Genetic Association of Deafness in Three Piebald Dogs

Introduction:

Congenital deafness in domestic dogs is known to be associated with the white pigment gene piebald, located on chromosome 20 or merle on chromosome 10, with further evidences supporting that the homozygosity of those alleles is correlated with more severe deafness in such breeds possessing piebald genotype [5]. However, there is little information about genes other than the ones having the piebald locus and merle locus and their potential interactions involved or related to congenital deafness in dogs as the importance of auditory functions of domestic dogs are easily overlooked compared to those of humans and other wild animals. However, auditory sense is a very basic yet essential function an organism requires as it is used as one of the primary means that it interacts with its environment. For instance, deafness of domestic dogs impedes their ability to serve as service dogs, ability to communicate with their owners in a household setting, and exposes themselves to physical dangers that could be avoided if they were hearing. As deeply as domestic dogs are incorporated into everyday human lives in today's society, improper function of auditory sense is not just a tragedy for themselves as organisms, but also for humans. Therefore, identifying the regions of genes linked to the causation of deafness in dogs and understanding their relationship should precede seeking potential solutions.

In this research, .bim, .bed, .fam, and phenotype text files of three canine breeds from the UK and North America(Dalmatians, Australian Cattle Dogs, and English Setters) produced by Hayward at al. [2] were used. In their experiment, their hearing ability was assessed using the brainstem auditory evoked response (BAER) [6] hearing test and phenotype for each sample was classified as unilaterally deaf, bilaterally deaf, and bilaterally hearing. In this research, principal component analysis(PCA) and genome-wide association study(GWAS) were performed on the provided database.

The results have revealed that there is no association located on the same chromosome as the piebald gene and merle gene, indicating the complexity of genetics that causes deafness in piebald dogs. Furthermore, there was no common association between the three dog breeds studied, which suggests the possibility that each breed has different genetic composition causing a piebald-associated deafness.

Methods:

Principal Component Analysis (PCA)

The data provided contains 304 Dalmatians, 120 Australian cattle dogs, and 79 English setters genotyped at 201,020 single nucleotide polymorphisms (SNPs). PCA was performed individually for each breed using the smartpca tool from the EIGENSTRAT software (EIGENSOFT v7.2.1 package) [3, 4].

In each iteration, smartpca takes in an input parameter file specifying a genotype file, a SNP file, an individual file specifying each breed, output file names for the resulting eigenvector and eigenvalue, and a population list file that specifies which breed/population PCA is specifically being run on. Because the given datasets were composed by PLINK, the genotype file input used the bed file included in the dataset and the SNP file input used the bim file. However, because smartpca requires specific file formats, it was renamed with the suffix .pedsnp. The unique individual file was created by modifying the 6th column of the fam file to specify the breed of each sample, and also renamed with the suffix .pedind. PCA for each breed was run with a unique population list file, containing a single line specifying the breed in order for smartpca to only consider samples of this particular breed. The parameter file is composed of each of these components (one parameter per line) along with the respective file names for each breed's eigenvector and eigenvalue result files. We also included the option altnormstyle: NO in order to use EIGENSTRAT's specific normalization method. This is run using the command: ./smartpca -p file.param.

The resulting eigenvector files for each breed were used to plot the PCA plots using Matplotlib in the Jupyter notebook.

Genome-wide Association Study (GWAS)

In order to identify single nucleotide polymorphisms most significantly linked to deafness in three dog breeds studied, a command line tool PLINK v1.90b6.9 [www.cog-genomics.org/plink/1.9/, 1] was used for each breed, Australian cattle dogs, English setters, and Dalmatians. Before running PLINK command, some data preprocessing steps were done: ID files for each breed were created so that GWAS could be performed separately on each breed, and phenotype text file was converted to .phen file by quantifying descriptive phenotype and changing format to PLINK .phen file where the first two columns represent family ID and individual ID and the last column represents quantified phenotype. Because of the clear

genetic variation between the population separation found between North American and UK dalmatians during our PCA, the dalmatian breed was separated by their origins in order to increase statistical significance while performing GWAS. The intermediate phenotype, unilateral hearing dogs were excluded from the input .phen and files containing ID's of each breed in order to observe stronger distinctions between SNPs of bilaterally deaf dogs and bilateral hearing dogs.

