Summer 2014

Machine learning approaches for predicting genotype from phenotype and a novel clustering technique for subgenotype discovery: an application to inherited deafness

Kyle Ross Taylor
University of Iowa

MACHINE LEARNING APPROACHES FOR PREDICTING GENOTYPE FROM PHENOTYPE AND A NOVEL CLUSTERING TECHNIQUE FOR SUBGENOTYPE DISCOVERY: AN APPLICATION TO INHERITED DEAFNESS

by

Kyle Ross Taylor

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Electrical and Computer Engineering
in the Graduate College of
The University of Iowa

August 2014

Thesis Supervisor: Professor Thomas L. Casavant

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Kyle Ross Taylor

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Electrical and Computer Engineering at the August 2014 graduation.

Thesis Committee: _____
Thomas L. Casavant, Thesis Supervisor

_____
Terry A. Braun

_____
Punam K. Saha

_____
Todd E. Scheetz

_____
Richard J.H. Smith

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

INTRODUCTION

In the era of High Throughput Sequencing (HTS), it is becoming increasingly clear that the downstream analysis of variants from sequencing is becoming more costly and challenging than sequence data generation itself [1]. The fundamental challenge has shifted from one of prioritization of candidate regions to interrogate via Sanger sequencing, to one of hypothesizing likely causative regions that can be used to refine the list of variations resulting from HTS. The main idea is to utilize phenotypes in the downstream analysis of variants to better-identify disease-causing mutations. The objective of linking phenotype to genotype is to make it possible to leverage a patient's phenotype to increase the diagnostic power of large-scale genetic screening – i.e., to ultimately determine a patient's molecularly-validated genotype upon which to base treatment decisions. The main goal of this thesis is to develop a framework for the prediction of genotype from phenotype – specifically in the case of Non-syndromic Hearing Loss (NSHL) – and develop a technique to discover novel sub-phenotypes and genotypes (subclasses) to improve the prediction and understanding of NSHL.

Patients with NSHL exhibit a large degree of diversity in observed phenotype which has been shown to associate with specific genetic causes [2]. For instance, patients with mutations in the DFNA2A loci have a typical hearing loss pattern that is more pronounced in the higher frequencies and progresses with age. In contrast, patients with mutations in the DFNA6/14/38 locus have a very different hearing loss pattern which only effects the lower frequencies (nominally below 2 kHz).

To exploit this heterogeneity and prioritize loci for screening, a method, software tool and website was developed called AudioGene [3]for ranking loci using the patient's hearing loss phenotype as reported in a standard audiogram. The accuracy of AudioGene for predicting the top three candidate loci was 68% when using an MI-SVM, compared to

44% using a Majority classifier for Autosomal *Dominant* Non-syndromic Hearing loss (ADNSHL). The ADNSHL dataset contained over 1,400 patients harboring mutations in 34 of the 64 known loci. The method was extended to predict the mutation type for patients with mutations in the Autosomal *Recessive* Non-syndromic Hearing Loss locus DFNB1, and had an accuracy of 83% compared to 50% for a Majority classifier when the classes were down-sampled to contain equal numbers of patients. The DFNB1 locus had a large class imbalance with one class containing approximately 93% of the patients; to reduce the effects of class imbalance the patients were down-sampled which resulted in better performance.

There are several challenges in predicting genotype from phenotype and these challenges likely impede further improvements in AudioGene. First, there exists a complex relationship between genotypes and phenotypes, and in cases of Mendelian diseases, a large variability in the clinical phenotype can often be seen [4]. With HTS, it is becoming cost-effective to explore complex models of genetic diseases, such as genetic factors that modulate the phenotype, that can be associated with the variability observed in the phenotype [5]. Second, the resolution at which the genotype is defined may not adequately reflect the way in which the phenotype actually segregates. For NSHL the genotype is typically defined by the genomic locus containing the putative mutation. In some cases, the locus level can be too coarse, where for instance differences in phenotype are caused by different mutation types within a single large locus and/or gene [6].

To allow the phenotype data to drive the discovery of subtypes (of phenotype and genotype) and identify better class labels, a novel clustering technique was developed called Hierarchal Surface Clustering (HSC), and has been shown to be particularly well-suited to clustering using audiometric data. Along with HSC, a novel 3D surface visualization technique was developed to allow a human expert to guide exploration of the sub-phenotype space, by observing the progression of the hearing loss with age. Using simulated data it was shown to perform better or have comparable performance to

K-means and spectral clustering. To evaluate HSC using real world data, a previously known difference in the phenotype among patients with mutations in the DFNA2A locus was observed for different mutation types [6]. A gold standard clustering assignment was generated by assigning patients in DFNA2A to clusters based on their mutation type – truncating or non-truncating. The Adjusted Rand Index was used to evaluate the clustering assignment found via the different clustering techniques versus the gold standard. HSC had the highest ARI with a value of 0.459 compared to 0.187 for spectral clustering and 0.103 for K-means clustering. When applying HSC to the DFNA8/12 locus, two clusters were found for DFNA8/12 that exhibited different phenotypes, which could not be attributed to any environmental or other known genetic causes. This demonstrates the ability for HSC to utilize phenotype data to drive the discovery of potential subgenotypes and genetic modifiers and to serve as a hypothesis-generating tool.

The two methods developed in this thesis, AudioGene and HSC, complement each other with AudioGene being a tool for prioritizing loci using phenotypic data and HSC developed as a tool to explore the phenotype to identify novel subclasses or genetic modifiers that could eventually be used to improve AudioGene. The general structure of the topics covered in the subsequent chapters are as follows: Chapter 2 contains the relevant clinical background and the necessary machine learning background, Chapter 3 discusses the method prediction of genotype from phenotype called AudioGene, Chapter 4 describes the hierarchal surface method developed for discovery novel subclasses along with a novel method for visualization progression of hearing loss, and finally Chapter 5 is the conclusion.

CHAPTER 2

BACKGROUND

2.1 Clinical Background

2.1.1 Measuring Hearing

An audiometer is a device used to measure hearing and produces tones of calibrated frequency and intensity. In order for an audiometer to be used for clinical diagnosis, it must meet the specifications listed in ANSI S3.6-2004 [7]. The American Speech-Language-Hearing Association (ASLHA) publishes guidelines for measuring pure-tone thresholds, but do not required them to be followed [8]. A threshold is defined as the lowest decibel hearing level at which responses are elicited in at least fifty percent of the measurements, when no other sound is present [8]. The conical set of pure-tones measured is: 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz and 8 kHz. The recommended order of frequencies measured is as follows: measure 1 kHz to 8 kHz, repeat 1 kHz, then measure 500 Hz down to 250 Hz, and then repeat 1 kHz. The repeated 1 kHz threshold measurement step is to validate that the patient is responding appropriately and that the measurements are consistent. It is recommend that if a difference of more than 20 dB between two octaves is found then interoctive frequencies should be measured (1.5 kHz, 3 kHz, and 6 kHz). After threshold values are obtained, the values are plotted on an audiogram with the frequencies measured on the x-axis and the hearing loss in dB loss, also referred to as dB Hearing Level, on the y-axis with 0 starting at the top. It has been observed that a test re-test variability of between 5-10 dB is present at every frequency. [9].

Interestingly, the first modern audiometer was developed at the University of Iowa by Carl E. Seashore in 1897 [10], and was designed to measure “keenness of hearing”. It consisted of an earpiece from a telephone receiver, battery, induction coil, galvanometer, resistance coil, and a few switches. The resolution of intensity was limited

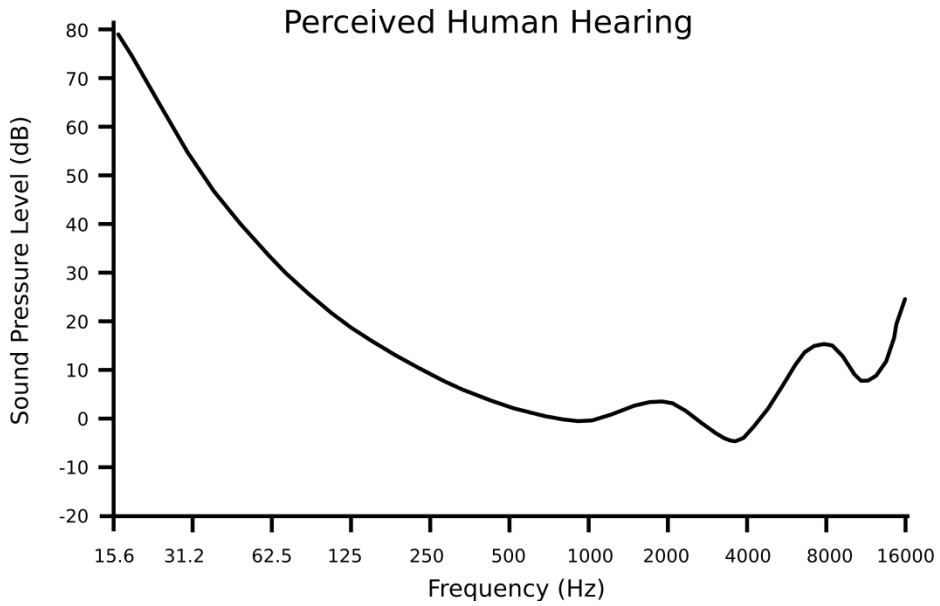


Figure 1. The absolute threshold of hearing (ATH)–minimum sound level at which the pure tone is perceived–plotted for frequencies ranging from 15 Hz to 16 kHz. The lower the sound pressure level, the quieter the sound that can be perceived. The hearing loss recorded on an audiogram is relative to the ATM.

to forty levels of sound intensity. The sound was generated by repeatedly turning a knob that interrupted a contact to create a clicking sound, or specific tones produced by connecting the audiometer to an external tuning fork apparatus.

Seashore's audiometer was limited to laboratory testing [11], and the first commercially available audiometer was the Western Electric 1-A and was developed in the 1922 with the advent of the vacuum tube [12]. The original procedure for the 1-A audiometer was to record the hearing loss as percentage of normal hearing at each frequency with 100% indicating normal hearing. The plot looks almost identical to the current audiogram with the exception of the y-axis being percent hearing instead of dB loss. In 1926, the standard audiogram that is still be used to day was finally developed with standard units of hearing loss or hearing level in decibels being defined [13]. The

decibel value used to measure hearing loss is relative to the absolute threshold of hearing (ATH)–minimum sound intensity that can be heard–of each frequency. The plot of the ATH versus frequency can be seen in Figure 1. Smaller sound pressure levels (SPL) values coincide with quieter sounds, and the most sensitive frequencies are between 1 kHz and 4 kHz.

2.1.2 Genetic Hearing Loss

Hearing loss is defined as the reduced hearing acuity during auditory testing. A person's hearing acuity is classified as normal when it falls within 20 dB of hearing loss, and hearing loss otherwise termed: mild (20-40 dB), moderate (41-55 dB), moderately severe (56-70 dB), severe (71-90 dB) or profound (>90 dB). Hearing loss can be further characterized as low frequency (<500Hz), mid-frequency (501-2000Hz) or high frequency (>2000Hz) [14]. Hearing loss can be caused by genetic or environmental factors. Genetic hearing loss is classified by the mode of inheritance: Recessive, Dominant, X-linked, and mitochondrial [4]. Autosomal Recessive Non-syndromic Hearing Loss (ARNSHL) accounts for approximately 77-93% of inherited hearing loss cases in neonates, and Autosomal Dominant Non-syndromic Hearing Loss (ADNSHL) accounts for about 10-20% of the cases [15]. Fractional percentages of cases are caused by mutations that are X-linked or in the mitochondria. For the scope of this thesis only non-syndromic hearing loss is considered, and therefore syndrome hearing loss, such as Ushers syndrome will not be discussed.

2.1.2.1 Autosomal Recessive Non-syndromic Hearing Loss

Mutations in 42 different genes have been reported as causes of Autosomal Recessive Non-syndromic Hearing Loss (ARNSHL). Of the 42 genes, mutations in GJB2, DFNB1 are the most prevalent cause of ARNSHL [15]. Previously, mutations in DFNB1 have been reported to account for between 20%-50% of all ARNSHL cases varying by population [16]. The most common mutation is a homozygous 35delG

mutation, and is estimated to account for up to 70% of the DFNB1 related cases in the Caucasian population [2]. The gap junction protein 2, also know as Connexin 26, is the protein product of GJB2 and is thought to be responsible for maintaining the correct K^+ ion levels in the inner ear [15]. Different mutations within GJB2 have been shown to have varying phenotypes with the majority being high frequency hearing loss [2]. When grouped by mutation type, the most profound hearing loss was seen in patients with homozygous truncation mutations, whereas homozygous non-truncating mutations had less profound hearing loss.

2.1.2.2 Autosomal Dominant Non-syndromic Hearing Loss

For autosomal dominant non-syndromic (ADNSHL), no single gene accounts for the majority of cases. There are currently 64 ADNSHL-mapped loci, with genes identified for only 34 of the loci. Current data suggest that of these 25 genes, mutations in WFS1 (DFNA6/14/38), KCNQ4 (DFNA2A), and COCH (DFNA9) are somewhat more common as causes of ADNSHL in comparison to the other 21 genes [16]. Interestingly, mutations in a few genes such as WFS1, COCH, and TECTA cause an easily recognizable hearing loss patterns [3].

The KCNQ4 gene or DFNA2A locus is a 695 amino acid protein expressed in the outer sensory hair cells and is responsive for recycling K+ ions after the stimulation of the hair cells [17]. The general phenotype for DFNA2A hearing impairment is progressive hearing loss at all frequencies with more attenuation at higher frequencies [18]. Patients with truncating mutations exhibit a distinctive hearing loss pattern with only high frequency hearing loss with no progression with age [6].

The COCH (DFNA9) gene encodes for the cochin protein, an extracellular protein, and has been found to be the most abundant protein in the inner ear [19]. It is not completely understood how mutations in DFNA9 cause hearing loss, but deposits in the inner ear that consist of cochin protein have been found in patients with DFNA9

mutations [20]. The clinical presentation of DFNA9 related hearing loss is a flat progression with age that is uniform.

DFNA8/12 contains the TECTA gene and has a clinical presentation that is described as "U"-shaped or "cookie bite" because the hearing loss is most pronounced in the mid frequencies [18]. The TECTA gene encodes for α-tectorin and is expressed in the tectorial membrane (TM), an extracellular matrix that runs the length of the cochlea. Mutations in different domains of the protein have been suggested to cause variations in the phenotype with mutations in the ZP domain showing mid-frequency hearing loss whereas mutations the ZA domain having high frequency hearing loss [21].

In contrast, DFNA6/14/38 is one of only three known loci to be associated with low-frequency hearing loss [22]. The gene within DFNA6/14/38 is WFS1 and is expressed in many other tissues such as the brain and pancreas. Mutations in WFS1 are also responsible for Wolfram syndrome type 1. The general hearing loss phenotype of patients with mutations in DFNA6/14/38 is low frequency hearing loss at frequencies at and below 2 kHz. Patients typically retained normal speech understanding and in some cases are unaware of their hearing loss until older age or when high frequency noise induced hearing loss occurs [23]. The current function of the WFS1 in hearing is still unknown, but its protein product wolframin has been localized to the endoplasmic reticulum–an organelle in many eukaryotic cells [23]. The low frequency hearing loss phenotype has been suggested through function studies to be caused by a reduction in woframin [24,25].

2.1.3 Diagnosis of Genetic Hearing Loss

With the advent of next generation sequencing platforms such as Illumina's HiSeq and ABI's SOLiD sequencers, the cost of sequencing the whole exome – all protein coding regions in the genome – has been significantly reduced [26]. The major challenge after sequence generation is the cost of post-sequencing analysis with some

proclaiming the $1,000 genome with a $100,000 analysis cost [1]. Many efforts have focused on better downstream filtering techniques, such as filtering using allele frequency within 1,000 Genomes data, using family structure for segregating mutations, and using scores of pathogenicity of variants to obtain fewer variants for validation [26]. Even with these standard techniques of filtering, around 2% of the average 24,000 identified variants via exome sequencing are found to be "novel" – mutations that have not be previously reported [26].

At the University of Iowa, OtoSCOPE [27] was developed as a cost effective genetic test to screen all genes related to hearing loss and has a current cost of $1,500 [28]. The general steps of OtoSCOPE include an optimized pipeline with targeted DNA capture, sequencing, and post sequencing analysis of all known hearing loss genes. Targeted capture of the exons for all hearing loss genes, including all known isoforms, was done using SureSelect solution based sequence capture. Sequencing is performed using the Illumina GA_{II} sequencing platform, and the resulting reads were mapped using BWA [29] and variant call was done using GATK [30]. If no previously reported deafness causing mutation were found, then the variants of unknown significance (VUS) were ranked by type of change (non-synonymous, splice-site, or frameshift deletion), and then based on allelic frequency within known populations including the 1000 Genomes [31] and the Exome Variant Server [24] database. Pathogenic scores were assigned to VUS using BLOSUM, SIFT, PolyPhen2, and Align-GVGD. Finally, if multiple family members were also sequenced, then VUSs were further filtered based on the expected segregation within the family based on the observed mode of inheritance. Finally, the algorithm that is described in Chapter 3 was used as a phenotypic filter to predict the likely genes and/or loci that harbor the putative mutation. Once a VUS is suspected to be the putative mutation, its is validated using Sanger sequencing [32].

2.2 Machine Learning Background

2.2.1 Supervised Machine Learning

Supervised machine learning deals with the task of training a mathematical model that can be used by a computational framework to predict the labels for a set of unlabeled instances. The true label of each model-training instance is assumed to be known. A single instance in either the training or unlabeled dataset consists of a vector of feature values. The feature vectors are composed of an ordered set of attributes that may include a mixture of categorical (e.g., textual labels) or continuous (e.g., numerical) data types. The label of the instance is called the class and is usually a categorical field but often denoted as -1 or 1 for binary class problems. Classifiers are not limited to predicting only two classes. Multiple classes can be predicted but usually require an indirect strategy to utilize binary classifiers in a procedure that integrates the results of a set of classifiers, if the classifier cannot be extended to predict multiple classes directly.

2.2.1.1 Classifiers

A classifier is a function that maps instances to discrete classes. Before a classifier can be used to predict the labels of unlabeled data, it must first be trained utilizing a training set for which the label is known for a representative set of instances. There are exceptions, such as K-nearest neighbors [33] but this is the exception, not rule. While there have been many classifiers proposed, implemented and described in the literature and in practice, within the scope of this thesis the following classifiers were used and will be described briefly in this section for completeness: Support Vector Machines (SVM), Random Forest (RF), Bagging, and Majority. The simplest classifier is the Majority Classifier, which predicts the majority class regardless of the features of the unlabeled instance and is used as a a sort of “base line” against which all other classifiers may be compared.

A Support Vector Machine (SVM) [34] is a linear classifier that finds the hyperplane that separates the classes in the feature space with the widest margin possible. In practice this margin is relaxed to a soft margin that allows for finding a hyper plane in the presence of classes that are not strictly linearly separable. The instance (features) that lie closest to the margin are called "support vectors" (SVs). The desired hyper plane can be defined in terms of the sum of the dot products of the SVs and the instance being classified, multiplied by parameters learned during training. The dot product can be replaced with an arbitrarily complex "kernel function" that allows the SVM to find a non-linear decision boundary. In order to guarantee that the SVM will converge towards a globally optimal solution, the kernel function must satisfy the Mercer Conditions [35]. Under the standard SVM formulation, the kernel function is often referred to as a linear kernel. Other example kernels include higher order and degree polynomials, and radial basis functions (RBFs). A polynomial kernel finds a hyper plane that is a polynomial of a specific degree, and an RBF kernel can be thought of as finding hyper spheres that separate the classes.

Bagging is an ensemble classification technique that is an acronym for steps of the algorithm *b*ootstrap *agg*regat*ing* [36]. An ensemble classifier combines the predictions of multiple (usually weaker) classifiers. Bagging generates predictions by treating the output of the multiple classifiers as equally weighted votes, in an "election" from which the final class predicted is the one that gains the majority "vote". The first step in training is called bootstrapping and is the process of generating a number of datasets that contain a random subset of the training set which are sampled with replacement. Each of the bootstrapped datasets are then used to train a different classifier, which is typically some form of a decision tree [37] but is not limited to decision trees. For unknown instances, the class is predicted based on which class is predicted most frequently from the trained classifiers.

The random forest classifier combines the bagging algorithm with random subspace mapping and uses decision trees as the classifier, hence the name random forest [38]. The bootstrapping step is performed in the same manner as bagging. The standard decision tree classifier is modified such that each node is determined using a random subset of all the features present. This prevents the forest of decision trees, in the presence a few highly informative features, from consistently using the same subset of features and leads to a set of diverse decision trees. Again, the prediction for an unknown instance is based on the most frequent class predicted by the decision trees.

All the classifiers described here are designed to be used for datasets with binary classes but can be extended to handle multiple classes. The two most-common methods are one-versus-all and one-versus-one classification which combine multiple binary classifiers to perform predictions for multiple classes [39]. One-versus-all is a multiclass strategy in which a classifier is trained for each class to distinguish that class from all other classes. The final class predicted is then based on which classifier had the highest response for a given class. Alternatively, the one-versus-one technique trains classifiers to distinguish between pairs of classes. The final class predicted is the class with the highest vote total. To obtain probabilities of class labels using one-versus-one, a method called pair-wise coupling [40] is used to combine the probabilities from all the pairs of classifiers into a single probability for each class.

2.2.1.2 Multi-instance Classifiers

For standard machine learning settings, each instance is a fixed feature vector. For some classification problems, however, a fixed feature vector is not sufficient to adequately capture the distinct essence of an instance for which multiple "sub-instances", or observations, are needed to describe a single instance. In these cases, multi-instance learning techniques are used that allow multiple instances to be grouped into *bags* that represent a single instance and are then given a single label [41]. The standard support

vector machine classifier can be modified to accept multi-instance datasets by changing the way in which the distance is computed in the kernel function by substituting the dot product with the sum of the kernel distances between each pair of instances within the bags [42]. The standard set of previously described kernels can still be used with no other changes needed to be made to the SVM. Standard classifier structures can still be applied to multi-instance datasets but require an appropriate method for combining the multiple feature vectors into a single feature vector. Often, this may involve averaging of feature vector values in the bag, but this may not be applicable to all multi-instance datasets.

2.2.1.3 Evaluating Classifier Performance

The standard approach used for estimating the performance of a classifier on unseen data is referred to as cross-validation, and involves splitting the data into *k* randomly stratified folds, training on a subset of *k*-1 folds, and then testing the recall capabilities on the one remaining set which was withheld from training. This is repeated for each of the folds, and the accuracy is calculated as the number of correctly classified instances divided by the total number of instances in the training set across all folds [39]. The value of *k* is typically set to 10, but if k is equal to the number of instances in the training set it is called "leave-one-out" analysis (LOOA). The reason for cross-validation is to avoid biasing of the performance metrics by allowing the data that is being predicted to be included in training set. This more closely simulates the real-world use-case in which the instance being predicted would not be included in the training set.

	Predicted Class	
Acutal Class	malignant	benign
malignant	75	25
benign	15	85

Figure 2. An example confusion matrix. The matrix shows the number of instances that were predicted as either the correct class of the incorrect class. The counts in the gray diagonal boxes indicate the number of instances that were correctly classified. In this example, there are 100 instances labeled "malignant" and 100 labeled "benign."

A confusion matrix can be generated from the predictions made during cross-validation where each column is the predicted class and each row is the true class label of the instances. An example of this matrix is shown in Figure 2. This matrix is useful for determining the classes that are most often being involved in misclassification events. It also gives insight into underlying causes of inaccuracy and other statistics that measure performance. A variety of metrics are useful to evaluate and compare the performance of different classifiers to choose the ones that perform best on a given dataset. Choosing the performance metric that should be used to determine which classifier is yielding the best performance is specific to each problem and dataset. Two common metrics are *accuracy* and receiver operator characteristic (ROC) [43].

The ROC curve is based on plotting the false positive rate versus the true positive rate over all possible decision thresholds [44]. For classification problems in which the classifier also produces probability values for class assignments, the decision threshold is the probability value (usually greatest) that implicates the final class assignment. For binary classification problems, the standard convention is that classes are named positive (+1) and negative (-1). The ROC curve is computed using the probabilities outputted from the classifier during cross-validation for each instance and the probability is typically defined as the probability that the instance belongs to the positive class (the probability of the instance belonging to the negative class is simply one minus that

value). A threshold is set that assigns the positive class label to all instances with a probably greater than or equal the threshold. The ROC curve is then generated by: rank ordering the instances with their probabilities in decreasing order, varying the threshold value form 0 to 100%, and computing the false positive and true positive rate for every possible threshold value. The false positive and true positive rates found for every possible threshold value are then used to plot points on an x-y plane, which forms the ROC curve. The area under the ROC curve (AUC) can then be calculated and has a range of values between 0 and 1. Random guessing of the class labels has an AUC value of .5 and an AUC value of 1 indicates that the classifier ranked, in terms of class probability, every instance of one class higher than the other. An intuitive explanation of the AUC value is that it is the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.

To compare statistics of different classifiers to determine which has the best performance for a given dataset, cross-validation can be repeated several times with different randomly sampled subsets and then a t-test can be performed to determine if the difference is statistically significant. Ideally, a LOOA would be performed and the statistics compared for the different classifiers. For a large dataset and classifiers that have complex training procedures, this can be computational prohibitive. Cross-validation can also be used to estimate parameters for a classifier such as the cost of miss-classification for an SVM–usually denoted as C–or the number of trees used in random forest.

2.2.2 Unsupervised Learning or Clustering

Unlike supervised classification, clustering techniques are used when the labels of a given dataset are unknown. The goal is then to partition the instances into groups or clusters that are more similar to each other rather than to instances from other clusters [45]. The three main applications of clustering are: finding underlying structure, natural

classification, and compression [46]. Finding underlying structure includes gaining insights from the data, generating hypotheses about the data, and finding outliers or anomalies. Natural classification deals with grouping organisms into a system of ranked taxa. Compression via clustering refers to methods that summarize the data (compress) by representing the instances as some combination of the clusters.

2.2.2.1 Clustering Techniques

Clustering techniques can be divided into three categories based on the way in which they model that data, and their definition of what constitutes a valid cluster. Clustering algorithms can be placed into one of two categories, partitioned and hierarchical [46]. Partition-based approaches seek to divide the data into a preset number of clusters, whereas hieratical clustering algorithms create a tree-like structure wherein instances or subgroups are merged until all the instances are in a single cluster.

The best-known partitioning clustering algorithm is the K-means algorithm [47], where K is defined as the desired number of clusters. The K-means algorithm begins by selecting K instances, then defines an initial centroid for each of K clusters to be the "location" of the instance itself. The K centroids are then repeatedly updated by alternating between assigning instances to the closest centroid, in terms of an appropriate distance metric, and then calculating the new centroid based on the arithmetic or geometric mean of the features of the all instances now assigned to the cluster. Typically, the Euclidian distance is used as the distance metric but other distance metrics have been used such as city block or Pearson correlation [48]. Variations on K-means exist, such as the aptly named K-median algorithm [49] that use the median instead of the mean in order to be more robust to outliers. Due to the gradient descent nature of K-means, the performance is dependent upon the initialization of the centroids [50]. Various methods for initialization have been developed, and a comprehensive analysis of the common initialization techniques can be found in [51].

