

2018

Differentiating Human Populations Based on k-mer Classification of Hand Bacteria

Thrisha Doppala

West Virginia University, thdoppala@mix.wvu.edu

Follow this and additional works at: <https://researchrepository.wvu.edu/etd>



Part of the [Bioinformatics Commons](#), and the [Other Electrical and Computer Engineering Commons](#)

Recommended Citation

Doppala, Thrisha, "Differentiating Human Populations Based on k-mer Classification of Hand Bacteria" (2018). *Graduate Theses, Dissertations, and Problem Reports*. 3720.

<https://researchrepository.wvu.edu/etd/3720>

This Thesis is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Thesis has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.


Graduate Theses, Dissertations, and Problem Reports

2018

Differentiating Human Populations Based on k-mer Classification of Hand Bacteria

Thrisha Doppala

Follow this and additional works at: <https://researchrepository.wvu.edu/etd>

 Part of the [Bioinformatics Commons](#), and the [Other Electrical and Computer Engineering Commons](#)

Differentiating Human Populations Based on k-mer Classification of Hand Bacteria

Thrisha Doppala

Thesis submitted to the Benjamin Statler college of Engineering and Mineral
Resources at West Virginia University

in partial fulfilment of the requirements for the degree of

Master of Science in Electrical Engineering

Jeremy M Dawson, Ph.D., Chair

Donald Adjero, Ph.D.

Stephen DiFazio, Ph.D.

Lane Department of Computer Science and Electrical Engineering
Morgantown, West Virginia
2018

Keywords: Hand bacteria, 16S rRNA, V3 hypervariable region, Population group,
Differentiating people, K-mer classification

Copyright 2018 Thrisha Doppala

Abstract

Bacterial communities found in and on the human body are not only used in studying human health conditions but are also effective in differentiating individuals due to their distinct profiles. Human palm regions harbor relatively more diverse bacterial communities and are indicative of population groups, life styles, geographic locations, age groups and health conditions. Sequences extracted from hypervariable region V3 of the 16S rRNA bacterial gene of hand bacterial samples from 9 different population groups were classified into Operational Taxonomic Units (OTU) with GreenGenes reference taxonomy using RDP (Ribosomal Database Project) classifier. Frequencies of identified OTUs were used to study dissimilarities between samples by calculating the Kullback-Leibler Divergence (KLD) between every two samples. In addition to OTU frequencies, the frequencies of nucleotide k-mers from each OTU sequence were used to study the dissimilarities between samples. Based on the structure, 65 nucleotides of V3 hypervariable region were mapped into 47 elements, and distribution of k-mers from these mapped elements were used to determine dissimilarities between samples. Furthermore, a new technique was applied to classify sequences where sequences were clustered based on their k-mer frequency profile and a unique signature is assigned to every cluster. Frequencies of these signature clusters were used to calculate the KLD between different samples. This method classifies the unknown sequences that were ignored in OTU based methods. Ensemble learning method is applied to each of the above case of k-mers to identify the population group of a given hand bacterial sample. Samples were identified with a range of 51-98 % accuracy for different cases of k-mer distribution. Samples were classified with greater accuracy with k-mer classified sequences than with OTU sequences. Though applied on a small group of samples, these results provide a basis for the use of k-mer distributions in classifying and identifying individuals which could perform better on a broader range of time-varying dataset from other regions of 16S rRNA or even other genes, such as 23S rRNA of bacteria

Acknowledgements

I would like to take this opportunity to thank my advisor Dr. Jeremy Dawson for giving me this opportunity. I am thankful for all the unending support and guidance I received for the past three years. I am grateful to Dr. Dawson for constantly encouraging me through my research which has been sometimes challenging and which I believe made me a better engineer and a better person.

I would like to thank Dr. Donald Adjero for his time and ideas. I am very much grateful for all the times he accommodated me into his busy schedule. I offer my sincere thanks to Dr. Stephen DiFazio for agreeing to be in my committee.

I would like to thank my friends/roommates Suha Reddy Mokalla and Silpa Beegala for their constant support and encouragement.

I would like to thank my friends from Christian Student Fellowship for their love and encouragement. Special thanks to Jared Zinn for helping me with my grammar.

I am grateful to my parents and brother for their endless love and support. I thank you from the bottom of my heart for always trusting me. I love you all very much

Most of all, I thank my God Almighty for being my hope and support.

Table of Contents

<i>Abstract</i>	<i>ii</i>
<i>Acknowledgements</i>	<i>iii</i>
<i>Table of Contents</i>	<i>iv</i>
<i>Table of Figures</i>	<i>vi</i>
<i>List of Tables</i>	<i>x</i>
Chapter 1 : Introduction	1
1.1 Executive Summary	2
1.2 Skin Microbiome	4
1.3 Factors influencing skin microbiome.....	6
1.4 Other significant microbiome of the human body.....	10
1.5 Advantages of skin microbiome and its applications	13
1.6 History of Human Microbiome study	14
1.7 Prokaryotic 16S rRNA	16
1.8 K-mers in DNA analysis:.....	19
1.9 Problem Statement.....	19
1.10 Conclusions and Future work	20
1.11 Thesis Organization.....	21
Chapter 2 : Theory	23
2.1 16S rRNA genome	24
2.2 Hyper Variable region V3.....	25
2.3 DNA Sequencing	27
2.4 QIIME	33
2.5 Kullback-Leibler Divergence.....	34
2.6 Principal Coordinate Analysis (PCoA).....	36
2.7 Unsupervised learning on OTU distribution and their k-mer frequencies.....	36
2.8 Structure based mapping of 16S rRNA V3 region	39
2.9 K-means Clustering	41
2.10 Ensemble Learning	43
Chapter 3 : Data Extraction and organization	45
3.1 Overview	46
3.2 Collection Process.....	46
3.3 Demographics	47
3.4. DNA Isolation and V3 region amplification.....	50
3.5 Classification and organization of raw sequence data:	51
Chapter 4 : Bioinformatics' Analysis	53
4.1 Taxonomic classification of Raw sequences	54
4.2 Verification of K-mer signatures for known OTUs	55
4.3 Unsupervised machine learning of population groups using OTU and k-mer frequencies in hand bacterial samples	57

4.4 Unique signatures of k-mer frequencies	77
4.5 Unsupervised machine learning of population groups using unique k-mer frequencies	81
4.6 Ensemble learning of samples using OTU and signature frequencies	88
Chapter 5 : Conclusions	103
5.1 Summary	104
5.2 Conclusions and Future work	109
References.....	111
<i>Appendix A: KLD analysis of OTU frequencies</i>	<i>118</i>
<i>Appendix B: KLD analysis of unweighted k-mer frequencies</i>	<i>119</i>
<i>Appendix C: KLD analysis of weighted k-mer frequencies</i>	<i>122</i>
<i>Appendix D: KLD analysis of unweighted k-mer frequencies from mapped sequences</i>	<i>125</i>
<i>Appendix E: KLD analysis of weighted k-mer frequencies from mapped sequences</i>	<i>127</i>
<i>Appendix F: KLD analysis of signatures of unique k-mer frequencies.....</i>	<i>129</i>
<i>Appendix G: KLD analysis of signatures of unique k-mer frequencies from mapped sequences</i>	<i>132</i>

Table of Figures

Fig 1.1 The Skin Microbiome.....	5
Fig 1.2 Skin sites and their nature.....	7
Fig 1.3 Factors that influence skin microbiome.....	9
Fig 1.4 Gut microbiome location in a human body	11
Fig 1.5 Oral microbiome location in a human body [35]	12
Fig 1.6 Structure of Nucleotide [52]	17
Fig 1.7 Structure of RNA vs structure of DNA [53].....	17
Fig 1.8 Ribosome in Prokaryotes.....	18
Fig 2.1 Structure of 16S rRNA and hypervariable regions [59]	24
Fig 2.2 Hypervariable regions in 16S rRNA [63].....	26
Fig 2.3 Phylum levels of classification	26
Fig 2.4 Illumina MiSeq Sequencer [66].....	28
Fig 2.5 Illumina DNA Sequencing [68].....	30
Fig 2.6 How data is stored and analyzed in BaseSpace [69]	31
Fig 2.7 FASTq file format	32
Fig 2.8 Screenshot of taxonomy assignment text file from QIIME.....	34
Fig 2.9 Illustration of Principal Coordinate Analysis	36
Fig 2.10 Nucleotide bond structure in 16S rRNA V3 region	40
Fig 2.11 Mapping of 65nt of 16S rRNA into 47 elements.....	40
Fig 2.12 Mutually exclusive clusters in k-means clustering.....	42
Fig 3.1 Population group of the participants in the hand bacterial sample collection	48
Fig 3.2 Population group and gender of the participants in the hand bacterial sample collection	49
Fig 3.3 Age groups of the participants in the hand bacterial sample collection	49
Fig 3.4 Left-hand and right-hand sample count of the participants in the hand bacterial sample collection.....	50
Fig 3.5 Work flow of 16S rRNA gene extraction from skin bacteria before DNA sequencing ...	51
Fig 4.1 Percentage of samples belonging to each population group.....	55
Fig 4.2 Number of left and right-hand samples in every population group.....	55
Fig 4.3 K-mer (k=1) frequencies in genus level OTU Staphylococcus from 4 different samples	56
Fig 4.4 K-mer (k=2) frequencies in genus level OTU Staphylococcus from 4 different samples	56
Fig 4.5 K-mer (k=3) frequencies in genus level OTU Staphylococcus from 4 different samples	57
Fig 4.6 PCoA plot from KLD analysis of OTU frequencies of 16 samples from 4 population groups.....	58
Fig 4.7 Reference Phylogenetic tree depending on geographical distance between population groups.....	59
Fig 4.8 Phylogenetic tree based on KLD analysis of OTU frequencies	60
Fig 4.9 PCoA plot from KLD analysis of unweighted k-mer (k=1) frequencies for 16 samples from 4 population groups.....	61
Fig 4.10 PCoA plot from KLD analysis of unweighted k-mer (k=2) frequencies for 16 samples from 4 population groups.....	62
Fig 4.11 PCoA plot from KLD analysis of unweighted k-mer (k=3) frequencies for 16 samples from 4 population groups.....	62
Fig 4.12 Phylogenetic tree based on KLD analysis of unweighted k-mer (k=1) frequencies	63
Fig 4.13 Phylogenetic tree based on KLD analysis of unweighted k-mer (k=2) frequencies	64
Fig 4.14 Phylogenetic tree based on KLD analysis of unweighted k-mer (k=3) frequencies	65

Fig 4.15 PCoA plot from KLD analysis of weighted k-mer (k=1) frequencies of 16 samples from 4 population groups.....	66
Fig 4.16 PCoA plot from KLD analysis of weighted k-mer (k=2) frequencies for 16 samples from 4 population groups.....	67
Fig 4.17 PCoA plot from KLD analysis of weighted k-mer (k=3) frequencies for 16 samples from 4 population groups.....	67
Fig 4.18 Phylogenetic tree based on KLD analysis of weighted k-mer (k-1) frequencies	68
Fig 4.19 Phylogenetic tree based on KLD analysis of weighted k-mer (k-2) frequencies	69
Fig 4.20 Phylogenetic tree based on KLD analysis of weighted k-mer (k-3) frequencies	70
Fig 4.21 PCoA plot from KLD analysis of unweighted k-mer (k-1 on the left; k-2 on the right) frequencies of 16 samples from 4 population groups	71
Fig 4.22 Phylogenetic tree based on KLD analysis of unweighted k-mer (k-1) frequencies from mapped sequences.....	72
Fig 4.23 Phylogenetic tree based on KLD analysis of unweighted k-mer (k-2) frequencies from mapped sequences.....	73
Fig 4.24 PCoA plot from KLD analysis of weighted k-mer (k-1 on the left; k-2 on the right) frequencies of 16 samples from 4 population groups	74
Fig 4.25 Phylogenetic tree based on KLD analysis of weighted k-mer (k-1) frequencies from mapped sequences.....	75
Fig 4.26 Phylogenetic tree based on KLD analysis of weighted k-mer (k-2) frequencies from mapped sequences.....	76
Fig 4.27 Average Silhouette values and number of negative silhouette values for different number of clusters for k-1	78
Fig 4.28 Average Silhouette values and number of negative silhouette values for different number of clusters for k-2.....	79
Fig 4.29 Average Silhouette values and number of negative silhouette values for different number of clusters for k-3.....	79
Fig 4.30 Average Silhouette values and number of negative silhouette values for different number of clusters for k-1 from mapped sequences	80
Fig 4.31 Average Silhouette values and number of negative silhouette values for different number of clusters for k-2 from mapped sequences	80
Fig 4.32 Phylogenetic tree based on KLD analysis on unique signatures of k-mer (k-1) frequencies	82
Fig 4.33 Phylogenetic tree based on KLD analysis on unique signatures of k-mer (k-2) frequencies	83
Fig 4.34 Phylogenetic tree based on KLD analysis on unique signatures of k-mer (k-3) frequencies	84
Fig 4.35 PCoA plot from KLD analysis of signatures from unique k-mer (k-1 on the left; k-2 on the right) frequencies from mapped sequences of 16 samples from 4 population groups.....	85
Fig 4.36 Phylogenetic tree based on KLD analysis on unique signatures of k-mer (k-1) frequencies from mapped sequences	86
Fig 4.37 Phylogenetic tree based on KLD analysis on unique signatures of k-mer (k-2) frequencies from mapped sequences	87
Fig 4.38 Confusion matrix from classifying samples using OTU frequencies.....	92
Fig 4.39 Confusion matrix from classifying samples using signature frequencies from unique k-mer (k-1) frequencies.....	94

Fig 4.40 Confusion matrix from classifying samples using signature frequencies from unique k-mer (k=2) frequencies	95
Fig 4.41 Confusion matrix from classifying samples using signature frequencies from unique k-mer (k=3) frequencies	97
Fig 4.42 Confusion matrix from classifying samples using signature frequencies from unique k-mer (k=1) frequencies from mapped sequences	98
Fig 4.43 Confusion matrix from classifying samples using signature frequencies from unique k-mer (k=2) frequencies from mapped sequences	100
Fig 4.44 Confusion matrix from application of relatively better performing signature k-mer frequencies for each population group.....	101
Fig 5.1 k-mer (k=2) frequencies in genus level OTU Staphylococcus from 4 different samples	105
Fig 5.2 Nucleotide bond structure in 16S rRNA V3 region	105
Fig 5.3 Mapping of 65nt of 16S rRNA into 47 elements.....	105
Fig A.1 PCoA plot from KLD analysis of OTU frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups.....	118
Fig B.1 PCoA plot from KLD analysis of unweighted k-mer (k=1) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	119
Fig B.2 PCoA plot from KLD analysis of unweighted k-mer (k=2) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	120
Fig B.3 PCoA plot from KLD analysis of unweighted k-mer (k=3) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	121
Fig C.1 PCoA plot from KLD analysis of weighted k-mer (k=1) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	122
Fig C.2 PCoA plot from KLD analysis of weighted k-mer (k=2) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	123
Fig C.3 PCoA plot from KLD analysis of weighted k-mer (k=3) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	124
Fig D.1 PCoA plot from KLD analysis of unweighted k-mer (k=1) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	125

Fig D.2 PCoA plot from KLD analysis of unweighted k-mer (k=2) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	126
Fig E.1 PCoA plot from KLD analysis of weighted k-mer (k=1) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	127
Fig E.2 PCoA plot from KLD analysis of weighted k-mer (k=2) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	128
Fig F.1 PCoA plot from KLD analysis of signatures from unique k-mer (k=1) frequencies from of all 69 samples (top left), 39 samples (top right), 26 samples (middle left), 22 samples (middle right) and 16 samples (bottom left).....	129
Fig F.2 PCoA plot from KLD analysis of signatures from unique k-mer (k=2) frequencies from of all 69 samples (top left), 39 samples (top right), 26 samples (middle left), 22 samples (middle right) and 16 samples (bottom left).....	130
Fig F.3 PCoA plot from KLD analysis of signatures from unique k-mer (k=3) frequencies from of all 69 samples (top left), 39 samples (top right), 26 samples (middle left), 22 samples (middle right) and 16 samples (bottom left).....	131
Fig G.1 PCoA plot from KLD analysis of signatures of unique k-mer (k=1) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	132
Fig G.2 PCoA plot from KLD analysis of signatures of k-mer (k=2) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups	133

List of Tables

Table 2.1 Example of Confusion matrix with 3 classes.....	44
Table 3.1 primers used the Amplification of 16S rRNA and V3 hypervariable region	51
Table 3.2 List of Population groups and their abbreviations	52
Table 4.1 Number of unique k-mer frequencies found in different cases of k	77
Table 4.2 Number of clusters (K) chosen for different cases of k.....	78
Table 4.3 Number of iterations to classify African with rest of the population groups	88
Table 4.4 Number of iterations to classify Turkish with rest of the population groups	89
Table 4.5 Number of iterations to classify Chinese with rest of the population groups	89
Table 4.6 Number of iterations to classify Hispanic with rest of the population groups	90
Table 4.7 Number of iterations to classify Middle Eastern with rest of the population groups ...	90
Table 4.8 Number of iterations to classify Caucasian with rest of the population groups	91
Table 4.9 Number of iterations to classify Asian with rest of the population groups	91
Table 4.10 Number of iterations to classify Asian Indian with African American	91
Table 4.11 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using OTU Frequencies	93
Table 4.12 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using signatures of k-mer (k-1) frequencies	94
Table 4.13 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using signatures of k-mer (k-2) frequencies	96
Table 4.14 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using signatures of k-mer (k-3) frequencies	97
Table 4.15 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using signatures of k-mer (k-1) frequencies from mapped sequences	99
Table 4.16 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using signatures of k-mer (k-2) frequencies from mapped sequences	100
Table 4.17 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using better performing signatures of k-mer frequencies for each population group	102
Table 5.1 Dissimilarity of resultant phylogenetic tree for different applications of KLD	106
Table 5.2 Dissimilarity of resultant phylogenetic tree for different signatures of k-mer	107
Table 5.3 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups in different cases	108
Table 5.4 Highest accuracy rates of each population group and overall accuracy from k-mer signatures	109

Chapter 1 : Introduction

1.1 Executive Summary

Human skin is a complex ecosystem with diverse groups of both stable and variable bacteria [1]. Stable bacteria are abundant in an individual and are less distinct among individuals, whereas the varying bacteria that occur in smaller proportions in an individual are significantly distinct among people, making skin bacteria a promising tool in distinguishing humans [2]. Skin bacterial composition is found to be unique even in twins, though with high similarity because of shared genetics and environment, making skin bacteria-based identification more promising [3]. High interindividual variability in human skin bacteria composition lead to skin bacteria being a potential supplement in forensic identification [4]. Furthermore, skin bacterial communities are stable in extreme environmental conditions, offering a better target in forensic investigations than human DNA, which is susceptible to extreme environments [5][6]. In addition to distinction among individuals, bacterial groups also vary across different body sites within an individual, owing to specific characteristics of the skin site [7].

A detailed study of the microbiome of various parts of the human body like skin, oral cavity, gastrointestinal tract, urogenital parts, blood, eye and airway parts was initiated with the Human Microbiome project in 2007, as an extension of the Human Genome project [8]. Researchers later conducted several experiments to learn the characteristics of skin and other microbiome including the features that affect them. Bacterial communities identified on human skin are observed to vary among individuals with environment, ethnicity, lifestyle, diet, age, gender, medication, birth process (natural or C-section) and personal hygiene habits [1][9][10][11][12][13][14]. Phylogenetics of bacterial groups is carried out by sequencing and classifying the variable regions of 16S rRNA gene (ribosomal Ribonucleic Acid) [15][16][17]. Species-level clustering and comparing of 16S rRNA gene sequences from 10 healthy adults over

20 different skin sites revealed that bacterial groups within an individual are more similar among the sites with similar physiological features [7]. A Similar study of the variable region V2 of 16S rRNA gene sequences from 27 body sites of 7-9 healthy adults over 4 occasions supports the inference that skin microbiome within an individual varies with the topography of skin [18].

Bacteria located on dry skin sites like hands (forearm, palms) are the most rich (highest number of distinct bacterial clusters) and evenly distributed i.e. size of clusters with 16S rRNA bacterial sequences are relatively equal [7][18]. Phylogenetic analysis of 16S rRNA genes collected from the palm surfaces of 51 young healthy adults finds that, on average, the hand surface harbors over 150 unique species level bacteria which is more than the unique bacterial types found on skin, or other human-associated microbial habitats like gut and mouth [9]. Bacterial communities identified on the hand skin of women, from two different populations, had considerable differences, which could be the result of biogeographical, genetic, cultural and behavioral dissimilarities [19]. Study of skin bacteria from 7 sites of 71 participants from different age groups, living conditions and genders verify that highest diversity of bacteria is found in samples from forearm and back of the hands [13]. PCoA plots of weighed UniFrac distances [20] between the same samples demonstrated clustering of samples from similar age group, gender, and living conditions [13]. Genus- level comparison between 200 skin samples from different ethnicities (Hong Kong, China, USA and Tanzania) reveals that skin bacteria is noticeably similar in individuals from the same ethnicity, with most diverse and relatively abundant bacteria in palm regions [21]. Taxonomic classification of variable region V3 of 16S rRNA gene sequences of hand bacterial samples from 9 different ethnicities (Caucasian, African, African American, Asian, Asian Indian, Hispanic, Turkish, Middle Eastern and Chinese) showed similar distribution of bacterial groups among people from the same ethnicity; moreover, PCoA plots on UniFrac weighed and

unweighted distances exhibited clustering of samples from similar descent, which indicates hand bacteria can be a potential biometric identifier [22].

Variations in hand bacterial 16S genome are largely studied by measuring the diversity and dissimilarity indices between species or genus-level operational taxonomic units (OTU) that are identified and classified using publicly available 16S reference genomes [9-14,18-21]. The main objective of this thesis is to develop a method to differentiate human populations based on the bacteria composition found on their palm regions. This will be accomplished by calculating the Kullback-Leibler Divergence between the frequencies of sequence clusters found in the variable region V3 of 16S rRNA gene extracted from hand bacterial samples of different individuals. Moreover, unclassified sequences that are ignored when studying OTUs are included here in this project by developing a classification method that considers the actual DNA symbols in terms of A, C, G and T (Adenine, Cytosine, Guanine and Thymine) profile of sequences to cluster sequences. Additionally, information from mapping actual RNA sequences to the structure of the 16S rRNA gene is considered to provide higher accuracy in determining the population group.

1.2 Skin Microbiome

Human skin is the largest organ of the human body, covering and protecting the internal organs and receiving sensory stimuli from the environment. The color of human skin has been a major factor in recording and recognizing the identities of human communities for ages. Human skin, in addition to shielding human organs, also acts as a habitat for various microorganisms – bacteria, algae, fungi and mites shown in Fig 1.1. Although all microorganisms are often perceived as toxic, many of the microbiota on human skin are harmless and, in some cases, have vital function beneficial to humans [1]. The colonization of microorganisms on human skin widely differs with the factors like body location, internal host factors, and external environmental factors. Human

skin is a cultural medium because of the fact that it's composition is influenced by human genetics, diet, lifestyle and the area we live in [3][13].