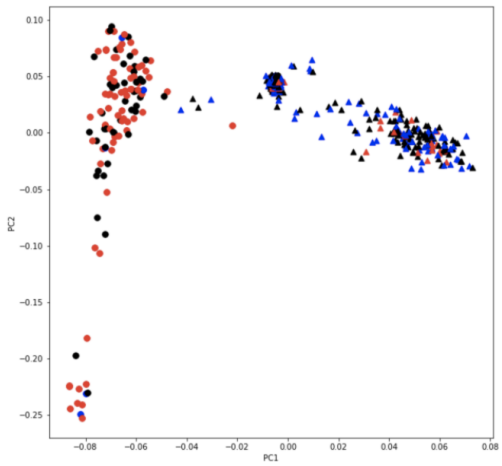
Then we performed linear regression on each breed mentioned above, hiding covariates and removing SNPs with less than 5% minor allele frequency in each group from the group's analysis. Also, phenotypes of individuals with ambiguous sex were included in the anaylsis since no gender difference is known to be observed in deafness of the dogs studied [7]. All phenotypes present in our .phen file were set to be subject to the association tests. We used the flag --keep-fam [family ID file] so that we could perform separate GWAS for each breed with the same .bim, .bed., and .fam files. The resulting files were .assoc.linear files containing chromosome numbers, SNPs, base pairs, and P-value of each expressed SNP. QQ plot and Manhattan plots were constructed from .assoc.linear files using Matplotlib.

Results:

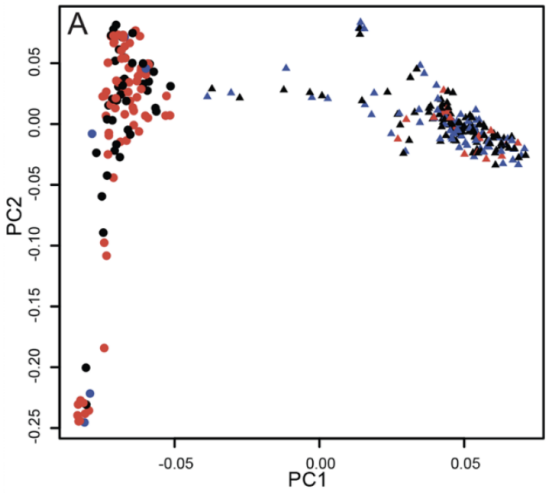
PCA Results

PCA for Dalmatians

Reproduction

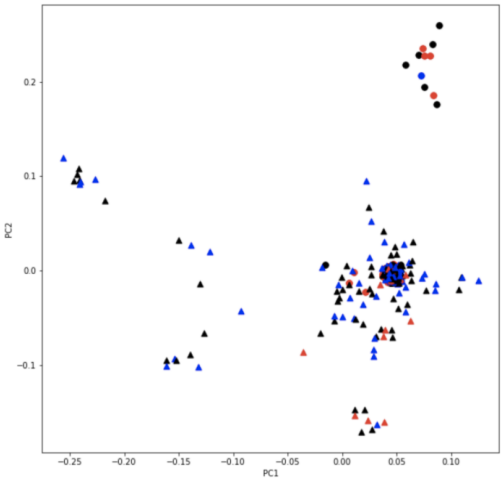


Original

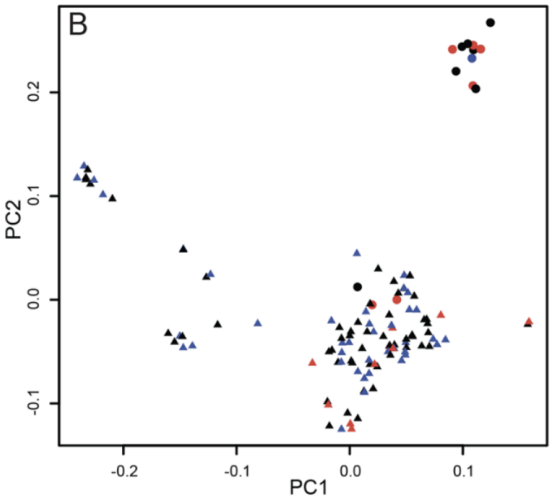


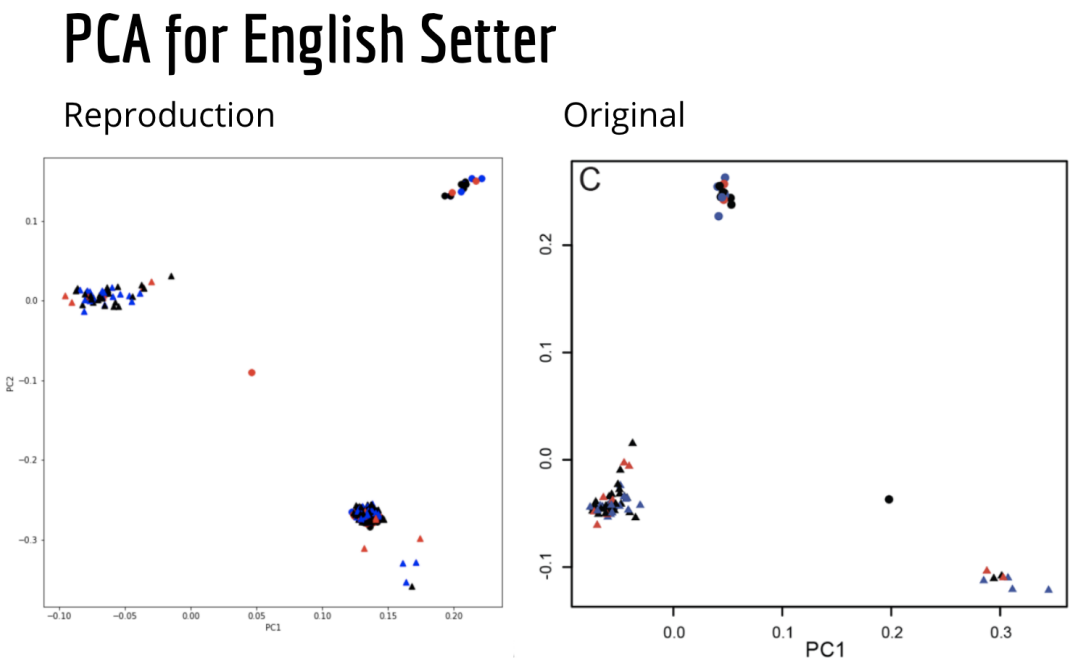
PCA for Australian Cattle Dog

Reproduction



Original

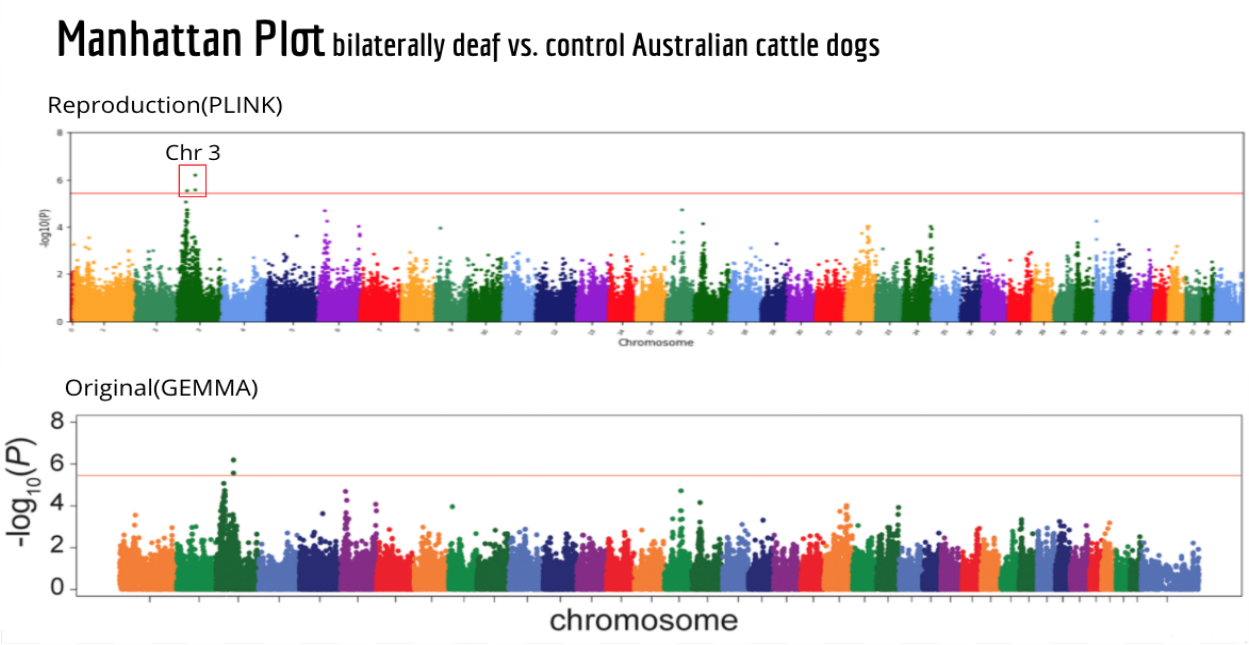




Samples from North America (circles) and UK (triangles) with bilaterally deaf dogs (red), unilaterally deaf dogs (blue), and control dogs (black).