Alternatively, expectation-maximization (EM) [52] assumes that the instances were generated from N probably distributions, and attempts to estimate the parameters of the N probability distributions to the data. Cluster membership is determined according to the probability distribution most likely to generate each instance. A common distribution that is used in EM is a mixture of a Gaussian model, which assumes that the probability distributions are multivariate Gaussians distributions [53]. The model parameters for Gaussian distributions (mean and variance) are considered latent variables, and the goal of EM is to compute maximum likelihood estimation of the parameters based on the data. EM clustering alternates between two steps: (E) expectation and (M) maximization. In the E step, the current estimation of the model parameters are used to compute the posterior probability–often referred to as the responsibility–of the model parameters for every instance, and in the M step the responsibility is used to re-estimate the model parameters. These two steps are repeated until the model parameters converge. A primer which contains several practical examples of expectation maximization can be found in [54].

Non-parametric approaches to clustering make different assumptions about the definition of clusters. For instance, density based clustering techniques define clusters as contiguous regions of high density separated by regions of low density [55]. Since these methods do not make assumptions about the underlying distribution of the data, they are able to define clusters of arbitrary shapes. The classical density based algorithm is DBSCAN [56], which takes as input two parameters MinPts and Eps. These parameters define the minimum points (MinPts) needed in a defined radius (Eps) for a point to be considered apart of a cluster. The results of DBSCAN are highly depended on the choice of the two parameters [57,58].

A recent non-parametric approach is called spectral clustering [59]. There are three main steps to spectral clustering: (1) computing an affinity matrix, (2) examining the eigenvectors of the affinity matrix, (3) clustering of the eigenvectors using K-means.

The main difference between spectral clustering algorithms is how the eigenvalues and vectors are used. The input parameters to spectral clustering are the number of clusters and sigma, which controls how the distances between the instances are computed. The standard equation for the distance is the Gaussian distribution, $e^{\frac{\|s_i - s_j\|}{\sigma^2}}$, where s_i and s_j are two instances. The affinity matrix contains the pairwise distances between each pair instances. The affinity matrix can also be viewed as a fully connected graph with the weights of each edge corresponding to entries in the affinity matrix. Then, the eigenvalues and eigenvectors of the affinity matrix are computed. In [60], the largest N eigenvectors are concatenated to form a matrix with the number of rows being equal to the number of instances by N, where N is the number of clusters. The rows are then used as features for K-means clustering, with K being equal to N. The *i*-th row of the combined eigenvector matrix corresponds to the *i*-th instance in the affinity matrix.

Alternatively, spectral clustering can be viewed as performing graph-based operations on the affinity matrix. Depending on the manner in which the affinity matrix is normalized, the method has been proven as being equivalent to computing the approximate solution to the normalized mincut (ncut) problem of a graph, which is NP-Hard [61]. The normalized mincut of graph G is defined as:

$$\mathrm{ncut}(A, B) = \frac{\mathrm{cut}(A, B)}{\mathrm{assoc}(A, V)} + \frac{\mathrm{cut}(A, B)}{\mathrm{assoc}(B, V)}$$

$$\mathrm{cut}(X, Y) = \sum_{u \in X, v \in Y} w(u, v)$$

$$\mathrm{assoc}(X, V) = \sum_{u \in X, t \in V} w(u, t)$$

With V being the set of verities of graph G, w(u,v) being the weight of the edge between vertex *u* and *v,* and A and B being two mutually exclusive sets of vertices whose union is the set of all vertices V. Similarly to density based clustering, the results of spectral clustering are largely dependent upon the choice of parameters, specifically sigma, and choosing the appropriate value is nontrivial [62]. Attempts have been made to

automatically choose the value of K, but appear to only work well if the clusters being found are reasonably separated [62].

The other category of clustering algorithms is hierarchically clustering techniques. Most hierarchal clustering algorithms can be considered a variant of three standard algorithms: single linkage, complete-linkage, and minimum variance [46]. Single linkage [63] initializes every instance as a separate cluster and combines the clusters based on the minimum distance between the clusters. The algorithm terminates once all the instances have been merged into a single cluster. The mergers are usually presented visually as a dendrogram, which shows every merger of the algorithm starting at the beginning with each instance as a single clustering. Complete linkage follows the same steps except it computes the maximum distance between clusters.

2.2.2.2 Evaluating Clustering Techniques

The two standard approaches of evaluating clustering techniques are intrinsic and extrinsic measures [64]. Intrinsic measures compute values based on cluster compactness and distance from other clusters. Extrinsic measures require a gold standard clustering assignment to be known. A commonly used choice for extrinsic measures is the Adjusted Rand Index (ARI) [65,66]. Other extrinsic measure include the F-measure, precision, and recall.

The ARI compares every pairwise combination of assignments for all instances and evaluates them against their corresponding cluster assignment in the gold standard. The ARI is based on the Rand Index (RI) [67], but considers the cluster assignments against assignments made by change. The ARI has a range from -1 to 1, whereas the RI has a value of 0 to 1. An ARI of 1 means that the two cluster assignments are in complete agreement. A rank of zero means that the assignments are equal to those made by randomly assigning instances to clusters, and less than 0 means that the cluster assignment are worse than those by chance but large negative values are less likely[68].

The ARI is also valid for comparing cluster assignments of different number of clusters [66].

The original Rand Index is computed as follows. Given a gold standard assignment of U and a resulting clustering assignment of V, four values are computed:

(a) number of pairs that were assigned to the same cluster in both U and V

(b) number of pairs that are in different clusters in U and in V

(c) number of clusters that were the same in U but were different in V

(d) number of clusters that were different in U but the same in V

Once those four values are computed, the RI is simply equal to (a+b)/(a+b+c+d). The numerator is considered the agreement between the two. The RI is converted to the ARI value by:

$$\text{ARI} = \frac{\text{RI-ExpectedIndex}}{\text{MaxIndex-ExpctedIndex}}$$

Where the ExpectedIndex is equal to value value of RI expected by chance, and the MaxIndex is maximum RI index expect by random chance.

2.2.3 Linking Phenotype and Genotype

Previous work on linking phenotype to genotype has focused on predicting in both directions. For predicting phenotype from genotype, predictions of hair, skin, and eye color of criminals from DNA evidence alone has been shown to be possible [69]. Using regression, it was shown that visual phenotypes of patients with compound heterozygous mutations in ABCA4 could be modeled additively in which each allele has a fixed contribution to the final phenotype. Using machine learning techniques, a model was developed to predict the drugs resistant of HIV-1 strains based on a set of 471 strains where the drug resistance was known [70].

Alternatively, predicting genotype from phenotype, a modified C4.5 decision tree was trained to predict the functional class of Open Reading Frames (ORF) in *S. cerevisiae* from phenotypic data [71]. In this case, the phenotype data was sensitivity or

resistance of a strain of *S. cerevisiae* to certain drugs with a known ORF deleted. Applying the trained C4.5 tree on unknown strains with deleted ORF, functional classes were predicted for the ORFs with unknown functional classes. A new type of analysis called Symptom- and sign-assisted genome analysis (SSAGA) [72] was developed for the prediction of a candidate disease gene set using clinical features (phenotypes) of 591 recessive diseases found in pediatric patients. Each of the 591 diseases was mapped to a subset of 227 clinical terms from nine symptom categories, and each gene was represented by, on average, 8 of these terms. Labels were assigned to a patient based the symptoms and signs of their clinical presentation, and genes that matched the terms were included in a candidate disease gene set. When retroactively testing on 533 children, a sensitivity of 99.3% was obtained based on the criteria that the correct gene was listed in the candidate disease gene set. On average 194 genes were nominated by the SSAGA analysis. The candidate disease gene set was then used to filter out variants identified in exome sequencing that were not within the list of genes in the candidate set. Another approach of predicting genotype from phenotype was demonstrated in the prediction of the genotype of patients with an inherited heart condition known as long QT syndrome using the measurements from a treadmill exercise routine [73].

CHAPTER 3

PREDICTING HEARING LOSS GENOTYPES FROM PHENOTYPES

3.1 Introduction

Hearing loss is the most common sensory deficit in Western societies [14]. In the United States, congenital hearing loss occurs three times more frequently than Down Syndrome, six times more frequently than spina bifida, and at least 50 times more frequently than phenylketonuria [74-76]. It is currently estimated that 1 child in 1,000 born suffer from some form of hearing loss and it is estimated that about half of those children have an inherited genetic cause [77]. This chapter is based on work published in "AudioGene: Predicting Hearing Loss Genotypes from Phenotypes to Guide Genetic Screening" [3].

Previous studies identified phenotypic differences in hearing loss that were dependent upon the genotype for non-syndromic hearing loss [2,78]. Using this observation, the goal was to create a tool that could predict the causative genotype (locus) from the phenotype (hearing loss pattern) for patients with non-syndrome hearing loss. Originally, the prioritization was designed to reduce cost and time of Sanger Sequencing by prioritizing the genes and loci that would be sequenced. In the era of next-generation sequencing, where all known protein encoding genes can be screened in parallel for considerably low cost, the method is still relevant as a method for prioritizing variants of unknown significance that are found during next generation sequencing. In this chapter, the method developed for predicting the genetic cause of patients with two forms of genetic hearing loss is described and an estimation of the performance of the method is also made. The system is named AudioGene and encompasses both the method and the publicly available web interface that allows the analysis to be performed on our servers.

3.2 Background

3.2.1 Hearing Loss Background

Hearing loss is defined as reduced hearing acuity during auditory testing. Hearing is measured in decibels hearing level (dB HL) with a frequency-specific normative threshold of 0 dB defining the level at which normal controls perceive a tone burst of a given intensity 50% of the time. A measurement of these thresholds across several frequencies is known as an audiogram. A person's hearing acuity is classified as normal when it falls within 20 dB of these defined thresholds, with hearing loss otherwise graded as mild (20-40 dB), moderate (41-55 dB), moderately severe (56-70 dB), severe (71-90 dB) or profound (>90 dB). Hearing loss can be further characterized as low frequency (<500Hz), mid-frequency (501-2000Hz) or high frequency (>2000Hz) [14].

3.2.2 Autosomal Dominant Non-syndromic Hearing Loss

Approximately 20% of the inherited hearing loss cases are caused by Autosomal Dominant Non-syndromic Hearing Loss (ADNSHL) [79]. ADNSHL is defined as hearing loss that is inherited and not associated with any other symptoms that are caused by a common genetic disorder such as Usher syndrome, which causes both hearing loss and progressive vision loss. There are currently 64 ADNSHL-mapped loci, with genes identified for only 34. Interestingly, mutations in a few genes such as WFS1, COCH, and TECTA cause an easily recognizable hearing loss pattern. This observation suggested that an automated tool could be developed for predicting hearing loss genotypes [2].

3.2.3 Autosomal Recessive Non-syndromic Hearing Loss

In 75-80% of inherited cases, both parents have normal hearing and the genetic cause is classified as Autosomal Recessive Non-syndromic Hearing Loss (ARNSHL). Currently, 95 ARNSHL loci have been found and approximately 50% of the genes for the loci have been found [80]. The most common cause of ARNSHL is the 35delG

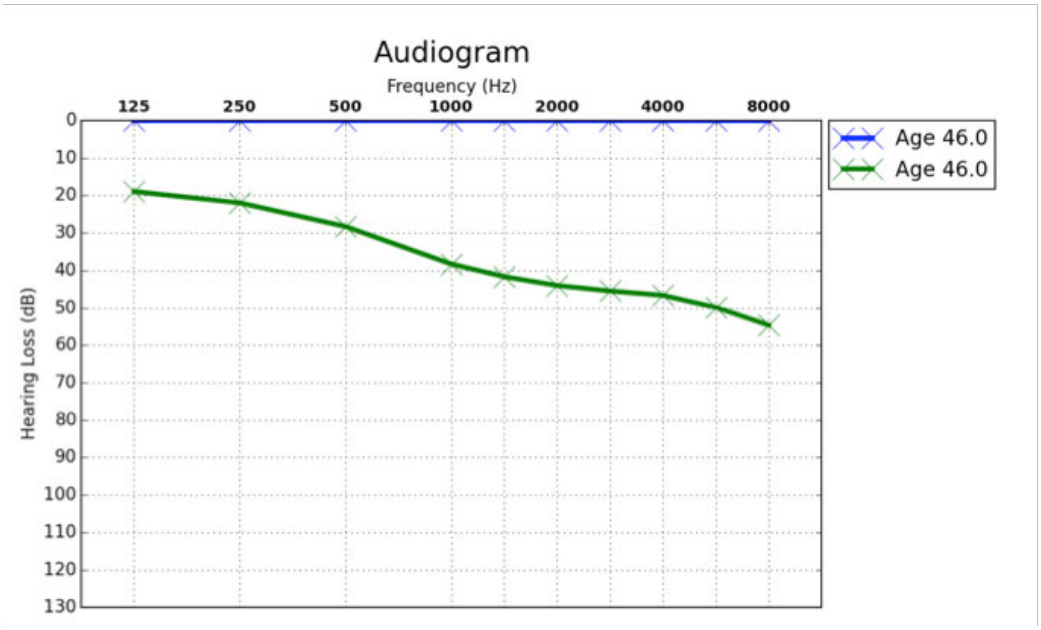


Figure 3. An example audiogram with the discrete frequencies measured along the x-axis and the amount of hearing loss in dBs along the y-axis. The blue audiogram (top) represents the expected audiogram for a normal patient with 0 dB of hearing loss across all frequencies. The green audiogram (bottom) is of a patient with slight hearing loss in the lower frequencies but has moderate hearing loss at higher frequencies and would be considered a down-slopping audiogram.

homozygous mutation in the GJB2 gene and it is believed to be responsible for approximately 70% of all ARNSHL cases in the Caucasian population [15]. It has also been shown the phenotype of ARNSHL varies based on the mutation and the mutation type [2].

3.2.4 Audiograms and Audioprofiles

An audiogram is a plot of a patients hearing loss with the *x*-axis being the discrete frequencies (250 Hz to 8 kHz) measured and hearing loss in dB loss along on the *y*-axis. The values on the y-axis are reversed, with normal hearing starting at the top at 0 dB loss and profound hearing loss at 130 dB. An example audiogram is shown in Figure 3. Different loci are known to have different patterns and progression of hearing loss with

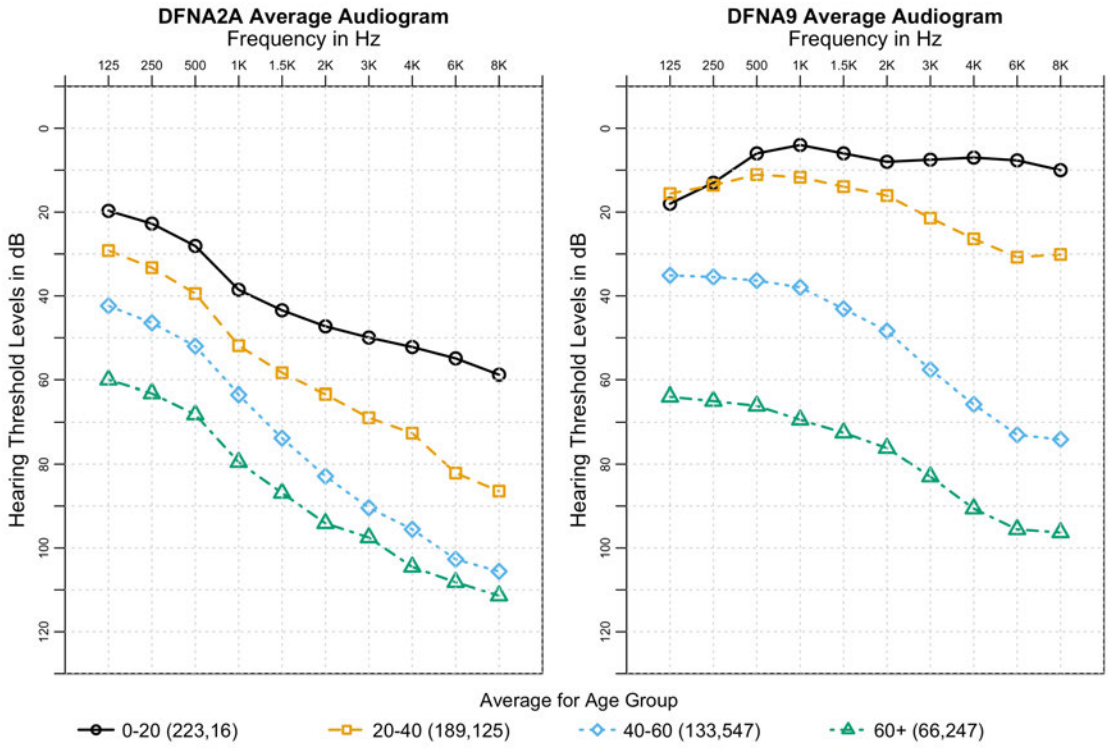


Figure 4. Sample audioprofiles from the averages of patients from DFNA2A and DFNA9 grouped into age groups spanning two decades. Average standard deviation across all ages and frequencies is 18.92 dB and 19.47 dB for DFNA2A and DFNA9, respectively. This same plot with error bars is shown in Supp. Figure 1. The number of audiograms for each age group is listed in parentheses in the legend, with the number of audiograms for DFNA2A listed first and then DFNA9. Both loci exhibit distinctly different shapes of hearing loss along with different rates of progression over time.

age. The progression does not have to be uniform for all frequencies and can be dependent upon the frequency. A visual representation of the progression with age, is called an audioprofile. An audioprofile consists of four audiograms that are the average of patients binned into twenty-year increments, i.e., 0-20, 20-40, and so forth. The audioprofile for the DFNA2A and DFNA9 locus can be seen in Figure 4. Both loci progress with age, but the initial hearing loss at birth is considerably different, and this difference forms the basis of the ability for AudioGene to predict the genotype. The same audioprofile with error bars can be seen in Appendix A, but is not very useful due to the large error bars.

3.3 Methods

The method that was developed for predicting genotype from phenotype consisted of several steps: preprocessing, training the classifier, and finally making predictions for unseen patients. The full pipeline can be seen in Figure 5. In order to demonstrate the effectiveness of the method, standard metrics were used to evaluate the accuracy and performance of this approach.

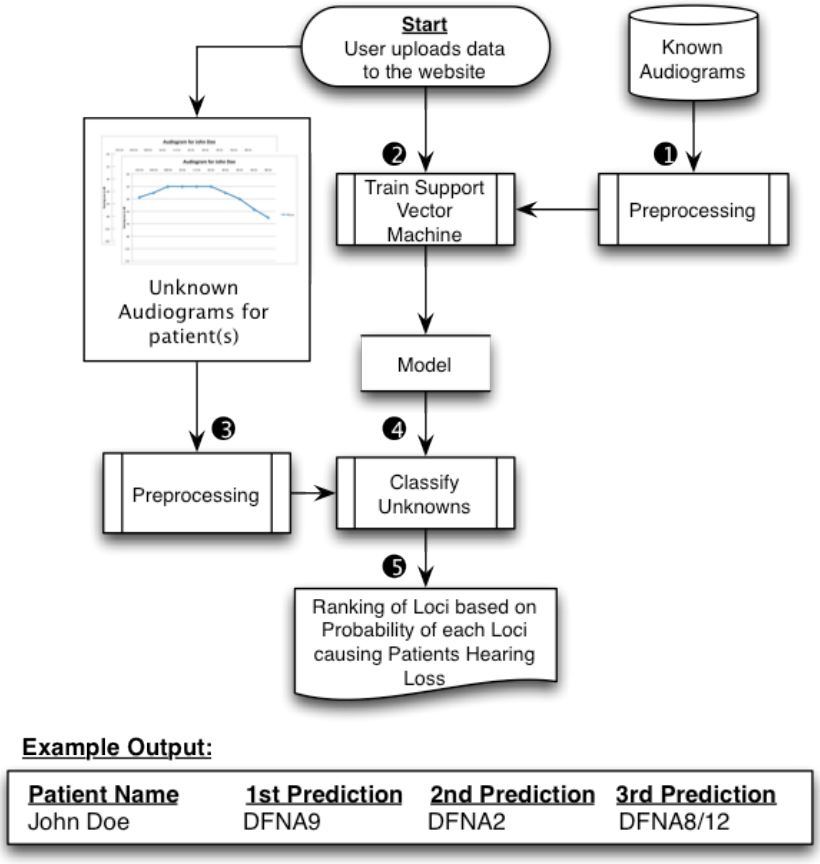


Figure 5. The final analysis pipeline of AudioGene used to make predictions for unknown patients. (1) The training set is preprocessed by filling in missing values and adding coefficients of fitted second and third order curves. (2) A Multi-Instance SVM is trained on the preprocessed training set from step 1. (3) Unknown patients' audiograms are preprocessed in the same manner as described in Step 1. (4) Probabilities for each locus are generated by the trained SVM model. (5) Loci are finally ranked by their probabilities, with results being displayed on the website and emailed to the user.

3.3.1 Audiometric Data

The dataset used to train AudioGene consists of audiograms collected from publications, original audiograms provided by authors of the publications, and by various otolaryngology and audiology clinics. Our dataset was comprised of 3,312 audiograms from 1,445 patients. The typical audiogram included data for six frequencies: 250Hz, 500Hz, 1kHz, 2kHz, 4kHz, and 8kHz. Measurements for frequencies at 1.5kHz, 3kHz and 6kHz were also present for a portion of the patients, but were less common than measurements for the six other frequencies. Audiograms with fewer than four frequencies measured were excluded from the dataset, and reduced the number of audiograms to 3,024 audiograms in the training. The total number of patients and audiograms for each locus is listed in Appendix B.

3.3.2 Preprocessing

Audiograms in the dataset were preprocessed prior to their use in prioritization or training. If available, audiograms from both ears that were taken at the same time were combined by retaining the minimum value (i.e. better acuity) at each frequency. This results in a composite audiogram that has the least amount of hearing loss at each frequency. Coefficients of second and third order polynomials were then fit to each audiogram and added as secondary features after the coefficients interpolation. Linear interpolation and extrapolation were used next to replace missing threshold values. Multiple audiograms for patients were grouped into a 'bag' for use with multi-instance classifiers, with a one-to-many relationship between patients and audiograms [81]. For classifiers that did not support multi-instance datasets, each bag was reduced to a single representative audiogram using the average of the audiograms in the bag.

3.3.3 Prioritization

Since the goal is to rank a set of loci for a patient for screening or to be used for determine phenotypic concordance of variants found via HTS, loci are ranked according

to the probabilities generated by a modified Support Vector Machine (SVM) using a linear kernel capable of utilizing multi-instance datasets. An implementation of the multi-instance SVM (MI-SVM) was used from the Weka machine learning toolkit [82,83]. SVM training was performed using the Sequential Minimization Optimization algorithm (SMO), in which a one-versus-one strategy is used to handle multiple classes in conjunction with pair-wise coupling to generate the probabilities for each locus [84]. Since probabilities of SVMs are not well calibrated, they are only useful in ranking. The multi-instance SVM processes the bagged audiograms at the kernel level, where the kernel distance between two patients is the sum of all pairwise kernel distances between all pairs of audiograms in each patient's bag. The loci/genes are then ranked in decreasing order of probability to produce a prioritized list of loci to inform genetic testing efforts. While these probabilities are useful for ranking they are not regularized, and are therefore only useful as relative probabilities.

3.3.4 Classifier Choice

Five classifiers were evaluated using two strategies: 1) Accuracy, area under ROC curves (AUC), precision and recall were computed for each classifier using ten 10-fold cross-validation experiments. AUC, precision and recall were then computed for each class using a weighted average based on the size of each locus. 2) leave-one-out analysis (LOOA) was performed of the aforementioned prioritization method using each classifier. Audiogram bags corresponding to each patient were removed from the training set one at a time, and the prioritization method was performed with the classifier trained on the dataset with the patient removed. For this analysis, patients were considered correctly classified if their locus was ranked amongst the top N loci, using the ranking method described in the previous section. SVM, Multi-instance SVM (MI-SVM), a Majority classifier, Random Forest [38], and Bagging [36] were each tested as classifiers. Both SVM implementations used a linear kernel and all of the classifiers were

derived from implementations in Weka [83]. The Majority classifier was considered the baseline against which the performance of all others was measured.

3.3.5 Validating Preprocessing

A leave-one out analysis of various combinations of preprocessing steps was performed on the training set. These permutations included combining only audiograms taken from different ears at the same age, combining and filling in missing (frequency) values, and adding the coefficients of fitted second- and third-order polynomials.

3.3.6 Noise Model and Robustness to Noise

A noise model was developed that represented real-world noise associated with the measurement and recording of audiometric data. This model was then used to perform a simulation to determine the robustness of our method in the presence of noise. The noise model takes into consideration a mis-calibrated audiometer and test-retest variability [9]. According to our model, a mis-calibrated audiometer could result in an additive (+/-) shift across an entire audiogram, and the test-retest variability of between 5 and 10 dB differences between measurements taken at two different times for the same patient. The noise model adds noise in the frequency domain. In other words, the added noise is based on treating the frequency values as values in the domain (x-axis) and the dB loss/gain as values along the range (y-axis). The Discrete Cosine Transform (DCT)[85] was used to transform the audiogram curves into the frequency domain. The DCT was chosen over the Fourier transform for simplicity, because all DCT components are all real-valued. The DCT transform function F is shown in Equation 3.1.

$$F_k = a_k \sum_{n=0}^{N-1} f_n \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right), k = 0, \ldots, N-1 \qquad (3.1)$$

$$\text{where } a_k = \begin{cases} \frac{1}{\sqrt{N}}, & k=0 \\ \sqrt{\frac{2}{N}}, & k>0 \end{cases}$$

Once in the frequency domain, noise is added in two parts. First, a random magnitude of noise is added to F_0 (the DC component) in order to shift the entire audiogram. This mimics the case where the calibration of the audiometer results in uniform inaccuracy for the entire measurement of the audiogram. Next, Gaussian noise is added to the other coefficients with a magnitude scaled by an exponential decay function. This simulates the test-retest variability discussed above. The exponential decay function effectively concentrates the noise in lower frequency components of the DCT and results in noisy audiograms that still retain their overall characteristic shape. With the addition of this noise, an inverse DCT was performed to recreate a time-domain audiogram. A few examples of the noise added to an audiogram are shown in Figure 6, with *ShiftScale* at 10 and *Scale* at 5. *ShiftScale* is defined as a scalar value that controls the magnitude with which the audiogram can be shifted, and *Scale* as a variable used to control the degree with which overall curve shape is changed. Lower values of both of these variables mean lower noise, and vice versa. The equations used to add noise to the DCT coefficients with parameters Scale and Shift scale are shown in Equation 3.2 and 3.3.

$$F'_0 = F_0 + \text{RandNorm}(0, \text{ShiftScale}) \qquad (3.2)$$

$$F'_k = F_k + \text{RandNorm}(0, \text{Scale}) e^{-\frac{k-1}{2}}, k > 0 \qquad (3.3)$$

Where RandNorm is a function that generates random numbers from a Gaussian distribution,n with the first parameter being the mean and the second parameter being the standard deviation. To determine the robustness of our prioritization method to noise, 5% of the patients were selected at random and removed from the training set. Noise was added to the removed audiograms using our noise model, with the value of *ShiftScale* always twice as large as the value of *Scale*. For a given level of noise *x, ShiftScale* and

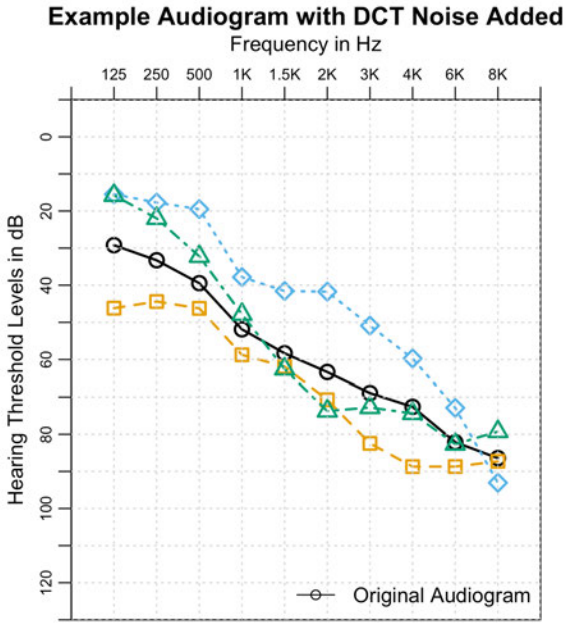


Figure 6. An audiogram with three examples of added noise, with a ShiftScale of 10 and Scale of 5. The overall characteristic shape of the audiogram still remains after noise is applied.