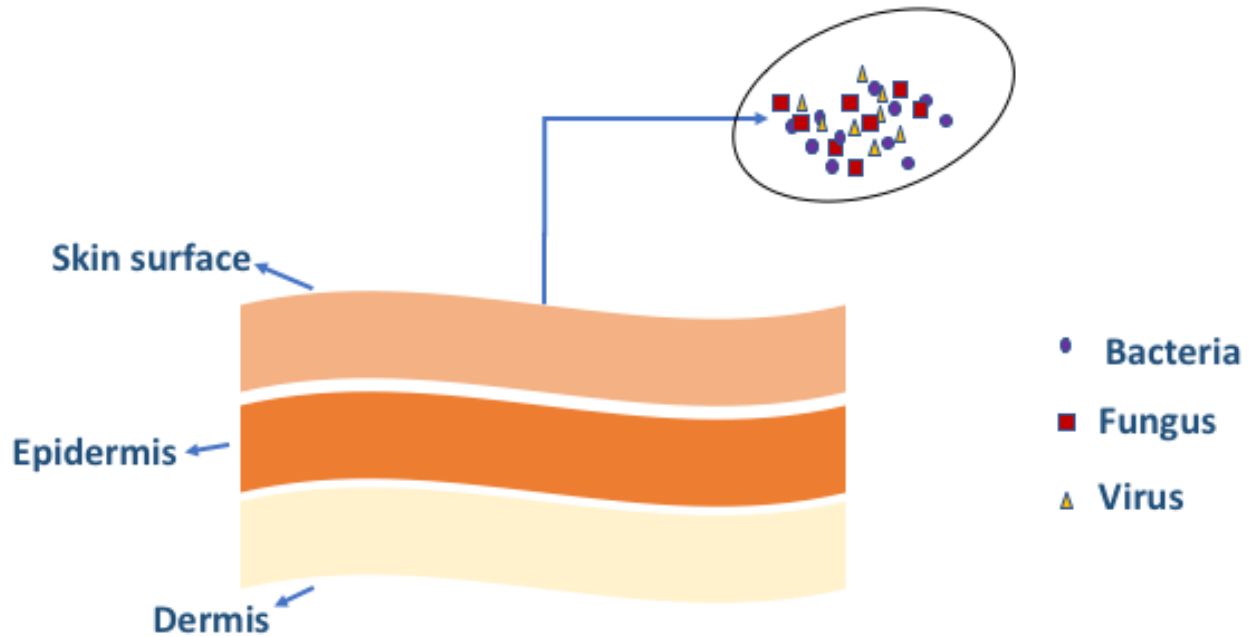


Fig 1.1 The Skin Microbiome

Skin in general is comprised of a fixed group of microorganisms, which are usually beneficial to the skin, and transient microorganisms, which arise from environment and last for hours to a lifetime [2][1]. The unique physiological and anatomical differences caused in an individual's skin, by hormone production, sweat rate, sebum production, surface pH, skin thickness, hair growth, frequent washing, overuse of antibiotics, and cosmetic use, influence skin microbiome, resulting in a significantly distinct skin microbiome [1][23]. Variation in the composition and characteristics of skin microbiome is observed to affect human skin condition and diseases [12]. In addition to host and external environmental factors, host-microbe and microbe-microbe interactions are revealed to have an important role in manipulating diversity patterns of skin microbiome [23].

1.3 Factors influencing skin microbiome

The following is a detailed discussion of the human factors that influence the composition of the skin microbiome.

Skin site: Human skin is classified into dry, moist, or sebaceous at various regions of the body. Different types of skin offer different types of environments for the existence and breeding of microorganisms, eventually causing variation among the bacterial communities over distinctive skin regions. Species-level analysis of 16S rRNA gene sequences over various skin sites shows that bacterial communities are more diverse in dry sites than in oily sites [7]. Dry skin sites, like forearms or palms, provide a better environment for the existence of bacteria, and thus, have more diverse bacterial communities when compared to sebaceous sites like upper back or skin behind the ear, which often exhibit less bacterial diversity [24].

After comparing the variable region V4 of 16S rRNA genes from 645 skin bacterial samples from three different sites of 110 men from 6 ethnic groups, it is observed that, body site is important in determining the bacterial communities, and bacterial communities are more diverse in dry skin sites like palms and forearm [25]. Hands have a more dynamic microbial community over time when compared with other skin sites. Bacterial phylotypes per individual are found the highest in palm skin compared to forearm or elbow skin [26]

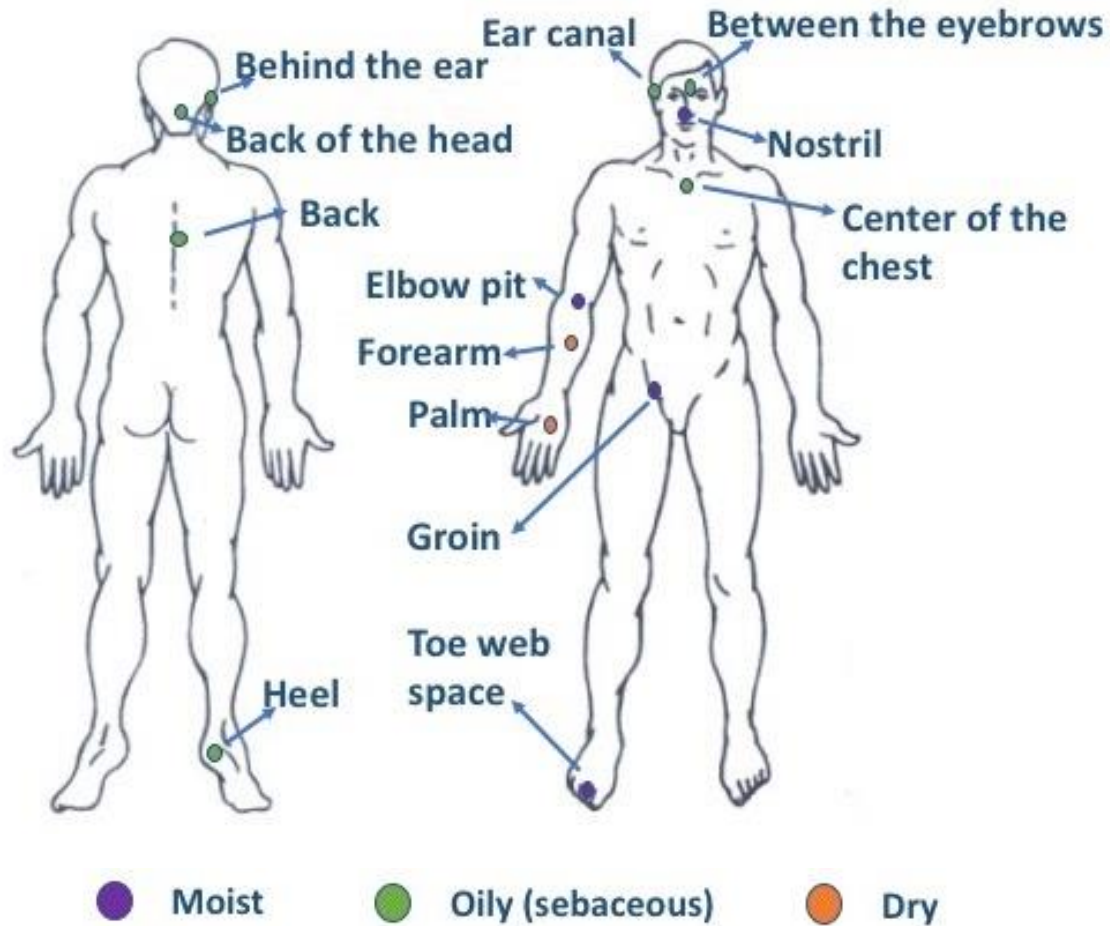


Fig 1.2 Skin sites and their nature

Ethnicity: People who can claim similar ancestry, language, culture or nation are considered an ethnic group. Humans from same ethnic group tend to have a similar gene pool. It is a known fact that the color of the skin varies over different ethnicities. In addition to the color of the skin, the bacterial composition of the skin at various sites is found to differ with ethnicity as well [21]. Hand skin microbiome, besides having the highest bacterial diversity, varies significantly over different populations [19]. Individuals of similar descent tend to have similar profiles of taxonomic groups on their palm regions [22].

Host Genetic factor: Human genetics not only affect physical appearance and individual traits, but also alter the skin microbial communities [3][1]. To get more insight into influence of genetics on skin microbiome, a study was conducted on 45 individuals, including monozygotic and dizygotic twins along with their mothers. Taxonomic classification of variable regions V2 and V3 of 16S rRNA gene from their skin microbial samples suggest that, samples with highest amount of shared genetics has most similar bacterial groups i.e., highest similarity is found in monozygotic samples, followed by dizygotic twins, mother-twin and unrelated samples [3].

Hygiene and Medication: Multiple hygiene products that are used on a daily basis on human skin have a noticeable effect on the diversity of skin microbial groups. Genus-level study and comparison of variable region V4 of 16S rDNA collected from the armpits of 7 individuals over a period of week suggest that, use of antiperspirants and deodorants result in more diverse bacterial composition on the armpit skin [14]. Antiperspirants are made of aluminum-based salts to reduce sweat by forming precipitates and, therefore, are believed to inhibit the growth of microbial communities causing higher density of bacteria, and rich bacterial species, unlike deodorants which are ethanol-based and are more water soluble and easily washed away [27][14]. Also, use of Antibiotics for dermatologic conditions influence the microbial composition on the skin [12]

Age Groups: Humans are born with a set of bacteria which evolves over time, with many new bacteria added, many of the existing bacteria diminished, or hybrid species formed from various combinations [24]. The microbial communities on skin are more diverse among different age groups than in the same age group, indicating that the composition of skin flora is more similar among same age group individuals, than across different age groups [16]. In a study of bacterial

samples from 190 volunteers, there was also another, rather surprising, observation that the children from a semi-nomadic population had more diverse bacterial community than the adults, suggesting a great deal of progression in bacterial community evolution from childhood [28]

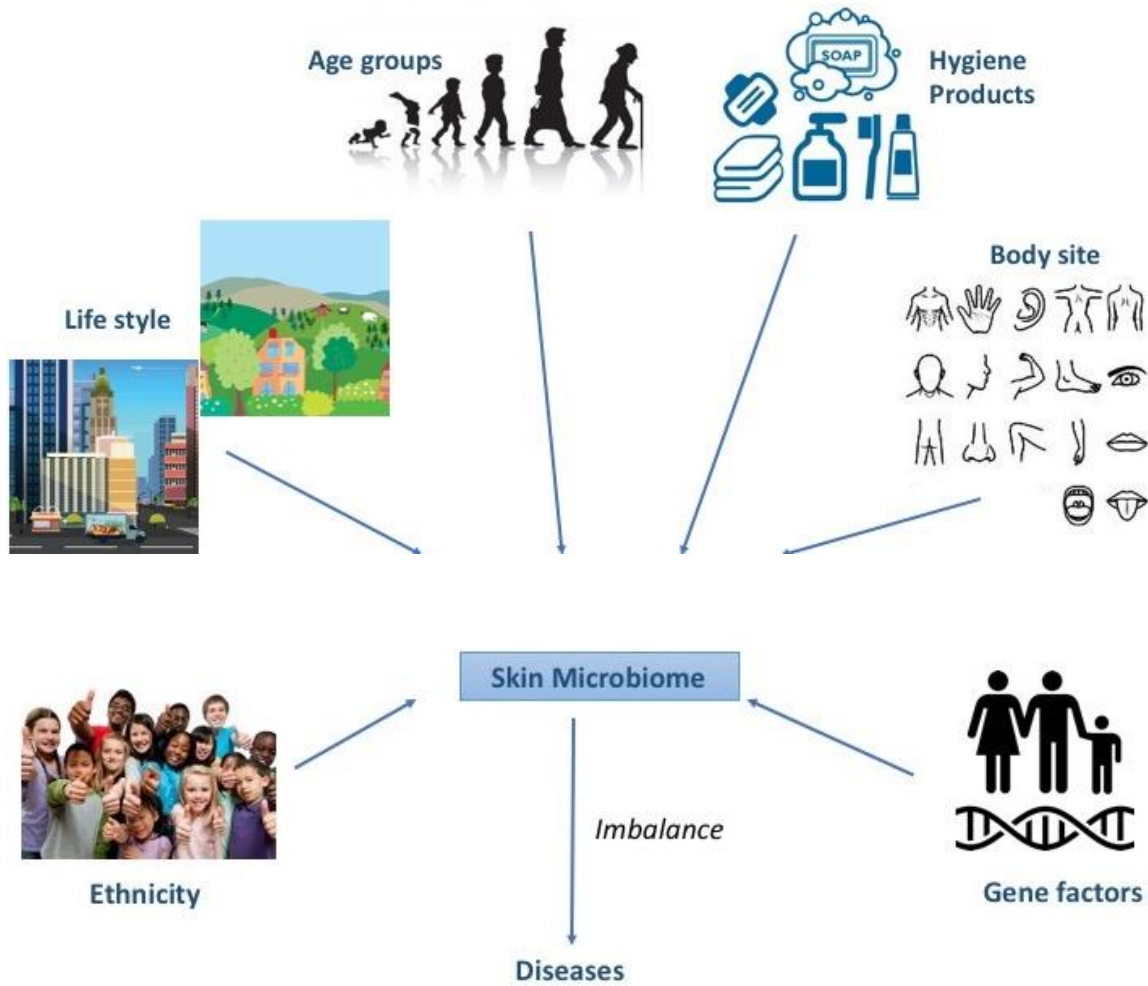


Fig 1.3 Factors that influence skin microbiome

Living Conditions: Living conditions are known to influence the health and well-being of humans and, therefore, also effect the skin microbiome of people. Urban and rural living conditions differ in various factors such as quality of food and water, pollution, lifestyle etc. Urban populations are composed of people who spend most of their time indoors and, therefore, most of their microbial

communities are human derived, and on the other hand, rural populations spend a significant amount of time outdoors, exposed to soil, dirt and the environment, resulting in a more diverse and rich microbiome on their skin [13]. Also, the intragroup variation among rural dwellers is higher than that found in urban populations [13]. Thus, comparison of skin-associated bacterial community structure, and composition, might help in deducing if the subject is from an urban or a rural environment.

1.4 Other significant microbiome of the human body

Gut Microbiome:

The human digestive tract, also known as the gut, is another region of the human body with a diverse microbiome. In the process of digestion, the gut is exposed to various microorganisms from food, drinks, and everything entering the body through the digestive track. The factors that affect the skin microbiome tend to affect the gut microbiome as well, in a similar way. The composition and the interactions between the microbial communities of the human gut vary over geographical locations and across age groups of individuals [29]. A study of gut microbiota obtained from fecal samples of 314 healthy young adults from 7 ethnicities throughout China indicates that similarities in gut microbiota exist more in samples from the same geographical/ethnic groups than in samples with similar lifestyles [30]. The composition of gut microbiota is relatively more stable in adults than in children. Microbial communities of older population from urban towns were more similar to the microbiome of children from urban towns than to the microbial communities observed in older populations from villages known for having higher life expectancy, indicating that the environment is a stronger determinant in diversity of gut microbiota than age [31].

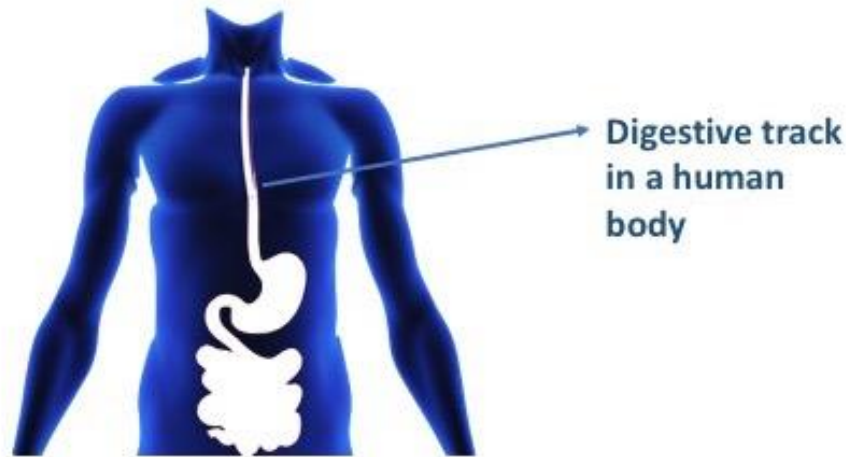


Fig 1.4 Gut microbiome location in a human body

Personal factors like gender, body mass index (BMI) and dietary habits have a significant effect on the gut microbiome [32]. Statistical analysis of taxonomic assignment to 16S rRNA genes of gut bacteria collected from 82 humans showed that a certain gut bacterial species was lesser in women than men and the association between BMI and overall gut bacteria was stronger in women than in men [32]. Fiber from different variety of foods such as beans, fruits and vegetables was associated with abundance of a certain bacterial species respectively and a better study and understanding of these relationships may lead to significant inferences for gastrointestinal health and disease prevention [32]. Composition of gut bacteria in obese and lean people are different at genus, species and phylum levels, supporting the idea of studying gut microbiome for the etiology of various human diseases [33].

Using an unconventional study design, gut bacteria from fecal samples, sewage of 71 different cities of the US was examined to see if they lend insight into the gut microbial community diversity between different ethnicities. Distribution patterns among municipal sewage communities reflected variation in the ethnicities, and the samples represented lean or obese populations with 81 to 89% accuracy, affirming that microbes found in sewage can be an indicator

of the fecal microbial communities in human populations, and thus, the traits of the human gut microbiome in different populations [34].

Oral Microbiome:

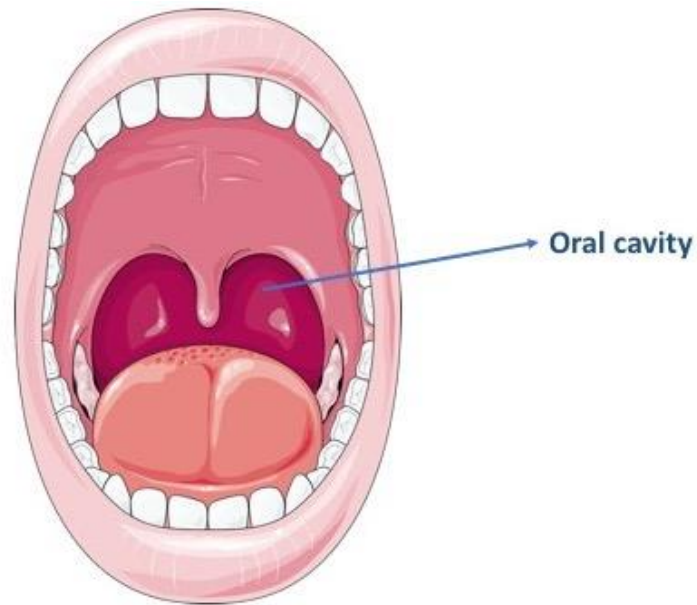


Fig 1.5 Oral microbiome location in a human body [35]

As the name suggests, the oral microbiome includes the microbial communities in the oral cavity, which is commonly called mouth. The oral microbiome is the second most diverse microbiome in the human body, highly specific at the species level [36], sheltering over 700 species of bacteria [37]. An individuals' diet, lifestyle and ethnicity play a vital role in the variability of oral bacteria among individuals [37][38]. Other habits like cigarette smoking can also influence the composition of oral microbiome [39]. Oral bacterial samples from Alaskans, Germans, and Africans reveal that the degree of diversity is significantly high, with relatively higher similarities between Alaskans and Germans, confirming the effect of ethnicity on the

composition of gut microbiome [38]. Oral microbiome plays a critical role in maintaining physiological, metabolic and immunological functions of the body [40][36].

1.5 Advantages of skin microbiome and its applications

Skin has always been related to beauty but, in fact, it plays a vital role in overall human health. Skin acts as a home to trillions of microorganisms which can be classified into diverse species of bacteria and fungi. Scientists now consider the skin microbiome as an important organ of our body that, when properly managed, contributes to our health and well-being[12]. The skin microbiome is found to relate to many fundamental health conditions like weight, mental health, immunity, diabetes, blood pressure, heart issues, and even cancer. Study of skin microbiome could greatly contribute to diagnosis and treatment of dermatological issues. The deep relationship between host and the skin microbiome contributes to its distinctive composition, potentially leading to human identification applications in the fields of biometric identification and forensic investigations.

Since its first use in 1986, human DNA fingerprinting has become widely used in forensic and criminal investigations. But, very often, criminal investigations are delayed due to lack of priority or too many cases to consider, causing human DNA evidence to become ineffective and eventually hindering the quality of investigation [41]. Even after solving hundreds of thousands of cases, there are still many more cases that couldn't be solved because of contamination or destruction of evidence, requiring the need for a better line of evidence to track the culprit [42]. One such line of evidence that came into light recently is the study of microorganisms in and on human body. The skin microbiome has shown uniqueness, especially at sub species level, making it a potential marker for human identification [42].

When samples collected over months and years were studied, it was observed that, in spite of constant exposure to a changing environment, the skin microbiome is relatively stable over time, making it a prospective target in forensic studies [6]. Humans leave a trace of millions of bacteria everywhere they go, or on things they touch, which makes the microbial sample collection moderately easier than looking for DNA samples at the crime scene [42]. Skin microbiome can be found on various surfaces, for example, keyboards, elevator buttons, telephones and shoes, and even from extreme conditions of -80°C , without much damage to the bacterial DNA [4][43][44][45]. Bacterial communities found on a fabric, or any surface in a crime scene can be used to compare to the skin microbiome or microbiome found on any personal belongings, and narrow down suspects in an investigation [46][43][47].

The use of hand bacterial samples in forensic investigation was tested and it was found that, the similarity index is higher among the samples from same individual over time, and also the clusters from different individuals could be distinguished even if they were collected under different conditions [48]. Bacterial communities found on the fingertips of individuals and the keys of their personal computer keyboards were more similar to each other, than they were with other keyboard keys or individuals' fingertips, indicating that, we can match an object to its owner using skin bacteria [4]. Analysis of hand bacteria is relatively faster, and helps in saving time and labor, by narrowing the number of suspects, thus increasing the efficiency of a criminal investigation [48]. Postmortem skin microbiome can also be used in forensic death investigations, to analyze the amount of time after death [49].

1.6 History of Human Microbiome study

To study human and non-human cells that exist within and upon the human body, the National Institute of Health (NIH) launched the Human Microbiome Project (HMP) in 2008 and

studied bacterial samples from multiple body sites of 250 individuals. The primary body sites included in the project are oral cavity, skin, gut, vagina and nasal/lung. The HMP took advantage of high-throughput gene sequencing technologies for a detailed study of the human microbiome, examining multiple factors of the human microbiome to associate its composition and variation with population, genotype, gender, disease, age, nutrition, medication, and environment. Goals of HMP included developing a reference set of sequences for microbial genomes, to explore the changes occurring in the microbiome with various diseases and vice-versa and to develop new technology and tools for the computational analysis over various microbial sequences [8]

Achievements of the Human Microbiome Project included:

- 10000 more species were discovered to live in human ecosystem and a new database was developed with ~99% of its genera identified
- The data acquired from the Human microbiome project led to numerous clinical researches, revelations and applications
- Pharmaceutical microbiologists were able to use the derived implications of various microorganisms from HMP data to enhance the production of pharmaceutical products

The human microbiome, an important feature of human physiology is affected in various health conditions becoming a supplementary data in the study of various diseases and their effects on the human body [50]. Apparent differences established in microbial communities in and on the human body of different individuals gives insight into how different and diverse they become over time. Other significant observations that were made in the Human Microbiome Project are that human survival is more strongly linked to microbial genes rather than human genes, and that bacterial protein coding genes are about 360 times more abundant than human

genes [51]. Though the human microbiome changes over time with disease and medication, it eventually arrives back at a baseline state, even with any change in the type of bacterial composition.

1.7 Prokaryotic 16S rRNA

RNA is a macromolecule that plays an essential role in various biological processes. RNA (Ribonucleic Acid) is a chain of nucleotides, but as a single-strand folded onto itself, rather than a paired double-strand (Fig 1.7). Nucleotides (Fig 1.6) are made of 5-carbon sugar molecule, nitrogenous base and a nucleobase. Adenine (A), Cytosine (C), Guanine (G), Thymine(T) and Uracil (U) are the five primary nucleobases that are fundamental units of the genetic code: A, C, G and U are found in RNA while A, C, G and T are found in DNA. In microbial ecology studies, scientists compare the bits of rRNA(Ribosomal Ribonucleic Acid) to the previously known reference microbes to classify the microbes or identify a new microbe. Ribosomes in all living beings act as the gene-translating machines. A gene from a piece of DNA is copied into a strand of messenger RNA (mRNA) and delivered from the cell nucleus into the cytosol where the ribosomes latch onto this mRNA and move along the mRNA strand, reading the code contained in its sequence of nucleotide bases (A, C, G & U) and stringing the right amino acids together based on the code to build protein chains. The slight changes in the genes of ribosomal RNA over the years provide clues as to how closely or distantly various organisms are related.