Although there are slight differences between the two sets of PCA results, clear geographic separation patterns and phenotypic patterns that appear in the original paper are also apparent in our findings as well. In the collected dalmatian population, there is a clear genetic difference between North American and UK populations seen in the two separate clusters along PC1. It is also clear that the North American sample is mostly composed of unilaterally deaf and control dogs, while the UK sample contained mostly bilaterally deaf and control dogs. The Australian cattle dog samples also show a clear geographic origin separation with North American and UK populations separated on PC2, but no clear phenotypic separation. Similarly, the English setter samples show vague geographic structures, yet again no clear biased phenotypic separation from the different locations.

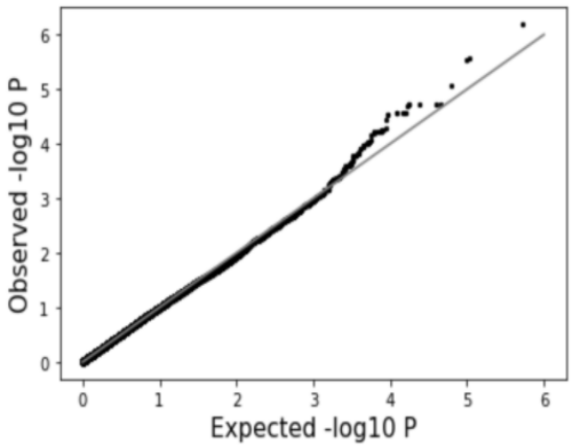
GWAS Results



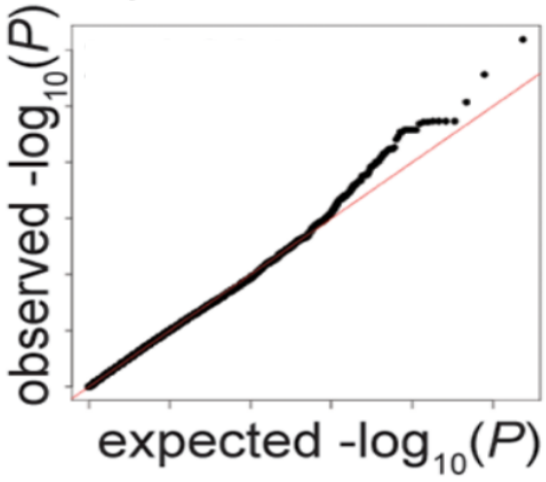
The Manhattan plot of the significant association using bilaterally deaf vs. control Australian cattle dogs above shows that our reproduction produced 3 significant SNPs on chromosome 3 while the paper's Manhattan plot had 2 significant SNPs on chromosome 3. Our Manhattan plot indeed has the same shape and distribution as the paper's plot. Our plots above do show that chromosome 3, 6, 9, 16, 17, 22, 24, 31, 32, and 33 have a larger range of variation of SNPs as the original Manhattan plot seems to show chromosome 3, 6, 9, 16, 17, 22 24 as the only chromosomes with a larger variation. This shows that opposed to the assumption based on previous studies that chromosome 10 and 20 have the top association with deafness, that multiple different chromosomes have an input in contributing for blindness where chromosome 3 seems to have the most significant other SNPs. This is confirmed as there are 3 SNP's over the Bonferroni correction threshold of 3.6×10^{-6}(alpha = 0.05), thus, we can see that chromosome 3 has multiple important SNP's important for piebald-associated deafness.

QQ plot bilaterally deaf vs. control Australian cattle dogs

Reproduction(PLINK)



Original(GEMMA)



The QQ plot of the significant association using bilaterally deaf vs. control Australian cattle dogs shows a strong correlation to the expected $-\log_{10}$(p-val) up to a p-value of ≈ 0.0005, $-\log_{10}(0.0005) \approx 3.3$. After this p-value, the QQ plot seems to deviate from the expected values, but they do share the same deviation shape of that of the QQ plot from the paper. There are slightly more SNP's in our QQ plot towards the deviated end of the QQ plot. The distribution of p-value shows a departure from the reference diagonal line, indicating that there exists a significant association.