Scale are typically 5*x* and 2.5*x*, respectively. The prioritization method was then trained on the remaining 95% of patients, and the 5% withheld subset was classified. This process was repeated 200 times with a different random sampling of patients on each repetition. The final accuracy for a given ranking requirement (*N*) is the sum of all the patients that were correct across all iterations divided by the total number of patients that were withheld over all 200 iterations. The value of *N* specifies that the locus/gene must be ranked amongst the top N loci/genes given by the prioritization method described in the Prioritization section.

3.3.7 Identifying Outliers

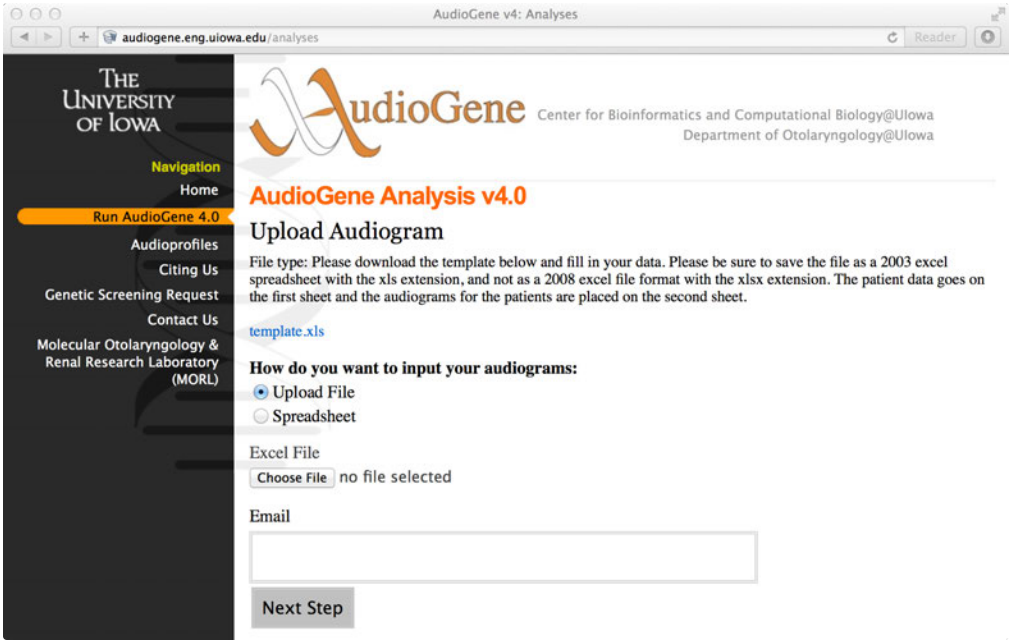
A variant of the leave-one out analysis was used to identify patients who are outliers to the classifier and are often misclassified. Each patient was removed and the classifier was retrained on the remaining patients. The noise model described in the

previous section was used to add noise to the removed patient's audiograms, with a noise Scale of 5. The patient was then classified with the retrained classifier, and the predicted locus was recorded. The classification was repeated 100 times with the noise model applied each time to the patient's original audiograms. If the correct locus was never predicted for any of the 100 iterations, the patient was considered an outlier.

3.3.8 Web Interface

AudioGene is accessible via a web interface (http://audiogene.eng.uiowa.edu) and all analyses are performed on secure servers managed by the Center for Bioinformatics and Computational Biology (CBCB) at the University of Iowa. An example of the upload page and the results page are shown in Figure 7. Audiometric data may be uploaded via a web-based spreadsheet form or by using a downloadable Excel™ spreadsheet provided on the website. After uploading data, audiograms are displayed as images to validate data entry. Once verified, the analysis can be completed using all available loci or a user-selected subset of these loci, an option that can be chosen when specific loci have already been excluded. Uploaded and verified data are submitted to a local computational cluster in the CBCB for analysis. When predictions are complete, results are made available to users online and by e-mail. Successful application of this website to genetic hearing loss has been demonstrated by the authors and others [86].

Uploading Web Interface



Results Interface

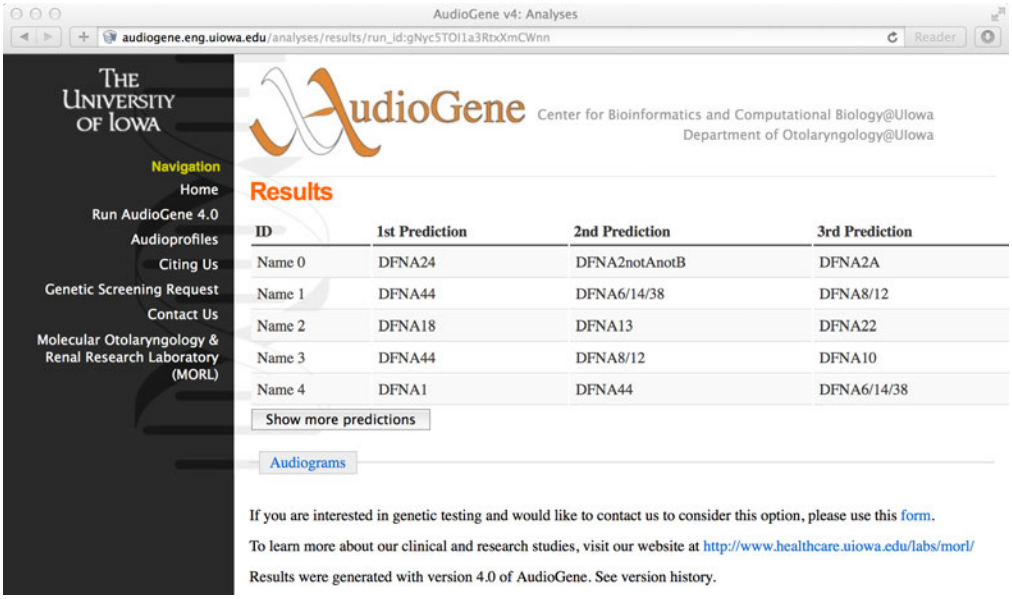


Figure 7. Screen captures of the web interface for AudioGene and is made available publically at http://audiogene.eng.uiowa.edu/. The upload interface has two methods for uploading data, the first is by uploading an excel sheet that is based on a template and the other is through the use of an online spreadsheet. The results are emailed to the user and also as a separate page. The top three predictions are displayed by default with an option to show additional predictions. The user can also compare the patients' audiograms with the audioprofiles of different loci by clicking on the audiograms button below the results.

3.4 Results

3.4.1 ADNSHL Classifier Choice and Performance

Bagging, Random Forest (RF), and Multi-Instance SVM (MI-SVM) had very comparable performance, shown in Table 1, and ROC curves for each class for these classifiers are shown in Appendix C. The bold value in Table 1 indicated the highest value, and multiple bold values indicate that they were statistically the same. AUC, precision, and recall were all computed as weighted averages based on the size of each locus. Asterisks indicate that values were not statistically significantly different from each other, and bold indicated the largest value. Based on these metrics any of the classifiers, except the SVM and Majority classifier, would obtain comparable performance. However, their performance differs when the number of guesses allowed is increased. The number of guesses allowed can be varied, and is termed as predicting the top N loci, and the accuracy can be determined when the top N loci are predicted. At higher values of N, when plotting accuracy versus the number of guesses, MI-SVM and Single-Instance SVM (SI-SVM) outperform all other classifiers, shown in Figure 8. The MI-SVM and SI-SVM have approximately equal accuracies at higher values of N but for lower values, the MI-SVM performs better. Both the Random Forest classifier (RF) and Bagging classifier perform as well as the MI and SI-SVMs at lower values of N, but at higher values of N, their accuracy reaches a maximum of approximately 91%, whereas

Table 1. Accuracy, AUC, precision and recall for all classifiers tested.

Classifier	Accuracy	AUC	Precision	Recall
Bagging	**43.86% (3.06)***	**0.82 (0.02)**	**0.37 (0.03)***	**0.44 (0.03)***
RF	**43.14% (3.16)***	0.71 (0.16)	**0.37(0.04)***	**0.43(0.03)***
MI-SVM	**42.95% (2.91)***	0.80 (0.16)	0.31 (0.03)	**0.43 (0.03)***
SVM	40.99% (2.57)	0.79 (0.17)	0.26 (0.02)	0.41 (0.03)
Majority	19.72% (0.34)	0.50 (0.00)	0.04 (0.00)	0.20 (0.00)

the MI- and SI-SVMs approach 100%. This difference is due to limitations in the training methods, since loci for which there are only a few audiograms are never predicted. Based on this analysis, the MI-SVM was chosen as the classifier for AudioGene. It has an estimated accuracy of 68% of including the correct locus/gene in the top 3 predictions. In contrast, the Majority classifier has an accuracy of only 44%. This measurement of performance is a good metric because it is similar to the intended use of AudioGene, where clinicians would sequence the predicted genes in an iterative fashion, often-times quite rapidly (days). This approach allows us to determine our accuracy in the event that multiple predictions are required before identifying the correct locus.

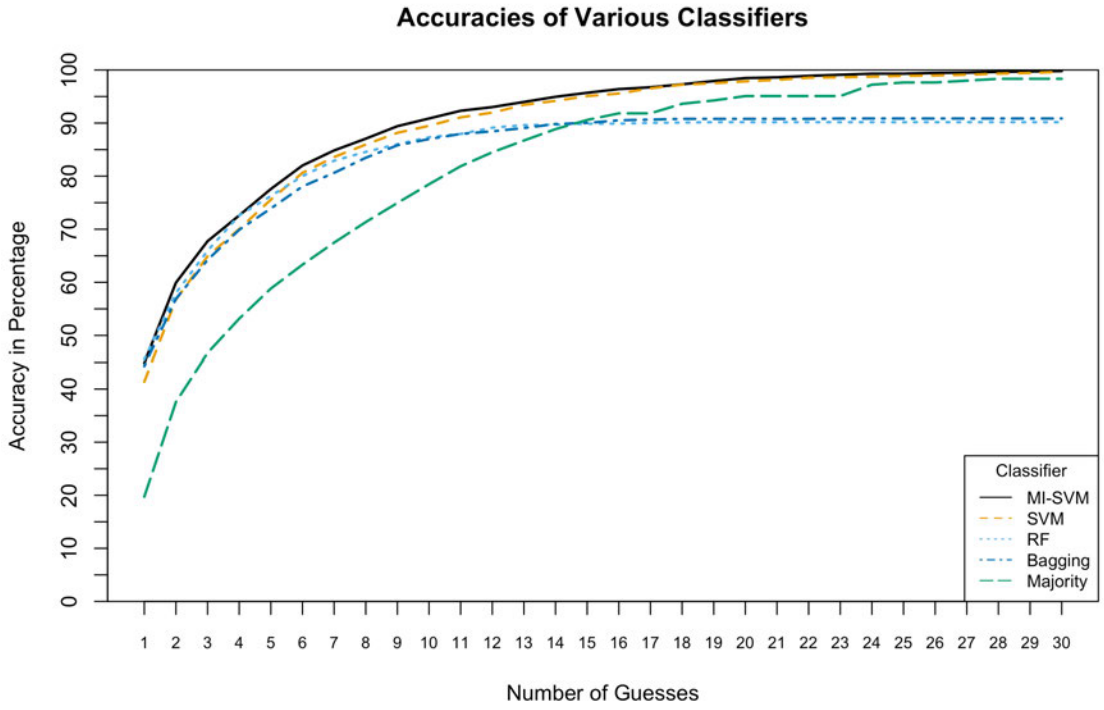


Figure 8. A comparative plot of the accuracy of the evaluated classifiers. This plots accuracy against N, where N represents whether or not the correct locus was ranked among the top N loci. Both SVMs outperform all other classifiers and the Multi-Instance SVM (MI-SVM) demonstrates the best accuracy of all.

3.4.2 Preprocessing Validation

When N (number of ranked predictions examined) is allowed to increase, the use of raw data (without any preprocessing) outperformed analyses which employed data which was preprocessed in any way It was hypothesized that this was due to a bias in which frequencies were measured and that measurements were not missing randomly, but rather were dependent upon their constitutive loci. This hypothesis was evaluated by converting audiograms into binary vectors in which the frequency values were coded as 1 if a threshold measurement was available or 0 if there was no measurement. A10-fold cross-validation was then run with an SVM and its accuracy was compared against a Majority classifier. Accuracies should have been similar if no information was contained in the missing frequency values, but the MI-SVM produced an accuracy of 33% while the Majority classifier had an accuracy of 20%. Therefore, it was concluded that filling in the missing values was necessary to eliminate this bias.

Based on the bias found by not filling in missing values, a further analysis in which polynomial coefficients were included, was not made. As Figure 9 shows, adding the coefficients has only a marginal effect on the accuracy. To prove statistical significance, 10-fold cross-validation experiments were performed as a follow-up to compare the addition of the coefficients. The accuracy of identifying the correct gene/locus within the first three predictions was 66.05% with the coefficient added, versus 65.22% without. This small gain improves performance and is computationally inexpensive to compute even though it was not statistically significantly different. Therefore the preprocessing step consists of adding the coefficients of fitted second and third order polynomials and filling in missing values.

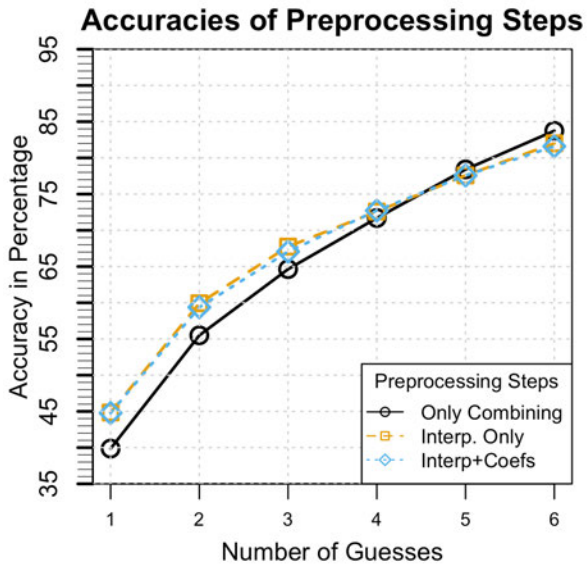


Figure 9. The accuracies of different combinations of preprocessing steps. While preprocessing with only combining audiograms taken at the same age but from different ears has greater accuracy as the number of guesses increase, it has been shown that this is due to a collection bias. Interpolating missing values is therefore necessary in order to remove this bias. Even though adding the coefficients of fitted second and third order polynomials produces marginal increase in performance, it has been shown in a follow-up experiment to be statistically significant.

3.4.3 DFNB1 Results

In addition to predictions for ADNHL loci, the ability to make predictions for ARNSHL loci was also investigated. Currently, the only dataset available contained patients from only the DFNAB1 locus and in particular mutations in the GJB2 gene. It had previously been reported that mutations in DFNB1 account for between 20%-50% of all ARNSHL cases, and this percentage varies based on the population studied [16]. That dataset consisted of 1,119 patients with 39 different mutations, with over 79% of the cases having the same 35delG homozygous mutation. In contrast, the next largest mutation only represented 3.6% of the dataset. Previous research observed differences in

phenotype for different mutations in GJB2 [2]. Attempting to make predictions for the exact mutation given the large class imbalance and relatively few number of patients for the majority of mutations in the dataset. However, using the current ARNSHL dataset the strongest phenotypic difference that could likely be predicted was between homozygous truncating mutations and other combinations of truncating and missense mutations [87]. The other combination of mutation types contained both homozygous missense mutations and heterozygous truncation and missense mutations.

The same method used for predicting the ADNSHL loci was also used for classification of the DFNB1 mutations with the exception of the target variable being mutation instead of locus. When attempting to predict the exact mutation using all 1,119 patients and 39 mutations the accuracy was equal to that of a majority classifier and every prediction was 35delG/35delG. Since there was a previously reported phenotypic difference between mutation types, the patients were grouped into two classes–homozygous truncating (T/T) mutations and other mutation types (Other)–consisting of non-truncating/truncating and non-truncating/non-truncating mutations. The accuracy and AUC values of the ROC curves are shown in Table 2. The T/T class was the largest class, 91.41% of the dataset, and an accuracy of 94.96% was obtained using the MI-SVM compared to 91.41% for a majority classifier. A statistically significant difference in AUC value was also observed with MI-SVM having a value of 0.83 compared to 0.47 for a majority classifier. The values are the averages from 100 runs of 10-fold cross-validation, and were all statistically significantly different with a p-value < 0.05. The withheld accuracy is the accuracy of the T/T patients being predicted by a classifier tried

Table 2. The accuracy and ROC values for both the original DFNB1 dataset and the results of downsampling the T/T class to be the same size as the "Other" class.

	Original Dataset		Downsampled Dataset		
	Accuracy	ROC	Accuracy	ROC	Withheld Accuracy
Majority	91.4% (0)	0.47 (0)	50% (0)	0.5 (0)	-
MI-SVM	95% (0.1)	0.76 (0.01)	83.3% (1.71)	0.83 (0.02)	87.7% (.01)

Table 3. The confusion matrix of both the original DFNB1 dataset and the downsampled dataset.

Correct Class	Original Dataset Predicted Class		Downsampled Dataset Predicted Class	
	T/T	Other	T/T	Other
T/T	1010	12	85	11
Other	42	54	22	74

on the downsampled dataset.

Examining the confusion matrix, shown in Table 3, from one of the cross-validation runs for the MI-SVM classifier, only about 56% of the patients in the "Other" class were correctly classified whereas 99% of the T/T patients were correctly classified. While the accuracy and AUC values improve over a majority classifier, the "Other" class is more important to predict correctly because the 35delG homozygous mutation would always be screened first for any patient with a presumed recessive inheritance because of its prevalence. However, in the case where a 35delG mutation is not found after screening a patient, then the predicted mutation type becomes more relevant.

In order to improve the prediction of the "Other" class of mutations and reduce the false positive rate, the T/T class was randomly down-sampled to contain the same number of patients for the "Other" class, equaling 96 in each class. Then cross-validation was performed on the down-sampled dataset and all patients that were removed during down-sampling for the T/T class were classified using a MI-SVM trained on the entire down-sampled dataset. This was done to determine how well the down-sampled dataset would generalize on the withheld T/T patients. The steps of down-sampling, cross-validation, and prediction of the withheld patients was repeated 100 times with different random down-sampled datasets to determine how consistent the performance was with different subsets of the T/T patients. The average accuracy, AUC, and accuracy of the withheld set can also be seen in Table 1, along with the standard deviation in parentheses.

While the accuracy does decrease from 94.96% to 82.99%, the AUC increase significantly from 0.76 to 0.83, which can be interpreted as indicating that the ability for the classifier to discriminate the two classes was improved. Viewing the confusion matrices for the original dataset and the down-sampled dataset, shown in Table 3, the number of patients being correctly predicted for the "Other" mutation type increased from 54 to 74. The accuracy of predicting the withheld patients was 87.7%, and had very little variation with a standard deviation of 0.01. The low standard deviation indicates that the phenotype for the T/T mutations was consistent within the dataset and do not vary much when using different down-sampled datasets. This implies that only a small number of patients are needed to represent the entire class, and that reasonable performance can be expected using the down-sampled dataset on T/T mutations. The higher accuracy for predicting the withheld patients was likely due to the fact that it contained only patients that were T/T–consisting of only patient with profound deafness. In contrast, the accuracy reported for the downsampled dataset contained both type of patients.

3.4.4 Robustness to Noise

AudioGene suffers minor performance degradation from its baseline performance when datasets contain modest amounts of noise (~3% at noise levels between 1 and 3, as defined previously), as shown in Figure 10. It is only with higher levels of noise that there is a significant loss of performance (~10%). This amount of noise would equate to a shift of the audiogram between 20 and 25 dB and substantial distortion to the original audiogram's shape.

3.4.5 Outlier Identification

A patient's audiogram was considered an outlier if the correct gene/locus was never predicted during any of 100 repetitions with the addition of high amounts of noise. The plot of outliers by loci is shown in Appendix D. From the results it can be inferred

that, to the classifier, patients who are outliers never appear similar to other patients from that locus.

As a general rule, smaller loci (in terms of the number of patients per locus) should contain a larger percentage of outlier patients and conversely, larger loci should contain fewer outliers. However, there exist some loci that have a larger number of outliers than expected. DFNA10 is an example in which 34 of 56 patients are labeled as outliers. Further investigation of these outliers is necessary to determine if they are truly outliers or are representative of the inherent variability of the audiograms for a particular locus. This variability could also be associated with unknown subclasses.

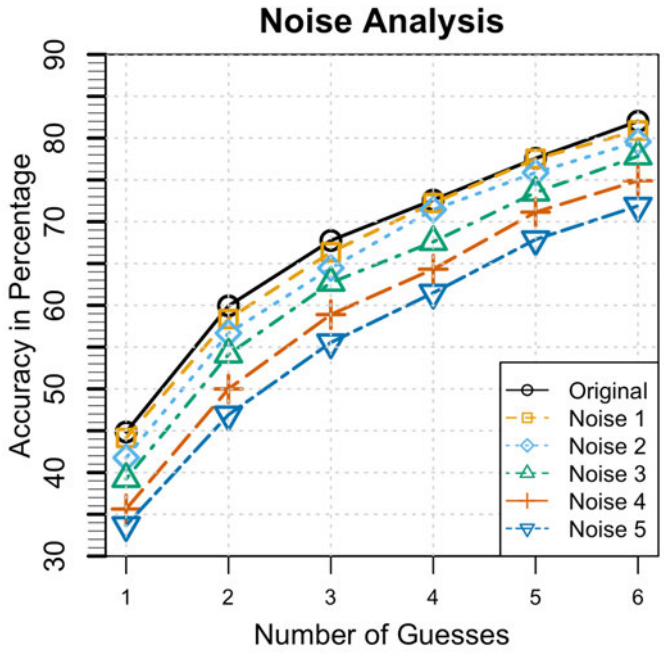


Figure 10. The accuracies of different combinations of preprocessing steps. While preprocessing with only combining audiograms taken at the same age but from different ears has greater accuracy as the number of guesses increase, it has been shown that this is due to a collection bias. Interpolating missing values is therefore necessary in order to remove this bias. Even though adding the coefficients of fitted second and third order polynomials produces marginal increase in performance, it has been shown in a follow-up experiment to be statistically significant.

3.5 Discussion

The results demonstrate that the genotypes can be predicted from the phenotype for patients with ADNSHL and mutations in DFNB1. The robust performance was achieved using a MI-SVM and had an accuracy of ~68% as compared to a Majority classifier, which has an accuracy of ~44% for predicting the top three loci. It was also shown that missing threshold values must be interpolated to guarantee an unbiased classifier that generalizes effectively to unknown data.

Applying the same method that was developed for ADNSHL to the DFNB1 dataset, it was initially unsuccessful in predicting the mutation but was able to predict the mutation type instead. To remove the effects of the large class imbalance, the T/T class was down-sampled to be of equal size the "Other" class. The AUC value increased from 0.75 to 0.83 after down-sampling, but the accuracy decreased from 95% to 83.3%. However, the number of patients correctly classified for the "Other" mutation class improved from 54 to 74. An accuracy of 88% was achieved for the withheld T/T patients using a classifier trained on the down-sampled dataset.

In some settings, missing data can serve as informative features. For example, a missing value from a "date of death" field implies that the patient is not deceased. In the case of an audiogram, missing frequency thresholds imply nothing about the genotype of the patient, but rather are normal variations in clinical practice between sites. Therefore, missing thresholds must be interpolated to guarantee an unbiased classifier; otherwise the classifier cannot generalize effectively to data collected at different clinics.

The results of applying the noise model to the method showed that the performance of the MI-SVM was robust to modest levels of noise that would be expected with measurement noise. Although it was attempted to employ a simple linear model to apply random amounts of noise independently at each frequency, this approach generated physically impossible audiograms and was abandoned. An example is a saw tooth-patterned audiogram produced by alternating +/-10 dB at each frequency. By applying

noise in the frequency domain with the DCT, the overall audioprofile shape was retained but produced audiograms that were shifted and/or stretched that are still physically possible. This noise model allowed for the determination of the robustness of the method to various amounts of noise and also enabled the identification of outliers.

The identification of outliers is particularly interesting, since genetic modifiers of hearing loss are known to exist. The method for identifying outliers is equivalent to selecting the patients who are not predicted correctly, even when allowing for large degrees of error in the data collection. This could be caused by inadequate training data for a given locus, inadequate separation between two phenotypically similar loci, an improperly assigned causative locus, or environmental and genetic modifiers that affect the patient's phenotype.

3.6 Conclusion

In summary, a method was developed for prioritizing genetic loci for ADNSHL screening based on a patient's phenotype. Using a leave-one-out analysis, AudioGene has an estimated accuracy of 68% for identifying the correct genetic cause of hearing loss within the top three predictions using a MI-SVM. Apply the same procedure to patients with mutations in the recessive DFNB1 locus, an accuracy of 83.3% was obtained by down sampling the larger class and predicting the mutation type as T/T or "Other". The method was shown to be robust to noise with a drop in accuracy only when large amounts of noise were applied. AudioGene is available as a web service at http://audiogene.eng.uiowa.edu. Originally developed for prioritizing loci for Sanger sequencing [88], as sequencing technologies have advanced, AudioGene has proven invaluable as a method of evaluating variants of unknown significance generated by targeted genomic capture and massively parallel sequencing, effectively linking a person's phenome to their genome [27,89].

CHAPTER 4

SUBCLASS DISCOVERY USING HIERARCHICAL SURFACE CLUSTERING

4.1 Introduction

The goal of supervised machine learning is to utilize labeled data to train a classifier that can be used to predict the class membership of unlabeled data. In most cases, the set of class labels is well defined. For instance, in the case of a bank predicting if a customer will default on a loan, there are two possible classes: *default* or *not default*. When attempting to predict genotype from phenotype, the genotypes used as class labels may not adequately reflect the manner in which the disease phenotype manifests. For example, using a large genomic locus containing one or more putative mutations as the genotype could be too coarse of a class label. A better genotype could be the mutation type, such as *truncating* versus *non-truncating* (perhaps in addition to genomic locus), if the phenotype was found to better correlate with the mutation type. Another issue is that the genotype itself can represent complex genetic factors and not just mutation, mutation type, or locus. The goal of the methods developed in this thesis is to identify possible subgenotypes or subclasses based on examination of the phenotype data for potentially-novel subclasses. Finding these subclasses can be difficult because the significance of the identified subclasses may be hard to determine. For instance, if clustering is used to identify potential subclasses, by the very nature of clustering, the class will be partitioned into groups. These groups however may represent subclasses or could be the result of the clustering algorithm being forced to partition the class and the significance of the partition must then be determined. The primary motivating application of the methods in this thesis is for finding subclasses in audiometric datasets. The main contribution therefore, is a novel hierarchical surface clustering technique accompanied by a novel visualization technique applied to simulated and measured audiometric data.