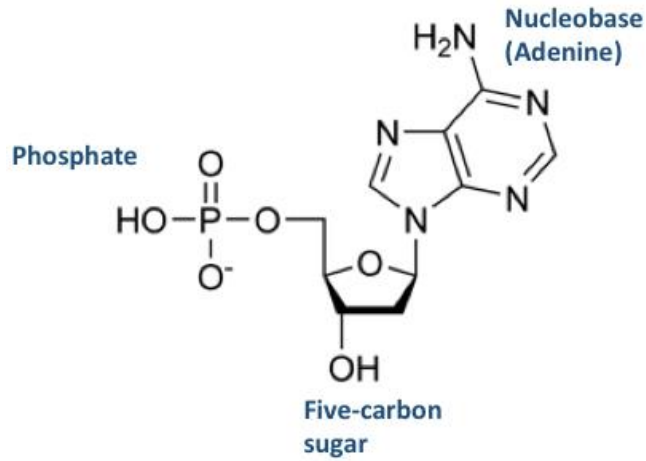


Fig 1.6 Structure of Nucleotide [52]

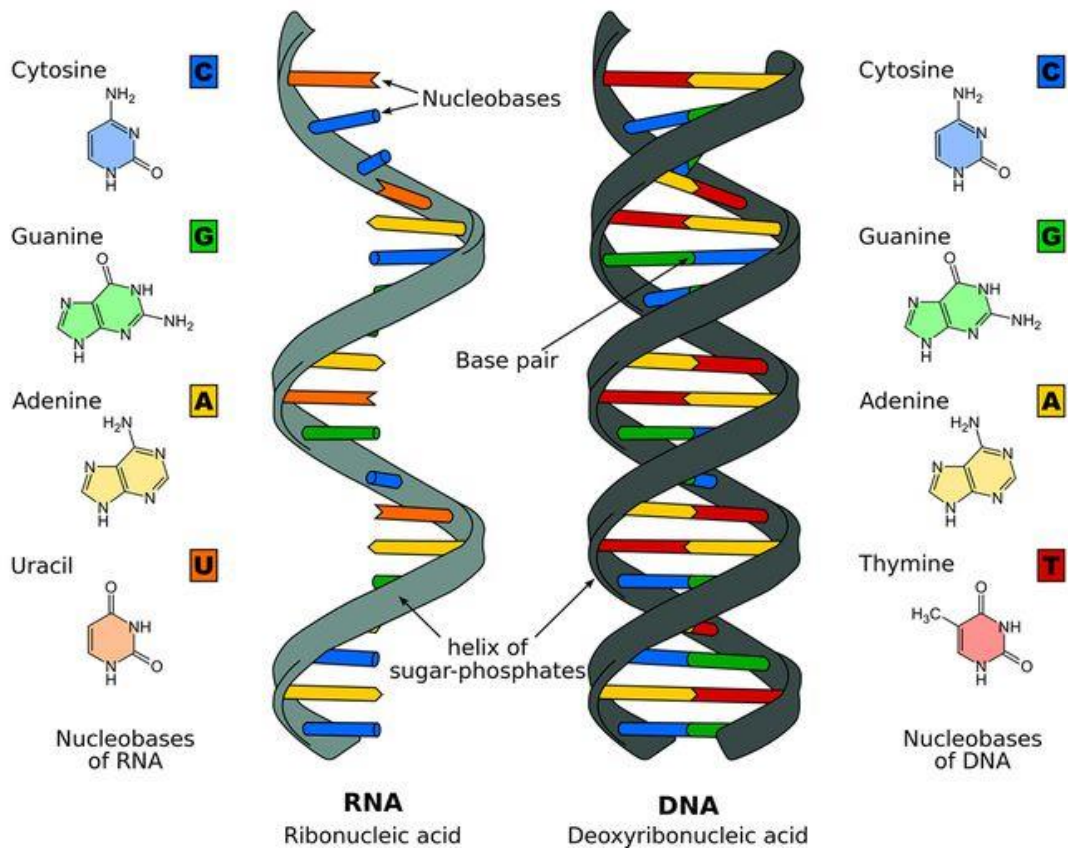


Fig 1.7 Structure of RNA vs structure of DNA [53]

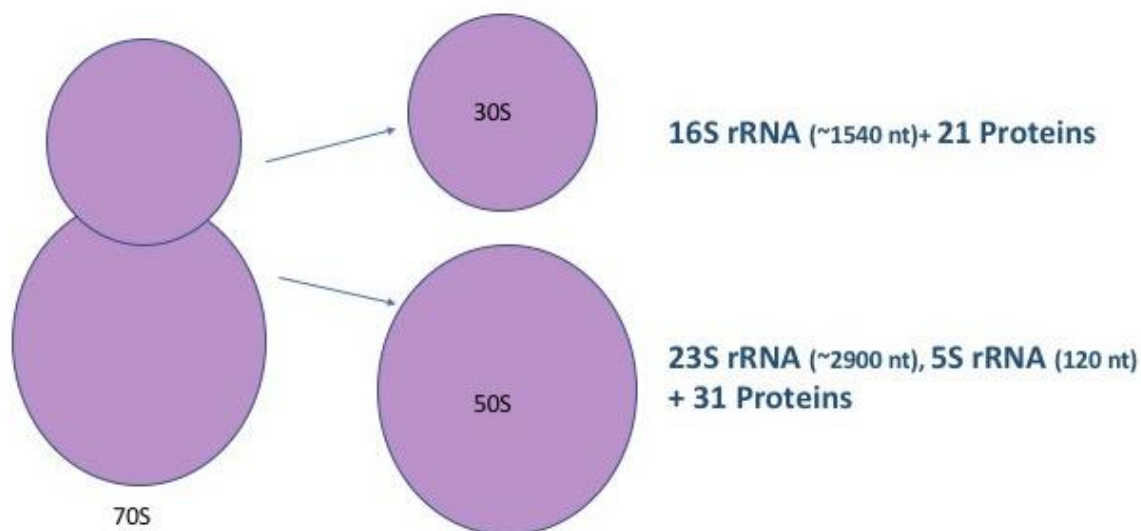


Fig 1.8 Ribosome in Prokaryotes

16S rRNA is a part of 30S (S for Svedberg unit) small subunit of 70S ribosomes (Fig 1.8) in prokaryotes. 16S rRNA gene is short with just 1,542 nucleotide bases making it easy and cheap to copy and sequence. When a sample is collected, it is cleaned, purified and the needed rRNA is pulled out from other RNA, DNA and extra unwanted fragments. Though 16S rRNA genes from different microbes have a few different nucleotides spread through the sequences, nucleotides at the very beginning or end of the gene are similar from organism to organism. Scientists use numerous copies of another bit of RNA called a primer, which is a mirror image of a short bit of RNA or single strand of DNA; that is, its sequence of nucleotides is the direct complement to the sequence of nucleotides in a known part of the target RNA or DNA.

In this research, the primer is the mirror image of the beginning or end of the 16S rRNA sequence. Since complementary nucleotides pair up into a bond, the primer enables the scientist to pull out the 16S rRNA in the sample. Then using Polymerase Chain Reaction (PCR), millions of copies of these genes are made to have enough 16S rRNA for its comparison to the libraries of

stored 16S rRNA genes from numerous known bacteria. The sequences which were classified into numerous genera were used to distinguish the samples over different population groups.

1.8 K-mers in DNA analysis:

Counting of k-mers in DNA sequence data has been an efficient way in bioinformatics to correct errors in sequences reads [54]. In an effort to minimize the memory issues that arise while storing k-mer counts of a large data set, a probabilistic data structure called bloom filter was designed to store all the observed k-mers with reduced memory requirements [55]. A similar filter was designed in a cache-efficient technique to reduce the experimental runtime[56]. K-mer analysis when aided with positional resolution showed correlation between k-mer frequencies and several genes, demonstrating similarities between classified and unclassified viruses, which may be of significant use in future taxonomic research [57]. Analysis of k-mer spectrum resulted in significant dissimilarity between the human gut metagenomes of different populations [58]. Moreover, dissimilarity measure based on k-mer analysis yields a better perspective than the techniques based on alignment against reference sequence sets[58].

1.9 Problem Statement

The main objective of this research is to study the variation of skin bacterial communities on the palm region of people for population group classification. For this particular research, 69 hand-swab samples collected from 39 people of 9 different population groups were used to analyze the third hypervariable region (V3) of the bacterial 16S ribosomal RNA (rRNA) gene. Using an Illumina MiSeq sequencer, PCR amplified 65nt long V3 region of 16S rRNA is sequenced and the data is stored in FASTQ files. The sequenced samples were then classified into various taxonomy levels using RDP (Ribosomal Database Project) classifier. Frequencies of nucleotides (A-Adenine,

C-Cytosine, G-Guanine and T-Thymine) and their combinations (AA, AC, AG, AT...TT, AAA, AAC, AAG, TTT) in sequences classified as the same OTU had similar profile charts. Therefore, the frequencies of A, C, G and T, AA, AC, AG, AT, ... TT and AAA, AAC, AAG, ... TTT are considered as k-mer signatures (k=1, k=2 and k=3 respectively) for all classified OTUs in a sample. Kullback-Leibler Divergence (KLD) between the frequencies of k-mers for Genus-level OTUs classified from all the samples is considered as the measure of dissimilarity between the samples. In an attempt to consider the unclassified sequences which are ignored while using known OTUs to measure the dissimilarity between samples, a novel classification method is developed to cluster the sequences according to their k-mer profiles. In addition to comparing the k-mer distribution of the V3 region of bacterial 16S rRNA genes, the structure of the V3 region and the bonds involved in building the structure were considered, to map the 65 nucleotide positions into 47 new elements. These new redistributed 47 elements of the V3 region are then considered as designated sequences and clustered into mutually exclusive groups. K-mer signatures (k=1 and 2) are assigned to each cluster and the KLD between distribution of these clusters is used to calculate the distance between different samples. Resultant distances between samples are used to build phylogenetic trees and Principle Coordinate Analysis (PCOA) plots, to study the grouping of samples from different population groups

1.10 Conclusions and Future work

The set goal of classifying population groups by studying the hypervariable region V3 of 16S rRNA genome of hand bacterial samples was achieved with 79.5 % accuracy. The highpoint of this thesis was, the development of a novel method to include unclassified sequences that are generally ignored in other OTU methods, through a novel k-mer classification model. Among 9 different population groups from all over the world, 6 population groups namely African, Turkish,

Chinese, Hispanic, Middle Eastern and Asian achieved accuracy greater than 75%, with African and Turkish achieving more than 90% accuracy. k-mer frequencies from mapped sequences resulted in comparatively better performance than the conventional OTU method, encouraging the idea of k-mer usage.

Nevertheless, hand bacterial samples that were included in this study does not specify other factors that could influence skin microbiome, for example, there is no record of when the individual last washed their hands before sample collection, or if the individual has lived in a country different from that of the origin of their population group. Also, among 9 hypervariable regions, only a single region was considered for the extraction and analysis of nucleotide sequences. Therefore, there is a scope for improving the methodology by considering other hypervariable regions of 16S rRNA. Considering multiple hypervariable regions could be beneficial while studying the structure of 16S rRNA and would allow using longer k-mers for classification, which was limited to k-3 in this research since we have only 65 nucleotides in the V3 region. Extending the k-mer study to other parts of 70S ribosome i.e. 23S RNA could offer better understanding of bacteria communities. Furthermore, increasing the number of samples from different population groups would improve the performance of the classification model by studying more and better patterns.

1.11 Thesis Organization

Following this introduction, this thesis is distributed into four chapters describing the various technologies and applications adopted in the study of bacterial 16S rRNA. Chapter 2 explains why the V3 hypervariable region is chosen for this project. Various technologies and platforms used, in culturing and studying the samples are also described in this chapter. Chapter 3 gives a detailed description of the data collection and the phases involved in preparing the data

for the bioinformatics analysis. Chapter 4 gives an account of the new classification technique applied in this project, and the observed performance of the technique. Chapter 5 summarizes the thesis and concludes with future work.

Chapter 2 : Theory

This chapter gives an insight into 16S rRNA genomics and other methods and platforms used in this thesis.

2.1 16S rRNA genome

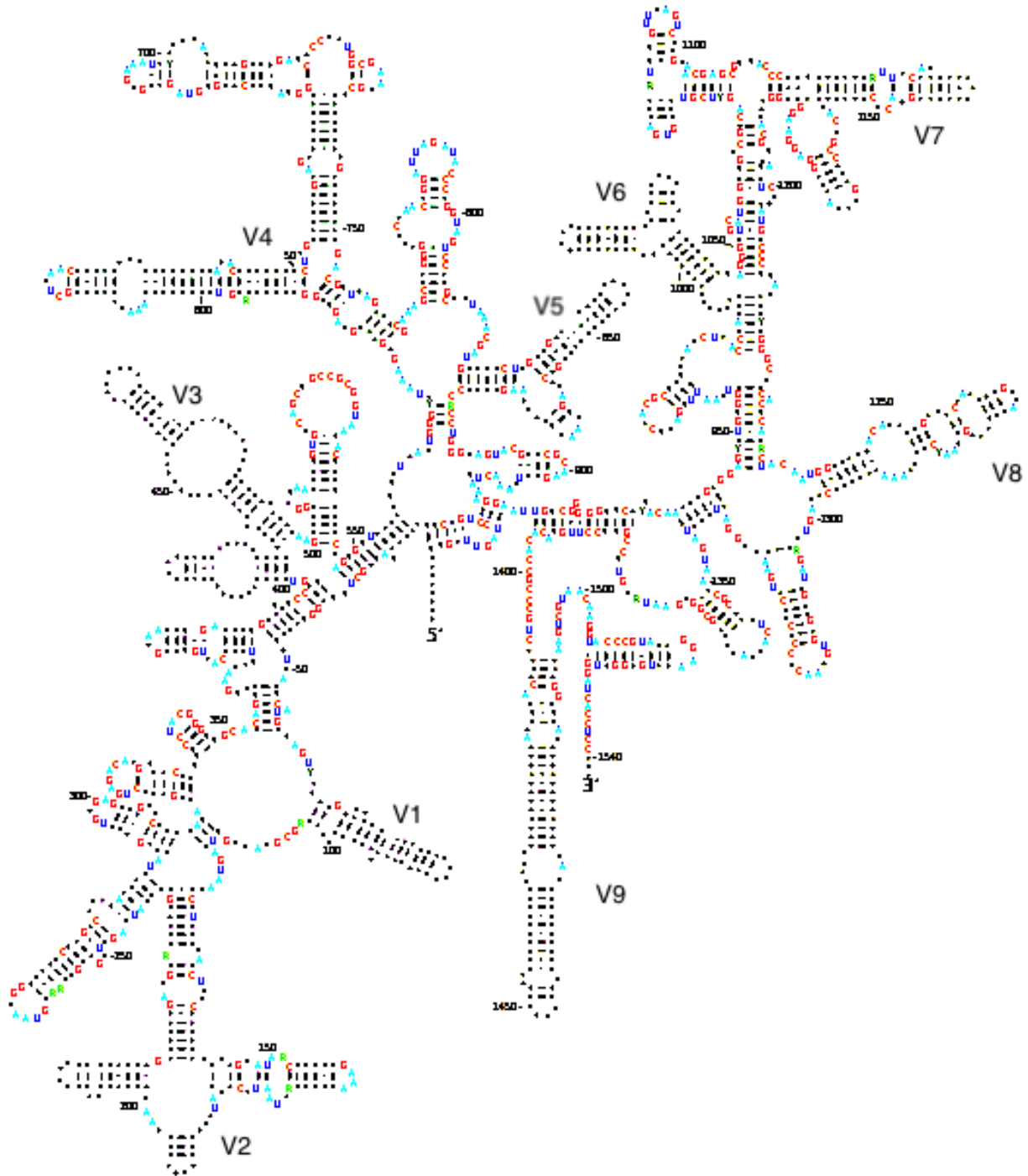


Fig 2.1 Structure of 16S rRNA and hypervariable regions [59]

Bacterial ribosome also called as 70S ribosome is composed of two subunits namely 50S large subunit and 30S small subunit. 16S RNA is a part of 30S (S for Svedberg unit) small subunit, while 23S RNA and 5S RNA are a part of 50S large subunit. Sequence analysis of 16S rRNA or 23S rRNA assists in understanding the phylogenies of prokaryotic bacteria. 16S rRNA, shown in Fig 2.1 is comparatively short with only 1,542 nucleotides and is easy and cheap to sequence; therefore, is chosen to study the hand bacterial samples in this thesis.

2.2 Hyper Variable region V3

While most of the bacterial 16S rRNA is conserved, the regions where sequences exhibit significant diversity among different bacteria is divided into 9 hypervariable regions (V1-V9, see Fig 2.2) and are used for taxonomic classification of bacteria. Sequences in a hypervariable region are specific to species, offering useful targets for numerous scientific investigations and diagnostic tests. A single region cannot distinguish among all bacteria, because of different degrees of sequence diversity; therefore, hypervariable regions are compared and combined depending on the relative advantage of each region for specific goals [60]. Comparing hypervariable regions of 16S rRNA gene sequences assists in distinguishing various organisms at genus level across all major phyla of bacteria. Classification is being done further than the genus level, what we now call the species and subspecies level. Species that are identified and clustered together at various taxonomic levels shown in Fig 2.3 based on sequence similarity are referred to as Operational Taxonomic Unit (OTUs). Though sequencing the entire 1500-bp 16S rRNA is necessary to distinguish between particular taxa or describing a new species, it is not required at clinical level, as the initial 500-bp sequence exhibits more diversity and is sufficient for differentiating and identifying taxa [61]. Association between species that are classified from regions of length 500-

bp and 1500-bp were similar for more than 100 organisms, encouraging the use of shorter sequence [62].

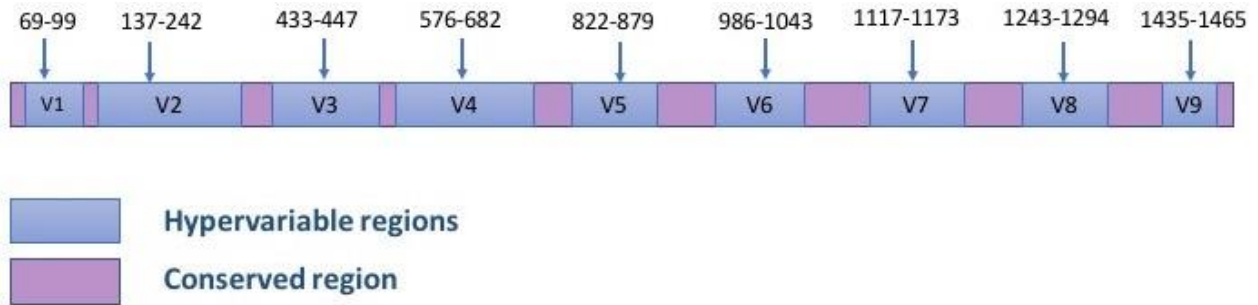


Fig 2.2 Hypervariable regions in 16S rRNA [63]

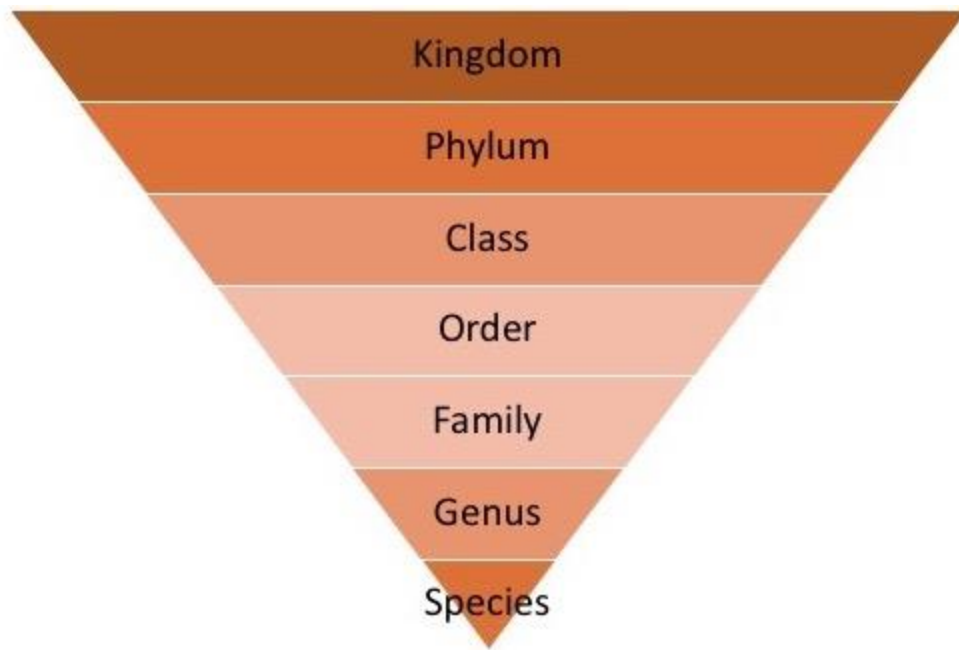


Fig 2.3 Phylum levels of classification

V2 and V3 regions of 16S rRNA produced relatively better results than the other regions in distinguishing 110 bacterial species up to genus-level, except for a few closely associated Enterobacteriaceae, encouraging the use of short V2 and V3 regions in phylogenetic and taxonomy studies [60]. Since the V3 region with 64-bp length is shorter than V2(105-bp), it is easier and

cheaper to sequence and therefore, chosen as the target in this research.

2.3 DNA Sequencing

Determining the sequence of nucleotide bases (Adenine, Cytosine, Guanine and Thymine) within a DNA strand is called DNA sequencing. The first ever DNA sequencing took place in the 1970's and many sequencing techniques emerged thereafter, spreading over 4 generations [DNA chain-terminating inhibitors]. Sequencing an entire genome is a complex task, where the entire DNA was required to be divided into various small segments to be sequenced. However, development of modern sequencing technologies made it faster and cheaper. The evolution from first generation to second generation DNA sequencing was very significant, where output increased to more than 5 orders of scale and cost falling to more than 5 orders of scale [64]. Sanger Sequencing can sequence up to 1000 bp for each run, and 384 sequences can be run in parallel in an automatic sequencer with a throughput of 80–100 kb per hour [65].

In 1985, reading a single base cost \$10. By 2005, the cost has fallen 10,000 lower. Second generation sequencing or the Next Generation Sequencing (NGS) platforms developed in the early 2000's can produce millions or even billions of reads in parallel. They also amplified the use of statistical methods and bioinformatics tools to analyze and manage the vast data generated. A few examples where NGS techniques are applied are, whole genome resequencing, targeted resequencing, de novo sequencing, gene expression analysis with whole transcriptome analysis, small RNA sequencing, methylation analysis, ChIP sequencing and nuclease fragmentation and sequencing [65]

Illumina sequencing:



Fig 2.4 Illumina MiSeq Sequencer [66]

Illumina sequencing is a Next Generation Sequencing technology that was first introduced by Bruno Canard and Simon Sarfati at the Pasteur Institute in Paris, later developed by Shankar Balasubramanian and David Clenerman and then acquired by illumina [67]. Illumina Next Generation sequencing technology basically works through four stages, namely Sample preparation, Cluster generation, Sequencing, and Data analyzing. Once the DNA is extracted from the samples, it is purified by removing all the unwanted debris before being cut into smaller pieces. These tiny fragments of DNA are given adapters on either sides and further introduced with sequencing binding sites, indices and regions complementary to flow cell oligos [68] .