Significant SNP's: original paper(GEMMA)

Breed	N	CFA	bp	SNP ID	af all (deaf/controls)	pve[a]	beta	*P*-value
Dalmatians North America	20 deaf, 91 controls	30	37,235,914	BICF2P1106247	0.104 (0.300/0.060)	0.167	-0.353	$7.25\times10^{-6*}$
		30	33,816,254	BICF2P113616	0.333 (0.625/0.269)	0.155	-0.222	$1.60\times10^{-5*}$
		23	48,506,877	BICF2G630365393	0.441 (0.725/0.379)	0.150	-0.220	$2.28\times10^{-5*}$
		30	22,647,163	BICF2G630405064	0.068 (0.200/0.038)	0.130	-0.408	$8.93\times10^{-5*}$
		37	27,255,309	BICF2G630132623	0.243 (0.450/0.198)	0.122	-0.245	$1.54\times10^{-4*}$
Dalmatians UK	72 deaf, 43 controls	38	21,626,523	BICF2G63068103	0.152 (0.083/0.267)	0.127	0.350	$8.22\times10^{-5*}$
Australian cattle dogs	16 deaf, 61 controls	3	37,793,043	BICF2G630338450	0.299 (0.656/0.205)	0.277	-0.313	$6.46\times10^{-7**}$
		3	17,067,881	BICF2G630703558	0.117 (0.344/0.057)	0.229	-0.408	$8.45\times10^{-6*}$
		16	36,220,138	BICF2P1229299	0.091 (0.281/0.041)	0.213	-0.453	$1.91\times10^{-5*}$
		6	10,527,823	BICF2S23125774	0.240 (0.500/0.172)	0.212	-0.330	$2.05\times10^{-5*}$
		17	18,275,241	chr17_18275241	0.110 (0.313/0.057)	0.187	-0.377	$6.96\times10^{-5*}$
		6	75,622,113	BICF2P481353	0.071 (0.219/0.033)	0.184	-0.505	$8.37\times10^{-5*}$
		22	48,747,165	BICF2G630335709	0.494 (0.188/0.574)	0.181	0.239	$9.66\times10^{-5*}$
		9	8,460,580	BICF2S23511312	0.130 (0.313/0.082)	0.178	-0.395	$1.09\times10^{-4*}$
		24	47,255,337	TIGRP2P322787_rs9139922	0.136 (0.344/0.082)	0.176	-0.344	$1.19\times10^{-4*}$
English setters	11 deaf, 39 controls	39	111,315,267	BICF2G6304357	0.220 (0.421/0.154)	0.192	-0.276	1.20×10^{-3}

Significant SNP's: reproduced table(PLINK)

Breed	CHR	SNP ID	BP	BETA	P	Rank*
Dalmatians North America	30	BICF2P1106247	37235914	-0.79066	0.000007	1
	30	BICF2P113616	33816254	-0.4430	0.000016	3
	23	BICF2G630365393	48506877	-0.4407	0.000023	4
Dalmatians UK	38	BICF2G63068103	21626523	0.6994	0.000082	1
Australian cattle dog	3	BICF2G630338450	37793043	-6.266000e-01	6.461000e-07	1
	3	BICF2G630703558	17067881	-8.161000e-01	8.448000e-06	4
	16	BICF2P1229299	36220138	-9.054000e-01	1.908000e-05	6
	6	BICF2S23125774	10527823	-6.591000e-01	2.050000e-05	11
	17	chr17_18275241	18275241	-7.500000e-01	6.950000e-05	29
English Setters	N/A	N/A	N/A	N/A	N/A	N/A

SNPs matching SNPs from the original table that were in the top 30 most significant SNPs.
*Rank refers to the ranks of significant SNPs from our result. e.g. rank = 1 most significant gene

The paper's top SNP table, made with GEMMA, consists of 16 different SNPs for various breeds of which there are 9 that match SNPs on our table. Our SNP table, made with PLINK, consists of 9 different SNP's of which all 9 match that of the paper's table. We see that most of the SNPs from the paper's table also come from the chromosomes we found to have almost significant SNPs from the Manhattan plot above. Most of the SNPs on our table that matched top SNPs of the original table were in approximately the top 10 range with the same order even though there were other SNPs in between that are not on the original table. SNPs of North American dalmatians, UK dalmatians, and Australian cattle dogs tend to match between our data and theirs with very similar p-values.