4.2 Background

4.2.1 Subclass Discovery Techniques

The main focus of previous subclass discovery work has been to improve classification accuracy [90-93]. Many of these techniques allow linear classifiers to approximate non-linear decision boundaries by performing clustering within existing classes to identify subclasses [34,93]. The labels for the subclasses are then mapped back to the original class labels following prediction [90-92]. In the case of Clustering in Classes (CIC) [90], subclasses were identified by performing K-means clustering on the instances of a given class, and then using the cluster assignments as new class labels. Any predictions made for the new (sub)class labels were then mapped back to the original

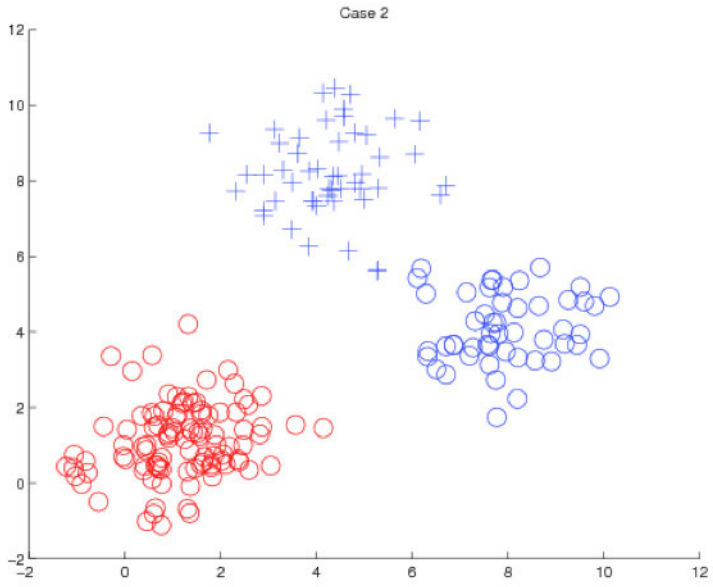


Figure 11. Example of the difficulty of using accuracy when finding subclasses within existing classes. Initially two classes are given the blue class (upper right union of "+" and "o" distributions) and the red class (lower left distribution). If the blue class was split with K-means, and then the accuracy were evaluated, there would be no increase in accuracy. Therefore, accuracy will not improve if the original class that is being split occupies a region of the feature space that is already separable from other classes.

class labels. Employing this approach, an increase in accuracy was observed for some datasets during cross-validation when using a Support Vector Machine [34] with a linear kernel. However, an increase in accuracy was not observed when using an SVM with a Radial Basis Function (RBF) kernel. An alternative approach, called multimodal softmax (MMS)[93], defines latent (soft) subclasses that may be inferred during training in an effort to increase accuracy.

All of these methods focus on increasing classification accuracy and therefore cannot find a hidden or missing subclass within a class if it does not increase classification accuracy. A trivial example of a class containing two subclasses, which after splitting does not increase accuracy, is shown in Figure 11. There are originally two classes given, red and blue. As depicted in Figure 11, the blue class consists of two subclasses–the plusses (+) and circles(o). The application of any of the previously described methods would not result in an increase in cross-validation accuracy, however it may correctly recover the two subclasses. Therefore, if a class and its subclasses occupy a region of the feature space that does not overlap other classes' regions, then these methods will fail to identify potential subclasses.

Other research has focused upon discovering disease sub-types from gene expression data [94-96]. These techniques are domain specific and deal with the added challenges of analyzing microarray data. The methods differ on how they find the most informative subset of genes and then how those are used to identify the disease sub-types. In one of the early cases, the class discovery approach the difference between two sub-types of leukemia (acute myeloid leukemia and acute lymphoblastic leukemia) [95].

4.2.2 Subclass Discovery Challenges in the AudioGene Dataset

For the case of AudioGene, attempting to identify subclasses solely by clustering can be difficult because many of the known loci reflect progressive hearing loss based on

age. Clustering of audiograms from a specific locus will likely result in sub-clusters based on progression but may not represent meaningful subtypes of the disease. This means that during clustering, the progression of hearing loss needs to be carefully considered in order to increase the likelihood of finding meaningful subclasses. Also, based on the CIC results [53,90] the increase in accuracy from splitting a class should not be used as the criterion for determining whether subclasses exist because it will likely increase because of the reasons CIC observed an increase in accuracy and not because a subclass was found.

Another problem is that the variability between individuals within a given locus can be quite large. In Figure 12, all the audiograms from the DFNA2A locus are shown. As can be seen, there is a large degree of variability among members of this locus. This can cause problems for certain clustering techniques, especially ones that cluster based on density. These techniques assume that clusters are regions of high density separated by regions of relatively low density. As can be seen from the audiograms from patients with a DFNA2A genotype (Figure 12), the degree of variability would likely cause problems in identifying regions of lower density and would be highly sensitive to parameter choice. More sophisticated clustering techniques could be used, such as spectral clustering [9,60,61,97], but these also have parameters that need to be chosen, and can drastically affect clustering and are not a silver bullet. Automated methods are available for choosing parameters for these approaches. However, the available methods are not robust when applied to datasets that contain clusters that are not well separated [62,65].

The audiograms from DFNA2A also illustrate the problem of visualizing the progression of hearing loss within the locus by age. Previously, progression of loci was described by an audioprofile (see Chapter 3). However, this falls short of taking advantage of all three dimensions that humans are capable of understanding (tonal frequency under test, dB loss at each frequency, and age at the time of testing). In the methods described below, the audiograms are mapped onto these three dimensions and a

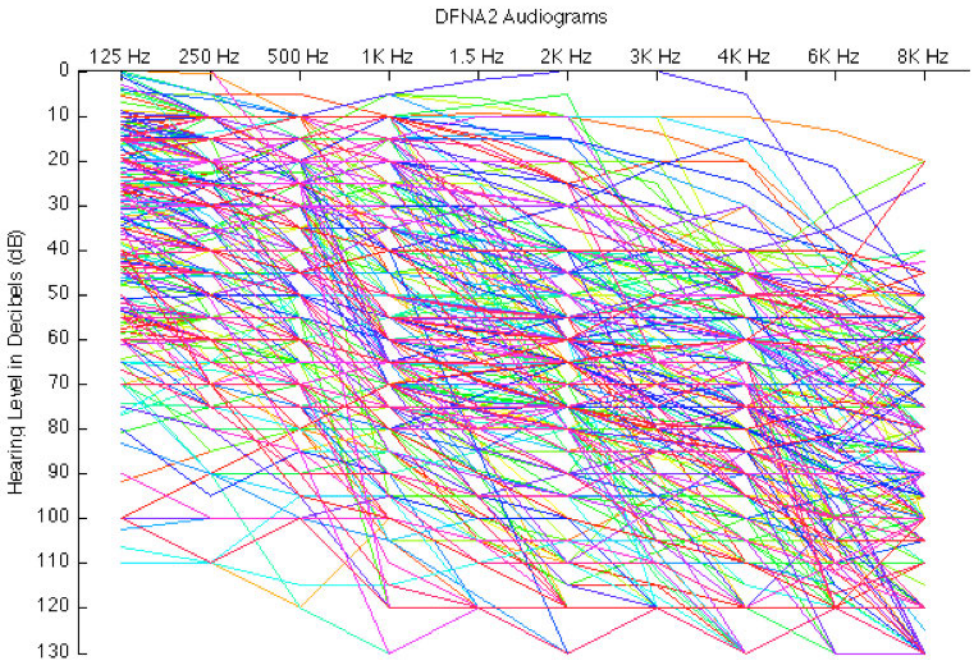


Figure 12. All the audiograms from the DFNA2A locus. The extreme degree of variability creates difficult in visualizing the overall trend with age, and also would cause other method of clustering difficulty, such as ones that rely on density estimation.

surface is fitted to them to aid in visualization and decision-making. This is useful because many loci exhibit a strong progression with age and therefore a 20 year old patient's audiograms will not likely not look the same as a 60 year old patient's audiograms taken from the same loci.

4.3 Methods

The general approach taken in this thesis for finding subclasses within a given locus is a combination of a novel visualization technique and a novel hierarchical surface-clustering algorithm. The novel visualization technique fits a surface to the audiograms by mapping the 2D audiograms into 3D space based on age and also allows these surfaces to be used during clustering. This novel hierarchical surface–clustering algorithm performs K-means clustering to initially cluster audiograms based on the

"shape" of the hearing loss. The clusters are then transformed into the 3-D surfaces that are clustered hierarchically based on similarity. The final clustering is then displayed as superimposed sets of cluster surfaces for a human domain expert to evaluate.

4.3.1 Algorithm

4.3.1.1 Fitting Audioprofile Surfaces

Audioprofile surfaces are fitted to audiograms by representing the audiograms in three dimensions with discrete frequencies (125, 250 Hz, etc) on the *x*-axis, the age at which the audiogram was measured along the *y*-axis, and the hearing loss in dB along the *z*-axis. Each audiogram is then transformed into 10 points (the 10 frequencies of the audiogram) in the three dimensional space with the x values corresponding the discrete frequencies (125=1, 250 Hz=2, etc), the y value for each point is the age at which the audiogram was measured, and the z value is the quantified hearing loss at the corresponding frequency. Using these points, a surface is fitted with a second-degree polynomial along the x-axis (frequency), and a third-degree polynomial along the y-axis (age), see equation 4.1. Other possible surface equations are possible, such as exponential or logarithmic, and the choice would be depended upon the data. For AudioGene, the surface equations were chosen that captures the expected progression and patterns of hearing loss that had previously been observed. Least squared regression with bi-squares robustness is used to fit the surface equation to the audiograms in 3D space [53]. Once the surface equation is fitted to the audiogram points, a synthetic audiogram can be generated for a specific age by fixing the age (y-value) and then iterating over the x values (frequencies). If there are fewer than ten, but more than five audiograms, the polynomial along the y-axis is reduced to second degree as shown in equation 4.2. Similarly, if a surface is being fit to a group of fewer than five audiograms, then the polynomial on the y-axis is further reduced to a first degree polynomial (equation 4.3). By reducing the degree of the polynomial along the y-axis (age) based on the number of

audiograms, the surface is effectively smoothed and is more useful during clustering when there are fewer number of audiograms. This is because there always must be at least one more audiogram than the degree that is being fitted to the audiograms along the age axis. The minimum values were chosen because they were approximately double the minimum number of audiograms needed for each of the surfaces. The full set of equations used for fitting the different surfaces are shown here (equations 4.1-4.3):

$$z = p_{1,0} + p_{1,0}x + p_{0,1}y + p_{2,0}x^2 + p_{1,1}xy + p_{2,1}x^2y + p_{1,2}xy^2 + p_{0,3}y^3 \quad (4.1)$$

$$z = p_{0,0} + p_{1,0}x + p_{0,1}y + p_{2,0}x^2 + p_{1,1}xy + p_{0,2}y^2 \quad (4.2)$$

$$z = p_{0,0} + p_{1,0}x + p_{0,1}y + p_{2,0}x^2 + p_{1,1}xy \quad (4.3)$$

A visualization of how the original 2D version of the audioprofile relates to the 3D audioprofile surface can be seen in Figure 13. The audioprofile is based on a set of example audiograms that were generated for illustrative purposes. When adding the average age of each the curves in the audioprofile as a new coordinate in 3D space and then performing a perspective change, the original audioprofile can be seen in three dimensions with the new axis being age. Plotting the four curves of the audioprofile in 3D is not very useful but if instead, a surface is fit to the audiograms then the entire progression with age becomes visible between the original four curves.

Visualization of 2D Audioprofile to 3D Audioprofile

Original Audioprofile in 2D

The Four Audioprofile Curves in 3D Based on Age

3D Audioprofile

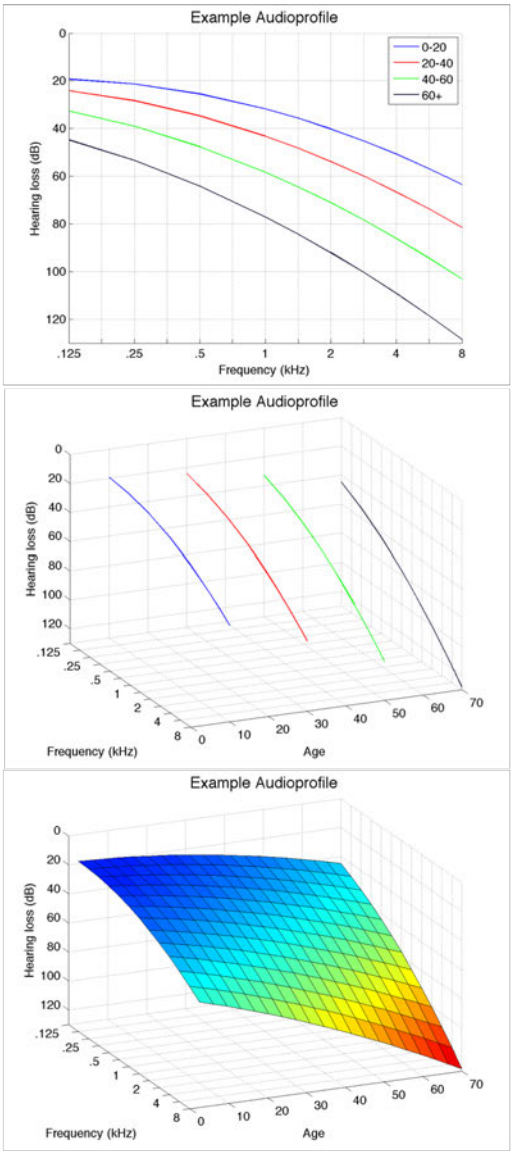


Figure 13. A visualization of how the 2D audioprofile relates to the 3D audioprofile. Starting in 2D, the new axis that represents age is going into the page. The second plot is of the four curves being plotting in 3D space with their respective ages as the value on the age axis. Finally, the 3D audioprofile surface is shown with the color representing progression of hearing loss in dB going from blue (0 dB) to red (130 dB).

4.3.1.2 Hierarchical Surface Clustering

All steps described here are depicted in Figure 14. The first step in hierarchical surface clustering (HSC) is to perform K-means clustering on the audiograms with a selected initial value of $K=K_0$[1]. Next, any spurious clusters are removed. A spurious cluster is defined as a cluster with less than S patients assigned to it[2]. The next step is to repeatedly merge the clusters based on their surface distance until a final clustering of K_f is found. The merging is performed as follows:

(1) A surface is fitted to the audiograms of each cluster using the method described in the previous section. This defines the audioprofile surface for each cluster *i*, denoted s_i, and is called a surface fragment. Using the surface fragment, s_i, an audiogram can be generated by using the coefficients fitted to the surface equation for a specific age, limited by the age range of audiogram data available for the cluster. These synthesized audiograms are sampled from each surface and used to compute the distance between them.

(2) The distance is computed between each of the pairs of surface fragments. This distance is defined as the minimum Euclidian distance between *n* sampled audiograms from the overlapping age range. If the two surfaces do not overlap, then the distance is defined as the Euclidian distance between the audiograms closest age in age that are still in the age range of the two clusters.

(3) The two surfaces with the smallest distance d_{ij} between them are merged into a single cluster by merging the audiograms from both clusters into a single cluster.

(4) Steps 1-3 are repeated until the number of surfaces remaining is equal to K_f.

---

[1] A typical value for K_0 used with the current AudioGene dataset is 15.

[2] S likewise had a nominal value of 4

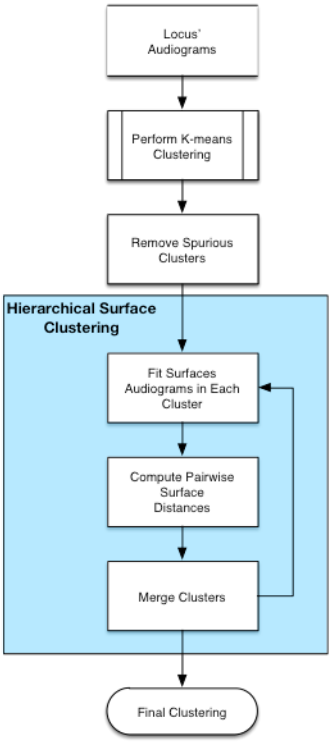


Figure 14. Steps of hierarchical surface clustering (HSC): (1) K-means clustering is performed with a $K=K_0$, (2) Clusters that contain less than S patients are considered spurious and are removed (3) Audioprofile surfaces are fitted to the audiograms in each cluster, (4) Pair-wise surface distances are computed between the surfaces, (5) The two closest surfaces (smallest Euclidean distance) are merged into a single cluster and the merger is stored; the algorithm terminates when only a single cluster remains; otherwise steps 2-4 are repeated (6) The final clustering for a given C (number of clusters), can be retrieved.

4.3.2 Investigating Discovered Subclasses

If the human hearing loss expert identifies possible subclasses within a locus, a series of hypotheses are evaluated to determine the likely cause of the subclasses. Shown in Figure 15 is a chart of the possible hypotheses that could be used to explain the identified subclasses. Before genetic causes are considered, it is important to rule out any environmental causes since they are easily to be ruled out. If there is no likely environmental cause, then the next plausible causes to investigate are genetic. These can

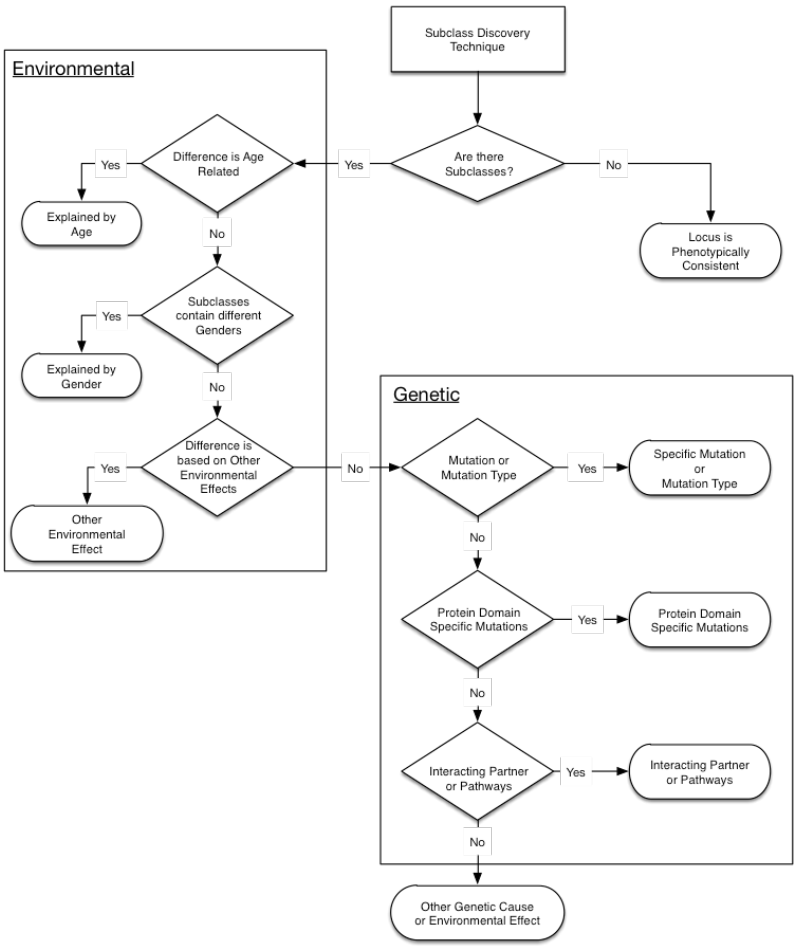


Figure 15. Flow chart of the various hypotheses evaluated after performing HSC. If subclasses are found, then the first set of hypotheses are based on the environment and are the most likely causes that explain the clusters. If environment is not the cause, then the next is genetic. The causes include different mutation types, such as truncating or non-truncation mutation showing different hearing loss pattern, or mutation in different protein domains.

be more time consuming to evaluate, so again, these are typically evaluated in ascending order of the difficulty involved in evaluating the hypothesis based on the data present.

4.3.3 Generation of Synthetic Datasets

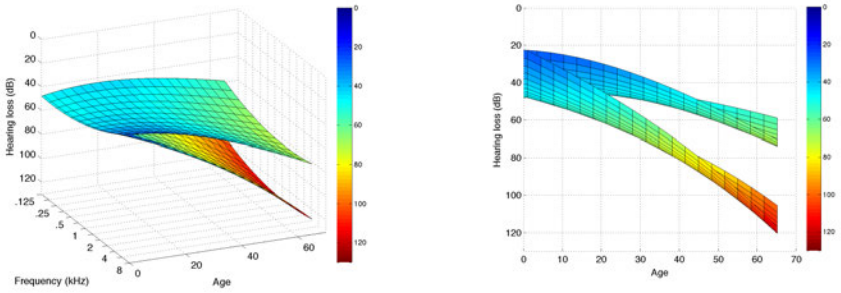
Synthetic datasets were generated to evaluate various aspects of the HSC algorithm and to compare it with existing clustering techniques. The first step in this process was to compute random surfaces based on coefficients similar to those of equations used to describe existing (deafness) loci surfaces. Several possibilities were explored for the form of these surface equations as seen in equations 4.1-4.3. The coefficients were generated randomly from a Gaussian distribution using empirical values for mean and standard deviation derived from the corresponding coefficients of all the surfaces fitted to each of the loci. For simplicity, and without significant loss of precision, the polynomial with quadratic terms along both the age and frequency axes were used as seen in equation 4.2. Coefficients were repeatedly generated until a surface was obtained that was within the range of possible hearing loss (0-130 dB) for all ages, and represented the phenotype of a plausible synthetic genetic cause, e.g. a surface in which hearing improved significantly with age would not be plausible. Modifying effects that would simulate genetic modifiers or other genetic effects that would modulate the phenotype (either increasing or decreasing the impact) could then be applied to the surface by modifying selected coefficients of the surface. For instance, to decrease the progression with age the $p_{0,1}$ coefficient could be gradually decreased. Using these surfaces, audiograms for the dataset could be sampled for patients of various ages for both cases. Finally, to simulate the test-retest variability of 5 dB [9], Gaussian noise was added that shifted the audiogram by a mean of 0 and a standard deviation of 2.5. The fraction of patients with the various genetic effects could also be controlled.

4.3.4 Evaluating Clustering

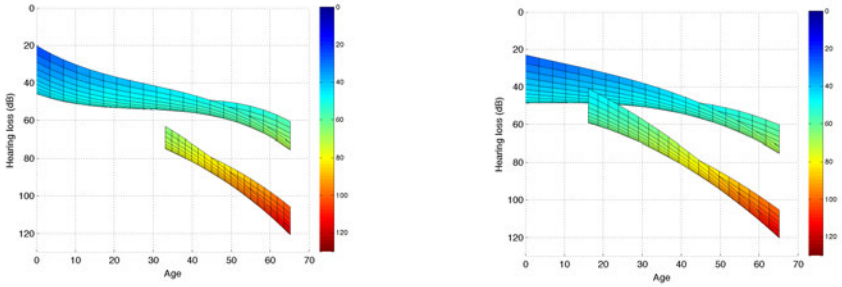
The performance of the new HSC algorithm was evaluated using the Adjusted Rand Index (ARI) for the cases in which a gold standard cluster assignment is known [65,68]. The ARI is an appropriate choice for comparing clustering assignments to a gold

standard [66]. It compares every pairwise combination of assignments for all instances and evaluates them against their corresponding cluster assignment in the gold standard. The ARI is based on the Rand Index (RI) [67], but also considers the cluster assignments against assignments made by random chance. The metric has a range from -1 to 1. An ARI of 1 means that the two cluster assignments are in complete agreement. A rank of zero means that the assignments are equal to those made by randomly assigning instances to clusters, and less than 0 means that the cluster assignment are worse than those made by chance. Large negative values are relatively unlikely in practice and only small negative values observed [68]. The ARI is also useful for comparing cluster assignments of different number of clusters [66]. A more detailed discussion of the Adjusted Rand index can be found in Chapter 3. All ARI values reported here are based on repeating the clustering algorithm 10 times, and was done to take into account the possible difference is cluster that would be the result of different initial clusters found via K-means.

Simulated Dataset Example 1 Original Surfaces

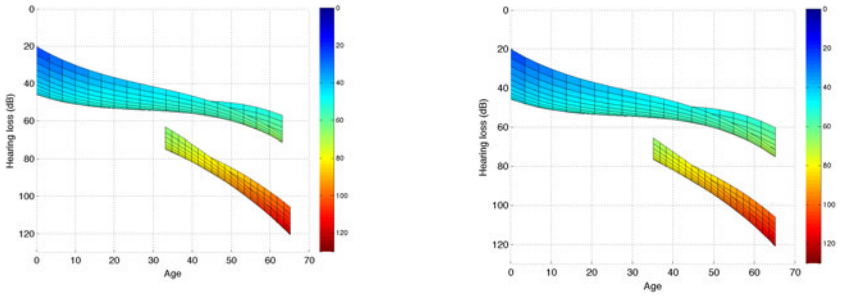


Hierarchal Surface Clustering (HSC) $K_f = 2$



Median (.2904 ARI) Best (.5162 ARI)

K-means $K_f = 2$ Spectral Clustering $K_f = 2$



Median (.19 ARI) Median (.218 ARI)

Figure 16. The original surfaces for the two simulated phenotypes for the synthesized dataset labeled example 1, and the resulting surfaces found with K_f set to 2. The ARI values are based on running the clustering algorithms 10 times. HSC had the highest average ARI value of .307, with the median shown for illustrative purposes.

4.4 Results

4.4.1 Synthetic Dataset

The first synthetic dataset generated (Example 1) represented a phenotype with a strong progression of hearing loss with age across all frequencies. A simulated genetic modifier was applied to a portion of the cohort that significantly reduced the progression with age. The two surfaces can be seen in Figure 16, and because HSC exhibited more variation in the resulting clustering assignment the median and best clustering assignment are shown. In practice, a domain expert would repeat the clustering algorithm multiple times and likely choose the best one. In contrast, both K-means and HSC found clustering assignment that varied only slightly in ARI. Two datasets were generated with different proportions of patients containing the genetic modifier. In the first dataset, the two portions were equal with a total of 200 patients, and in the second dataset only 20% of the patients contained the modifier within a total of 250 patients.

When K_f was set equal to 2, applied to the datasets with equal proportions, HSC had the highest ARI value and therefore obtained clusters that were, on average, more similar to the gold standard. The resulting ARI values can be seen in Table 4. For the dataset in which the affected proportion was 20%, both spectral clustering and HSC showed an increase in ARI but K-means showed worse performance according to the ARI; results also shown in Table 4. The cause of the decrease in ARI for K-means was likely due to an inherent bias in K-means wherein the clusters found tended to be skewed towards clusters of equal size; a phenomenon known as the uniform effect [98]. To investigate the effect of using values of K_f larger than the optimal number of clusters, K_f was increase to 3, and the resulting ARI values calculated; as shown in Table 2. When increasing the value of K_f to 3, the additional cluster for both K-means and spectral clustering was composed of a combination of patients from both phenotypes whereas HSC found a surface representing an additional cluster of patients from one of the two

genotypes. Because the additional cluster found by both K-means and spectral clustering contained patients from both genotypes, the ARI was significantly decreased. The difference was not as significant for HSC, which had the highest ARI of 0.438 when K_f was equal to 3 (see Table 5).