Clustering is the isothermal amplification of these altered DNA fragments and it takes place in a glass slide with lanes called flow cell. Two types of oligonucleotides (short, synthetic pieces of DNA) are attached along the surface of the flow cell. DNA fragments are loaded onto the flow cell and hybridization takes place as one of the two oligos is the complementary sequence to the adapter region of one of the DNA fragments. Complementary sequence of the hybridized DNA

fragment is synthesized with the help of polymerases, and the original DNA strand is denatured and washed away. Now the strand is clonally amplified by bridge amplification where the strand bends over and the adapter region of the strand is hybridized to the second type of oligo. Polymerases generate a complementary strand forming a double stranded bridge which denatures into two separate DNA strands on the surface of the flow cell (forward and reverse) [68] as shown in Fig 2.5. The process is repeated simultaneously for millions of clusters causing clonal amplification of all the DNA strands. Then all the reverse strands are washed away, and 3-prime ends are blocked to avoid unwanted priming. The forward strands are sequenced in cycles with the help of the first sequencing primer where a nucleotide with a fluorescent tag is added to the growing chain according to the nucleotide in the template. Then the sequence reads are excited by a light source and the wavelengths of the fluorescent signals emitted by the tags on the nucleotides are noted to determine the nucleotides. In a given cluster, all the identical DNA strands are read simultaneously. Millions of clusters are sequenced in a parallel process. Then the indexed primers are hybridized to the template to generate index 1 read which would be useful to separate the sequences according to the sample during data analysis. Then the 3-prime ends are deprotected and the strand bends over to hybridize with second oligo. Index 2 reads are generated in the same manner and are extended by polymerases to form a double stranded bridge. The bridge denatures, and the forward strands are washed away. Then the reverse reads are sequenced in the same way. The process is repeated until the full DNA molecule is sequenced. Through massive parallel sequencing, thousands of reads throughout the whole genome can be sequenced at once.

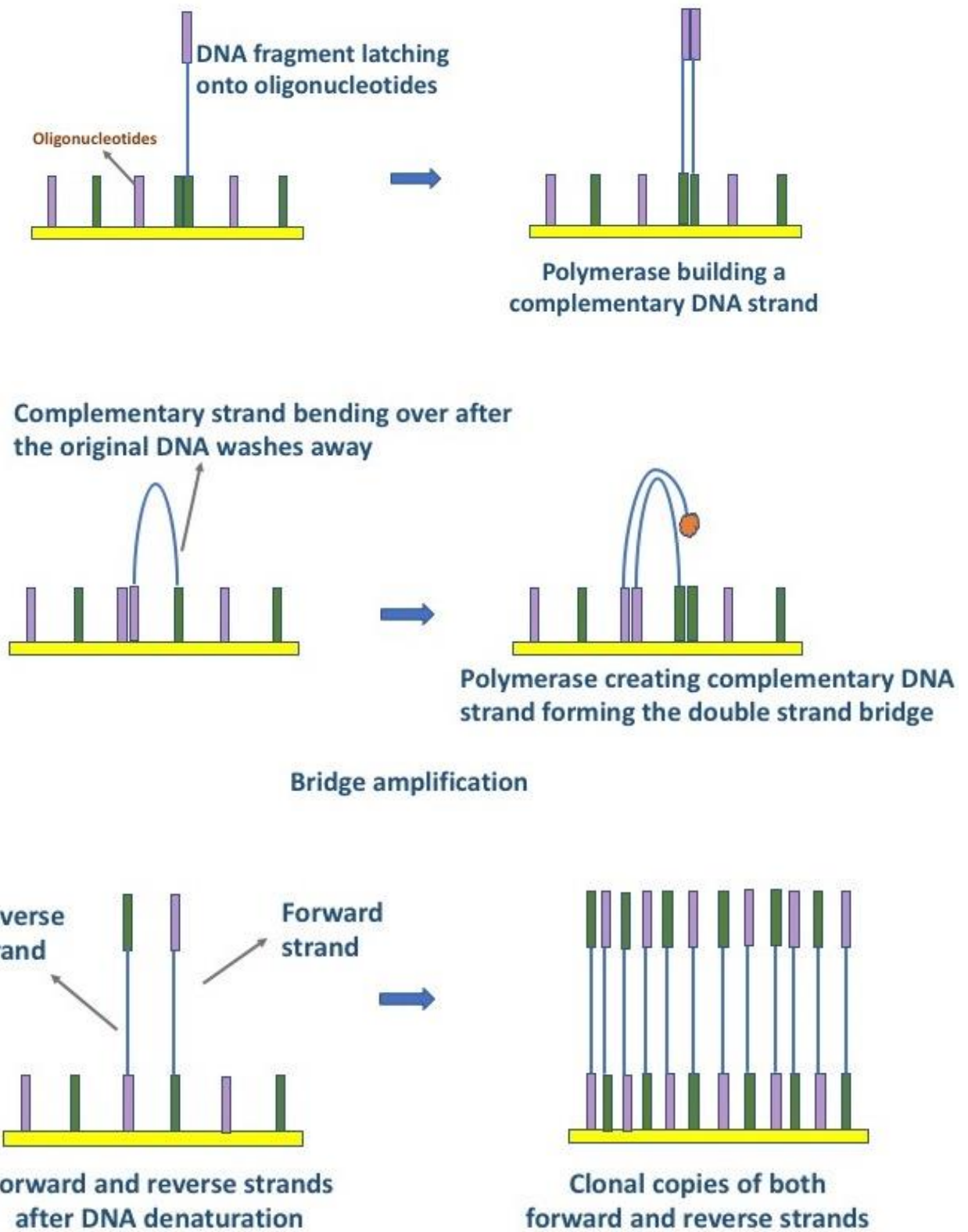


Fig 2.5 Illumina DNA Sequencing [68]

BaseSpace:

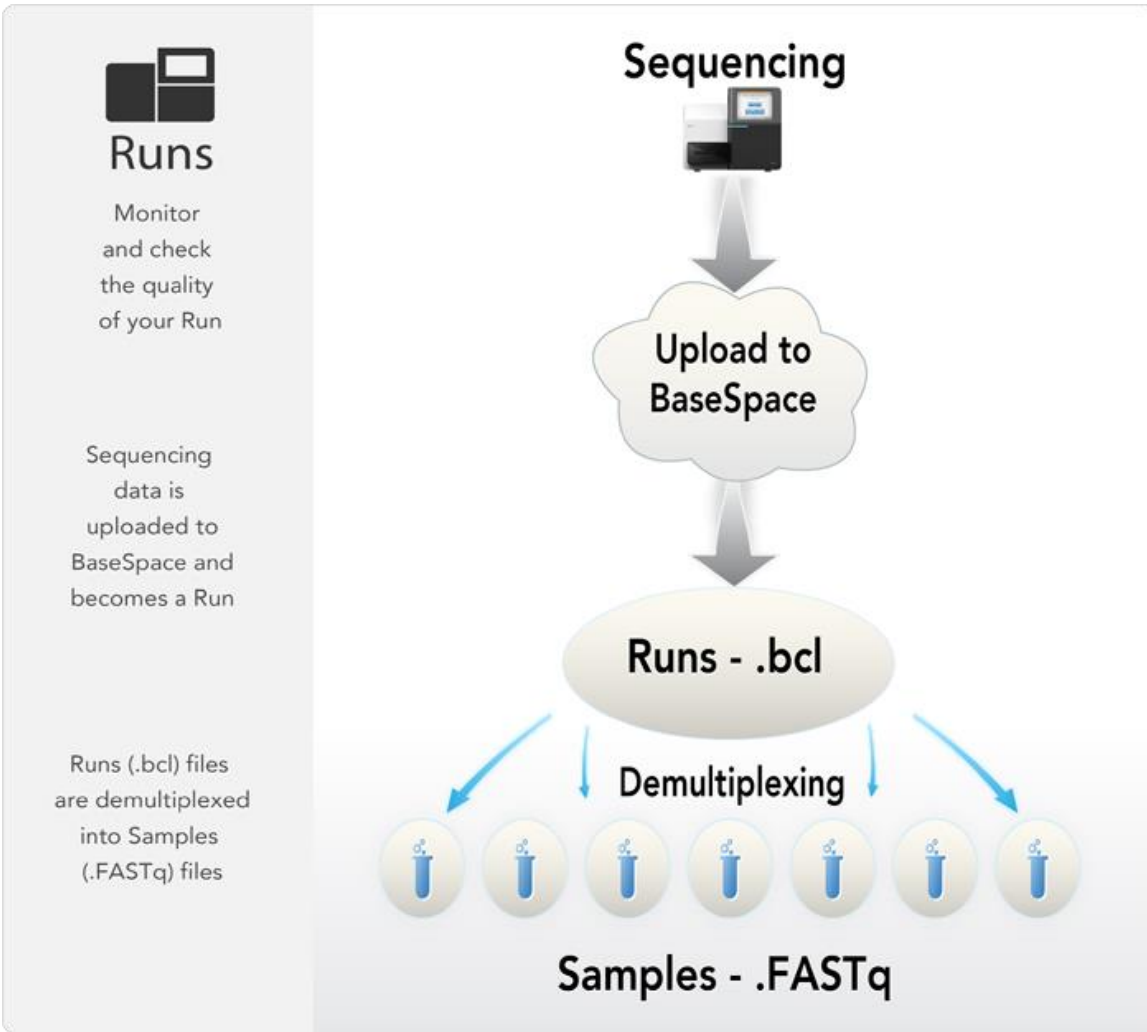


Fig 2.6 How data is stored and analyzed in BaseSpace [69]

Illumina in addition to sequencing the data, also provides the platform to analyze and share the sequenced data. BaseSpace is the built-in genomic computing platform in all the Next Generation Sequencing instruments of NextSeq, MiSeq (see Fig 2.4) and HiSeq. Illumina offers BaseSpace in both online cloud and offsite set-ups. We chose the online cloud for this work to make it easy for the transfer of the data between biology and genomic departments.

The MiSeq instrument converts all the data after sequencing into base call files (.bcl file) and then sends them to the allotted user space on the BaseSpace cloud. Now the .bcl file containing

all the sequenced data requires to be sorted into individual sample files to identify and classify the sequenced data. Therefore, BaseSpace converts and demultiples the data into individual sample FASTq files (Fig 2.6). In this research, one sample *. FASTq file represents the bacteria either from right or left hand of an individual. FASTq format is a text-based format that stores both the genomic sequence (nucleotides) and its respective quality score. A single ASCII character is used to encode both the nucleotide and its quality to make it brief. A FASTq file shown in Fig 2.7 usually has four lines per sequence

Line 1: It starts with '@' character and is followed by sequence identifier and an optional description.

Line 2: Raw sequence.

Line 3: It has a '+' sign and sometimes followed by optional sequence identifier or any other description.

Line 4: It has the same number of characters as in line 2 that represent the quality of base calls.

```
@Sequence ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((( (**+))%%%+)(%%%).1***-+*'))**55CCF>>>>>CCCCCCC65
```

Fig 2.7 FASTq file format

BaseSpace has a metagenomics workflow which is used in the taxonomic classification of 16S rRNA sequences. Once the data is converted into FASTQ files, BaseSpace compares the sequence reads against the GreenGenes reference database and classify them to the species [69]. The algorithm adopted in taxonomic classification is a Ribosomal Database Project (RDP) naïve Bayesian algorithm [70]. Once the sequences are compared against GreenGenes reference sequences, the reads for various taxonomy levels classified for every sequence is recorded.

2.4 QIIME

QIIME (Quantitative Insights into Microbial Ecology) is an open-source bioinformatics pipeline designed to perform microbiome analysis on raw DNA sequence data [71]. QIIME helps the users to generate graphical and statistical analysis from raw sequencing data generated from Illumina or other platforms suitable for publications. QIIME does various microbiome analysis through python (.py) scripts. Tasks that can be accomplished by QIIME include:

- Demultiplexing and quality filtering.
- OTU picking.
- Taxonomic assignment and phylogenetic reconstruction.
- Diversity analyses and visualizations

Raw sequences imported from outside sources are denoised to by either detecting and correcting sequences or truncating low quality sequence reads. Unique sequence features and their frequencies in each sample are identified using any OTU picking techniques available. OTUs are later assigned taxonomy with the help of multiple classifiers and reference databases. Resultant taxonomic assignment files can be used to build phylogenetic tree and perform several diversity analyses. Fig 2.8 is an example of Taxonomic assignment text file from QIIME listing the sequence names, taxonomy and their quality score

```

denovo1124 k_Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales 0.890
denovo580 k_Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales 0.880
denovo1170 k_Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Xanthobacteraceae;g__Ancylobacter;s__ 0.570
1066621 k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__melanigenica 1.000
denovo523 k_Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae 0.690
4344213 k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__ 0.630
denovo1030 k_Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__s__ 0.690
664697 k__Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__Flavobacteriaceae;g__Capnocytophaga;s__ 0.940
denovo852 k_Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales 0.990
denovo1569 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales 1.000
851938 k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Bulleidia;s__moorei 0.890
denovo2479 k__Bacteria;p__Firmicutes;c__Bacilli;o__Gemellales;f__g__s__ 0.570
denovo897
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Actinobacillus;s__parahaemolyticus 0.940
denovo2015 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas 0.950
denovo2619 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae;g__Propionibacterium;s__acnes
0.820
denovo2065 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales 0.970
denovo1061 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae 0.680
denovo2328 k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__ 0.520
denovo1259 k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales 0.620
denovo1078 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae;g__Propionibacterium;s__acnes
0.930
denovo1272 k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Planococcaceae;g__Staphylococcus;s__saprophyticus 0.500
denovo1797 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales 1.000
257278 k__Bacteria;p__Fusobacteria;c__Fusobacteriia;o__Fusobacteriales;f__Leptotrichiaceae;g__Leptotrichia;s__ 1.000
denovo2654 k__Bacteria;p__Firmicutes;c__Bacilli;o__Gemellales;f__Gemellaceae;g__s__ 0.810
12560 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces;s__ 1.000
560779 k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales 0.850
858535 k__Bacteria;p__Actinobacteria;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;g__Atopobium;s__ 0.710
denovo2371 k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Lautropia;s__ 0.930
denovo207 k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales 0.740
588414 k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Oribacterium;s__ 0.780
denovo848 k__Bacteria 0.920
1950608 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Corynebacteriaceae;g__Corynebacterium;s__ 0.590
4448346 k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Neisseriales;f__Neisseriaceae;g__Neisseria;s__subflava 0.570
815169 k__Bacteria;p__Cyanobacteria;c__Chloroplast;o__Streptophyta;f__g__s__ 0.980
1093959 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae;g__Propionibacterium;s__acnes
0.950
denovo1063 k__Bacteria 0.940
254469 k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Carnobacteriaceae;g__Trichococcus;s__ 0.820
denovo2635 k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Veillonella;s__dispar 0.580
251317 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Psychrobacter;s__ 0.550
denovo788 k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rickettsiales;f__mitochondria 0.570
denovo2601 k__Bacteria;p__Firmicutes;c__Bacilli;o__Gemellales;f__g__s__ 0.800
559030 k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales 1.000
693411 k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Corynebacteriaceae;g__Corynebacterium;s__ 0.690
130002 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas;s__viridiflava

```

Fig 2.8 Screenshot of taxonomy assignment text file from QIIME

2.5 Kullback-Leibler Divergence

The Kullback-Leibler Divergence was introduced in 1951 by Solomon Kullback and Richard Leibler to measure how one probability distribution diverges from a second probability distribution [72]. In simple words, a Kullback-Leibler divergence of 0 means that the two distributions are more similar, and their similarity decreases as the divergence values increases. The Kullback-Leibler Divergence between two distributions B and A denoted by $D_{KL}(A||B)$ is defined by the equation [73]

$$D_{KL}(A||B) = \sum_i A(i) \log \frac{A(i)}{B(i)} \quad \text{Equation 2.1}$$

In other words, Kullback-Leibler Divergence from B to A is the expectation of the logarithmic difference between the distributions A and B, while the expectation is taken from the distribution of A. Kullback–Leibler divergence is defined only if $B(i) = 0$ implies $A(i) = 0$ for all ‘i’ (absolute continuity). If $A(i) = 0$, the contribution of the i^{th} term is interpreted as zero because

$$\lim_{x \rightarrow 0} x \log(x) = 0.$$

Properties of Kullback-Leibler Divergence are [72][74]:

1. Kullback-Leibler Divergence is always non-negative

$$D_{KL}(A||B) \geq 0 \quad \text{Equation 2.2}$$

2. Kullback-Leibler Divergence for independent distributions is additive.

$$D_{KL}(A||B) = D_{KL}(A_1||B_1) + D_{KL}(A_2||B_2) \quad \text{Equation 2.3}$$

3. Kullback- Leibler divergence is non-symmetric [problem of KLD]

$$D_{KL}(A||B) \neq D_{KL}(B||A) \quad \text{Equation 2.4}$$

Therefore, another form of divergence, called Jensen-Shannon divergence, is obtained by averaging the two unequal KLDs to get a symmetric Kullback-Leibler Divergence which can be called the distance between the two distributions.

$$Distance_{A,B} = D_{JS} = \frac{1}{2} [D_{KL}(A||C) + D_{KL}(B||C)] \quad \text{Equation 2.5}$$

$$C = \frac{1}{2} (A + B) \quad \text{Equation 2.6}$$

2.6 Principal Coordinate Analysis (PCoA)

Principal Coordinate Analysis is a type of multidimensional scaling to visualize the dissimilarities between individual items of a data set. PCoA is different from Principal Component Analysis (PCA), which is a statistical procedure that analyzes the collection of differences between observations by converting possibly correlated variables into linearly uncorrelated variables, while PCoA analyzes the dissimilarities between observations [75]. PCoA takes a distance matrix containing dissimilarities between pairs of items and gives a coordinate matrix with the positions of the items. In Fig 2.9, distance matrix (a) with 4 data points is given as the input for PCoA and resultant Coordinate matrix (b) with coordinates of 4 data points is used to plot PC1 vs PC2 (c).

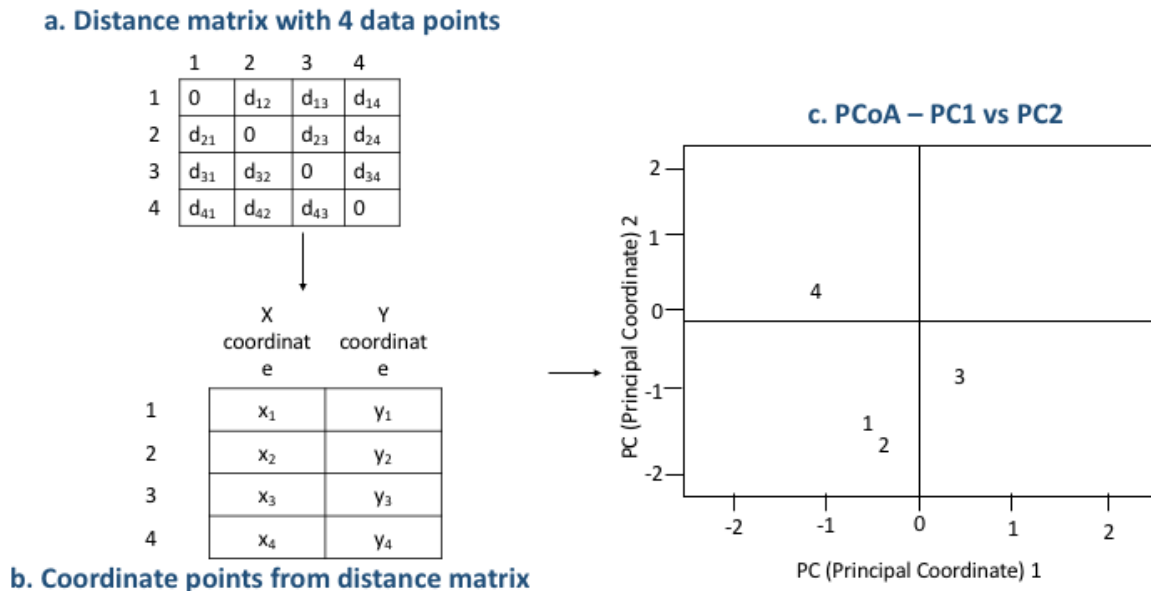


Fig 2.9 Illustration of Principal Coordinate Analysis

2.7 Unsupervised learning on OTU distribution and their k-mer frequencies

A union of 702 OTUs (denoted as g) identified by QIIME from 69 samples (denoted as n) are stored in taxonomy assignment files. Frequencies of g OTUs $\{Q_1^{1-g}, Q_2^{1-g}, \dots, Q_n^{1-g}\}$ for n

samples $\{S_1, S_2 \dots S_n\}$ are calculated. K-mer frequencies $\{P_1^{1-g}(z), P_2^{1-g}(z), \dots P_n^{1-g}(z)\}$ were derived by calculating the frequencies of A, C, G and T for k-1 where $z = \{A, C, G \& T\}$, the frequencies of AA, AC, AG, ... TT for k-2 where $z = \{AA, AC, AG, \dots TT\}$ and the frequencies of AAA, AAC, AAG, ... TTT for k-3 where $z = \{AAA, AAC, AAG, \dots TTT\}$. Multiple ways were adopted to perform unsupervised and supervised learning on the OTU distribution and their k-mer frequencies.

