Nonparametric Modelling of Genetic Variations using Fragmentation-Coagulation Processes. *Journal of Machine Learning Research*.

TEH, Y. W. and JORDAN, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics: Principles and Practice* 158–207.

THE AMERICAN SOCIETY OF HUMAN GENETICS (2008). Ancestry Testing Statement. `http://www.ashg.org/pdf/ASHGAncestryTestingStatement_FINAL.pdf`. Accessed: 2013-08-01.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

TSAI, H.-J., CHOUDHRY, S., NAQVI, M., RODRIGUEZ-CINTRON, W., BURCHARD, E. G. and ZIV, E. (2005). Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. *Human genetics*, **118** 424–433.

TSUDA, K., KIN, T. and ASAI, K. (2002). Marginalized kernels for biological sequences. *Bioinformatics*, **18** S268–S275.

VAN OVEN, M., VERMEULEN, M. and KAYSER, M. (2011). Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution. *Investigative genetics*, **2** 1–14.

VAYSSE, A., RATNAKUMAR, A., DERRIEN, T., AXELSSON, E., ROSENGREN PIELBERG, G., SIGURDSSON, S., FALL, T., SEPP, E. H., HANSEN, M. S. T., LAWLEY, C. T., KARLSSON, E. K., BANNASCH, D., VILA , C., LOHI, H., GALIBERT, F., FREDHOLM, M., HAGGSTROM, J., HEDHAMMAR, A., ANDRE, C., LINDBLAD-TOH, K., HITTE, C., WEBSTER, M. T. and CONSORTIUM, T. L. (2011). Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. *PLoS Genet*, **7** e1002316.

VERT, J.-P., SAIGO, H. and AKUTSU, T. (2004). Local alignment kernels for biological sequences. *Kernel methods in computational biology* 131–154.

VIA, M., ZIV, E. and BURCHARD, E. G. (2009). Recent advances of genetic ancestry testing in biomedical research and direct to consumer testing. *Clinical genetics*, **76** 225–235.

VILÀ, C., SAVOLAINEN, P., MALDONADO, J., AMORIM, I., RICE, J., HONEYCUTT, R., CRANDALL, K., LUNDEBERG, J. and WAYNE, R. (1997). Multiple and ancient origins of the domestic dog. *Science*, **276** 1687.

VISSCHER, P. M., HILL, W. G. and WRAY, N. R. (2008). Heritability in the genomics era-concepts and misconceptions. *Nature Reviews Genetics*, **9** 255–266.

VISSCHER, P. M., MEDLAND, S. E., FERREIRA, M. A., MORLEY, K. I., ZHU, G., CORNES, B. K., MONTGOMERY, G. W. and MARTIN, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by=descent sharing between full siblings. *PLoS genetics*, **2** e41.

VOITH, V., INGRAM, E., MITSOURAS, K. and IRIZARRY, K. (2009). Comparison of Adoption Agency Breed Identification and DNA Breed Identification of Dogs. *Journal of Applied Animal Welfare Science*, **12** 253–262.

VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, **17** 395–416.

VONHOLDT, B., POLLINGER, J., LOHMUELLER, K., HAN, E., PARKER, H., QUIGNON, P., DEGENHARDT, J., BOYKO, A., EARL, D., AUTON, A. ET AL. (2010). Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature*, **464** 898–902.

WADE, C. M. (2011). Inbreeding and genetic diversity in dogs: Results from DNA analysis. *The Veterinary Journal*, **189** 183–188.

WAGNER, J. K., COOPER, J. D., STERLING, R. and ROYAL, C. D. (2012). Tilting at windmills no longer: a data-driven discussion of DTC DNA ancestry tests. *Genetics in Medicine*, **14** 586–593.

WAGNER, J. K. and WEISS, K. M. (2012). Attitudes on DNA ancestry tests. *Human Genetics*, **131** 41–56.

WASSERMAN, L. (2006). *All of nonparametric statistics*. Springer Science+ Business Media.

WAYNE, R. (1993). Molecular evolution of the dog family. *Trends in Genetics*, **9** 218–224.

WAYNE, R. and VONHOLDT, B. (2012). Evolutionary genomics of dog domestication. *Mammalian Genome*, **23** 3–18.

WAYNE, R. K., GEFFEN, E., GIRMAN, D. J., KOEPFLI, K. P., LAU, L. M. and MARSHALL, C. R. (1997). Molecular systematics of the Canidae. *Systematic biology*, **46** 622–653.

WEBB, A. R. (2003). *Statistical pattern recognition*. John Wiley & Sons.

WEBB, K. and ALLARD, M. (2010). Assessment of minimum sample sizes required to adequately represent diversity reveals inadequacies in datasets of domestic dog mitochondrial DNA. *Mitochondrial DNA*, **21** 19–31.

WEIR, B. S., CARDON, L. R., ANDERSON, A. D., NIELSEN, D. M. and HILL, W. G. (2005). Measures of human population structure show heterogeneity among genomic regions. *Genome research*, **15** 1468–1476.

WEIR, B. S. and COCKERHAM, C. C. (1984). Estimating F-statistics for the analysis of population structure. *evolution*, **38** 1358–1370.

WESSEL, J. and SCHORK, N. J. (2006). Generalized genomic distance–based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*, **79** 792–806.

WILCOX, B. and WALKOWICZ, C. (2010). *Atlas of dog breeds of the world.* 6th ed. TFH Publications Inc.

WILEY, E. and LIEBERMAN, B. (2011). *Phylogenetics: Theory and Practice of Phylogenetic Systematics*. Wiley-Blackwell.

WILSON, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22** 209–212.

WOLD, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *Perspectives in Probability and Statistics, In Honor of MS Bartlett* 117–144.

WONG, A. K., RUHE, A. L., DUMONT, B. L., ROBERTSON, K. R., GUERRERO, G., SHULL, S. M., ZIEGLE, J. S., MILLON, L. V., BROMAN, K. W., PAYSEUR, B. A. ET AL. (2010). A comprehensive linkage map of the dog genome. *Genetics*, **184** 595–605.

WU, B., LIU, N. and ZHAO, H. (2006). PSMIX: an R package for population structure inference via maximum likelihood method. *BMC bioinformatics*, **7** 317.

WU, M. C., KRAFT, P., EPSTEIN, M. P., TAYLOR, D. M., CHANOCK, S. J., HUNTER, D. J. and LIN, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, **86** 929–942.

YAEGER, R., AVILA-BRONT, A., ABDUL, K., NOLAN, P. C., GRANN, V. R., BIRCHETTE, M. G., CHOUDHRY, S., BURCHARD, E. G., BECKMAN, K. B., GORROOCHURN, P. ET AL. (2008). Comparing genetic ancestry and self-described race in african americans born in the United States and in Africa. *Cancer Epidemiology Biomarkers & Prevention*, **17** 1329–1338.

YAN, X. and SU, X. G. (2009). *Linear regression analysis: theory and computing*. World Scientific Publishing Company.

YENIAY, O. and GOKTAS, A. (2002). A comparison of partial least squares regression with other prediction methods. *Hacettepe Journal of Mathematics and Statistics*, **31** 99–101.

ZHANG, J. (2010). Ancestral informative marker selection and population structure visualization using sparse Laplacian eigenfunctions. *PloS one*, **5** e13734.

ZHANG, Y. and YEUNG, D.-Y. (2010). A Convex Formulation for Learning Task Relationships in Multi-Task Learning. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*. AUAI Press, Corvallis, Oregon, 733–742.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67** 301–320.