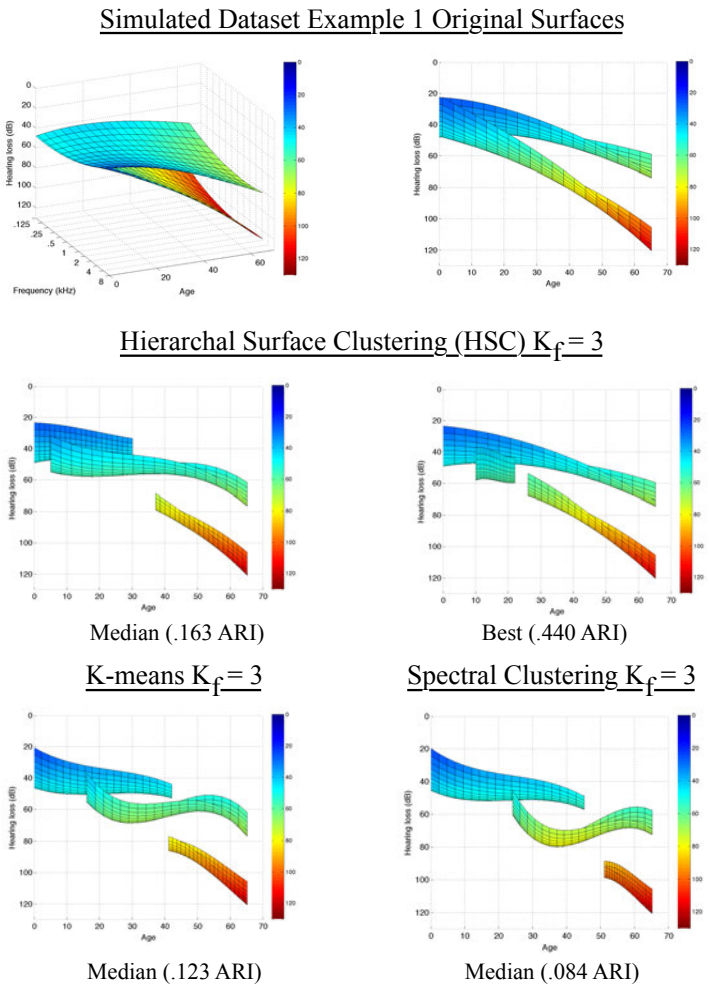


Figure 17. The result of setting the value of K_f, final number of clusters, to a value larger than the optimal number of clusters with K_f set to 3. Both K-means and spectral clustering have two surfaces that represent the two different genotypes but have an additional surface that is an amalgamation of subsets of the two true genotypes. In contrast, HSC finds similar surfaces to those found when the value of K_f was set to 2 with an additional surface that contains patients from either of the genotypes.

Table 4. The Adjusted Rand Index for the synthetic dataset labeled Example 1 both with equal proportions of the patients being affected by the genetic modifier and a skewed proportion with the genetic modifier only affecting 20% of the patients.

	Equal Size (100/100) K_f=2				Skewed Size (200/50) Kf=2			
Adj Rand Index	Avg	Std Dev	Min	Max	Avg	Std Dev	Min	Max
K-means	0.19	0	–	–	0.116	–	–	–
Spectral Clust	0.218	0	–	–	0.409	0.019	0.393	0.393
HSC	**0.307**	0.232	0.066	0.516	**0.438**	0.111	0.311	0.57

Table 5. Results for the first synthetic dataset, Example 1, when the value of the K_f is increased to 3.

	Equal Size (100/100) K_f=3			
Adj Rand Index	Avg	Std Dev	Min	Max
K-means	0.1226	–	–	–
Spectral Clust	0.0836	–	–	–
HSC	**0.2273**	0.1148	0.1175	0.4402

Note: Bold values indicate the algorithm that performed the best and the difference was statistically significant.

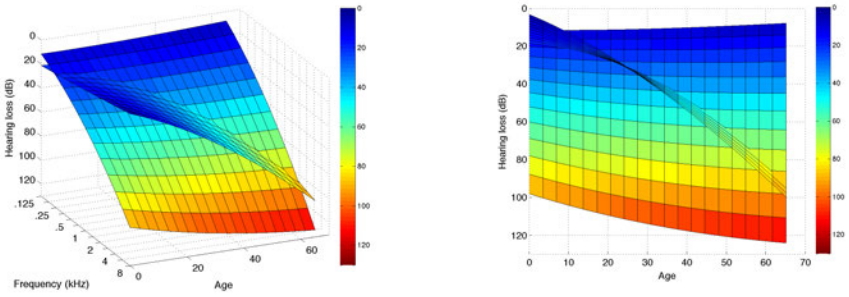
The second synthetic dataset, Example 2, simulated a different type of case in which a hypothetical genetic modifier created a significant difference in the observed phenotype and was not simply an attenuation of the progression with age. The two phenotype surfaces can be seen in Figure 18. The surfaces were generated by first computing the mean and standard deviation of the coefficients using surfaces fitted to each of the loci separately, and then randomly sampling the coefficients independently using a Gaussian distribution with the respective mean and standard deviation. Two surfaces were generated that were within the bounds of the a phenotype that was clinically feasible, i.e. surfaces were rejected that had hearing improving with age and were outside the bounds of the decibels measure by an audiogram. On average spectral clustering had an ARI of 1, which means that the clustering assignment found was

identical to the gold standard. Both K-means and HSC produced average ARI values of 0.491 and 0.408, respectively. However, the maximum ARI value obtained during the 10 runs for HSC was 1 and the minimum value was 0.006. By examining the standard deviation of HSC (.51) and the ARI values from each of the 10 runs, it was determined that HSC alternated between a score of 1 and 0.006, with the latter clustering being found slightly more often. In practice, HSC (or any clustering technique) would be performed multiple times on a given dataset to determine an acceptable clustering assignment based on a human domain expert. While spectral clustering did consistently produce the best result, HSC was able to produce the gold standard clustering assignment during at least a few of the runs and the best result, in this case, would have been chosen by a domain expert if presented with the results from multiple runs as shown in Table 6. Therefore, based on the average ARI value spectral clustering performed the best, but because HSC was able to find the correct assignment in at least a few of the runs it would be appropriate to say that it has comparable performance when paired with a human domain expert. This is not the case for K-means, which never finds the gold standard clustering assignments.

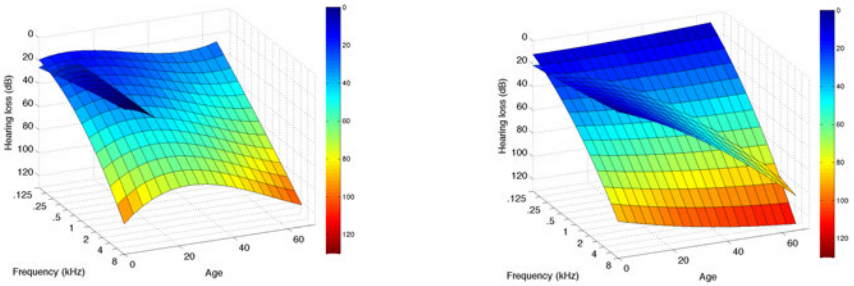
Table 6. Results for the second synthetic dataset, Example 2, with spectral clustering having the highest average ARI value.

	Equal Size Subclasses (100/100) K_f=2			
Adj Rand Index	Avg	Std Dev	Min	Max
K-means	0.4906	0.0145	0.4737	0.5018
Spectral Clustering	**1**	–	–	–
HSC	0.408	0.51	0.006	1

Simulated Dataset Example 2 Original Surfaces

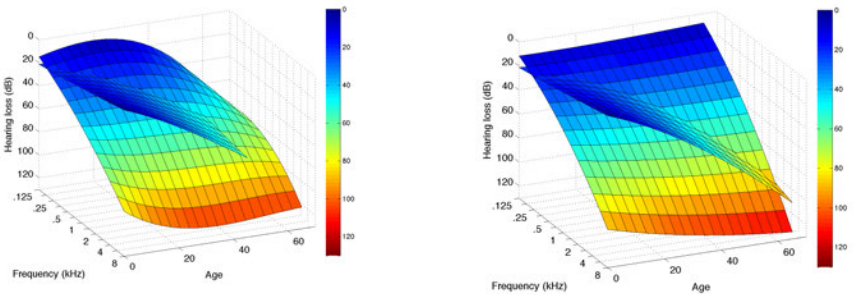


Hierarchal Surface Clustering (HSC) $K_f = 2$



Median (.017 ARI) Best (1 ARI)

K-means $K_f = 2$ Spectral Clustering $K_f = 2$

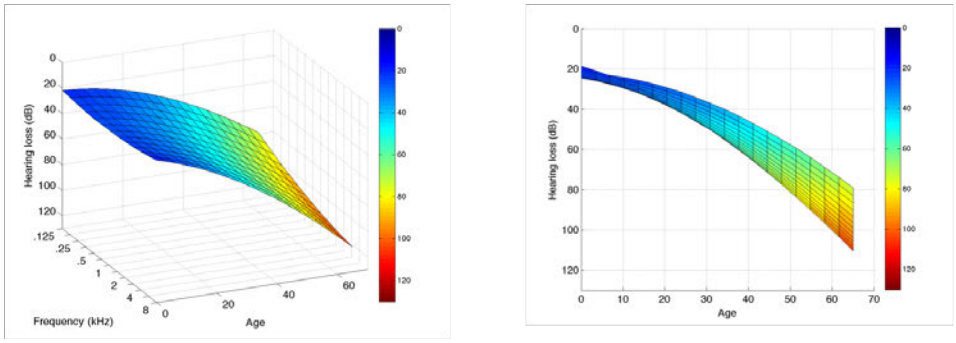


Median (.495 ARI) Median (1 ARI)

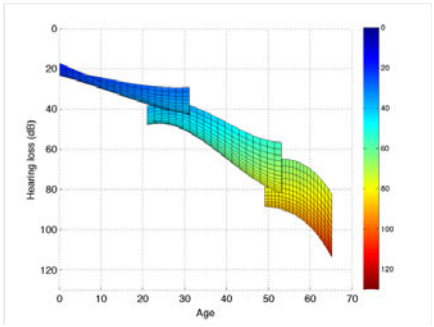
Figure 18. The results of the three different clustering algorithms on the second simulated dataset with two different distinct phenotypes, labeled Example 2, with K_f set to 2. Spectral clustering consistently finds the perfect clustering assignment across all 10 runs, but HSC appears to alternate between the perfect clustering assignment and a poor clustering assignment. K-means does not find a perfect cluster and has an ARI of 0.495.

The third synthetic dataset created, Example 3, was used for evaluating the case in which no genetic modifiers were present and the phenotype was consistent for all patients. The surface can be seen in Figure 19 along with the clusters found for each of the three clustering algorithms. The ARI value for each of the three clustering algorithms was zero and the resulting clusters were highly similar. The value of 0 is a side effect of there being only one logical and valid cluster assignment and the way in which the ARI value is calculated. The surfaces are shown for the case when K_f is equal to 3. The interesting feature to note is the "stair stepping" pattern that represents a progression corresponding to age, along with the partial overlapping regions of the surfaces. The overlapping regions represent the variability of the phenotype among the patients.

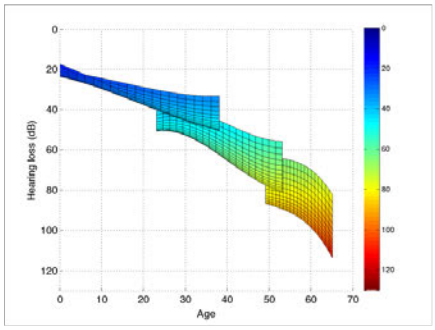
Simulated Dataset Example 3 Original Surfaces



Hierarchal Surface Clustering (HSC) $K_f = 3$



K-means $K_f = 3$



Spectral Clustering $K_f = 3$

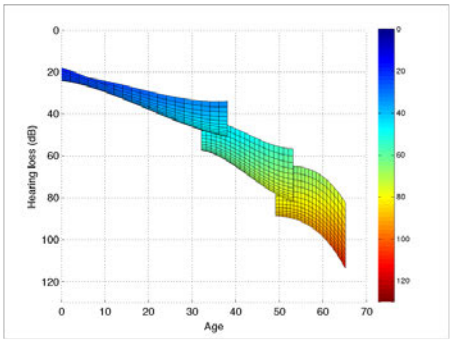


Figure 19. The third simulated dataset, labeled Example 3, which contained only a single phenotype and genotype. The three clustering algorithms find very similar surfaces with equal ARI values of 0 (the effect of having only a single cluster). The interesting characteristic is the "stair stepping" pattern is based on the progression with age and the overlapping region that corresponds to the variability.

4.4.2 DFNA9 Locus

An audioprofile surface was fitted to the audiograms in the DFNA9 locus and the resulting surfaces found via HSC can be seen in Figure 20 with a K_f set equal to 3. The hearing loss pattern for DFNA9 is a "down-sloping" hearing loss pattern that progresses approximately uniformly with age and frequency across all ages and frequencies. There is an apparent downward turn of the surfaces at the lower age grouping of 0-20 years, but this is likely due to the small number of younger patients and the inter-patient variability of their hearing loss. The surfaces found via HSC exhibit the now familiar stair stepping pattern, with each stair step indicating greater hearing loss as age increases. Using the hypothesis flow chart (Figure 15) to examine the cause of the clustering, the first hypothesis to examine is age. Using an unpaired t-test, the difference in age between each of the cluster pairs is statistically significant (p-value< 0.05), and therefore the clustering is dominated by age. A plot of the age distribution for each cluster is shown in Figure 21. The p-values can be seen in Appendix E. If the value of K_f was increased then the resulting surfaces were simply additional stair steps that segregate by disease progression and age (also shown in Appendix E). This is consistent with the simulated case in which the phenotype is consistent.

While age is the dominant cause of the clustering found for DFNA9, if the progression were uniform for each patient, invalidating an hypothesis based on environmental effects, then the resulting cluster surfaces would not overlap and there would be no discontinuity between the surfaces. However, for DFNA9 the surfaces do overlap and there is a large discontinuity between them. The overlapping regions indicate the variability of the progression with age, and are much larger than in the simulated case. For instance, there are patients between the ages of 40 and 55 in each of the three clusters and therefore the degree of hearing loss ranges from slight to profound deafness. While much of this can be attributed to environmental effects, the variability could later be used as a means for identifying genetic factors that may be attenuating the observed

hearing loss. The results from the DFNA9 locus show an example of how these techniques can be used to generate and evaluate molecular discovery hypotheses.

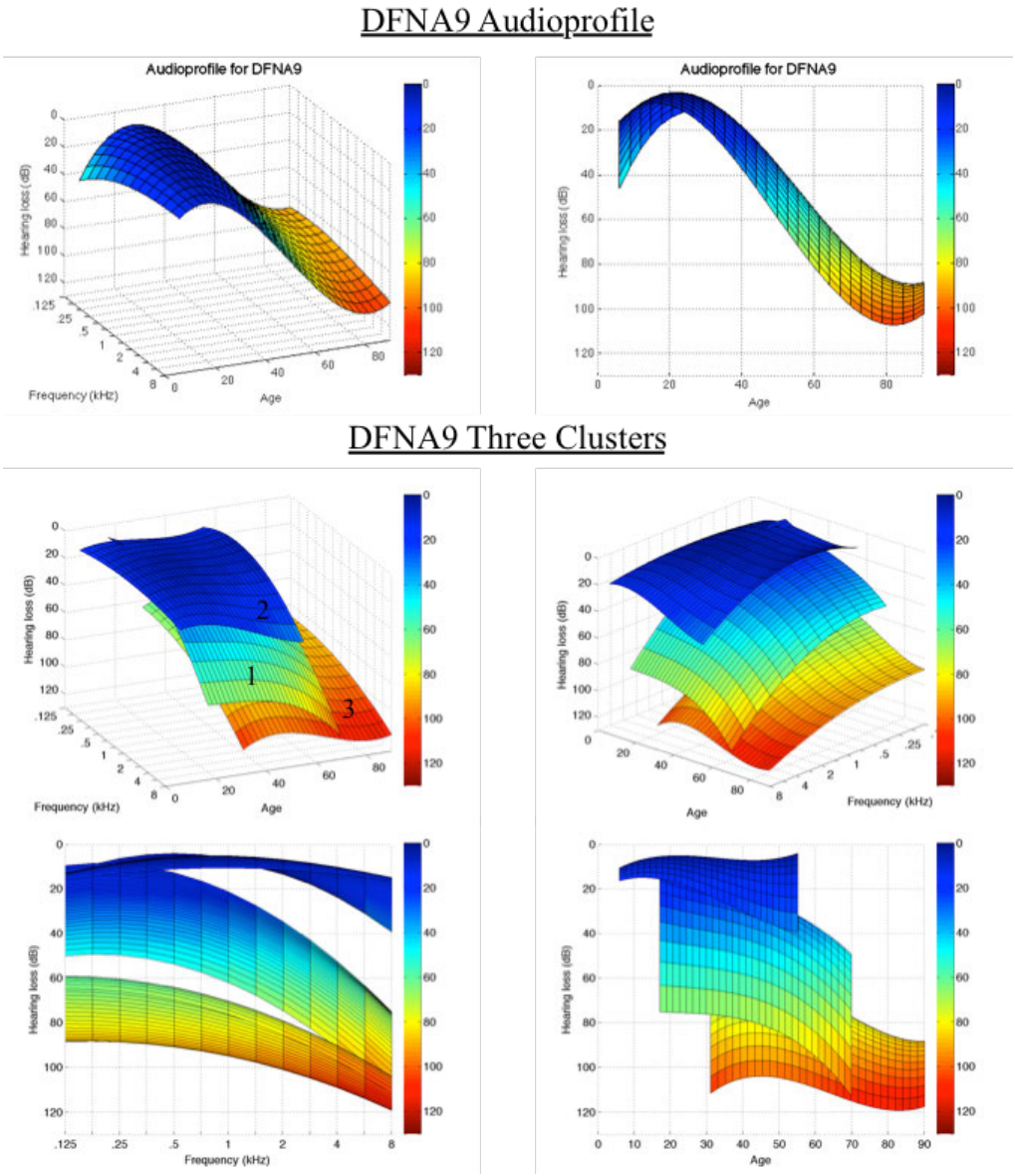


Figure 20. The audioprofile and the three surfaces identified by HSC for DFNA9. The surfaces exhibit a "stair stepping" pattern, and this means that the overall progression with age of the hearing loss is consistent amongst all the patients

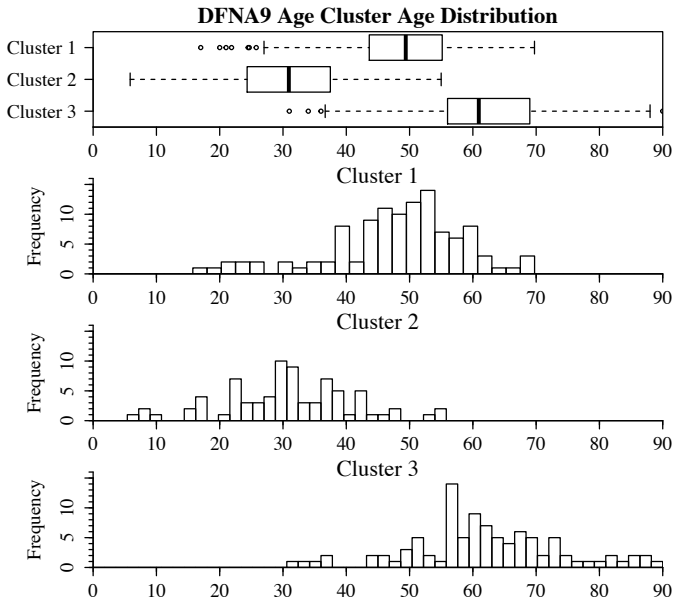


Figure 21. Age distributions of the three clusters found for DFNA9. The distribution indicates that the surfaces are clustering primarily based on progression with age.

4.4.3 DFNA2A Locus

The audioprofile surface of DFNA2A is shown in Figure 22 along with the results of clustering with K_f set to 3. The overall pattern of the hearing loss for DFNA2A also shows a "down sloping" hearing loss pattern that exhibits progressively worse hearing loss with age that is more pronounced in the higher frequencies. The final value of K_f was chosen based on examining the results of setting K_f to a range of 2 to 5 and visually evaluating the clusters. The clustering results for the different values of K_f can be seen in Figure 23. When K_f is set to 2, one cluster represents elderly patients with profound hearing loss and the other group consists of patients with a mild progression of hearing loss. When K_f is increased to 3, the additional surface is distinctly different from the progression of the audioprofile as compared to the other surfaces (surface 2 in Figure 22). This also modulates the surface that represented milder progression of the disease in the

higher frequencies. The additional surfaces that are found with increasing values of K_f are similar to existing cluster and represent only minor differences of profoundness and not a unique pattern of hearing loss. Therefore, the value of 3 was chosen for K_f .

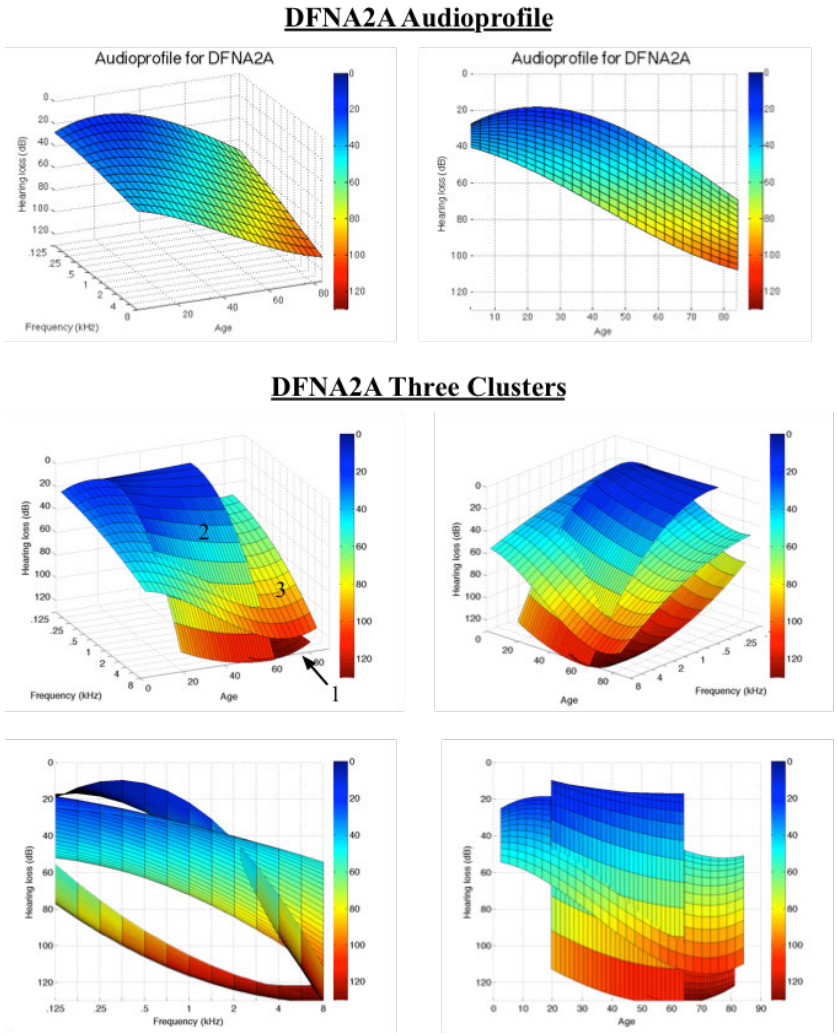


Figure 22. The three audioprofile surfaces found after applying HSC to the audiograms in DFNA2A. Surfaces 1 and 3 capture the progression of the hearing loss with age in DFNA2A, but surface 2 is drastically different from the other surfaces. Upon further investigation, it was determined that surface number 2 represents the patients with truncating mutations compared to the other audioprofile surfaces corresponding to missense mutations.

With K_f set to 3, two of the three groups (1 and 3) are likely due to the variably of the disease and also differences caused by environmental effects. Surface 1 only contains 8 patients or 3% of the DFNA2A locus, and all of the patients in the surface exhibit a perplexing pattern of hearing loss with the hearing loss values oscillating between 120 dB and 130 dB. This strange oscillation is likely due to an artifact from the collection of the audiograms or from the interpolation among a small number of audiograms. Regardless of the cause, these audiograms should be treated as outliers. The remaining surface has a distinctly different progression of hearing loss typified by severe loss of hearing at high frequencies. Based on comparing the audioprofile to other surfaces, surface 1 appears to represent the typical progression of DFNA2A, but surface 2 seems to be distinctly

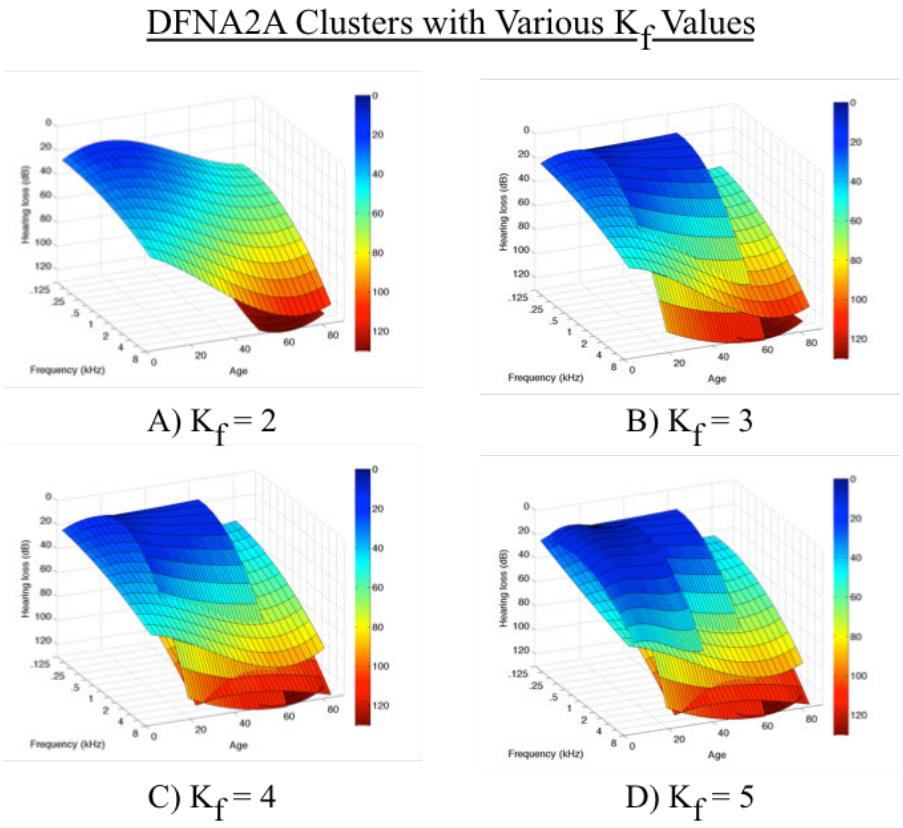


Figure 23. The surfaces found by increasing the value of K_f. The surfaces that are found segregate based on progression with the exception of the surface that represents the patients with truncating mutations.