2.7.1 KLD analysis of OTU frequencies in each sample:

OTU frequencies $\{Q_1^{1-g}, Q_2^{1-g}, \dots Q_n^{1-g}\}$ were used to calculate the Kullback-Leibler Divergence D_{KLD} between two samples using Eq. 2.6. Symmetric distance (d_{ij}) between every two samples of 69 samples was calculated by taking the average of KLD at respective OTUs using Eq. 2.7. Resultant distance matrix was normalized and used to build a phylogenetic tree and perform Principal Coordinate Analysis. Depending on the geographical location of different ethnic groups, a reference distance matrix was created in such a way, that its resultant phylogenetic tree accommodates samples of one population group at a single node. This reference phylogenetic tree was used to compare with resultant trees, and dissimilarity between reference and resultant distance matrices was considered to measure the performance.

$$D_{KLD}^{otu}(S_i||S_j) = \sum_{o=1}^g Q_i^o \log \frac{Q_i^o}{Q_j^o} \quad \text{Equation 2.6}$$

$$d_{ij}^{otu} = \frac{1}{2} [D_{KLD}^{otu}(S_i||S_j) + D_{KLD}^{otu}(S_j||S_i)] \quad \text{Equation 2.7}$$

where $i=1,2,\dots,n, j=1,2,\dots,n$

2.7.2 KLD analysis of unweighted k-mer frequencies:

k-mer frequencies $\{P_1^{1-g}(z), P_2^{1-g}(z), \dots, P_n^{1-g}(z)\}$ of genus level OTU sequences were calculated using MATLAB commands. Average of k-mer frequencies over g number of OTU sequences was considered as the unweighted frequency sign for each sample. For example, for sample 'n', in case of k-1, k-mer frequencies

$$P_n = \begin{bmatrix} P_n^1(A) & P_n^1(C) & P_n^1(G) & P_n^1(T) \\ P_n^2(A) & P_n^2(C) & P_n^2(G) & P_n^2(T) \\ \vdots & \vdots & \vdots & \vdots \\ P_n^g(A) & P_n^g(C) & P_n^g(G) & P_n^g(T) \end{bmatrix} \quad \text{Equation 2.8}$$

$$P_n^{uw} = \frac{1}{g} \left[\sum_{o=1}^g P_n^o(A) \quad \sum_{o=1}^g P_n^o(C) \quad \sum_{o=1}^g P_n^o(G) \quad \sum_{o=1}^g P_n^o(T) \right] \quad \text{Eq. (4.4)}$$

...Equation 2.9

where 'uw' denotes unweighted

A distance matrix was built by calculating the symmetric KLD distance between the $P_n^{unweighted}$ of every two samples using equations 2.10 and 2.11, and phylogenetic tree and PCoA plots were generated from the matrix.

$$D_{KLD}^{uw}(S_i||S_j) = \sum_z P_i^{uw}(z) \log \frac{P_i^{uw}(z)}{P_j^{uw}(z)} \quad \text{Equation 2.10}$$

$$d_{ij}^{uw} = \frac{1}{2} [D_{KLD}^{uw}(S_i||S_j) + D_{KLD}^{uw}(S_j||S_i)] \quad \text{Equation 2.11}$$

where $i=1,2,\dots,n, j=1,2,\dots,n$

2.7.3 KLD analysis over weighted k-mer frequencies:

Frequencies of Genus level OTU sequences $\{Q_1^{1-g}, Q_2^{1-g}, \dots, Q_n^{1-g}\}$, and their respective k-mer frequency values $\{P_1^{1-g}(z), P_2^{1-g}(z), \dots, P_n^{1-g}(z)\}$ were combined to get the weighted

frequency sign for each sample, which were used to calculate the symmetric pair-wise distances using the KLD. For example, for sample ‘n’, in case of k=1,

$$\text{K-mer frequency values for } g \text{ OTUs } P_n = \begin{bmatrix} P_n^1(A) & P_n^1(C) & P_n^1(G) & P_n^1(T) \\ P_n^2(A) & P_n^2(C) & P_n^2(G) & P_n^2(T) \\ \vdots & \vdots & \vdots & \vdots \\ P_n^g(A) & P_n^g(C) & P_n^g(G) & P_n^g(T) \end{bmatrix} \quad \text{Equation 2.12}$$

$$\text{Frequencies of } g \text{ OTUs } Q_n = [Q_n^1 \quad \dots \quad Q_n^g] \quad \text{Equation 2.13}$$

$$Q_n \times P_n = \begin{bmatrix} Q_n^1 \times P_n^1(A) & Q_n^1 \times P_n^1(C) & Q_n^1 \times P_n^1(G) & Q_n^1 \times P_n^1(T) \\ Q_n^2 \times P_n^2(A) & Q_n^2 \times P_n^2(C) & Q_n^2 \times P_n^2(G) & Q_n^2 \times P_n^2(T) \\ \vdots & \vdots & \vdots & \vdots \\ Q_n^g \times P_n^g(A) & Q_n^g \times P_n^g(C) & Q_n^g \times P_n^g(G) & Q_n^g \times P_n^g(T) \end{bmatrix} \quad \dots \text{Equation 2.14}$$

$$P_n^w = \frac{1}{g} \left[\sum_{o=1}^g Q_n^o \times P_n^o(A) \quad \sum_{o=1}^g Q_n^o \times P_n^o(C) \quad \sum_{o=1}^g Q_n^o \times P_n^o(G) \quad \sum_{o=1}^g Q_n^o \times P_n^o(T) \right] \quad \dots \text{Equation 2.15}$$

where ‘w’ denotes weighted

$$D_{KLD}^w(S_i || S_j) = \sum_z P_i^w(z) \log \frac{P_i^w(z)}{P_j^w(z)} \quad \text{Equation 2.16}$$

$$d_{ij}^w = \frac{1}{2} [D_{KLD}^w(S_i || S_j) + D_{KLD}^w(S_j || S_i)] \quad \text{Equation 2.17}$$

where $i=1,2,\dots,n, j=1,2,\dots,n$

2.8 Structure based mapping of 16S rRNA V3 region

Hypervariable region V3, which is 65 nucleotides long (position 433 to 497) was split into its RNA secondary structure elements strand, stem, bulge, internal loop and loop (see Fig 2.10). The short chain of 4 nucleotides in the beginning with no bonds was termed as strand followed by a series of bonded nucleotides, which was labelled as stem. The single unbonded nucleotides that

occur randomly along the stem were named as bulge, where as the sequence of unbonded nucleotides along the stem were named as internal loops. The remaining sequence of nucleotides acting as a bridge between a pair of bonded nucleotides was labelled as loop. The pair of nucleotides that share a bond in the stem were together considered as a single element, for example nucleotide 5 and 65 together were considered as single element 5 and therefore 65 nucleotides were mapped into 47 elements as shown in Fig 2.11.

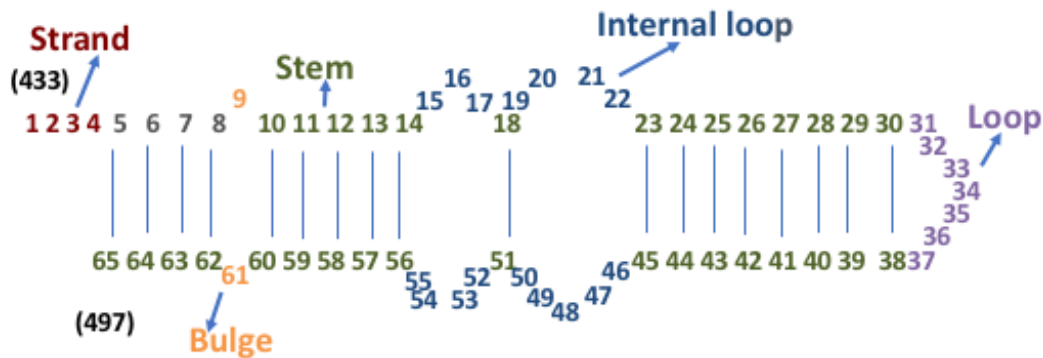


Fig 2.10 Nucleotide bond structure in 16S rRNA V3 region

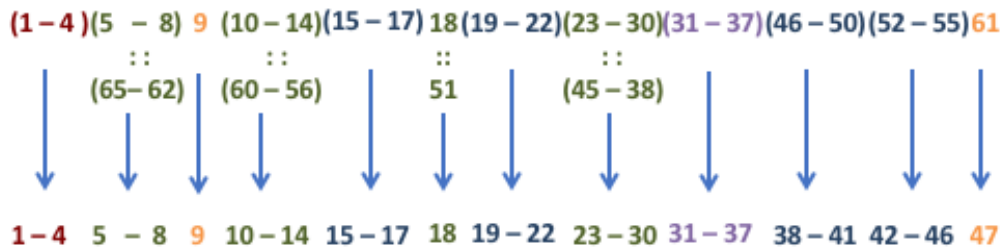


Fig 2.11 Mapping of 65nt of 16S rRNA into 47 elements

The mapped sequences result in 20 (m) different nucleotide possibilities like A, C, G, T, AA, AC, AG, ...TT; therefore, for $m=20$, the number of k -mers in mapped OTU sequences are $z=m^k=20$ for $k=1$ and $z=m^k=20^2=400$ for $k=2$. For example, for sample 'n', in case of $k=1$,

$$\text{K-mer frequency values for mapped OTU sequences } P_n = \begin{bmatrix} P_n^1(1) & P_n^1(2) & \dots & \dots & P_n^1(20) \\ P_n^2(1) & P_n^2(2) & \dots & \dots & P_n^2(20) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_n^g(1) & P_n^g(2) & \dots & \dots & P_n^g(20) \end{bmatrix}$$

...Equation 2.18

$$\text{Frequencies of } g \text{ OTUs } \quad Q_n = [Q_n^1 \quad \dots \quad Q_n^g] \quad \text{Equation 2.19}$$

$$Q_n \times P_n = \begin{bmatrix} Q_n^1 \times P_n^1(1) & Q_n^1 \times P_n^1(2) & \dots & \dots & Q_n^1 \times P_n^1(20) \\ Q_n^2 \times P_n^2(1) & Q_n^2 \times P_n^2(2) & \dots & \dots & Q_n^2 \times P_n^2(20) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Q_n^g \times P_n^g(1) & Q_n^g \times P_n^g(2) & \dots & \dots & Q_n^g \times P_n^g(T) \end{bmatrix} \quad \text{Equation 2.20}$$

$$P_n^{uw} = \frac{1}{g} [\sum_{o=1}^g P_n^o(1) \quad \sum_{o=1}^g P_n^o(2) \quad \dots \quad \dots \quad \sum_{o=1}^g P_n^o(20)] \quad \text{Equation 2.21}$$

$$P_n^w = \frac{1}{g} \left[\sum_{o=1}^g Q_n^o \times P_n^o(1) \quad \sum_{o=1}^g Q_n^o \times P_n^o(2) \quad \dots \quad \dots \quad \sum_{o=1}^g Q_n^o \times P_n^o(20) \right]$$

...Equation 2.22

In case of mapped sequences, k-mer analysis is stopped with k=2 because k=3 resulted in $z=m^k=20^3=8000$ k-mers where the majority of the k-mers could not be found in the mapped sequences and the execution was time-consuming.

2.9 K-means Clustering

K-means clustering is a technique used to compress large dataset with 'n' number of observations into 'k' mutually exclusive clusters, where each observation in a given cluster is more similar to the observations in the same cluster than they are to the observations in other clusters. For example, if $X_1, X_2, X_3, \dots, X_n$ are n number of observations, k-means clustering divides the n observations into k ($\leq n$) clusters S_1, S_2, \dots, S_k by applying the formula

$$\operatorname{argmin}_S \sum_{i=1}^k \sum_{x \in S_i} \|X - \mu_i\|^2 = \operatorname{argmin}_S \sum_{i=1}^k |S_i| \operatorname{Var} S_i \quad \text{Equation 2.23}$$

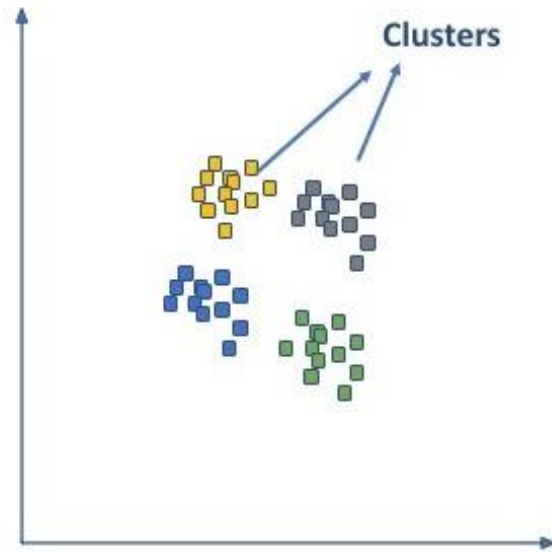


Fig 2.12 Mutually exclusive clusters in k-means clustering

An observation is placed into a cluster with the nearest mean, in an effort to minimize the variance within a cluster. Fig 2.12 shows an example clustering. The chief limitation of k-means clustering is its model, where the number of clusters is picked by the user. In order to determine the appropriate number of clusters, multiple diagnostic checks and comparisons are required to be made.

Silhouette:

Silhouette is a method to validate the number of clusters in a given data set [76]. It is a measure of how similar the observation is to its own cluster, and how different it is from other clusters. Silhouette values provide a concise illustration of how well an observation is placed in a certain cluster. Silhouette values range from -1 to +1, where higher number indicates higher similarity with its own cluster and higher dissimilarity with other clusters. Therefore, a greater

number of higher silhouette values affirm proper clustering.

2.10 Ensemble Learning

Ensemble learning is a machine learning prototype that uses several learning algorithms to obtain high-quality predictive performance. Some of the algorithms can be applied only for classification ensembles, and some algorithms apply only for regression ensembles. Different from other machine learning methods where one hypothesis is learned from training the data, ensemble learning tries to construct multiple hypotheses and use them to predict the class. In order to avoid over-fitting from training with multiple models, a bagging technique is applied while training the data. For this research, MATLAB R2017b was used to train an ensemble for classification with a bootstrap aggregation (bagging) method and a decision tree learner.

Bootstrap Aggregation:

Bootstrap aggregation, referred to as bagging, is one of the ensemble learning algorithms that was proposed by Leo Breiman in 1994 to improve classification by combining classifications of randomly generated training sets from within the given training set [77]. Bootstrap aggregation is aimed to improve the stability and accuracy of the learner models. Bootstrap aggregation is a model averaging approach, that reduces variance and assist in avoiding overfitting of the model. Bagging is applied usually on decision tree models though it can be used with any other model.

Leave one out cross validation:

Leave-one-out cross validation (Loo-CV) is a K-fold cross validation technique, where K is equal to the number of data points in the given set. Therefore, the model is trained N different

times on all the data except for one data point and prediction is done on that left-out data point. Then the average accuracy is calculated to evaluate the model. Leave-one-out is also a special case of Leave-p-out cross validation where $p=1$. Unlike Leave-p-out cross validation, Loo-CV doesn't take much computational time since $C_1^N = 1$

Confusion matrix:

Confusion matrix or error matrix is a table that is used to evaluate the performance of a classification model. In confusion matrices presented in this thesis, rows represent the actual class i.e. true population group, while the columns represent the predicted class. The values in the diagonal of confusion matrix are the accurately predicted samples and the sum of all columns in a row gives the total number of samples belonging to the population group represented by that particular row. Table 2.1 is a model of confusion matrix with 3 classes. For example, in pigeon class, 4 samples out of 7 samples (4+2+1) were predicted accurately as pigeon, while 2 samples were predicted as parrot and 1 sample was predicted as peacock.

Table 2.1 Example of Confusion matrix with 3 classes

		Predicted class		
		Pigeon	Parrot	Peacock
Actual class	Pigeon	4	2	1
	Parrot	1	3	0
	Peacock	0	0	5

Chapter 3 : Data Extraction and organization

Chapter 3 describes the procedure of collecting the samples and extracting the data from the samples

3.1 Overview

Data collection was done under the name Human DNA and Facial features with approval from the Institutional Review Board (IRB #H-23693). The data collection was funded by US Department of Justice/ Office of Justice programs/National Institute of Justice. Data collection included taking the Human genomic DNA samples and Hand swab samples. The hand bacterial project mentioned in this thesis is a part of a larger research project with two significant goals. One aim is to study the Human genomic DNA and encode facial features like eye color, hair color, nose, ear lobe *et cetera*. The other is to analyze the bacterial communities found on human hands and determine if they can be used to differentiate people from different population groups. Blood samples, hand swabs, facial images and medical histories of around 200 individuals both male and female from diverse population groups and age groups were collected over a time period of three months for the project.

3.2 Collection Process

Data collection took place in September 2012 and lasted for approximately three months. Participants were approached through email they provided for future contact in any previous studies they partook and were given a website and contact number to make a reservation with the collection team. The collection was carried out in West Virginia University's Health Sciences Center. This location is connected to Ruby Memorial Hospital, which is the largest medical complex in the state of West Virginia, so that any medical assistance needed during the sample collection could be provided with ease. Before taking the samples, the participant was asked to go

through a consent form and was explained the procedure of data collection by a co-investigator. Then, the participant was registered into a database to record the details about their physical traits and was assigned with a random number in the demographics to maintain confidentiality. The participant was also asked to fill out the forms with their medical history, handedness and hand washing.

Firstly, 2-D facial images were collected with a commercial digital camera and a gray backdrop. The images were taken from five different angles (-90° , 45° , 0° i.e. front pose, 45° , 90°) with no facial expressions. Then in a connecting room, hand swab samples were collected. An end of a cotton swab is sterilized prior to use by being wrapped in an aluminum foil and autoclaved. It is then dipped in a double distilled water solution with 0.15 M NaCl and 0.1% Polysorbate 20 (aka Tween 20) as it works as a non-toxic cleaner and helps in lifting the bacteria from the skin surface [78]. The cotton swab was then used to collect the hand bacterial sample by swabbing the entire palm by rotating the cotton tip. The head of the cotton was placed in a 2ml bead solution tube of an Ultraclean Plan DNA Isolation kit (MO BIO laboratories Carlsbad, CA, USA) and cut with a sterilized scissors. Then, the procedure was repeated for the other hand using a different cotton swab. The tubes were then stored at -80°C until DNA extraction. Once both bacterial samples and blood samples were collected, participants were given \$40 after the data collection.

3.3 Demographics

The number of individuals that took part in the collection was 255. Though it was initially planned to have 200 individuals for the study, blood samples from around 30 people couldn't be collected because either their veins were small or were not close enough to the skin surface to be found. The number of individuals that participated in the hand bacterial collection was 51. The

following charts provide a detailed break-down of the demographics regarding the 51-people considered for the hand bacterial study [22].

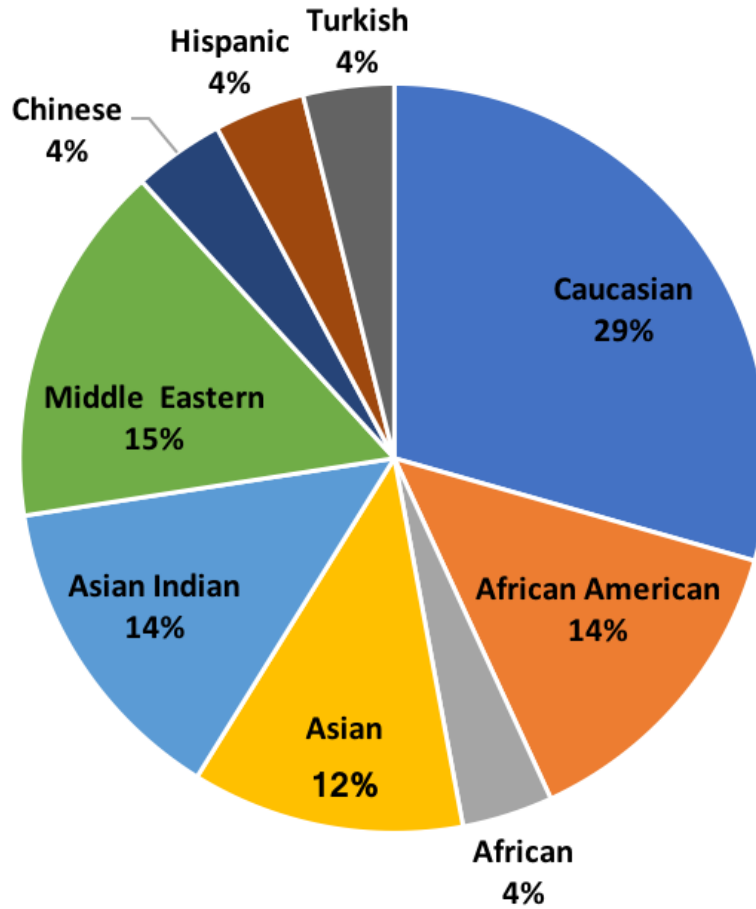


Fig 3.1 Population group of the participants in the hand bacterial sample collection

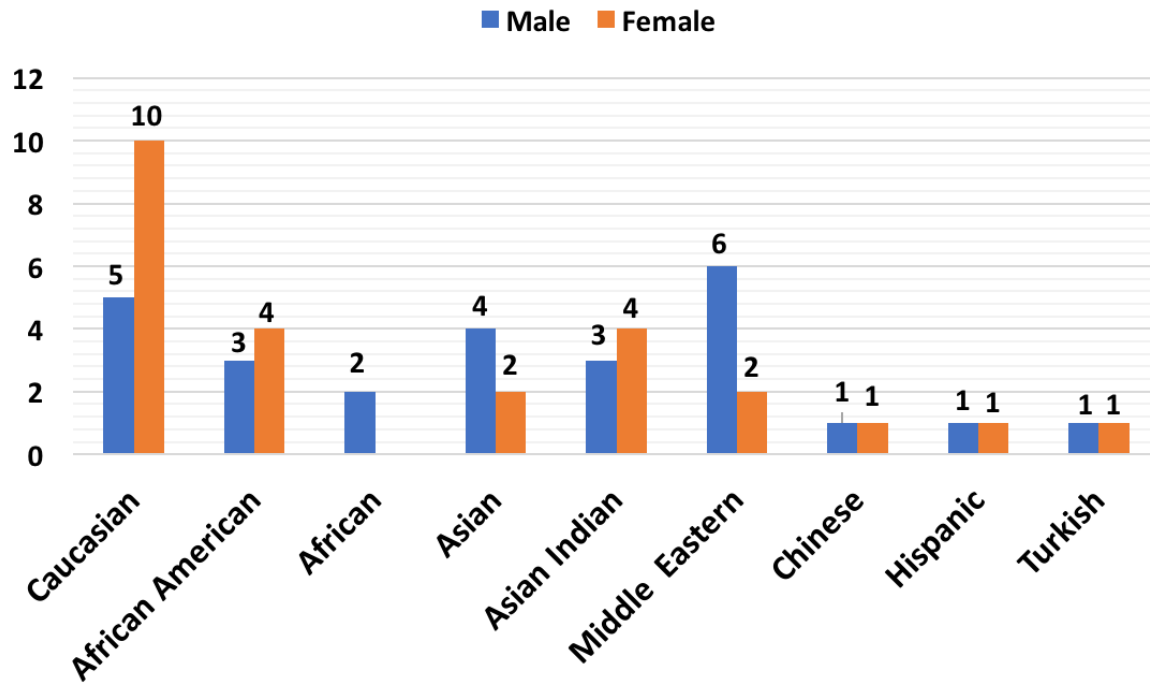


Fig 3.2 Population group and gender of the participants in the hand bacterial sample collection

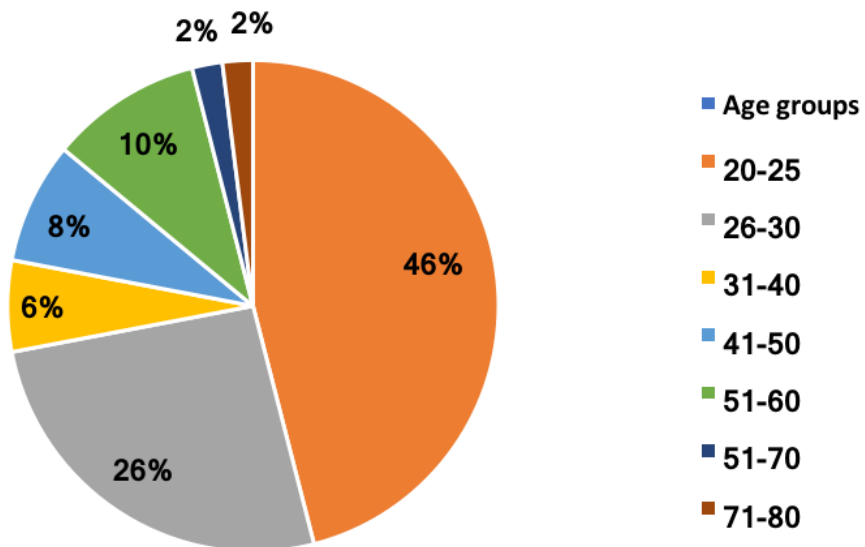


Fig 3.3 Age groups of the participants in the hand bacterial sample collection

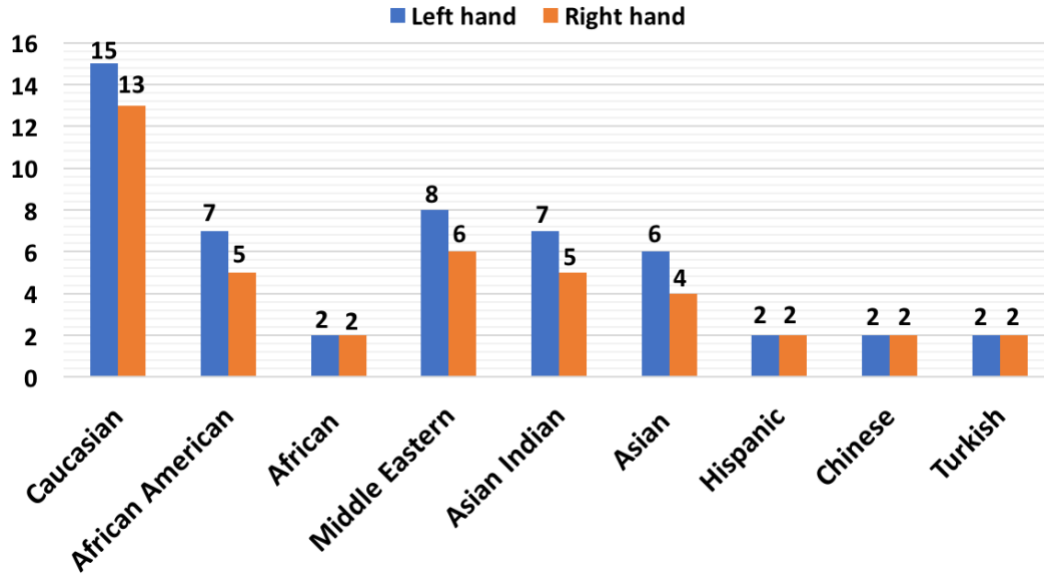


Fig 3.4 Left-hand and right-hand sample count of the participants in the hand bacterial sample collection

3.4. DNA Isolation and V3 region amplification

DNA was first isolated from the hand swab samples before PCR Amplification. 16S region of the bacterial DNA was amplified using the primers E8F and E1541R (see Fig 3.5 and Table 3.1) [79]. PCR amplification was repeated to amplify the V3 hypervariable region with primers 341F and modified 518R synthesized by Eurofins Genomics [80] [81]. Sequences obtained from amplification of V3 region were sent to Illumina's BaseSpace next-generation sequencing cloud, for automatic analysis and storage [82]. Sequence reads of length 151 and their quality scores for each sample were written to two FASTQ format files; one file for the forward run and the other for the reverse run.