Discussion:

During the process of reproducing the figures of this paper, we came across many issues that hindered our ability to perfectly reproduce their result. The first issue we faced was that the paper mentions that they used EIGENSTRAT(EIGENSOFT v5.0.1 package), but EIGENSOFT has since been updated many times such that v5.0.1 is not available to download and use anymore. As a result we had to download the EIGENSOFT v7.2.1 package to perform

PCA analysis. In addition, the paper removed SNP's with <0.05% minor allele frequency, but we could not find a flag to remove these values through the smartpca of EIGENSTRAT. Thus we returned .evec and .eign files with no SNPs filtered.

We also found that PLINK's PCA command is the same algorithm for EIGENSTRAT such that we tried to see if the pca results from EIGENSTRAT and PLINK matched as PLINK has the --maf 0.05 command. Nonetheless, the PLINK PCA results had Australian Cattle dogs needing to have flipped values along y=x, Dalmatians needing to be flipped along y=x, and English Setters needing to be flipped along x = -0.15. These results conflict with the EIGENSTRAT PCA results as all but English Setters seem to follow the same shape as those in Figure 1 of the paper. This led us to conclude that either the --maf 0.05 flag of PLINK, or the pca algorithm from plink differ. As a result, our PCA plots do not exactly match, but have relatively the right shape for PC1 vs PC2 across all breeds. We also believe that if we had the non-binary data from the .bed file we would better be able to pinpoint and reproduce the PCA plots as we would know how the data is organized and which SNPs may need to be filtered.

The next source of error is seen through their 2nd and 3rd GWAS step which included needing to use GEMMA to perform the next step of the GWAS. Having been limited with time, we could not reproduce results with GEMMA, but through the use of the .assoc.linear file from our GWAS part 1 step using PLINK, we were able to create a Manhattan and QQ plot. The results are quite consistent with the exception of 3 snp's above the $-\log_{10}$(p-val) compared to their 2, as well as a slight difference in the QQ plot at the upper end of the $-\log_{10}$(p-values). The main difference in GEMMA vs PLINK for the third GWAS step is seen in the SNP table above as only some of the top significant snp's matches between tables which we believe comes down to the differences between GEMMA vs PLINK for conducting this GWAS analysis step.

Nonetheless, through the difficulties of software and data formats, we did manage to reproduce most figures to a rather high degree to a level in which we were able to still pull out meaningful SNPs based on the breed of the dogs. Future pipeline analysis would ideally start from a .vcf file if given and then create our own .bed file with specific SNP data tailored to our analysis. In the case that we were not given the .vcf file, as we were not, we would potentially run two parallel methods such that we would have a PLINK vs EIGENSTRAT + GEMMA pipeline in which at the end of such analysis we can merge and compare results to see if there are meaningful results that each pipeline has that the other does not. This would provide a more comprehensive understanding of the study, not limited by the power of one tool. This project has also further showed versatility of PLINK is as a diverse tool that can not only run PCA analysis

but also produce manhattan plots, QQ plots and find the top expressed SNPs while EIGENSTRAT and GEMMA as two completely different softwares were needed to produce the results.

References:

1. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4.
2. Hayward JJ, Kelly-Smith M, Boyko AR, et al. A genome-wide association study of deafness in three canine breeds. PLoS One. 2020;15(5):e0232900. Published 2020 May 15. doi:10.1371/journal.pone.0232900
3. Patterson, Nick, Alkes L. Price, and David Reich. "Population structure and eigenanalysis." PLoS genetics 2.12 (2006): e190.
4. Price, Alkes L., et al. "Principal components analysis corrects for stratification in genome-wide association studies." Nature genetics 38.8 (2006): 904-909.
5. Strain, George. (2012). Canine Deafness. The Veterinary clinics of North America. Small animal practice. 42. 1209-24. 10.1016/j.cvsm.2012.08.010.
6. Strain GM. Brainstem auditory evoked response (BAER). In Strain GM, ed Deafness in dogs and cats. CAB International, Wallingford, UK; 2011. pp. 83–107.
7. Strain GM. Deafness prevalence and pigmentation and gender associations in dog breeds at risk. Vet J. 2004;167: 23–32. Pmid:14623147

Appendix:

Neel performed PCA using PLINK, made PCA plots from PLINK, pre-processed/edited input files.
Soyeon preprocessed files for GWAS and performed GWAS using PLINK and constructed the table.
Michelle performed PCA using EIGENSTRAT and constructed respective plots, and constructed the QQ plot and the Manhattan plot from the PLINK GWAS output file.