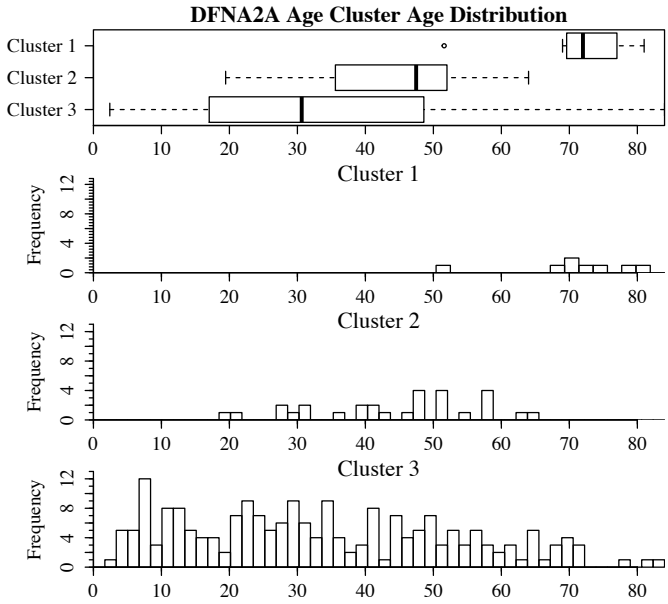


Figure 24. The age distribution of the three surfaces found for DFNA2A. With the exception of cluster 1, the difference between the surfaces cannot be attributed to different ages.

different. This age distribution of the surfaces is shown in Figure 24. A previous study had found that truncating mutations in DFNA2A caused only high frequency hearing loss with lower frequency hearing largely preserved compared to missense mutations [6]. The number of missense and truncating mutations assigned to each cluster is shown in Table 7. By performing a chi-squared test, it was determined that clusters were segregating significantly based on either truncating or missense mutations (p-value < 0.05). Therefore by using the HSC technique, a novel surface was found and upon further investigation a previously known phenotypic difference was found without prior knowledge [6].

Table 7. Number of patients by mutation type assigned to each of the three clusters for DFNA2A.

	Surface Clustering		
	Number For Mutation Type		
Cluster	Missense	Truncating	Total
1	7	1	8
2	5	25	30
3	45	14	59
Total	57	40	97

4.4.4 DFNA8/12 Locus

The audioprofile for DFNA8/12 is shown in Figure 25. The hearing loss for DFNA8/12 is milder overall than for the DFNA2A locus, and does not share a similar progression as either DFNA2A or DFNA9. The results of performing HSC with a value of K_f equal to 2 are shown in Figure 25. The two surfaces represent patients with mild or moderate hearing loss (sub-clusters 1 and 2, respectively). The surfaces do not exhibit the familiar stair stepping pattern, as was seen in DFNA2A, DFNA9 or the simulated cases, but the difference in age was statistically significant with a p-value of 0.0021 using an unpaired t-test. If the age range is matched to a range of 0 to 40, then the difference was not statistically significant with a p-value of 0.1158. The distribution of ages of the surfaces is shown in Figure 26. The larger age range of the second cluster could be the result of a collection bias; i.e., this locus only contains 61 patients (2 were removed as spurious clusters). Alternatively, if it is hypothesized that the two clusters are the result of different genetic causes and the milder cluster's genetic cause exhibited a progression with age, then the older patients with that genetic cause would likely be assigned to the second cluster. Because of these two possibilities, the difference cannot be entirely attributed to age. Other environmental factors could also contribute to the difference in observed hearing loss, but because the patients are younger and there does not appear to

be any difference attributable to gender, this leads to the logical possibility of a genetic cause.

The first hypothesis to evaluate is that the clusters segregate based on mutation type. The DFNA8/12 locus consisted only of missense mutations, with the exception of a family with a compound truncating and missense mutation. The hypothesis that the clusters were assigned based on the mutation type was not found to be statistically significant using a Fisher's exact test (p-value<0.05). Next, we consider the resolution at the level of the mutations. The mutations, along with their domains, and the number of patients assigned to each cluster are shown in Table 5. Performing a Fisher's exact test, it was found that the difference was not statistically significant with a p-value of .1754. Therefore, the clusters were not significantly different in terms of mutation domain. Based on the hypothesis chart in Figure 15, all but the last two genetic causes should be ruled out. Interestingly, the DFNA8/12 locus contains the TECTA gene and all the mutations in the dataset were from the TECTA gene. TECTA is expressed as α-tectorine and is important component ofthe tectorial membrane (TM). There other non-colligen proteins that form the TM are ß-tecotrian and otogelin [99]. There have been no reported mutations in TECTB, which encodes for ß-tecotrian, and only recently has OTOG, which encodes for otogelin, been identified as a likely cause of deafness in a single family [100]. Since these proteins are known to interact with each other to form the TM, one possible cause of the observed difference in phenotype are mutations with in TECB or OTOG. There also four collagen types that are expressed in the TM, types II, V, IX, and XI [101]. Further *in-vivo* studies are needed to attempt to identify any mutations these other genes that could be truly causing the observed difference in the phenotype.

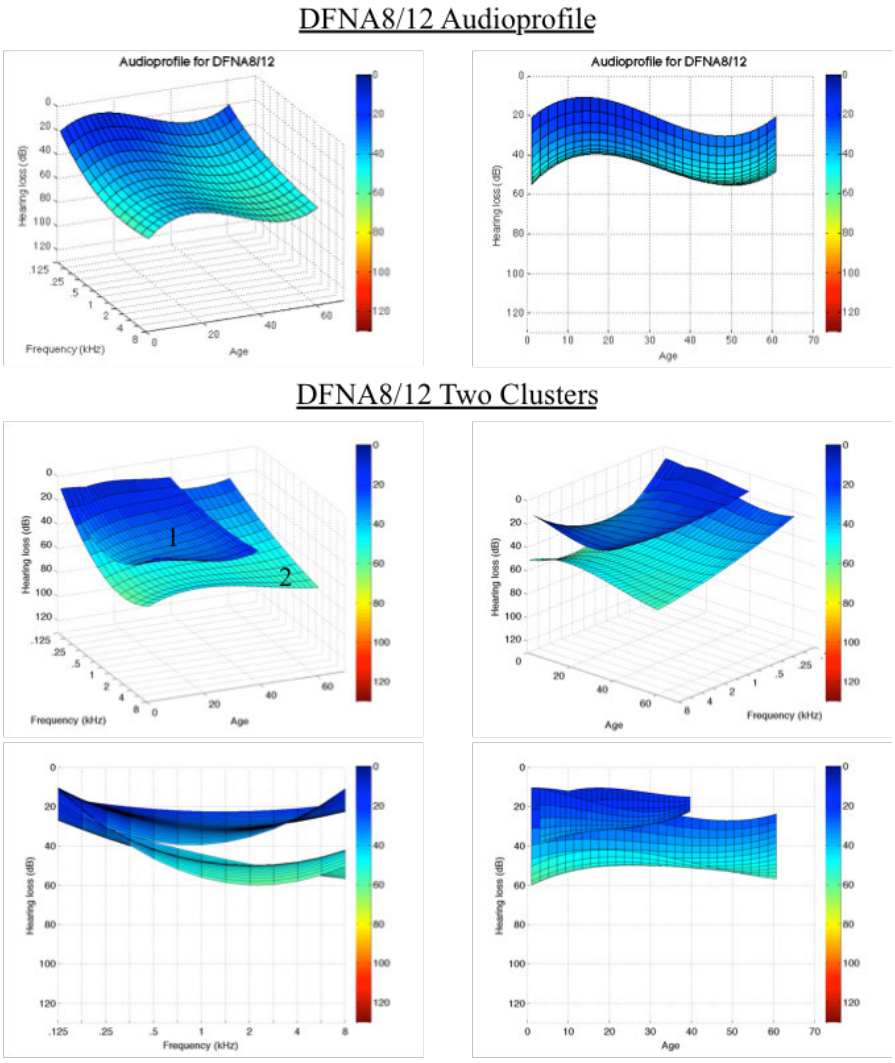


Figure 25. Audioprofile for DFNA8/12 along with the results of applying HSC with K_f set to 2. The audioprofile shows only slight hearing loss in the low and high frequencies with a valley of mild hearing loss in the mid-frequencies. There is a slight progression with the hearing loss going from slight to mild hearing loss in the higher frequencies. The surfaces found for the DFNA8/12 locus cluster based on severity of hearing loss.

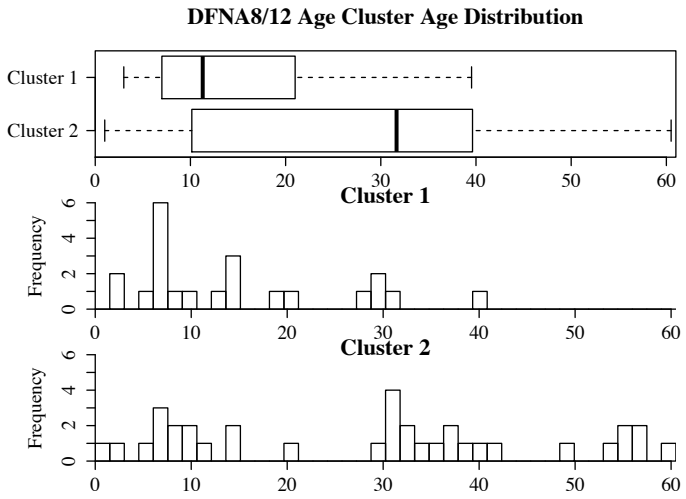


Figure 26. Age distribution of the two DFNA8/12 surfaces, and when matched for the age range there is no difference that can be attributed to age.

Table 8. DFNA8/12 clustering assignments based on mutation and domain of mutation.

Mutation	Domain	1	2	Total
p.C1057S	Trypsin inhibitor-like repeat	2	0	2
p.C1837G	ZP	1	2	3
p.C1837R	ZP	1	4	5
p.C1898R	ZP	1	0	1
p.D197N	Partial entactin G1 domain	0	3	3
p.P1791R	Interdomain sequence	0	1	1
p.R1890C	ZP	11	4	15
p.R2021H	ZP	2	2	4
p.S1758Y/G1759_N1795del	Interdomain sequence	3	12	15
p.S362C	vWF type D repeat	0	1	1
p.T1866M	ZP	0	4	4
p.T1866R	ZP	0	1	1
p.T562M	Interdomain sequence	1	0	1
p.T815M	vWF type D repeat	1	0	1
p.V317E	Interdomain sequence	0	1	1
	ZP	16	17	33
	Other	7	18	25
	Total	23	35	58

Note: The clustering does not associate based on domain of the protein (p-value=.175).

4.5 Discussion

Based on the analysis of the simulated datasets and the actual data from three different loci above, HSC appears to be an effective clustering technique for discovering subclasses within a useful category of datasets. The results for DFNA9 illustrate an example of a phenotype that is relatively consistent with the resulting cluster surfaces exhibiting a stair stepping pattern based on age and progression. This conforms to the expected results found for the example 3 simulated dataset. Clustering of the DFNA2A locus verified the capabilities of HSC to find a known subclass that was previously reported[6]. Applying HSC to DFNA8/12 demonstrated the ability of the method to be utilized as a hypothesis-generating tool. All possible hypotheses of both environmental and genetic causes of the clustering that were easily tested were ruled out, leaving possible genetic causes such as interacting genetic components as likely explanations that could then be followed up with manual wet-lab or in-silico verification.

4.5.1 Generalization

In general terms, the hierarchical surface-clustering algorithm described is considered a variation on single-linkage clustering because the criteria for merging clusters is defined as the minimum distance between the surfaces. While the application described in this thesis was focused on audiometric data with surfaces in three dimensions, it could be extended to other applications. For higher dimensional data where the surfaces would be hyperplanes, the merging of clusters would be done in the same manner as surfaces in three dimensions. The limiting factor would be the visualization of the clusters in three dimensions which could be accomplished through various dimensionality reduction techniques such as principle component analysis [102]. Other distance metrics could be used in place of the Euclidean distance, and one such distance metric is the Mahalanobis distance [103] that considers the variable of the data along with the Euclidian distance .

4.5.2 Comparing Against Other Clustering Techniques Using a Known Subclasses

While this method appears to be effective at finding novel subclasses in the AudioGene dataset and performs better in two out of the three simulated cases than K-means and spectral clustering. Ideally, the performance of HSC would be compared using cases of known subclasses within the AudioGene dataset. In order to evaluate the performance, ground truth must be known. For the AudioGene dataset, there is little ground truth for evaluating subclasses. The only currently available ground-truth in the dataset is based on the results of DFNA2A, which showed that the phenotypes for truncating and missense mutations were significantly different and had previously been reported in the literature [6]. The ground truth of the mutation type for each DFNA2A patient is known and can be used as the gold standard to compare cluster assignments from different clustering techniques. Patients with truncating mutations are assigned to one cluster and the remaining patients assigned to a different cluster.

The average ARI for 10 runs with K_f set to a range of 2 to 5 for HSC, K-means, and spectral clustering for DFNA2A can be seen in Table 9. The resulting surfaces for the various values of K_f for each of the clustering algorithms are shown in Figure 27. Overall, HSC with K_f set to 3 had the highest ARI value of 0.459, and was significantly better than the other clustering algorithms. The ARI value for both K-means and spectral clustering significantly increases when K_f is set to 4 because the surface that represents the patients with the truncation mutations is finally found. However, the ARI values are still significantly less than HSC even when K_f is also equal to 4 for HSC.

Visual inspection of the clusters in Figure 27 reveals that HSC finds the truncating phenotype surface when K_f is set to 3, and the same surface appears when K_f is set to 4 for both K-means and spectral clustering. This explains the previous result when comparing ARI between the various clustering techniques and values for K_f. Spectral clustering appears to be somewhat sensitive to the progression of hearing loss based on

age, and exhibits a similar stair stepping pattern to the clusters of DFNA9. The initial cluster assignments by HSC and spectral clustering were the same, but diverged as K_f was increased. Both HSC and K-means find a continuous surface that spans almost the entire range of ages but a similar surface does not appear for spectral clustering until K_f equals 5. Another interesting difference is noted when K_f is equal to 5. The new surface found by K-means displays approximately the same shape as the existing surface but indicates approximately 5 dB less hearing loss. In contrast, HSC reveals a surface that has a different shape with patients having less pronounced hearing loss in the mid frequencies as does the corresponding cluster in the case of K-means.

Overall, based on quantitative and qualitative measures, HSC performs better than K-means and spectral clustering for the case of DFNA2A and the simulated dataset. While this is not an exhaustive comparison between all clustering methods, it does appear that HSC improves upon K-means and spectral clustering. It has been demonstrated, however, that HSC is capable of generating hypotheses and has the potential to perform better than other commonly used clustering algorithms.

Table 9. The results of varying K_f for the DFNA2A truncating versus missense mutation evaluation set.

	Number of Clusters			
	2	3	4	5
Surface Clustering	-0.015 (0.01)	**0.459 (0)**	0.417 (.055)	0.286 (.081)
K-means	0.052 (0)	0.018 (0.001)	0.103 (.04)	0.092 (.01)
Spectral Clustering	-0.01 (0)	0.068 (0)	0.187 (0)	0.131 (.001)

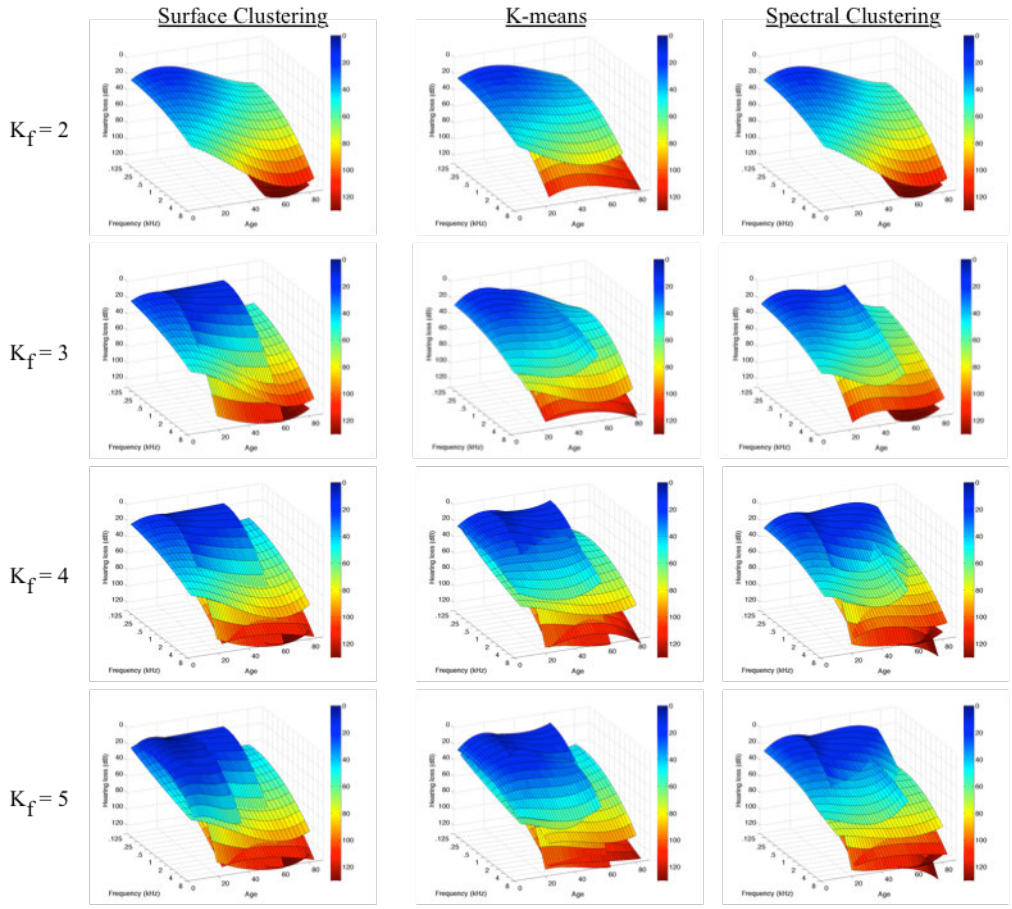


Figure 27. Comparison of the surfaces found by both HSC and K-means for the DFNA2A locus. Overall, the surfaces found are very similar with the exception to the first surface shown. For K-means, the first surface is very similar to the last surface of K-means, whereas HSC identifies a very different surface from others found.

4.5.3 Audioprofile surfaces

No previously known attempts have been made to fit three-dimensional surfaces to audiograms based on age. The current 2D form of the audiogram with decibels represented in descending order on the vertical y-axis and frequencies along the horizontal x-axis has remained relatively unchanged since it was first developed and refined in the mid 1920's [13,104]. The previous method for displaying progression of

hearing loss with age was the audioprofile, which superimposed average audiograms over 20 year intervals. By adding the third dimension to the audiogram and generating audioprofile surfaces, the progression of frequency based upon age becomes more apparent. As can be seen from the results, there are clear inter-loci differences in progression of the audioprofiles.

Another advantage of the audioprofile surface is the ability to visualize the resulting clusters and generate hypotheses that would have been difficult to generate otherwise. For instance, in the results for DFNA9 with K_f set to 3 it was observed that the surfaces had a stair stepping pattern in which the surfaces overlapped considerably. For surfaces that represent clusters 1 and 2 the overlap spanned 40 years (see Figure 9). While this can been seen from the age distribution, it also can be readily seen in the audioprofile surfaces along with the maximum difference between the two surfaces of approximately 40 dB. These patients that lie in these overlapping regions show the variability in the progression of the disease but could be further investigated for genetic modifiers that significantly modulate the hearing loss, assuming that environmental causes can be ruled out.

4.5.4 Parameter Choice for Surface Clustering

The behavior of the HSC algorithm can be tuned using three parameters: K_0, K_f, and S. K_0 defines the number of initial clusters, K_f is the final number of clusters and S is the minimum number of audiograms that a cluster must contain without being considered spurious and removed. The value of K_0 is bounded by N – the number of patients in a dataset–and K_f. Setting K_0 to be equal to K_f reduces the algorithm to K-means. If the value of K_0 is too large, then a significant portion of the patients would become singleton or small clusters and be removed because of S. Even if S were equal to 0, the singleton or small clusters would have distinct shapes, by virtue of being spurious, and remain throughout the merging steps of HSC and contaminate the results. On the other side, if K_0

is too small then the uniform effect of K-means will likely cause it to not find subclasses that consist of only of a relatively small portion of the class, similarly to the skewed case for the first synthetic dataset. Ideally, the value of K_0 should be large enough to overcome the uniform effect but small enough that any truly spurious clusters would be removed by S and at least one or more of the remaining cluster would represents any subclasses.

Another way to interpret K_0 is as the number of "prototype" surfaces that are initially generated. After each iteration of the HSC, the two most similar surfaces are merged. Ideally, K_0 is large enough to find interesting prototype surfaces, but small enough that only a few or none of the clusters can be attributed to noise while novel cluster surfaces still remain. K_0 also controls indirectly the smoothness of the initial prototype surfaces. Setting K_0 to higher values increases the likelihood of finding spurious clusters that are distinctively shaped but which may show discontinuities. Because of their distinctive shape, they may be resistant to being merged with other surfaces and likely persisting throughout most iterations of the algorithm. Therefore, the parameter S is needed to remove clusters that do not have much support in terms of the number of patients. After the initial K-means clustering is performed, the similarity of the surfaces is based purely upon shape, while the size of clusters is no longer considered. Through trial and error, the value of 15 was chosen for K_0 because it resulted in interesting clusters for most loci studied, while few spurious clusters were found that appeared to be truly spurious upon inspection with S set to 4.

4.5.5 Attributes of Surface Clustering

HSC outperformed or had comparable performance to both K-means and spectral clustering in both the simulated cases and the DFNA2A test case. The reason for the improvement is likely due a few features of the HSC algorithm itself. First, representing the audiograms in each cluster as surfaces has a smoothing effect on the audiograms but the surface captures the general pattern of the hearing loss. As an example, the 3D

audioprofile of a cluster with the audiograms superimposed can be seen in Figure 28. The cluster came from the initial clustering of K-means during HSC for DFNA2A. For this cluster, all the audiograms are similar in progression but have some slight variability–but the surface is capturing the overall shape. This also has a secondary benefit of reducing majority class effects of a large cohort of patients having a similar genotype and a relatively small number of patients with a genetic modifier that has a significant effect on the phenotype. Therefore the overall shape becomes the overriding feature. In contrast, K-means exhibited poor performance when a disproportionate number of patients had a genetic modifier. This was observed in the synthetic dataset example 1 where the proportion of patients with the modifier was skewed (see Table 4).

One aspect that still needs further research is in how best to handle multiple audiograms. For the analysis of the AudioGene loci, the audiograms were averaged for patients with multiple audiograms but this likely creates an age bias that could be limiting performance. Note that not all patients contained multiple audiograms and over 63% of the patients had only a single audiogram. Even for the patients with multiple audiograms the average age span of the audiograms was 13 years. To handle multiple audiograms, K-means would need to be modified to cluster multi-instance datasets, but the remaining steps of HSC would remain relatively unchanged.

Another aspect of HSC is that it explicitly considers age whereas spectral clustering and K-means only use it as one of many equally weighted features. By computing the distance from only the overlapping region or for patients of similar ages between the two surfaces, only the region that is most likely to be similar is considered. This means that if there is a strong progression with age, then the regions of the surfaces that are considered during the distance computation are only the regions that are likely to have similar hearing loss. For example, in the case of DFNA9 with a strong progression with age, surface 2 in Figure 9 represents the younger patients and surface 3 represents the older patients. The patients on the extreme of the ages (around age 0 and 70) would

Example Cluster With Audiograms

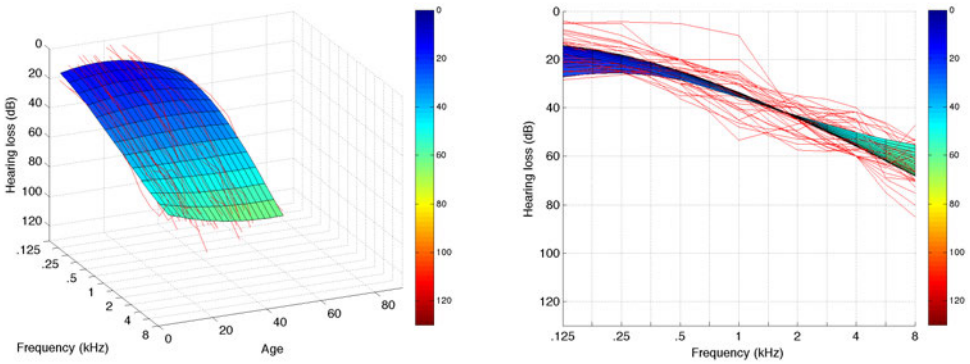


Figure 28. Example cluster with the surface and audiograms assigned to that cluster superimposed. The surface and clusters are from the initial clustering with K_0 set to 15 for DFNA2A, before surfaces are merged. As can be seen, the surface captures the general shape of the audiograms.

have drastically different hearing loss and it would not make sense to compare them to each other. With spectral clustering the difference this causes during the clustering can be seen in the results of DFNA2A, Figure 16, for values of K_f greater than or equal to 2 with the stair stepping pattern of the middle surface. In contrast, both K-means and HSC have a continuous surface for the patients with the standard pattern of hearing loss for DFNA2A, surface 3 in Figure 11.

4.5.6 Incorporating DFNA2A Result into AudioGene

From the results of DFNA2A a phenotypic difference was found between patients with missense and truncating mutations. Based on this difference, the DFNA2A patients in the AudioGene dataset were split into two subclasses DFNA2A_miss and DFNA2A_trunc. Ten rounds of 10-fold cross-validation were ran in order to determine the accuracy and ROC values, as in same major as the previous chapter. The results of the

original data and the new dataset with the DFNA2A subclasses can be seen in Table 10. While the average accuracy did decrease slightly from 42.89% for the original dataset to 41.84% with the new subclasses, the difference was not statistically significant (p-value < 0.05). However, the average weighed AUC value was equal for both datasets and had a value of 0.81. Interestingly, by examining the confusion matrix from one of the cross-validation runs, the number of correctly classified DFNA2A_trunc was 24 and 16 were incorrectly classifier. Comparing the number correctly classified for DFNA2A_trunc to the number of truncating patients found in cluster 2 during the HSC, shown in Table 7, the number of patients is only differs by 1. This similarity between HSC and the classifier used in AudioGene, indicates that approximately 25 patients with truncating mutations had a strong phenotypic difference but for the remaining 16 other the difference was not sufficient enough to be able to distinguish. This furthermore indicates that HSC likely found the largest cluster of DFNA2A truncating that could be found given the current DFNA2A patients. Finally, this result also reinforces the initial observation that accuracy is a not an effective method to determine if a class should be spit, and had accuracy been the metric it would have been concluded that the class should not have been split. Also, more informative predictions can now be made for DFNA2A without sacrificing accuracy.

Table 10. Comparison of accuracy and AUC values for the original dataset and the dataset with the DFNA2A subclasses.