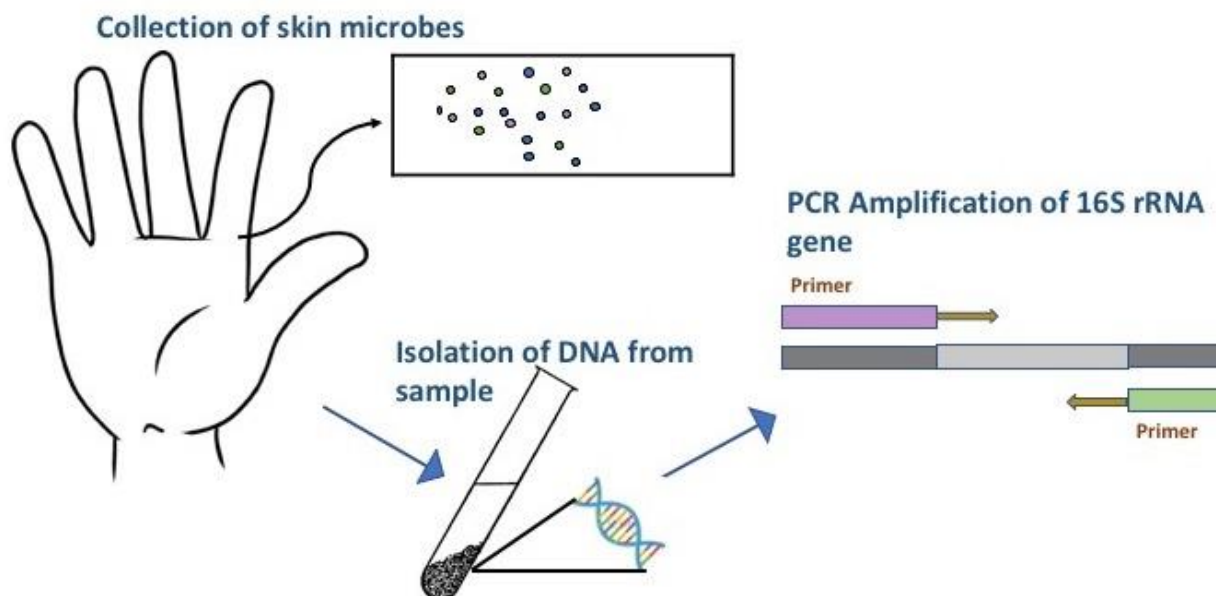


Fig 3.5 Work flow of 16S rRNA gene extraction from skin bacteria before DNA sequencing

Table 3.1 primers used the Amplification of 16S rRNA and V3 hypervariable region

Forward and Reverse Primer list
<p>16S rRNA E8F: AGAGTTTGATCCTGGCTCAG E1541R: AAGGAGGTGATCCANCCRCA</p>
<p>V3 region 341F: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGACGCTCTTCCGATCTCCTACGGAGGCAGCAG 518R: CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTG TGCTCTCCGATCTATTACCGCGGCTGCTGG</p>

3.5 Classification and organization of raw sequence data:

Forward and Reverse Fastq files from Illumina's sequencing cloud were joined into a single fastq file using join_paired_ends.py script in QIIME. Joined fastq files were then converted to fasta files. OTUs were picked by clustering sequences using pick_open_reference_otu.py script with default otu picking method and reference sequences. Representative sequences for OTUs

were picked using pick_rep_set.py script where a single sequence is assigned for every OTU. Representative sequences were classified using RDP (Ribosomal Database Project) classifier with GreenGenes database by applying assign_taxonomy.py script. Output from assign_taxonomy.py is given in the form of text files with sequence identifiers and their assigned taxonomy. Frequencies of OTUs were calculated from these resultant text files using MATLAB commands. Representative sequences were aligned with align_seqs.py script. Nucleotides in 65 positions of V3 hypervariable region between positions 2095 and 2159 of aligned sequences fasta file were separated to apply multiple Bioinformatics' techniques explained in chapter 2. Each sample is named after its gender, hand, n and date of birth. For example, F_L_MidE_55 represents Female, Left hand, Middle eastern and born in 1955. Abbreviations used for all the population groups presented in this thesis are listed in the following Table 3.2

Table 3.2 List of Population groups and their abbreviations

Population group	Abbreviation
African	Af
Turkish	Tur
Chinese	Chin
Hispanic	Hisp
Middle Eastern	MidE
Caucasian	Ca
Asian	Asian
Asian Indian	AsInd
African American	AfAm

Chapter 4 : Bioinformatics' Analysis

Chapter 4 illustrates results from application of different statistical and bioinformatics' techniques, that were explained in chapter 2.

4.1 Taxonomic classification of Raw sequences

V3 hypervariable region of 16S rRNA is amplified from isolated DNA and sequenced using Illumina MiSeq sequencer. Sequences are then classified into OTUs using RDP (Ribosomal Database Project) classifier and identified with GreenGenese reference database. Representative consensus sequences for genus-level OTUs were separated from each sample to apply various bioinformatics and statistical methods. Out of 104 samples, only 69 samples had both taxonomic and sequence data and rest of the samples could not be used due to lack of enough information. Demographics of 69 samples that were used in this study are illustrated in the below graphs Fig 4.1 and Fig 4.2. Nucleotides at 65 positions of 16S rRNA V3 hypervariable region of genus level OTUs, which were 702 from 69 samples were separated, and their A, C, G and T frequency profiles (K-mers) are calculated.

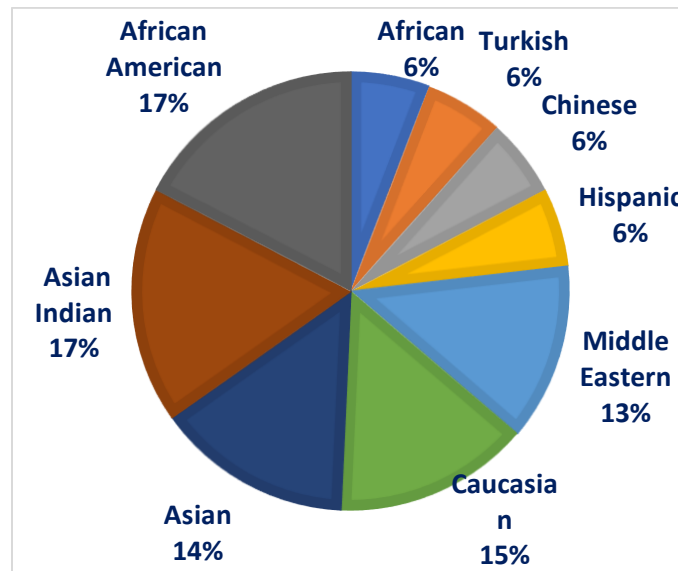


Fig 4.1 Percentage of samples belonging to each population group

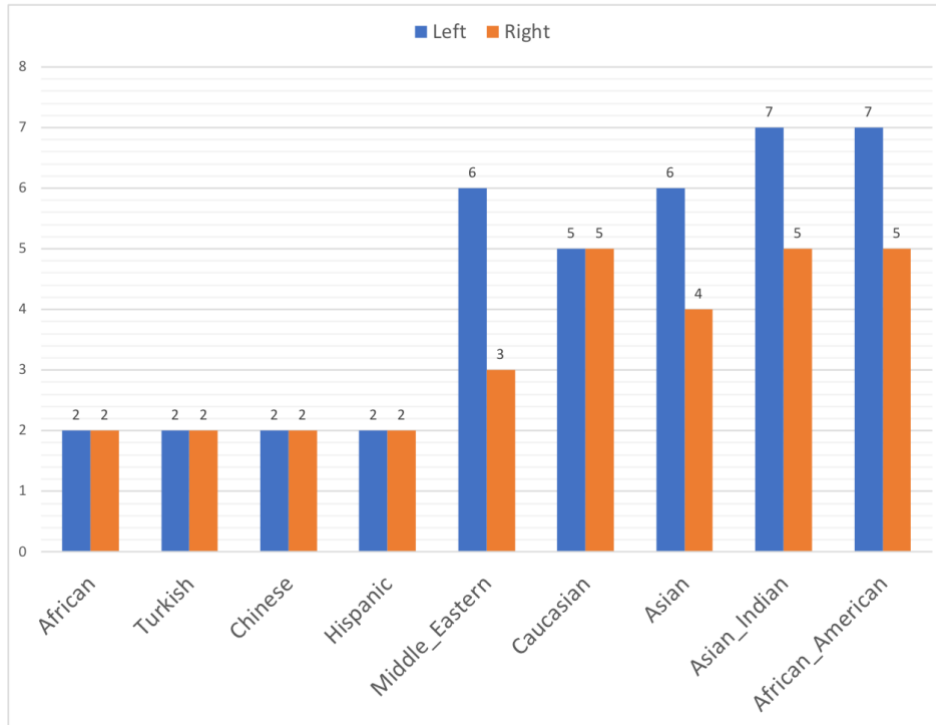


Fig 4.2 Number of left and right-hand samples in every population group

4.2 Verification of K-mer signatures for known OTUs

K-mer frequencies of nucleotides of genus-level OTU sequences for $k=1, 2$ and 3 were verified by comparing their graphs from two different samples. Same OTU from two different samples had similar patterns confirming that k-mer profiles can be used as a feature to compare and cluster different samples. Figures 4.3, 4.4 and 4.5 demonstrate k-mer frequencies illustrating similar patterns for the genus level OTU *Staphylococcus* from 4 different samples for $k-1, k-2$ and $k-3$ respectively.

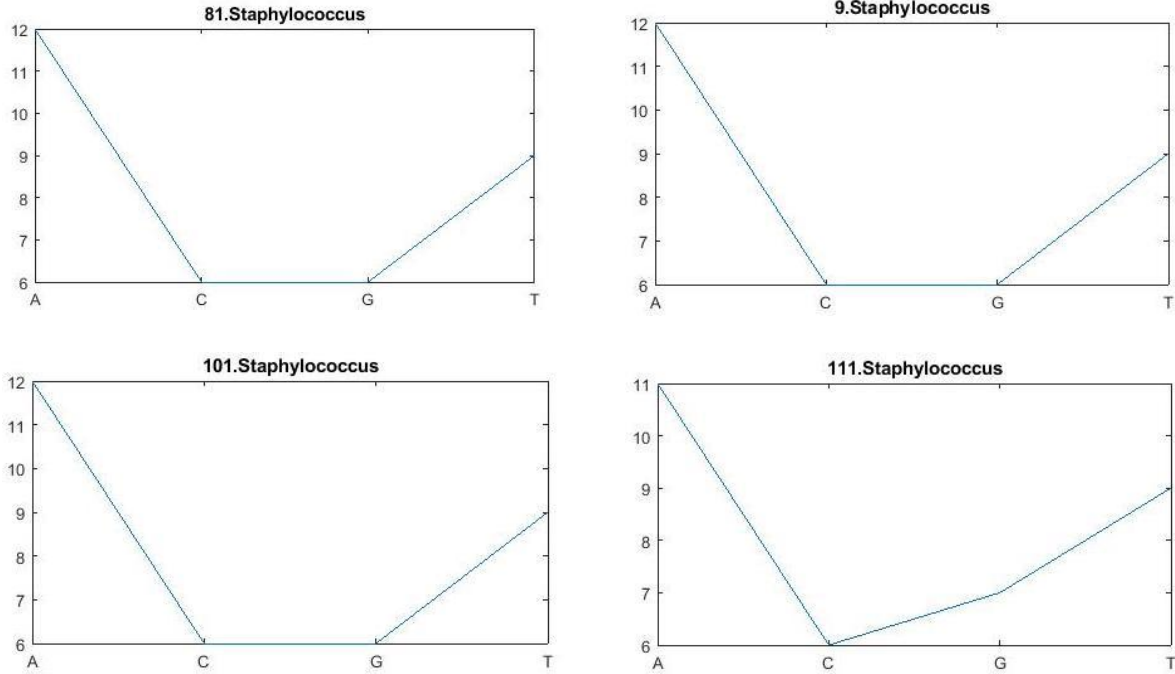


Fig 4.3 K-mer (k=1) frequencies in genus level OTU Staphylococcus from 4 different samples

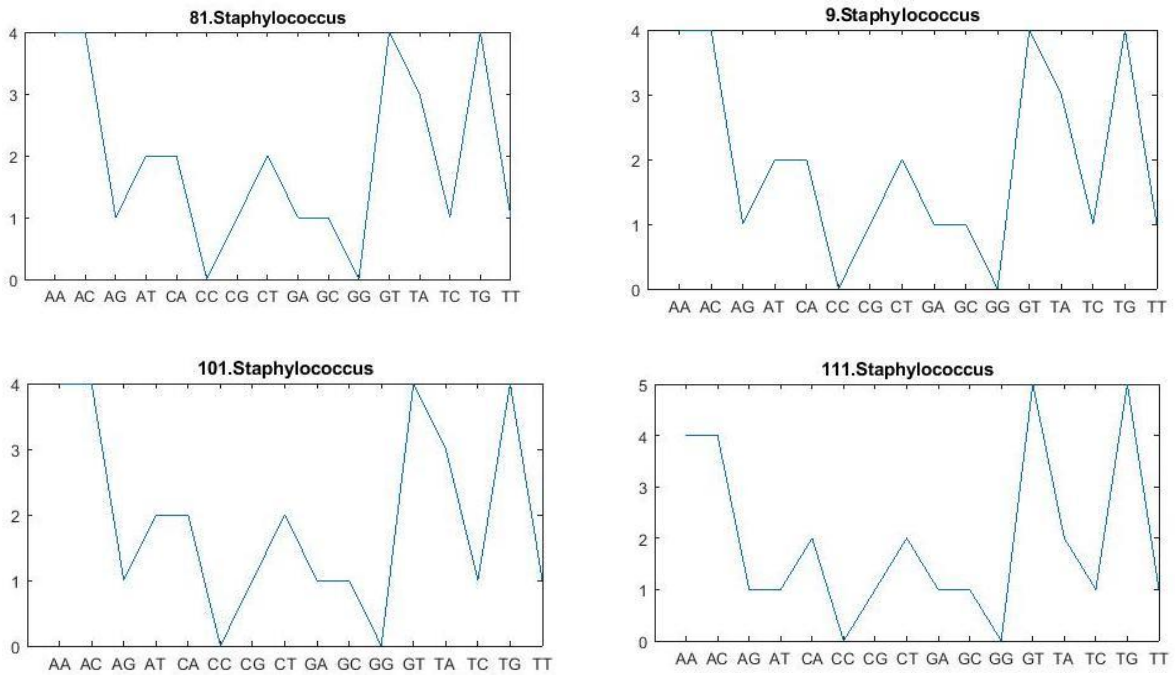


Fig 4.4 K-mer (k=2) frequencies in genus level OTU Staphylococcus from 4 different samples

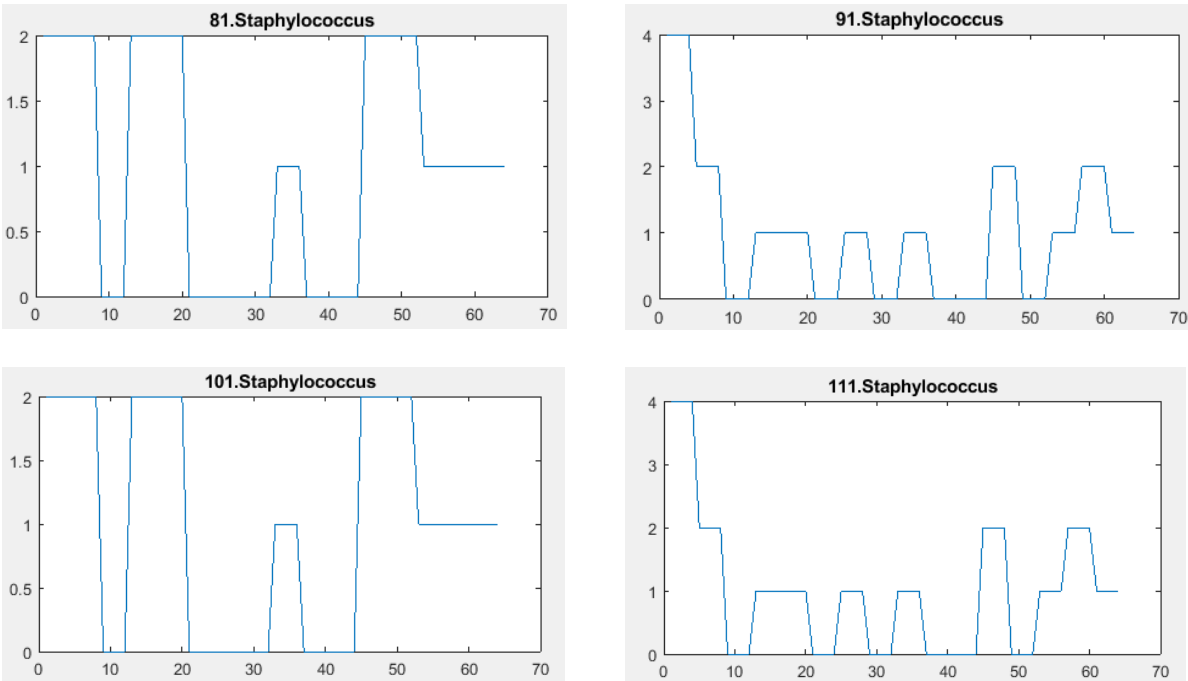


Fig 4.5 K-mer (k=3) frequencies in genus level OTU Staphylococcus from 4 different samples

4.3 Unsupervised machine learning of population groups using OTU and k-mer frequencies in hand bacterial samples

4.3.1 KLD analysis of OTU frequencies:

Like explained in section 2.7.1, OTU frequencies were used to calculate the Kullback-Leibler Divergence D_{KLD} between two samples. Symmetric distance (d_{ij}) between every two samples of 69 samples was calculated by taking the average of KLD at respective OTUs. Resultant distance matrix was normalized and used to built phylogenetic tree and perform Principal Coordinate Analysis. Depending on the geographical location of different population groups, a reference distance matrix was made in such a way, that the resulting phylogenetic tree accommodates samples of one population group at a single node. Fig 4.7 is the reference phylogenetic tree that was used to compare with resultant trees, and dissimilarity between reference and resultant

distance matrices was considered to measure the performance. Fig 4.8 is the resultant phylogenetic tree after applying KLD analysis on OTU frequencies. Samples from same person or same population group or even similar age groups were observed to often share the same nodes in the tree. To study the dissimilarity between reference and resultant distance matrices, average pairwise Euclidean distance between two distance matrices was calculated and found to be 0.3276. Furthermore, Principle Coordinate Analysis was implemented on different population groups to see which population groups are more similar. PCoA plot Fig 4.6, there is certain clustering of African and Turkish samples and they seem to be closer to each other than they are to Hispanic and Chinese samples reflecting their geographical locations. PCoA plots on other population groups are listed in the Appendix A section of this thesis.

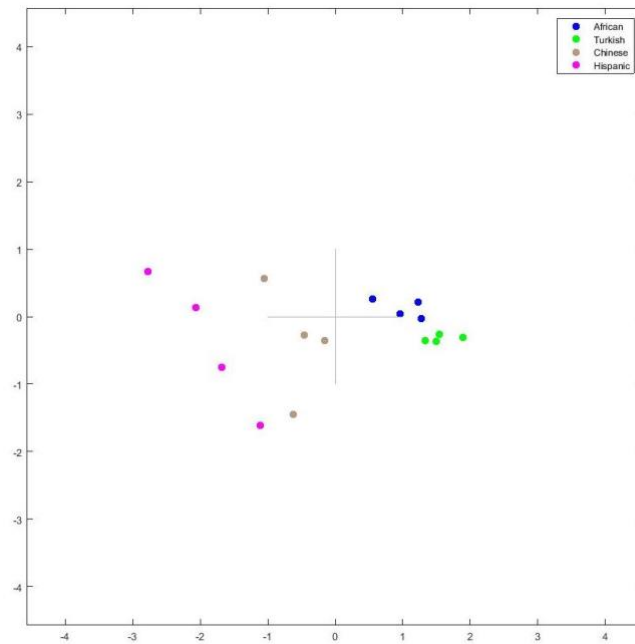


Fig 4.6 PCoA plot from KLD analysis of OTU frequencies of 16 samples from 4 population groups

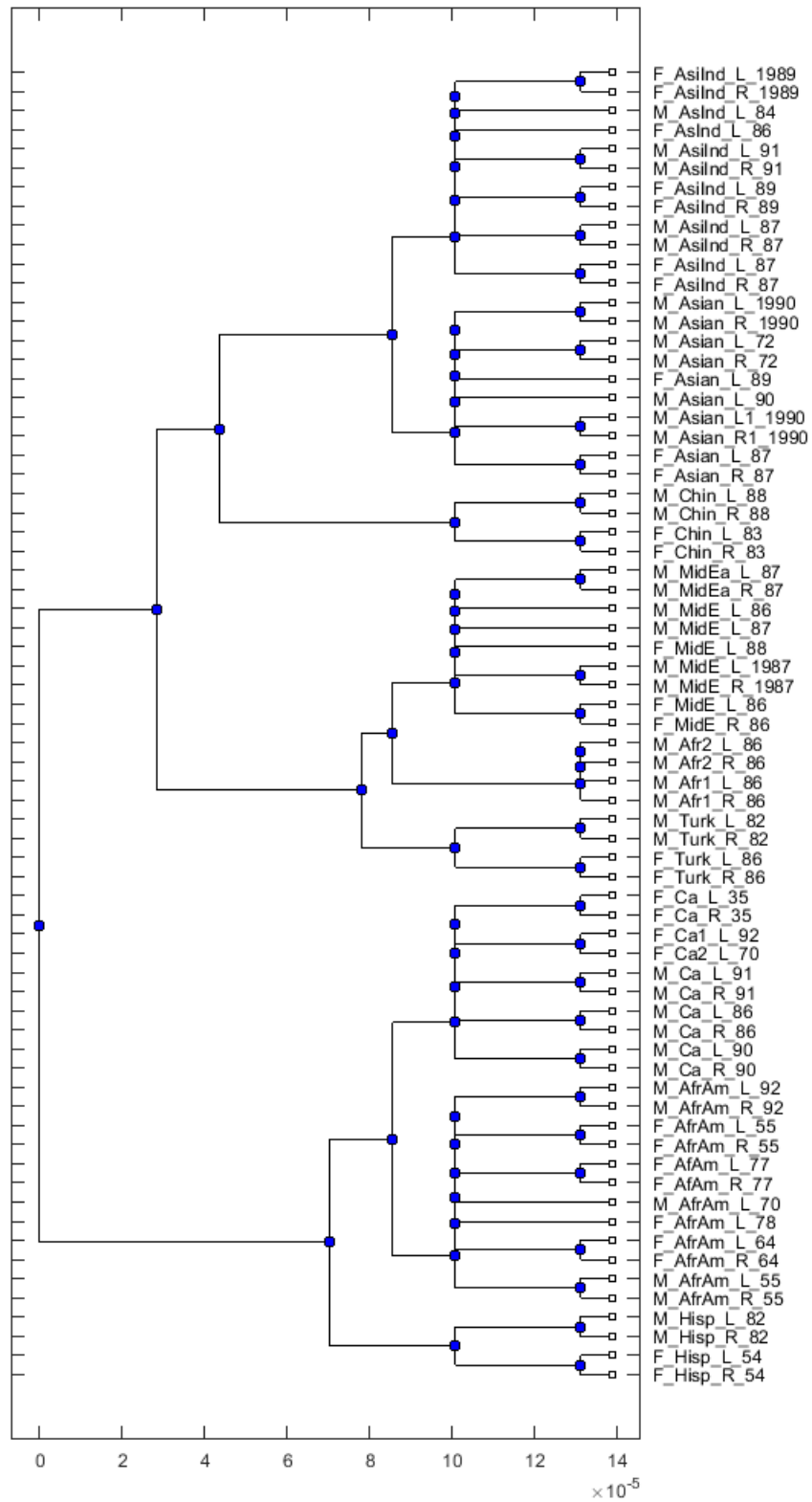


Fig 4.7 Reference Phylogenetic tree depending on geographical distance between population groups