Dataset	Accuracy	AUC
Original Dataset	42.89 (2.99)	0.81 (0.02)
Dataset with DFNA2A Subclasses	41.84 (2.74)	0.81 (0.02)

4.6 Conclusion

In this chapter, a new clustering technique was described called hierarchal surface cluster (HSC), and was shown to perform better or have comparable performance to two existing clustering technique when clustering hearing loss data. Simulated datasets were used to initially evaluate the performance, and then using the results of DFNA2A as a gold standard it was further shown that HSC performed better than the other clustering techniques evaluated. To visualize the clusters of audiograms found, a technique of plotting the audiograms as surfaces in 3D was also developed. This technique is useful for plotting the results of HSC, but is also useful for showing the audioprofile in 3D to visualize the progression with age of different frequencies that is different for each locus.

The results of DFNA2A demonstrated that the clustering technique could identify subgenotypes based on the phenotype. The identified subgenotypes were then used to define new subclasses in the AudioGene dataset. By performing cross-validation using with the new subclasses for DFNA2A, no statistically significant difference in accuracy or AUC was found. For the case of DFNA9, no subclasses were identified but clusters appeared to be based on the age of the patient and exhibited a stair stepping pattern based on progression and age, which was similar to the third synthetic dataset. Finally, the use of HSC as a hypotheses generating tool was shown based on the results of DFNA8/12. Two clusters were found which appear phenotypically different but no easily testable

cause was identified. This leaves other possible causes such as mutations in interacting partner proteins that will require additional *in vivo* studies to be undertaken.

Simulated datasets were generated to evaluate and test different cases of genetic modifiers affecting the phenotype and allowed for a quantitative comparison between the different clustering techniques. In the first synthetic dataset, HSC obtained the highest ARI value, while spectral clustering consistently found the correct clustering assignment for the second synthetic dataset. HSC alternated between the correct clustering assignment and an incorrect assignment that had a low ARI value, and since in practice the clustering algorithm would be repeated it can be said to have comparable performance to spectral clustering. Using the results of DFNA2A to create a gold standard clustering assignment with the missense and the truncating patients in separate clustering, HSC had the highest ARI value of 0.459.

CHAPTER 5

CONCLUSION

The main goal of the thesis was to apply machine-learning techniques for predicting the genetic cause of patients with a genetic disorder based on the phenotype of the disease, and leverage the phenotype to discover subgenotypes - specifically for Non-syndromic Hearing Loss (NSHL). The problem was tackled using two different approaches; first using supervised learning techniques and then developing an unsupervised technique called Hierarchal Surface Clustering (HSC) to explore the phenotype space to infer undiscovered genetic causes.

In Chapter 3 supervised machine learning techniques were used to develop a pipeline, called AudioGene, for the prediction of Autosomal Dominant Non-syndromic Hearing Loss (ADNSHL) loci. The phenotypic features used were audiograms, which are plots of the patient's hearing loss. The audiograms were preprocessed before use in AudioGene by filling in missing values and using the coefficients of second and third degree polynomials as secondary features. The accuracy of predicting the top three candidate loci was 68% when using an MI-SVM, compared to 44% using a Majority classifier. A noise model was developed using the discrete cosine transform (DCT) based on the test-retest variable that is expected during the measurement of the audiograms. Using the noise model it was shown that AudioGene did not suffer significant performance degradation under the expected test-retest variably.

The techniques developed for predicting ADNSHL loci were also applied to Autosomal Recessive Non-syndromic Hearing Loss (ARNSHL), however the only dataset available contained only mutations for patients with mutations in DFNB1 (GJB2). A large class imbalance also exists in this dataset, with patients with the 35delG homozygous mutation accounting for approximately 91% of the dataset. To reduce the effects of the class imbalance, the mutations were relabeled as homozygous truncating

(T/T) or "Other" which included patients with heterozygous truncating/missense and homozygous missense mutations. The T/T patients were down-sampled to be of equal size to the "Other" class and the average accuracy of 100 runs of cross-validation was 83% for the MI-SVM compared to 50% for a Majority classifier. The accuracy of predicting the patients removed during down-sampling using a classifier trained on the down-sampled dataset was 88%.

While AudioGene was originally developed for the prioritization of loci and genes for Sanger sequencing, it is equally applicable when High Throughput Sequencing (HTS) is used. When using HTS, even after post-sequencing filtering steps, the number of variants of unknown significance (VUS) remaining can be quite large. AudioGene can serve as either a phenotypic filter tool or as a phenotypic concordance check. The latter case applies in the cases where increased evidence of a VUS being suspected as the putative mutation is needed. The pipeline is available to the pubic at http://audiogene.eng.uiowa.edu, and all analyses are performed on a CBCB webserver and do not require any special software to be downloaded by the user.

In Chapter 4, an unsupervised technique called Hierarchal Surface Clustering (HSC) was developed to further explore the relationship between phenotypes of NSHL and genotypes. Particularly of interest were subgenotypes that caused differences in the manifestation of the disease phenotype such as different mutation types and mutations in interacting partners of the gene. The technique uses 3D surfaces fitted to clusters of audiograms, which are based on an initial clustering by K-means, and then repeated merging using their surface distances. Using simulated datasets where the effect of the subgenotype and phenotype are known, it was shown that HSC performed better or had comparable performance when evaluating using the Adjusted Rand Index (ARI) metric.

Applying HSC to three loci (DFNA2A, DFNA8/12 and DFNA9) demonstrated its usefulness as a means to finding subgenotyes and as a hypothesis-generating tool. For DFNA2A, a cluster was found that was determined to be representing patients with

truncating mutations. This subtype was previously known in the literature but was found without prior knowledge of the previously published literature. Two clusters were found for DFNA8/12 that represented differences in hearing loss that could not be attributed to any known causes and further *in vivo* studies are needed to determine the genetic cause of the phenotypic difference. No subgenotypes were found for DFNA9, but an interesting stair stepping pattern was observed that was related to the progression of the hearing loss with age.

Based on the results of DFNA2A analyses, a gold standard clustering assignment was created based on the mutation types of the patients and used to compare the different clustering techniques using real world data. HSC had the highest ARI with a value of 0.459 compared to 0.187 for spectral clustering and 0.103 for K-means clustering. HSC also found the surface that represented the patients with truncating mutations when K_f was set to 3, compared to both spectral clustering and K-means that successfully found this surface when K_f was set to 4. Finally, the performance of AudioGene was evaluated when using the subgenotypes as new class labels for DFNA2A and the accuracy found to not decrease significantly going from 44% to 43%. By using the new class labels DFN2A, AudioGene can make more informative predictions for DFNA2A without sacrificing accuracy.

The main goal of this thesis was to apply machine-learning techniques for predicting the genetic cause of a patients genetic disorder based on the phenotype of the disease – specifically NSHL. The problem was tackled using two different approaches; first using supervised learning techniques and then developing an unsupervised technique called Hierarchal Surface Clustering (HSC) to explore the phenotype space to infer undiscovered genetic causes.

REFERENCES

1. Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Med. 2010;2:84.

2. Snoeckx RL, Huygen PLM, Feldmann D, Marlin S, Denoyelle F, Waligora J, Mueller-Malesinska M, Pollak A, Ploski R, Murgia A, Orzan E, Castorina P, Ambrosetti U, Nowakowska-Szyrwinska E, Bal J, Wiszniewski W, Janecke AR, Nekahm-Heis D, Seeman P, Bendova O, Kenna MA, Frangulov A, Rehm HL, Tekin M, Incesulu A, Dahl H-HM, Sart du D, Jenkins L, Lucas D, Bitner-Glindzicz M, Avraham KB, Brownstein Z, Del Castillo I, Moreno F, Blin N, Pfister M, Sziklai I, Toth T, Kelley PM, Cohn ES, Van Maldergem L, Hilbert P, Roux A-F, Mondain M, Hoefsloot LH, Cremers CWRJ, Löppönen T, Löppönen H, Parving A, Gronskov K, Schrijver I, Roberson J, Gualandi F, Martini A, Lina-Granade G, Pallares-Ruiz N, Correia C, Fialho G, Cryns K, Hilgert N, Van de Heyning P, Nishimura CJ, Smith RJH, Van Camp G. GJB2 mutations and degree of hearing loss: a multicenter study. Am. J. Hum. Genet. 2005 Dec;77:945–957.

3. Taylor KR, Deluca AP, Shearer AE, Hildebrand MS, Black-Ziegelbein EA, Anand VN, Sloan CM, Eppsteiner RW, Scheetz TE, Huygen PLM, Smith RJH, Braun TA, Casavant TL. AudioGene: Predicting Hearing Loss Genotypes from Phenotypes to Guide Genetic Screening. Human mutation. 2013 Apr;34:539–545.

4. Gorlin RJ, Toriello HV, Cohen MM. Hereditary Hearing Loss and Its Syndromes. 2nd ed. Oxford University Press; 2004.

5. Raffan E, Semple RK. Next generation sequencing--implications for clinical practice. British Medical Bulletin. 2011 Sep;99:53–71.

6. Kamada F, Kure S, Kudo T, Suzuki Y, Oshima T, Ichinohe A, Kojima K, Niihori T, Kanno J, Narumi Y, Narisawa A, Kato K, Aoki Y, Ikeda K, Kobayashi T, Matsubara Y. A novel KCNQ4 one-base deletion in a large pedigree with hearing loss: implication for the genotype-phenotype correlation. J. Hum. Genet. 2006;51:455–460.

7. ANSI. S3. 6-2004, Specification for audiometers. American National Standards Institute. 2004.

8. Guidelines for Manual Pure-Tone Threshold Audiometry. Rockville, MD: American Speech-Language-Hearing Association; 2005.

9. Schmuziger N, Probst R, Smurzynski J. Test-retest reliability of pure-tone thresholds from 0.5 to 16 kHz using Sennheiser HDA 200 and Etymotic Research ER-2 earphones. Ear Hear. 2004 Apr;25:127–132.

10. Seashore CE. New Psychological Apparatus. University of Iowa studies in psychology. Iowa City, Iowa: The University of Iowa; 1897;2:153–163.

11. Sterne J. MP3: The Meaning of a Format. Duke University Press; 2012.

12. Fowler EP, Wegel RL. Audiometric methods and

the application. Trans Am Laryngol Rhinol Otol Soc. 1922;:98–132.

13. Fletcher H. Audiometric measurements and their uses. Transactions of the College of Physicians of the College of Pysicians of Philadelphia. 45 ed. 1923;:489–501.

14. Smith RJ, Bale JF Jr, White KR. Sensorineural hearing loss in children. The Lancet. 2005 Mar;365:879–890.

15. Duygu Duman MT. Autosomal recessive nonsyndromic deafness genes: a review. Frontiers in bioscience : a journal and virtual library. NIH Public Access; 2012;17:2213.

16. Hilgert N, Smith RJH, Van Camp G. Forty-six genes causing nonsyndromic hearing impairment: which ones should be analyzed in DNA diagnostics? Mutat. Res. 2009 Mar;681:189–196.

17. Kubisch C, Schroeder BC, Friedrich T, Lütjohann B, El-Amraoui A, Marlin S, Petit C, Jentsch TJ. KCNQ4, a novel potassium channel expressed in sensory outer hair cells, is mutated in dominant deafness. Cell. 1999 Feb;96:437–446.

18. Cremers CWRJ, Smith R. Genetic hearing impairment: its clinical presentations. Karger Medical and Scientific Publishers; 2002.

19. Ikezono T, Omori A, Ichinose S, Pawankar R, Watanabe A, Yagi T. Identification of the protein product of the Coch gene (hereditary deafness gene) as the major component of bovine inner ear protein. Biochim. Biophys. Acta. 2001 Mar;1535:258–265.

20. Robertson NG. Cochlin immunostaining of inner ear pathologic deposits and proteomic analysis in DFNA9 deafness and vestibular dysfunction. Hum. Mol. Genet. 2006 Feb;15:1071–1085.

21. Hildebrand MS, Morín M, Meyer NC, Mayo F, Modamio-Hoybjor S, Mencía A, Olavarrieta L, Morales-Angulo C, Nishimura CJ, Workman H, Deluca AP, Del Castillo I, Taylor KR, Tompkins B, Goodman CW, Schrauwen I, Van Wesemael M, Lachlan K, Shearer AE, Braun TA, Huygen PLM, Kremer H, Van Camp G, Moreno F, Casavant TL, Smith RJH, Moreno-Pelayo MA. DFNA8/12 caused by TECTA mutations is the most identified subtype of non-syndromic autosomal dominant hearing loss. Human mutation. 2011 Apr;32:825–834.

22. Toth T, Pfister M, Zenner H-P, Sziklai I. Phenotypic characterization of a DFNA6 family showing progressive low-frequency sensorineural hearing impairment. Int. J. Pediatr. Otorhinolaryngol. 2006 Feb;70:201–206.

23. Gonçalves AC, Matos TD, Simões-Teixeira HR, Pimenta Machado M, Simão M, Dias OP, Andrea M, Fialho G, Caria H. WFS1 and non-syndromic low-frequency sensorineural hearing loss: a novel mutation in a Portuguese case. Gene. 2014 Apr;538:288–291.

24. Exome Variant Server [Internet]. [cited 2014 May 1]. Available from: http://evs.gs.washington.edu/EVS/.

25. Hofmann S, Philbrook C, Gerbitz K-D, Bauer MF. Wolfram syndrome: structural and functional analyses of mutant and wild-type wolframin, the WFS1 gene product. Hum. Mol. Genet. 2003 Aug;12:2003–2012.

26. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. 2011 Nov;12:745–755.

27. Shearer AE, Deluca AP, Hildebrand MS, Taylor KR, Gurrola J, Scherer S, Scheetz TE, Smith RJH. Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing. Proc. Natl. Acad. Sci. U.S.A. 2010 Dec;107:21104–21109.

28. MORL Testing Menu [Internet]. [cited 2014 May 14]. Available from: http://www.medicine.uiowa.edu/morl/otoscopecost/.

29. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009.

30. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20:1297–1303.

31. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov;491:56–65.

32. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. 1977 Dec;74:5463–5467.

33. Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician. 1992 Aug;46:175–185.

34. Vapnik V. The nature of statistical learning. Springer; 1995.

35. Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. Kluwer Academic Publishers; 1998;2:121–167.

36. Breiman L. Bagging predictors. Machine Learning. 1996 Aug;24:123–140.

37. Iba W, Langley P. Induction of One-Level Decision Trees. Citeseer; 1992;:233–240.

38. Breiman L. Random Forest. Machine Learning. 2001;45:5–32.

39. Tan P-N. Introduction to Data Mining. Addison-Wesley; 2006.

40. Hastie T, Tibshirani R. Classification by pairwise coupling. The Annals of Statistics. Institute of Mathematical Statistics; 1998 Apr;26:451–471.

41. Amores J. Multiple instance classification: Review, taxonomy and comparative study. Artificial Intelligence. 2013 Aug;201:81–105.

42. Xu X. Statistical learning in multiple instance problems. [Hamilton, New Zealand]: The University of Waikato; 2003.

43. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier; 2011.

44. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997 Jul;30:1145–1159.

45. Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. New York: Springer; 2010.

46. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 2010 Jun;31:651–666.

47. MacQueen J. Some methods for classification and analysis of multivariate observations. California, USA; 1967;1:14.

48. de Hoon MJL, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics. 2004 Jun;20:1453–1454.

49. Bradley PS, Mangasarian OL, Street WN. Clustering via Concave Minimization. Advances in Neural Information Processing Systems. MIT Press; 1997. pp. 368–374.

50. Agha El M, Ashour WM. Efficient and Fast Initialization Algorithm for K-means Clustering. International Journal of Intelligent Systems and Applications(IJISA). 2012 Feb;4:21.

51. Celebi ME, Kingravi HA, Vela PA. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications. 2013.

52. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal statistical Society. Series B; 1977;39:1–38.

53. Bishop CM. Pattern Recognition and Machine Learning. Secaucus, NJ, USA: Springer-Verlag New York, Inc; 2006.

54. Do CB, Batzoglou S. What is the expectation maximization algorithm? Nature biotechnology. 2008.

55. Kriegel H-P, Kröger P, Sander J, Zimek A. Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery [Internet]. John Wiley & Sons, Inc; 2011;1:231–240. Available from: http://onlinelibrary.wiley.com/doi/10.1002/widm.30/full.

56. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD. 1996.

57. Zhou H, Wang P, Li H. Research on Adaptive Parameters Determination in DBSCAN Algorithm⋆. Journal of Information & Computational Science. 2012.

58. Karypis G, Eui-Hong Han, Kumar V. Chameleon: hierarchical clustering using dynamic modeling. Computer. IEEE; 1999;32:68–75.

59. Luxburg Von U. A tutorial on spectral clustering. Stat Comput. Springer; 2007;17:395–416.

60. Ng AY, Jordan MI, Weiss Y. On Spectral Clustering: Analysis and an algorithm. In: Dietterich T, Becker S, Ghahramani Z, editors. MIT Press; 2001. pp. 849–856.

61. Meila M, Shi J. A random walks view of spectral segmentation. Citeseer; 2001.

62. Luxburg U. A tutorial on spectral clustering. Stat Comput. Springer US; 2007 Aug;17:395–416.

63. Sibson R. SLINK: An optimally efficient algorithm for the single-link cluster method. The Computer Journal. 1973 Jan;16:30–34.

64. Liu Y, Li Z, Xiong H, Gao X, Wu J. Understanding of Internal Clustering Validation Measures. 2010 IEEE 10th International Conference on Data Mining (ICDM). IEEE; 2010. pp. 911–916.

65. Hubert L, Arabie P. Comparing partitions. Journal of Classification. Springer-Verlag; 1985 Dec;2:193–218.

66. Santos JM, Embrechts M. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. Artificial Neural Networks–ICANN 2009. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. pp. 175–184.

67. Rand WM. Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association. Taylor & Francis Group; 1971 Dec;66:846–850.

68. Fritsch A, Ickstadt K. Improved criteria for clustering based on the posterior similarity matrix. Bayesian Analysis. International Society for Bayesian Analysis; 2009 Jun;4:367–391.

69. Valenzuela RK, Henderson MS, Walsh MH, Garrison NA, Kelch JT, Cohen-Barak O, Erickson DT, John Meaney F, Bruce Walsh J, Cheng KC, Ito S, Wakamatsu K, Frudakis T, Thomas M, Brilliant MH. Predicting phenotype from genotype: normal pigmentation. J. Forensic Sci. 2010 Mar;55:315–322.

70. Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, Korn K, Selbig J. Diversity and complexity of HIV-1 drug resistance: A bioinformatics approach to predicting phenotype from genotype. Proc. Natl. Acad. Sci. U.S.A. 2002 Jun;99:8271–8276.

71. Clare A, King RD. Machine learning of functional class from phenotype data. Bioinformatics. 2002 Jan;18:160–166.

72. Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, Alnadi NA, Andraws N, Patterson ML, Krivohlavek LA, Fellis J, Humphray S, Saffrey P, Kingsbury Z, Weir JC, Betley J, Grocock RJ, Margulies EH, Farrow EG, Artman M, Safina NP, Petrikin JE, Hall KP, Kingsmore SF. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci Transl Med. 2012 Oct;4:154ra135.

73. Wong JA, Gula LJ, Klein GJ, Yee R, Skanes AC, Krahn AD. Utility of Treadmill Testing in Identification and Genotype Prediction in Long-QT Syndrome. Circulation: Arrhythmia and Electrophysiology. 2010 Apr;3:120–125.

74. Stierman L. Birth defects in California: 1983–1990. California Department of Health Services; 1994.

75. Leonard D, Shen T, Howe H, Egler T. Trends in the prevalence of birth defects in Illinois and Chicago 1989 to 1997. Springfield, IL: Illinois Department of Public Health; 1999.

76. White KR. The current status of EHDI programs in the United States. Ment Retard Dev Disabil Res Rev. 2003;9:79–88.

77. Van Camp G, Willems PJ, Smith RJ. Nonsyndromic hearing impairment: unparalleled heterogeneity. American journal of human …. 1997.

78. Hildebrand MS, Tack D, McMordie SJ, DeLuca A, Hur IA, Nishimura C, Huygen P, Casavant TL, Smith RJH. Audioprofile-directed screening identifies novel mutations in KCNQ4 causing hearing loss at the DFNA2 locus. Genetics in Medicine. 2008 Nov;10:797–804.

79. Cryns K, Van Camp G. Deafness Genes and Their Diagnostic Applications. Audiol Neurootol. Karger Publishers; 2004;9:2–22.

80. Ali Mosrati M, Schrauwen I, Ben Saiid M, Aifa-Hmani M, Fransen E, Mneja M, Ghorbel A, Van Camp G, Masmoudi S. Genome-wide analysis reveals a novel autosomal-recessive hearing loss locus DFNB80 on chromosome 2p16.1-p21. J. Hum. Genet. Nature Publishing Group; 2012 Dec;58:98–101.

81. Auer P. On Learning From Multi-Instance Examples: Empirical Evaluation of a Theoretical Approach. Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1997;:21–29.

82. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics. 2004 Oct;20:2479–2481.

83. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. SIGKDD Explor. Newsl. 2009 Nov;11:10.

84. Platt JC. Sequential minimal optimization: A fast algorithm for training support vector machines. Advances in Kernel MethodsSupport Vector Learning. Citeseer; 1998;208:1–21.

85. Ahmed N, Natarajan T. Discrete Cosine Transform. IEEE Trans. Computers. 1974;:90–93.

86. de Heer A-MR, Schraders M, Oostrik J, Hoefsloot L, Huygen PLM, Cremers CWRJ. Audioprofile-directed successful mutation analysis in a DFNA2/KCNQ4 (p.Leu274His) family. Ann. Otol. Rhinol. Laryngol. 2011 Apr;120:243–248.

87. Chan DK, Schrijver I, Chang KW. Connexin-26-associated deafness: phenotypic variability and progression of hearing loss. Genet. Med. 2010 Mar;12:174–181.

88. Hildebrand MS, Deluca AP, Taylor KR, Hoskinson DP, Hur IA, Tack D, McMordie SJ, Huygen PLM, Casavant TL, Smith RJH. A contemporary review of AudioGene audioprofiling: a machine-based candidate gene prediction tool for autosomal dominant nonsyndromic hearing loss. Laryngoscope. 2009 Nov;119:2211–2215.

89. Eppsteiner RW, Shearer AE, Hildebrand MS, Taylor KR, Deluca AP, Scherer S, Huygen P, Scheetz TE, Braun TA, Casavant TL, Smith RJH. Using the Phenome and Genome to Improve Genetic Diagnosis for Deafness. Otolaryngol Head Neck Surg. 2012 Jul.

90. FRADKIN D. Clustering Inside Classes Improves Performance of Linear Classifiers. 2008 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI). IEEE; 2008. pp. 439–442.

91. Japkowicz N. Supervised Learning with Unsupervised Output Separation. International Conference on Artificial Intelligence and Soft Computing. 2002;3:321–325.

92. Luo Y. Can subclasses help a multiclass learning problem? IEEE; 2008;:214–219.

93. Ahmed N, Campbell M. On estimating simple probabilistic discriminative models with subclasses. Expert Systems with Applications. 2012.

94. Ben-Dor A, Friedman N, Yakhini Z. Class discovery in gene expression data. RECOMB '01. New York, New York, USA: ACM Press; 2001. pp. 31–38.

95. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999 Oct;286:531–537.

96. Roth V, Lange T. Bayesian class discovery in microarray datasets. IEEE Trans Biomed Eng. 2004 May;51:707–718.

97. Shi J, Malik J. Normalized cuts and image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on. IEEE; 2000;22:888–905.

98. Wu J. The Uniform Effect of K-means Clustering. Advances in K-means Clustering. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. pp. 17–35.

99. Goodyear RJ, Richardson GP. Extracellular matrices associated with the apical surfaces of sensory epithelia in the inner ear: molecular and structural diversity. J. Neurobiol. 2002 Nov;53:212–227.

100. Schraders M, Ruiz-Palmero L, Kalay E, Oostrik J, del Castillo FJ, Sezgin O, Beynon AJ, Strom TM, Pennings RJE, Seco CZ, Oonk AMM, Kunst HPM, Domínguez-Ruiz M, García-Arumi AM, del Campo M, Villamar M, Hoefsloot LH, Moreno F, Admiraal RJC, Del Castillo I, Kremer H. Mutations of the gene encoding otogelin are a cause of autosomal-recessive nonsyndromic moderate hearing impairment. Am. J. Hum. Genet. 2012 Nov;91:883–889.

101. Richardson GP, Lukashkin AN, Russell IJ. The tectorial membrane: one slice of a complex cochlear sandwich. Curr Opin Otolaryngol Head Neck Surg. 2008 Oct;16:458–464.

102. Jolliffe IT. Principal Component Analysis. Second. Springer; 2002.

103. Mahalanobis, P. C. On the generalised distance in statistics. 1936. pp. 49–55.

104. Fowler EP, Wegel RL. Audiometric methods and their applications. Trans Am Laryngol Rhinol Otol Soc. 1922;28:98–132.

APPENDIX A. AVERAGE AUDIOGRAMS WITH ERROR BARS

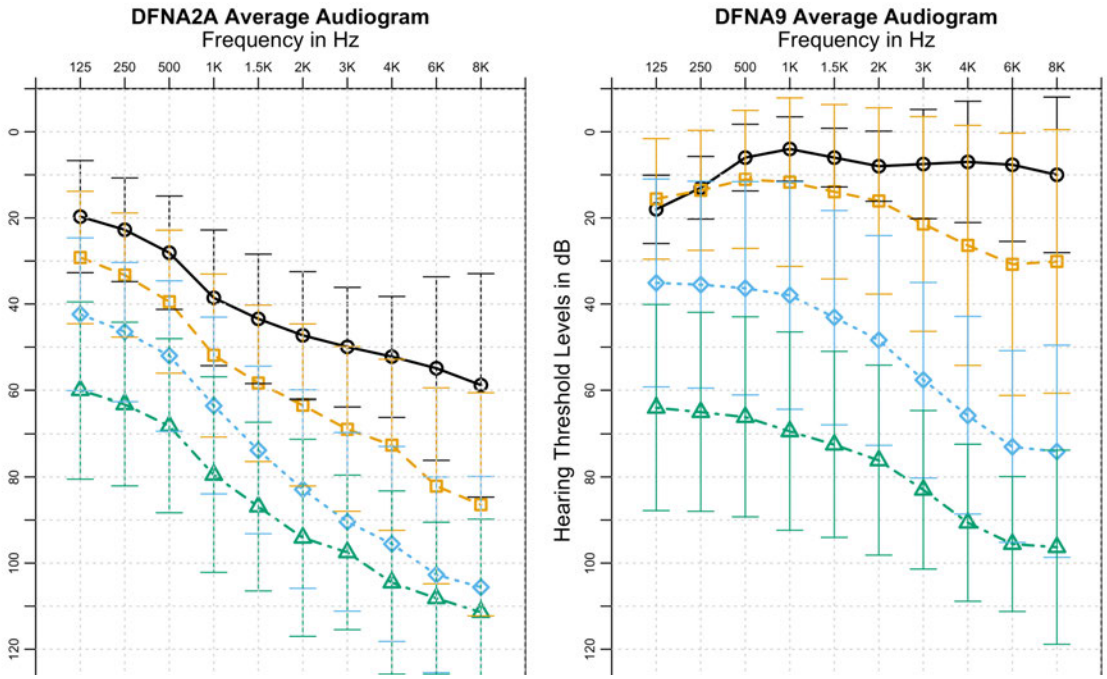


Figure A1. The average audiograms from Figure 1 with error bars representing one standard deviation added.

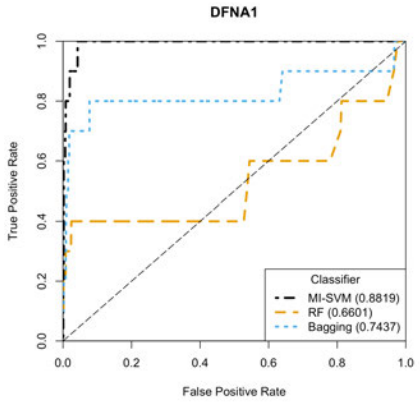
APPENDIX B. DATASET COMPOSITION

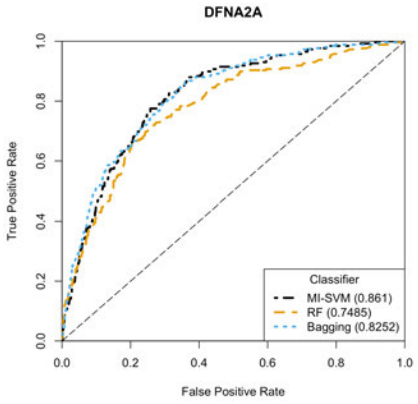
Table B1. Number of patients and audiograms for each locus before preprocessing.

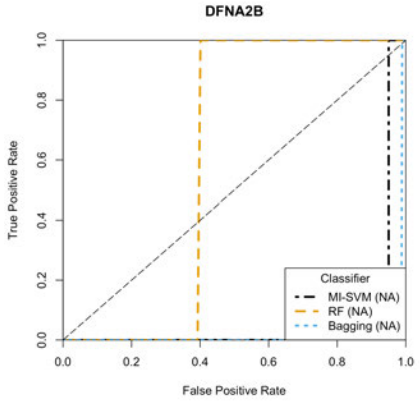
Locus	Num. of Patients	Num. of Audiograms
DFNA9	285	935
DFNA2A	258	611
DFNA6/14/38	132	324
DFNA2notAnotB	93	94
DFNA5	82	343
DFNA20/26	65	155
DFNA8/12	60	137
DFNA10	56	96
DFNA11	52	113
DFNA22	51	52
DFNA15	49	70
DFNA13	38	64
DFNA3A	32	41
DFNA36A	31	57
DFNA24	25	34
DFNA31	18	35
DFNA17	13	23
DFNA25	13	16
DFNA27	11	17
DFNA1	10	10
DFNA57	8	8
DFNA50	8	10
DFNA28	8	13
DFNA33	7	9
DFNA41	6	6
DFNA44	5	5
DFNA43	5	5
DFNA36notA	4	6
DFNA59	4	4
DFNA16	4	4
DFNA18	4	4
DFNA4A	4	4
DFNA4B	3	6
Total	**1445**	**3312**

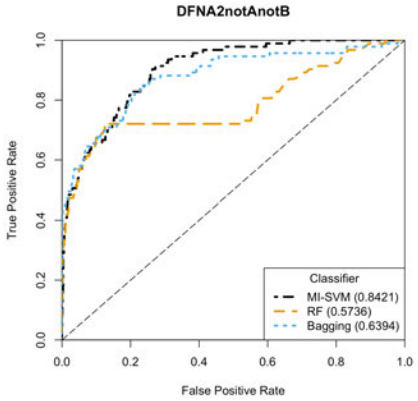
Note: Patients can have multiple audiograms taken at different ages.

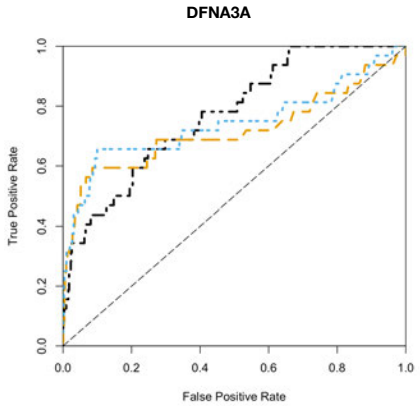
APPENDIX C. AUDIOGENE ROC CURVES

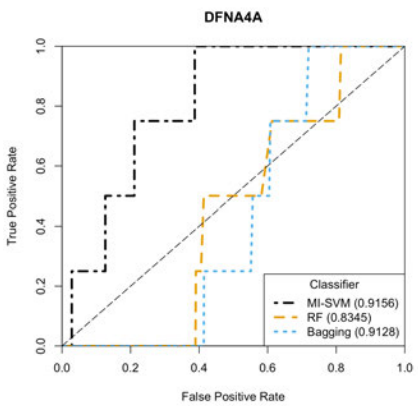


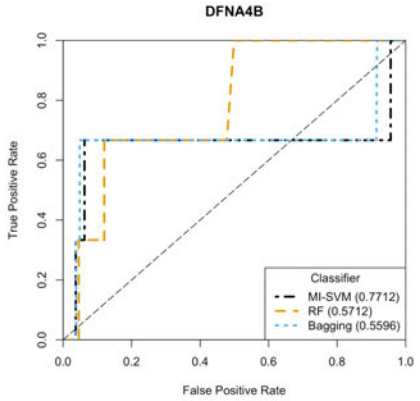


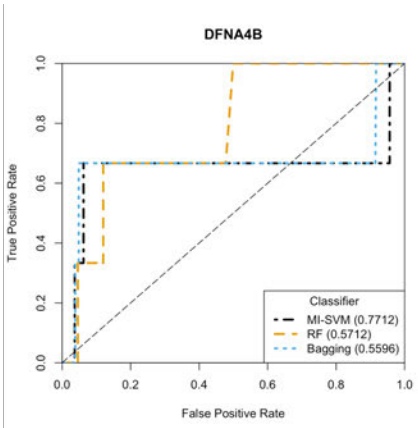


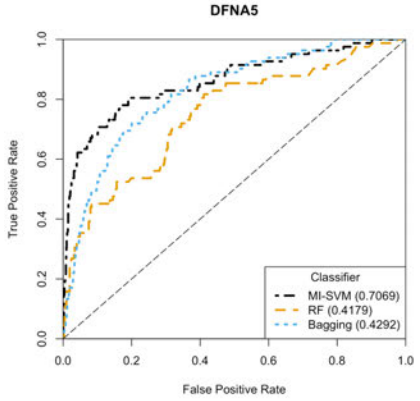


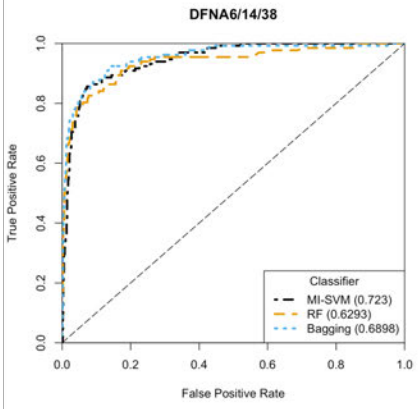


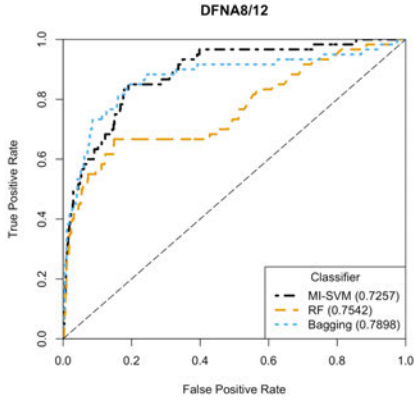


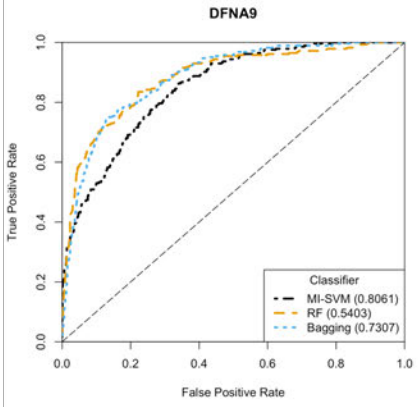


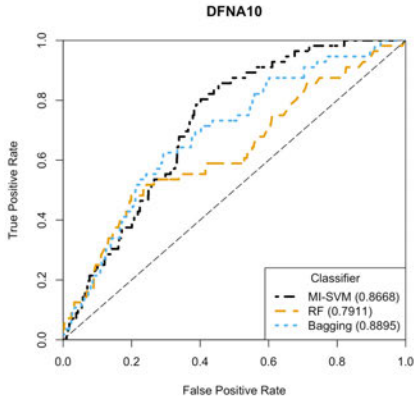


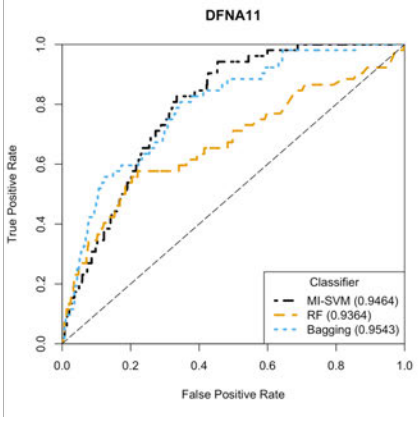


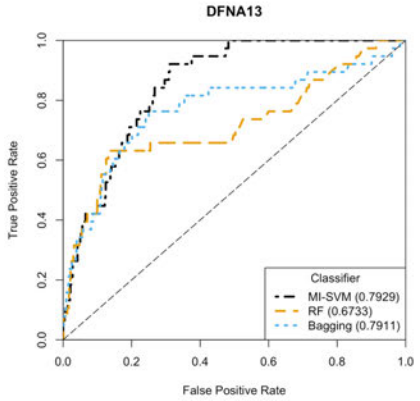


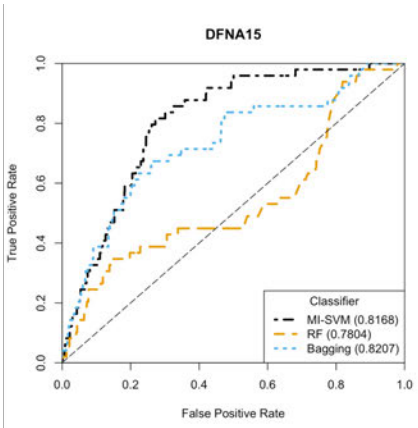


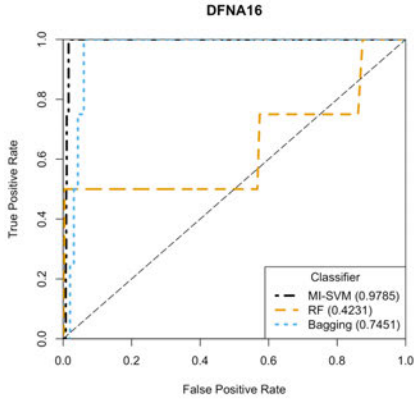


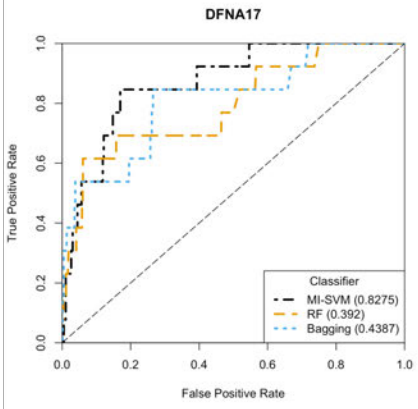


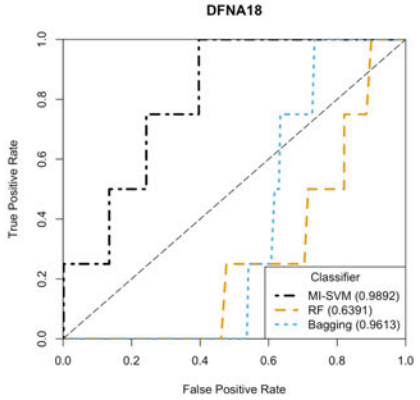


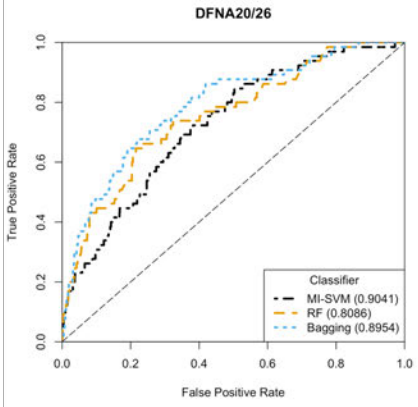


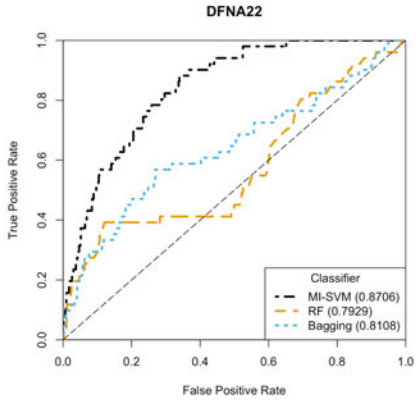


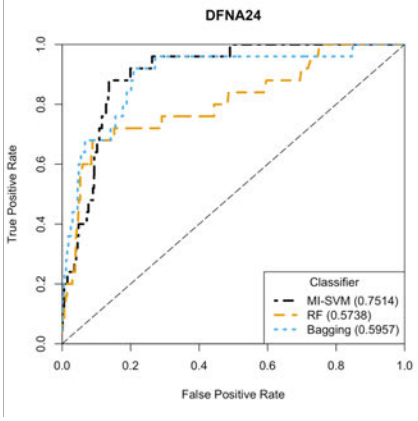


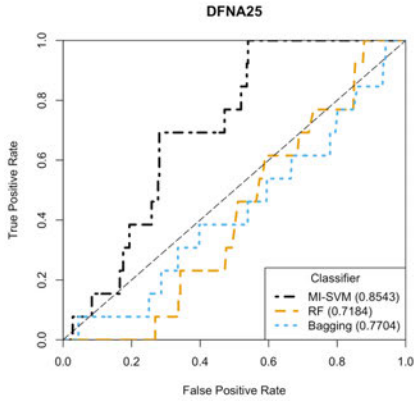


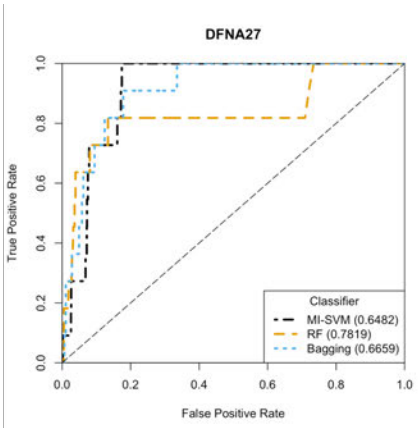


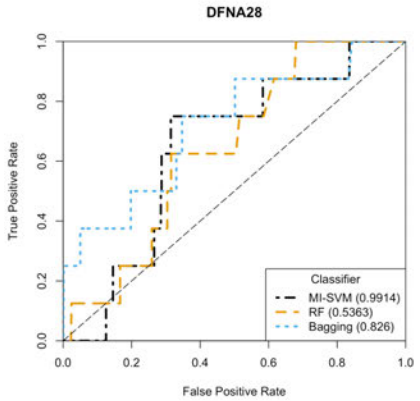


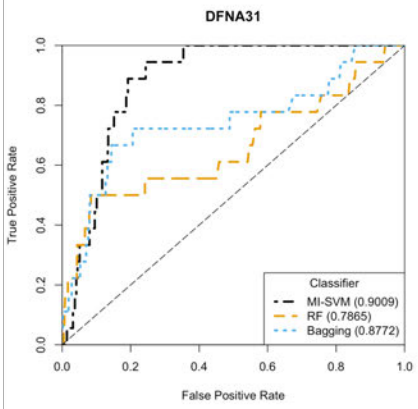


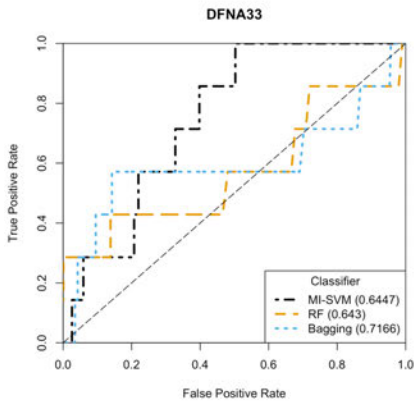


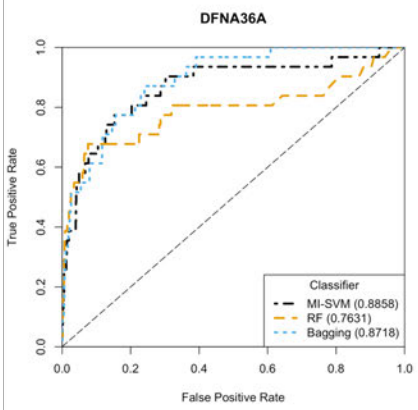


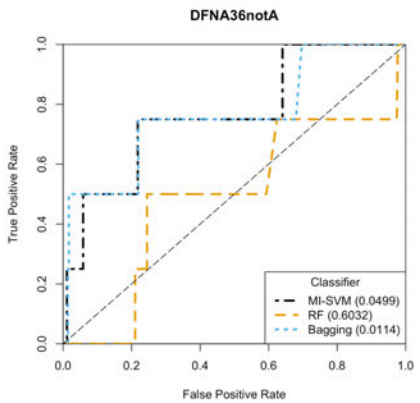


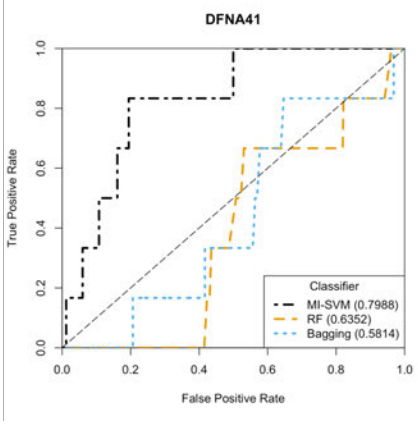












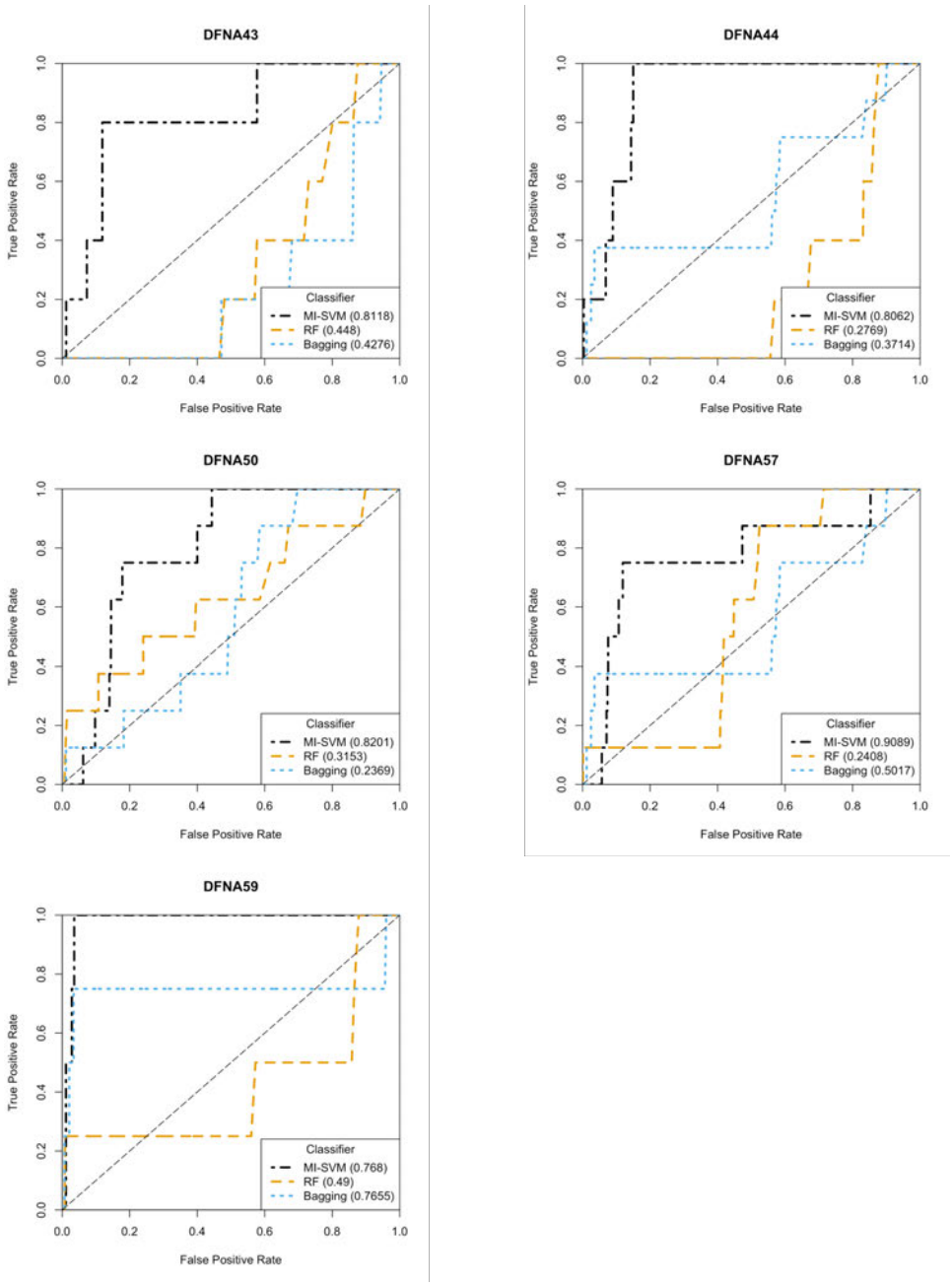


Figure C1. ROC curves for each locus for each classifier generate from a single 10-fold cross validation. AUC values are shown in the parentheses for each classifier.

APPENDIX D: AUDIOGENE OUTLIERS DISTRIBUTION

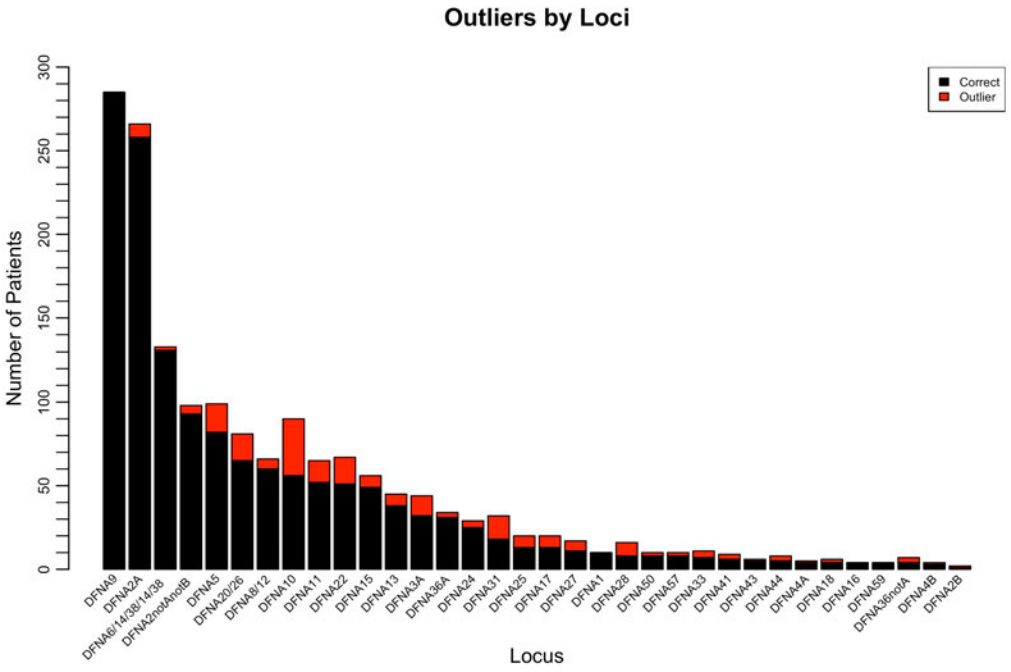


Figure D1. The chart illustrates the number of patients for each locus that might be considered an outlier (red) along with the number of patients that were not considered outliers (black).

APPENDIX E. ADDITIONAL HSC RESULTS FOR DFNA9

Table E1. P-Values of using an un-paired t-test for comparing the ages of the clusters found for DFNA9 when using HSC with K_f set to 3.

Cluster A	Cluster B	P-Value
1	2	3.35E-22
1	3	6.81E-16
2	3	5.63E-42

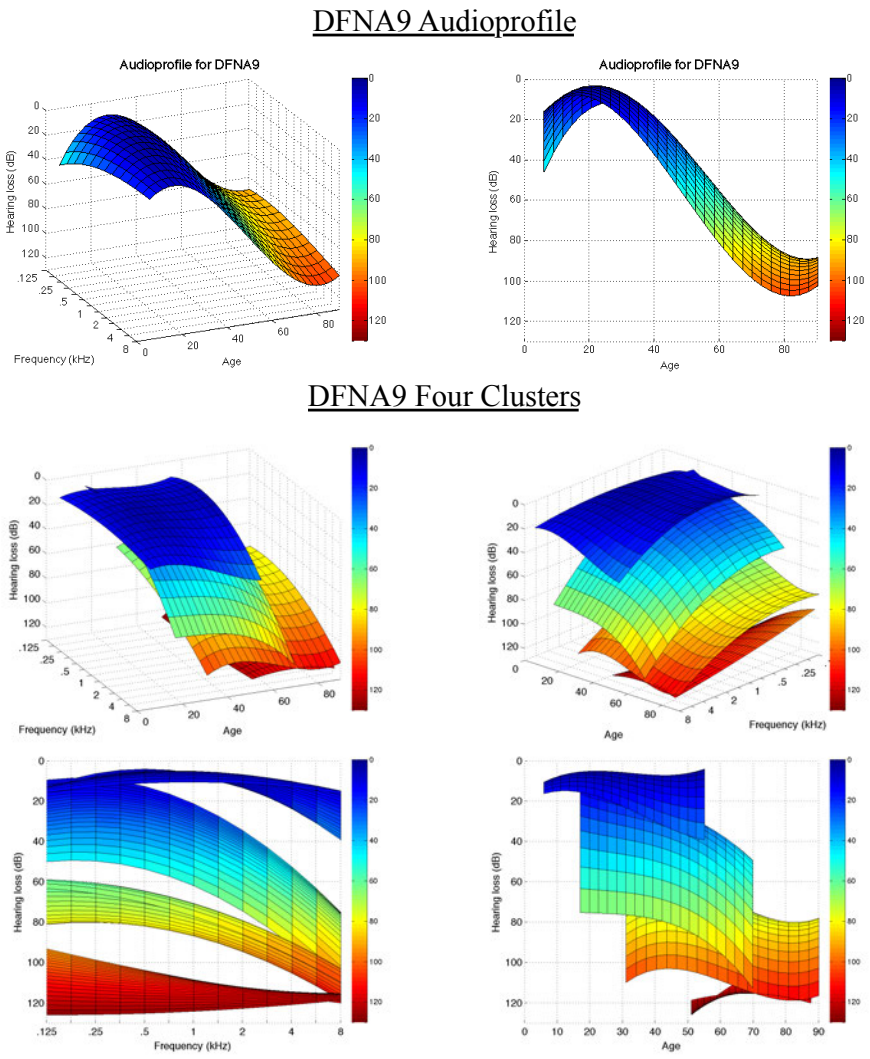


Figure E1. The surfaces when K_f is set to 4. Even with four clusters, the surfaces from a stair stepping pattern that segregate based on age.