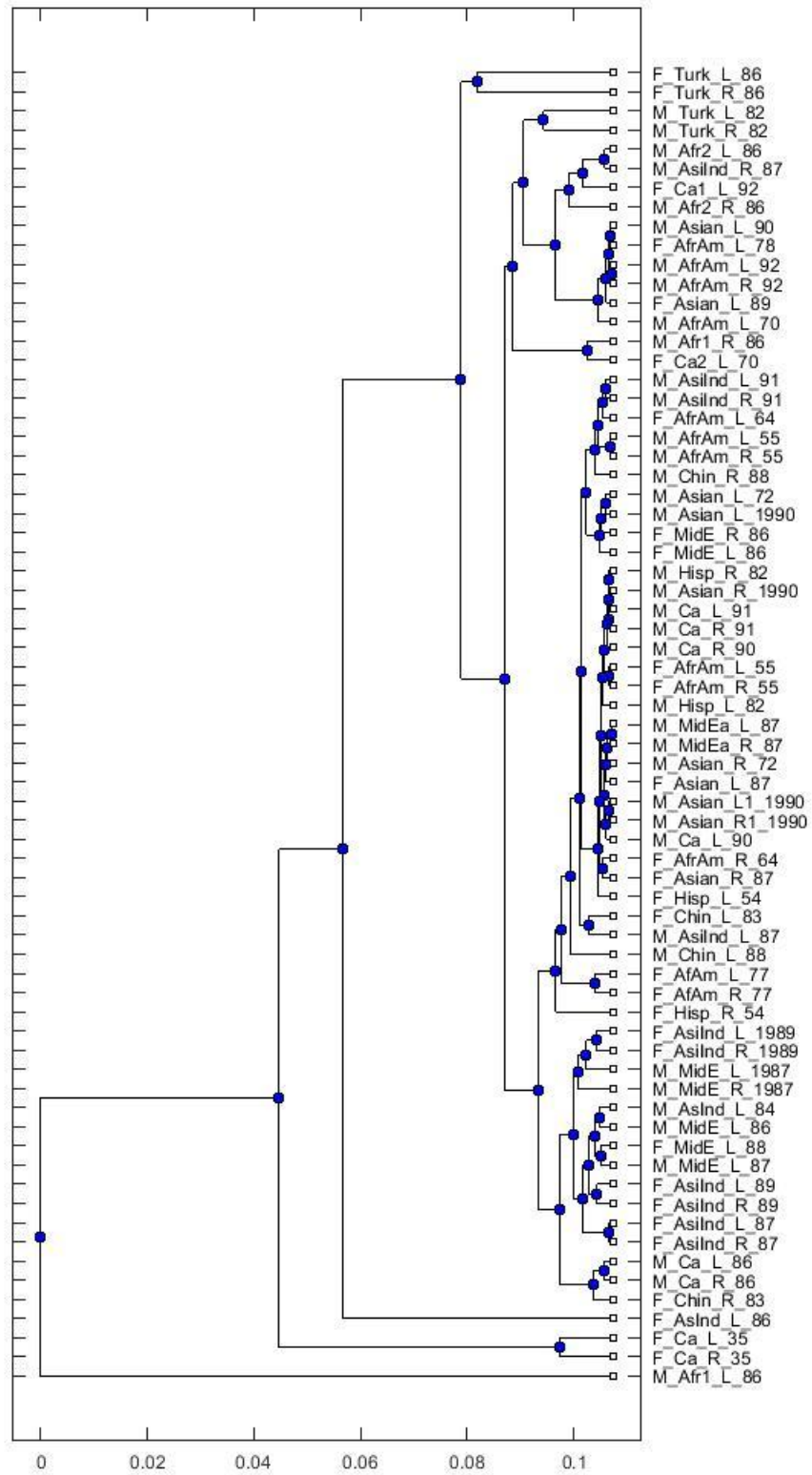


Fig 4.8 Phylogenetic tree based on KLD analysis of OTU frequencies

4.3.2 KLD analysis of unweighted k-mer frequencies:

KLD analysis of unweighted k-mer frequencies as explained in section 2.7.2 was applied for three cases of k=1, 2 and 3. Average Euclidean pairwise distance between reference and resultant distance matrices were 0.3806, 0.3479 and 0.3602 for k=1, 2 and 3 respectively. K-mer (k=2) frequencies produced less dissimilarity from reference distance matrix, hinting that phylogenetic tree in Fig 4.13 is more similar to reference phylogenetic tree than those in Fig 4.12 and 4.14. Moreover, longer branches in Fig 4.13 indicate that k=2 frequencies resulted in more widely spread cluster i.e., samples were comparatively more distinct in case of k=2. In the PCoA plots compared in Fig 4.9, 4.10 and 4.11, Hispanic samples are clustered distinct from rest of the population groups. Moreover, axes lengths suggest that k-mer frequencies for k=2 can be more distinctive than k=1 and k=3. PCoA plots on other population groups are listed in the Appendix B section of this thesis.

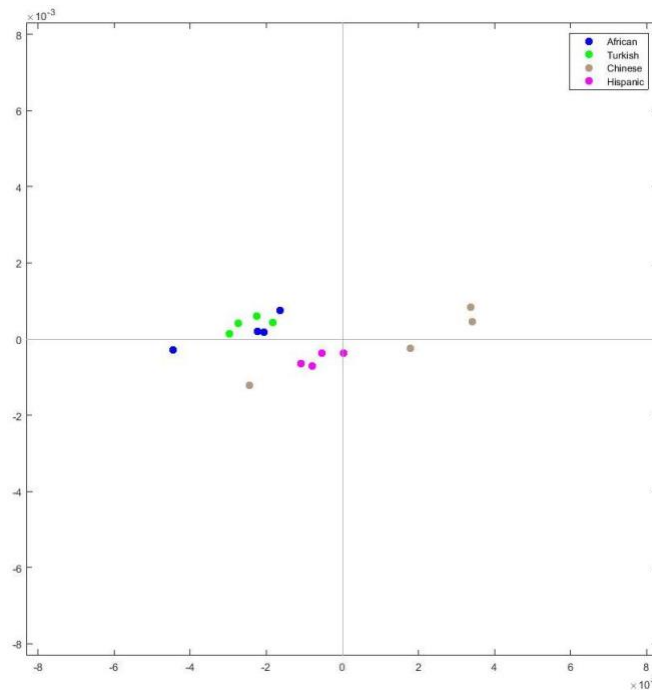


Fig 4.9 PCoA plot from KLD analysis of unweighted k-mer (k=1) frequencies for 16 samples from 4 population groups

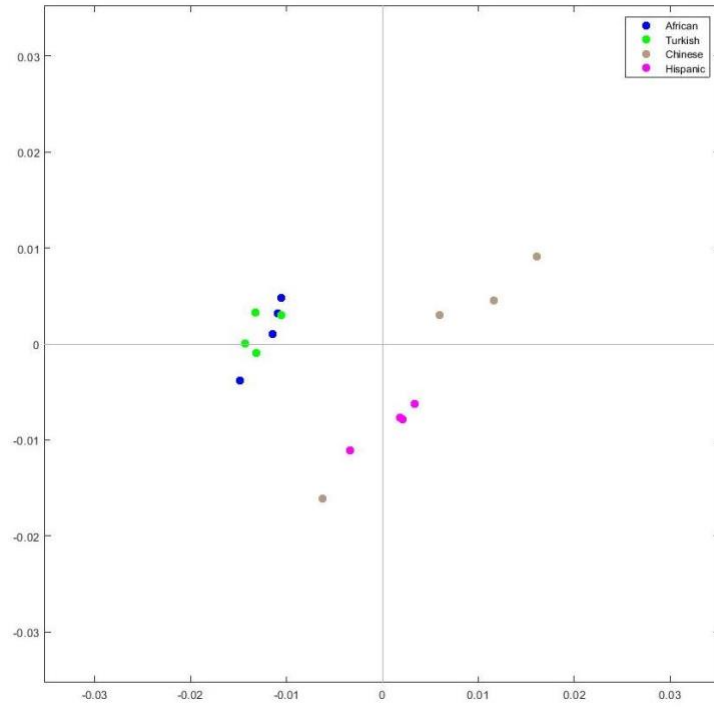


Fig 4.10 PCoA plot from KLD analysis of unweighted k-mer (k=2) frequencies for 16 samples from 4 population groups

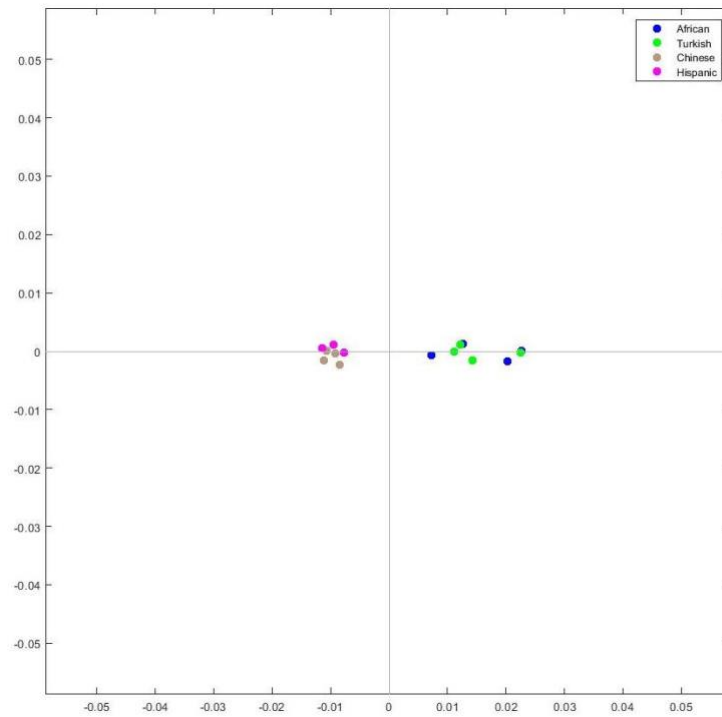


Fig 4.11 PCoA plot from KLD analysis of unweighted k-mer (k=3) frequencies for 16 samples from 4 population groups

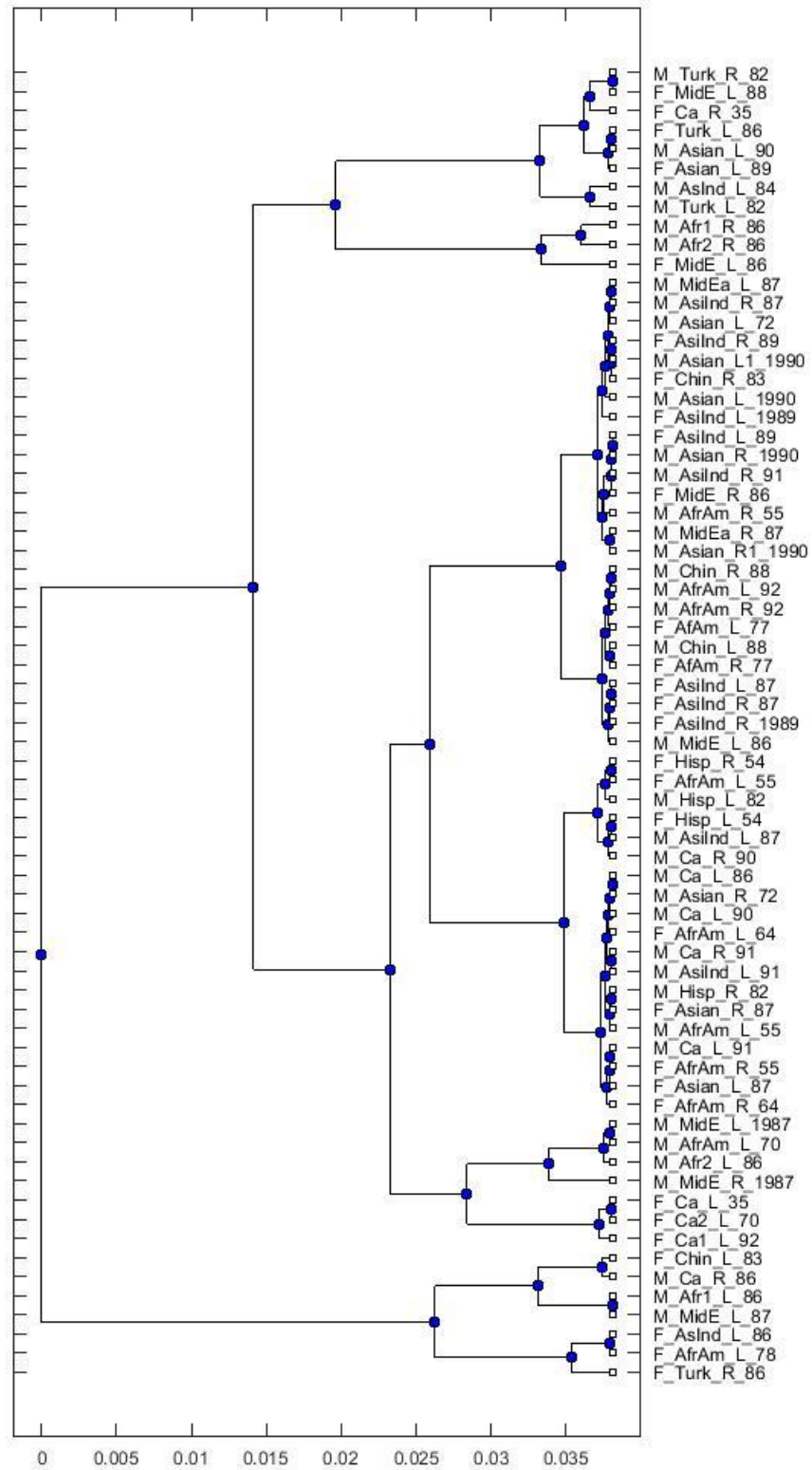


Fig 4.12 Phylogenetic tree based on KLD analysis of unweighted k-mer (k-1) frequencies

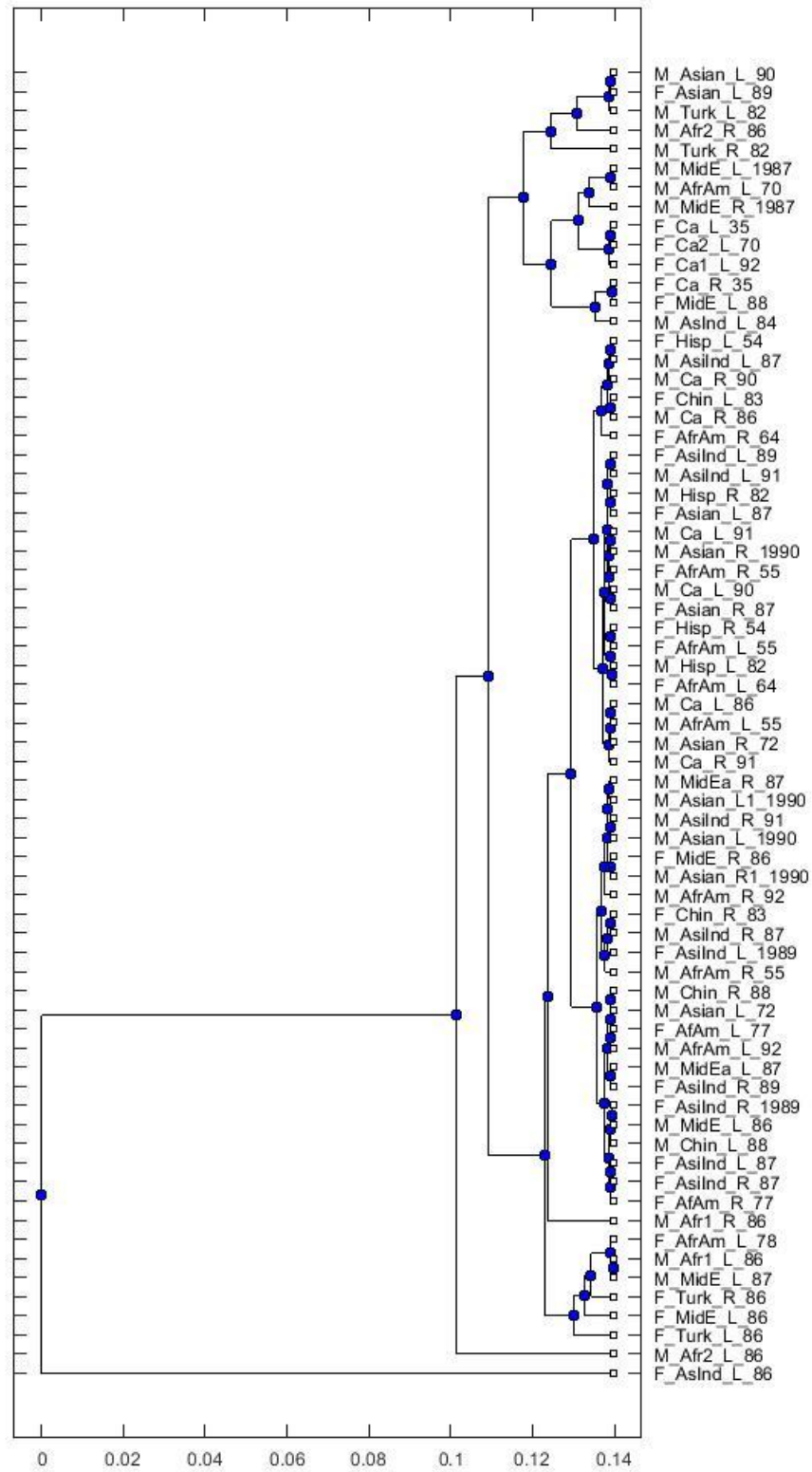


Fig 4.13 Phylogenetic tree based on KLD analysis of unweighted k-mer (k=2) frequencies

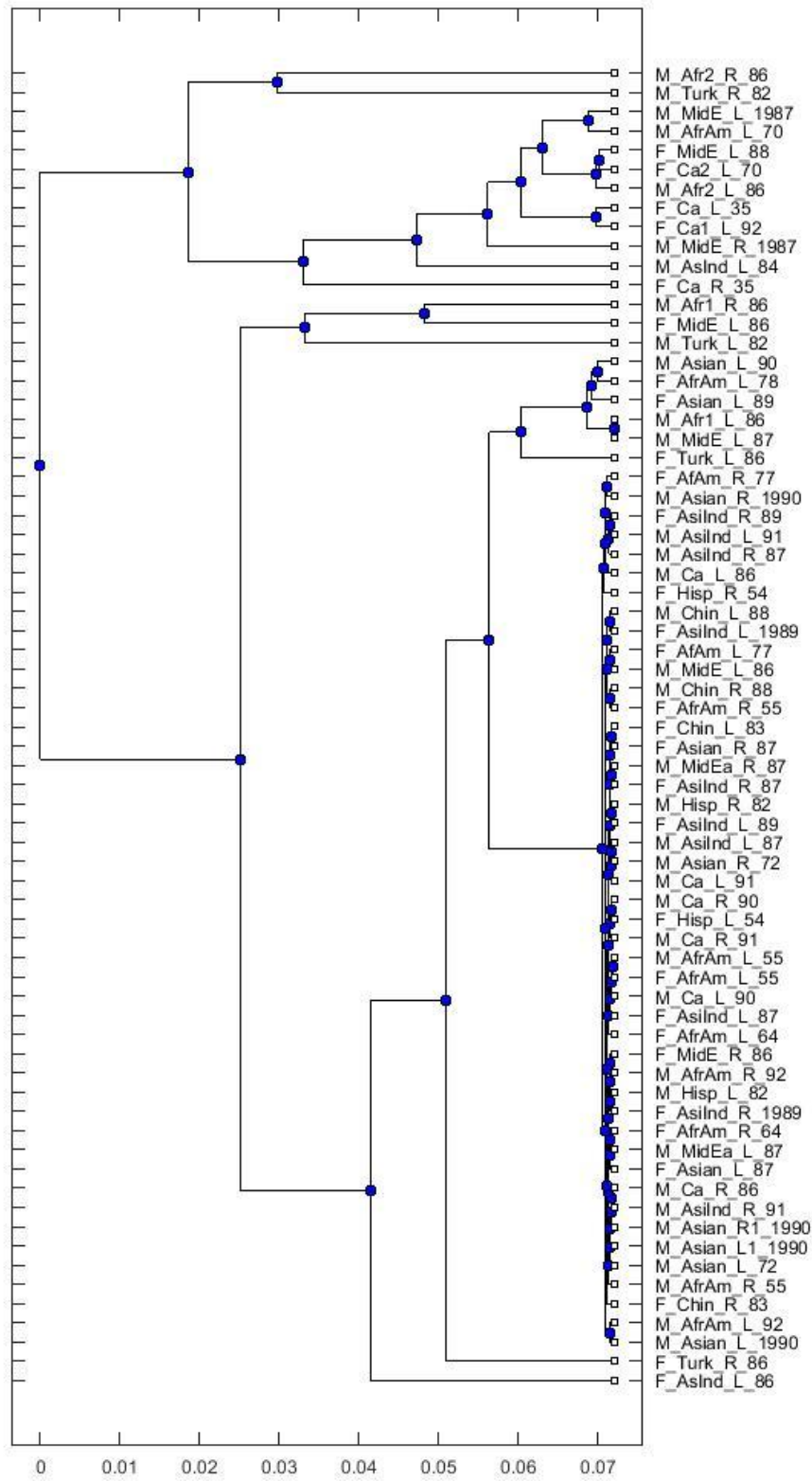


Fig 4.14 Phylogenetic tree based on KLD analysis of unweighted k-mer (k=3) frequencies

4.3.3 KLD analysis over weighted k-mer frequencies:

KLD analysis of weighted k-mer frequencies as explained in section 2.7.3 was applied for all three cases of k=1, 2 and 3. Average Euclidean pairwise distance between reference and resultant distance matrices were 0.3718, 0.3425 and 0.3630 for k=1, 2 and 3 respectively. Similar to unweighted frequencies, distance matrix from KLD analysis on k-mer (k=2) frequencies resulted in less dissimilarity from reference distance matrix. Longer branch lengths in phylogenetic tree shown in Fig 4.19 suggest that k-mer frequencies with k=2 can be more effective in distinguishing samples than k=1 (Fig 4.18) and k=2 (Fig 4.20). When compared to unweighted frequencies, weighted frequencies produced less dissimilarity between reference and resultant distance matrices. Fig 4.16 shows that samples in case of k-mer frequencies with k=2 are more widely spread than in case of k=1 (Fig 4.15) and k=3 (Fig 4.17), indicating that k=2 frequencies are more distinguishing. PCoA plots on other population groups are listed in the Appendix C section of this thesis.

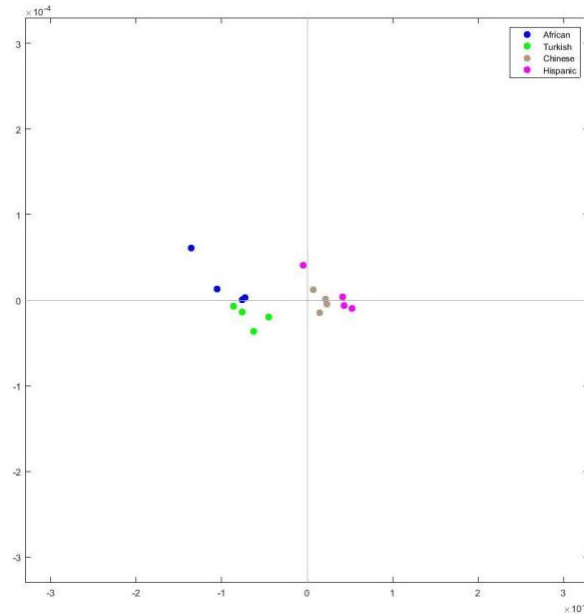


Fig 4.15 PCoA plot from KLD analysis of weighted k-mer (k=1) frequencies of 16 samples from 4 population groups

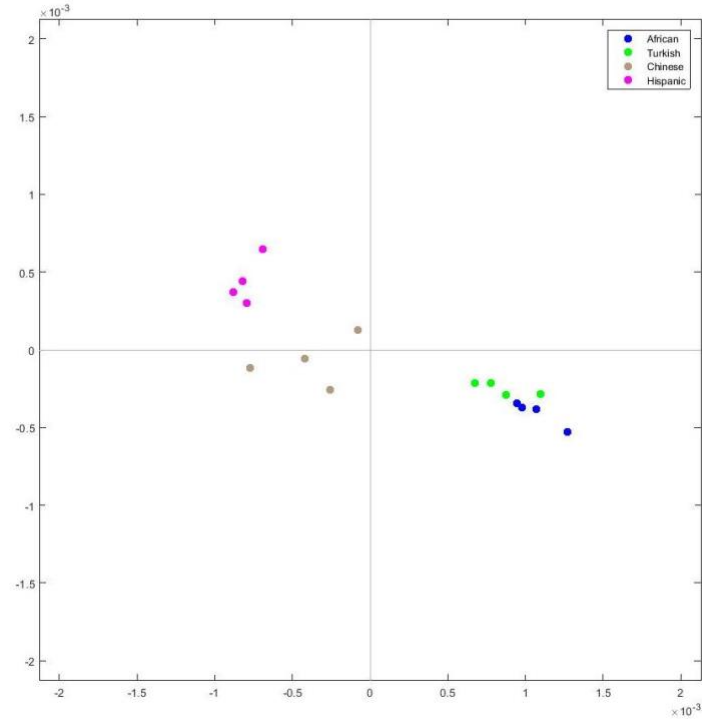


Fig 4.16 PCoA plot from KLD analysis of weighted k-mer ($k=2$) frequencies for 16 samples from 4 population groups

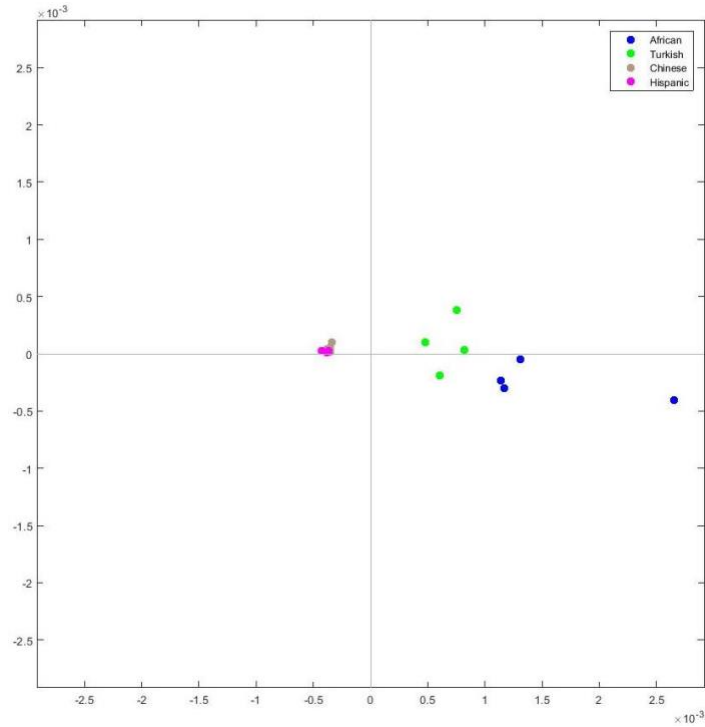


Fig 4.17 PCoA plot from KLD analysis of weighted k-mer ($k=3$) frequencies for 16 samples from 4 population groups

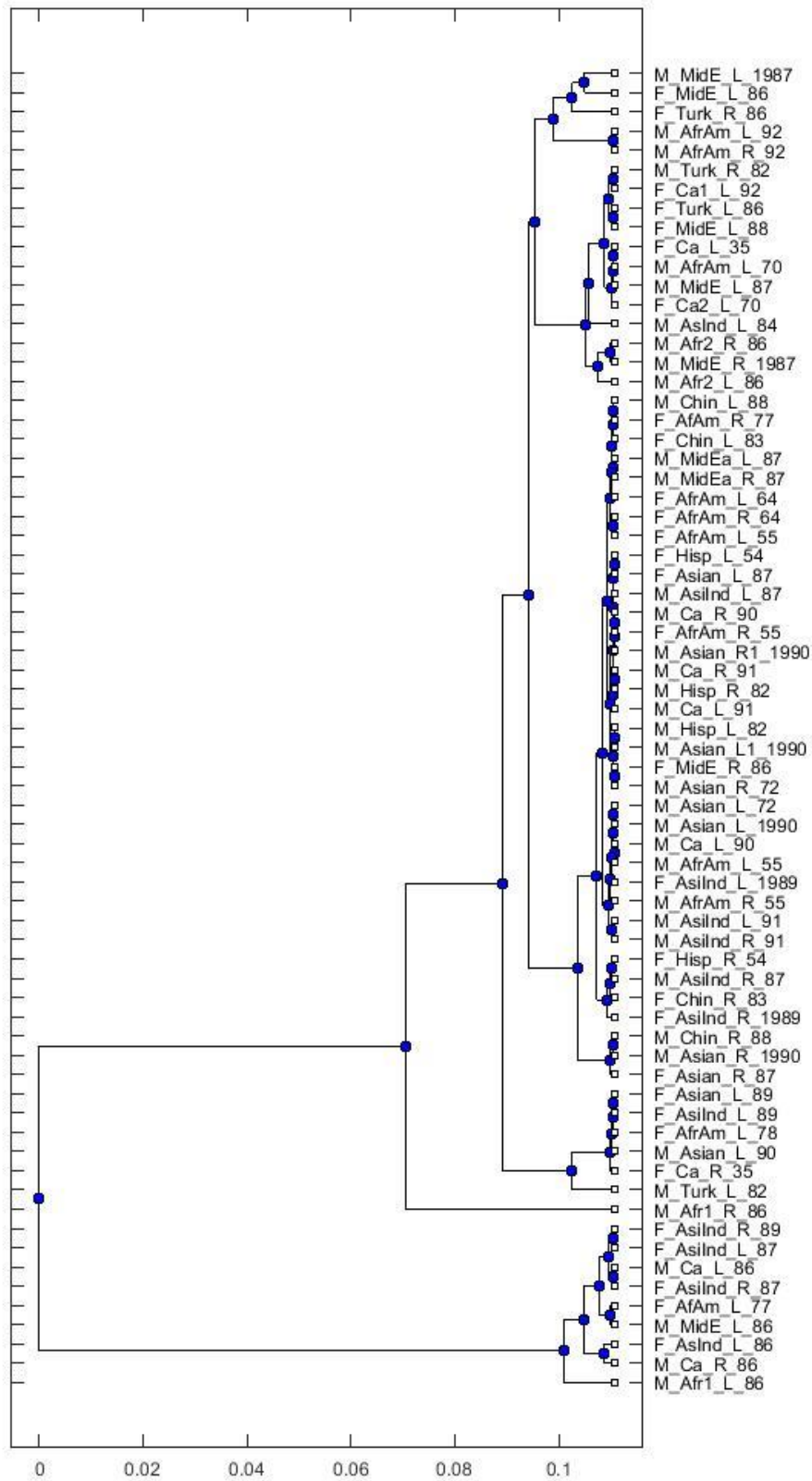


Fig 4.18 Phylogenetic tree based on KLD analysis of weighted k-mer (k-1) frequencies

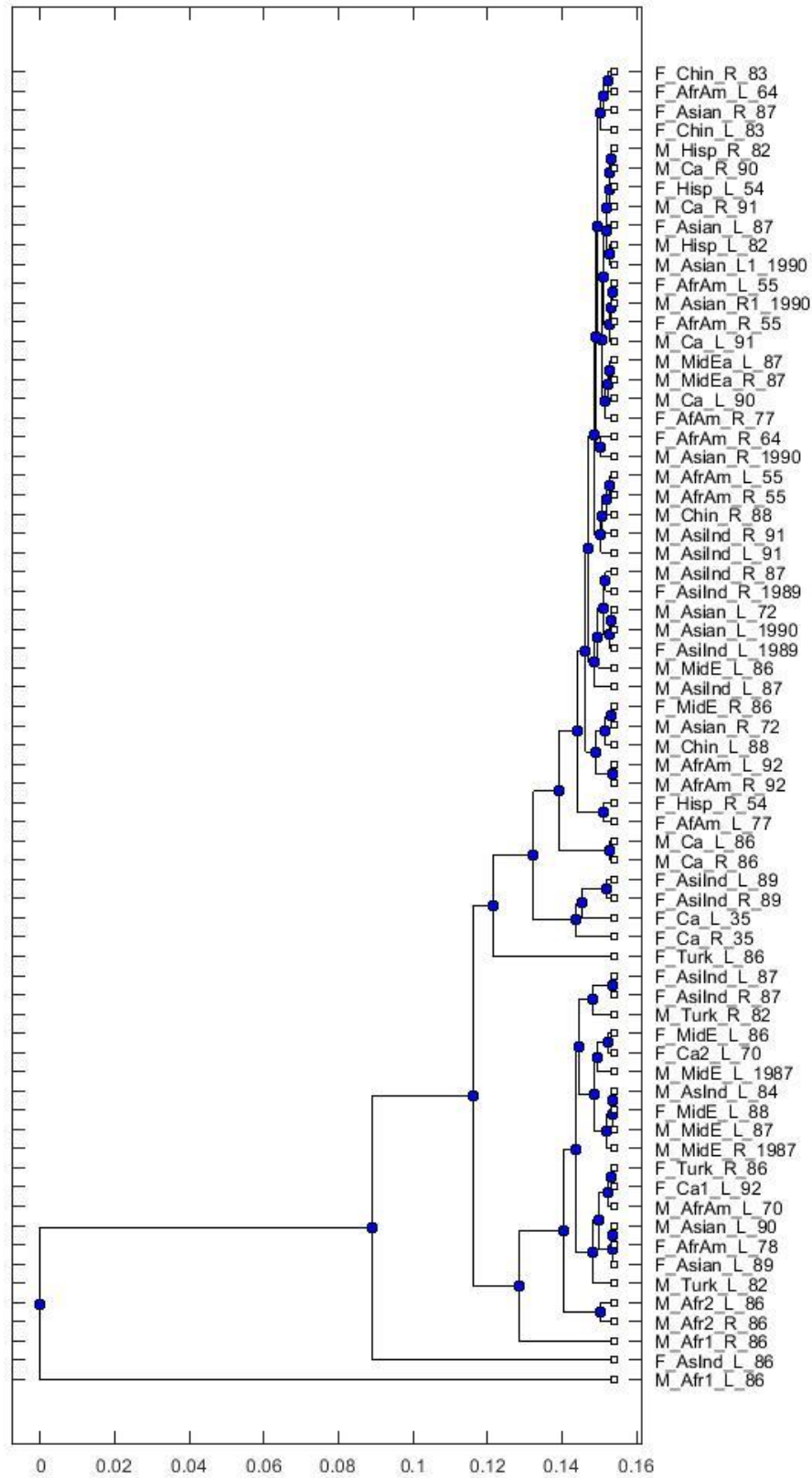


Fig 4.19 Phylogenetic tree based on KLD analysis of weighted k-mer (k=2) frequencies

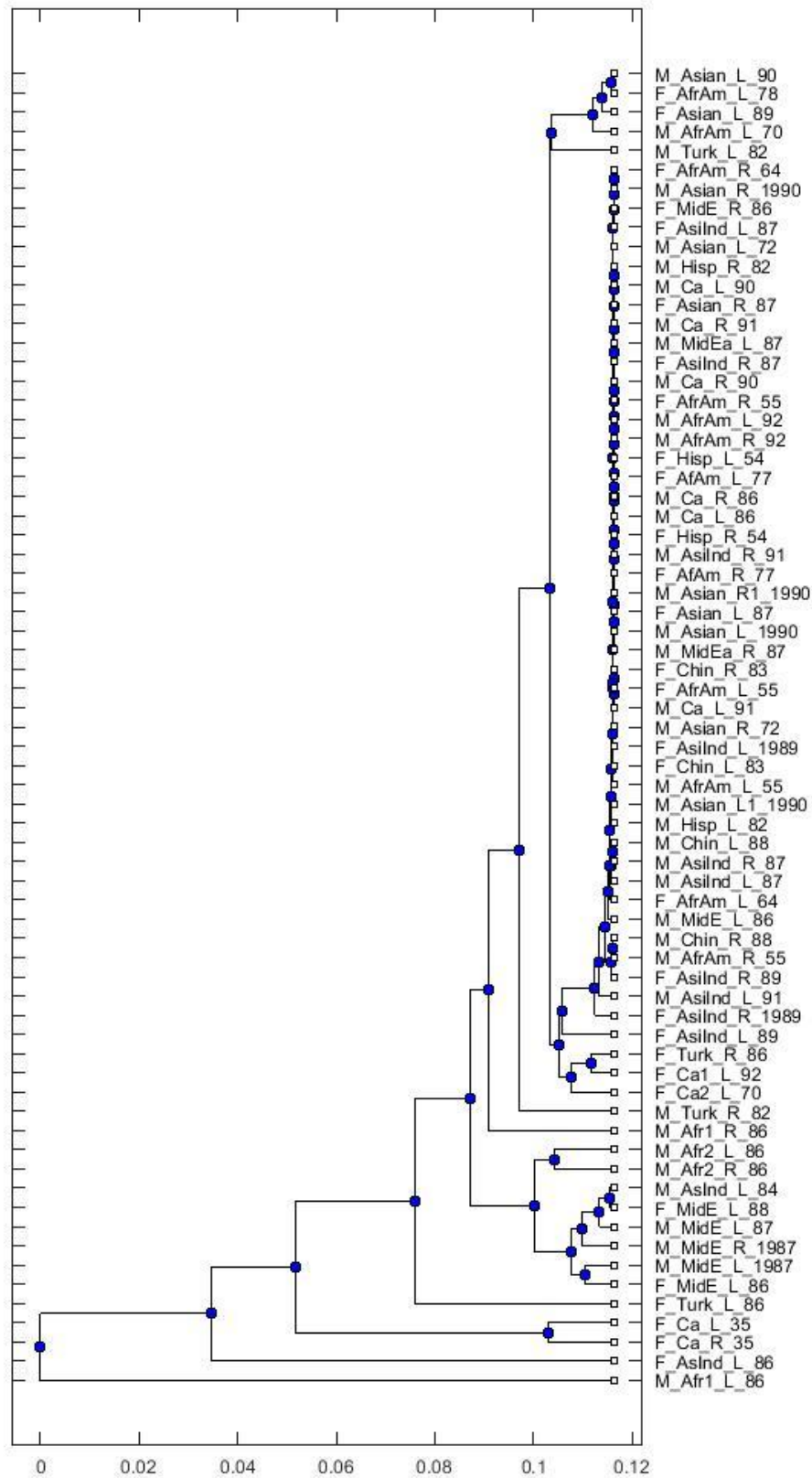


Fig 4.20 Phylogenetic tree based on KLD analysis of weighted k-mer (k=3) frequencies

4.3.4 KLD analysis over unweighted k-mer frequencies from mapped sequences:

After mapping sequences from V3 hypervariable region as explained in section 2.8, KLD analysis was repeated on unweighted and weighted frequencies of mapped OTU sequences using equations from sections 2.7.2 and 2.7.3 respectively. Figures 4.22 and 4.23 are the phylogenetic trees from KLD on unweighted k-mer frequencies from mapped OTU sequences. 0.3797 and 0.3363 were the average Euclidean pair-wise distances between resultant and reference distance matrices for k-1 and k-2 respectively. Comparatively, Fig 4.23 has longer branches, indicating that unweighted mapped k-2 frequencies are better in distinguishing in samples. Both PCoA plots in Fig 4.21 show distinct clustering of Hispanic samples but there is no significant record of which case of k-mer yields better distinction among samples. PCoA plots on other population groups are listed in the Appendix D section of this thesis.

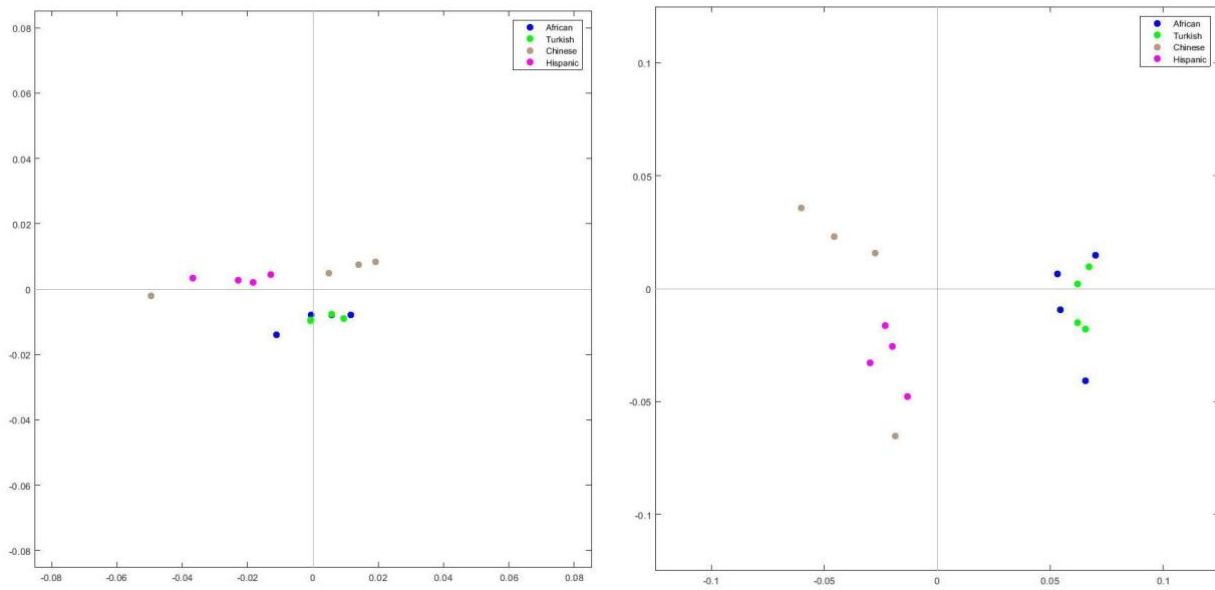


Fig 4.21 PCoA plot from KLD analysis of unweighted k-mer (k-1 on the left; k-2 on the right) frequencies of 16 samples from 4 population groups

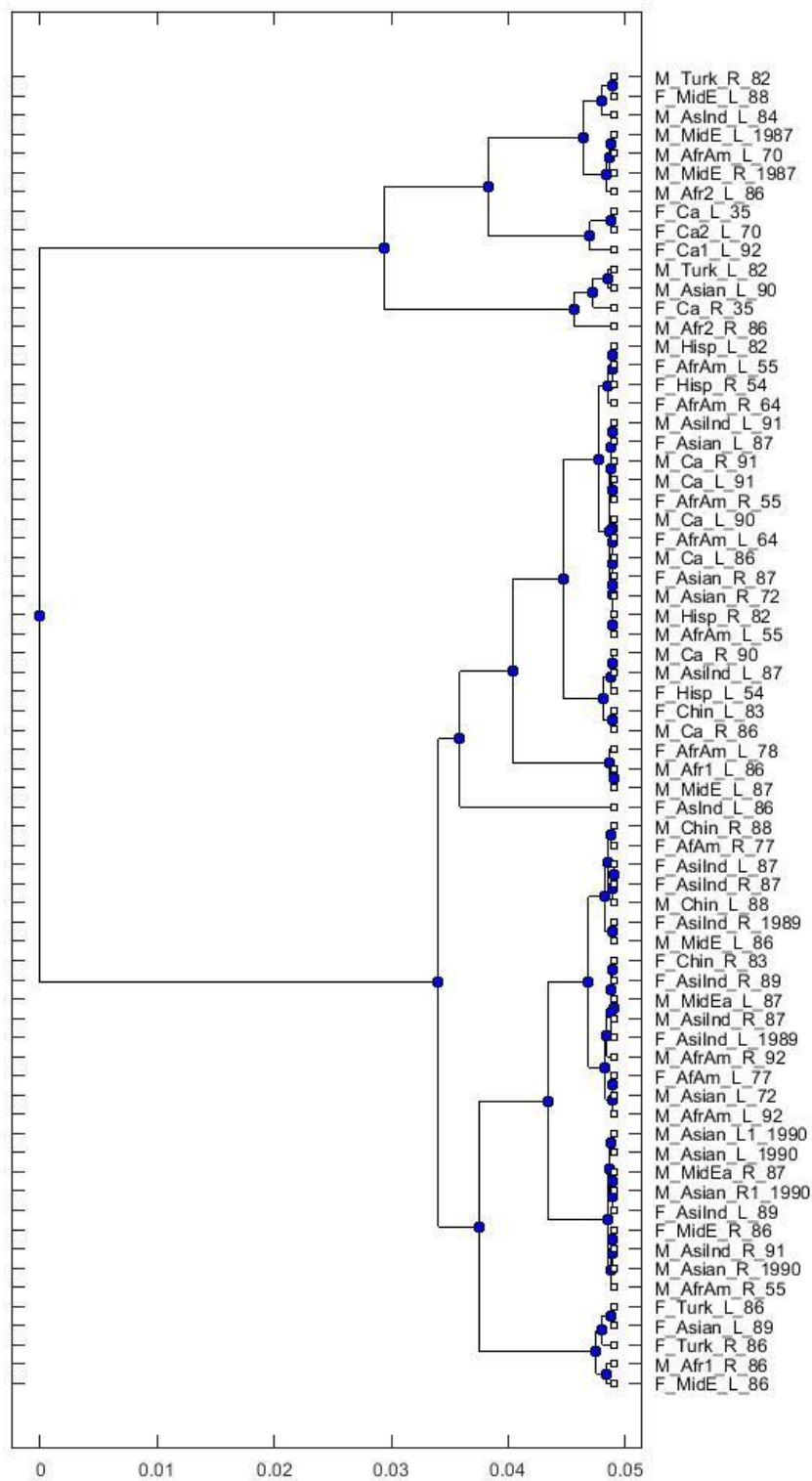


Fig 4.22 Phylogenetic tree based on KLD analysis of unweighted k-mer (k-1) frequencies from mapped sequences

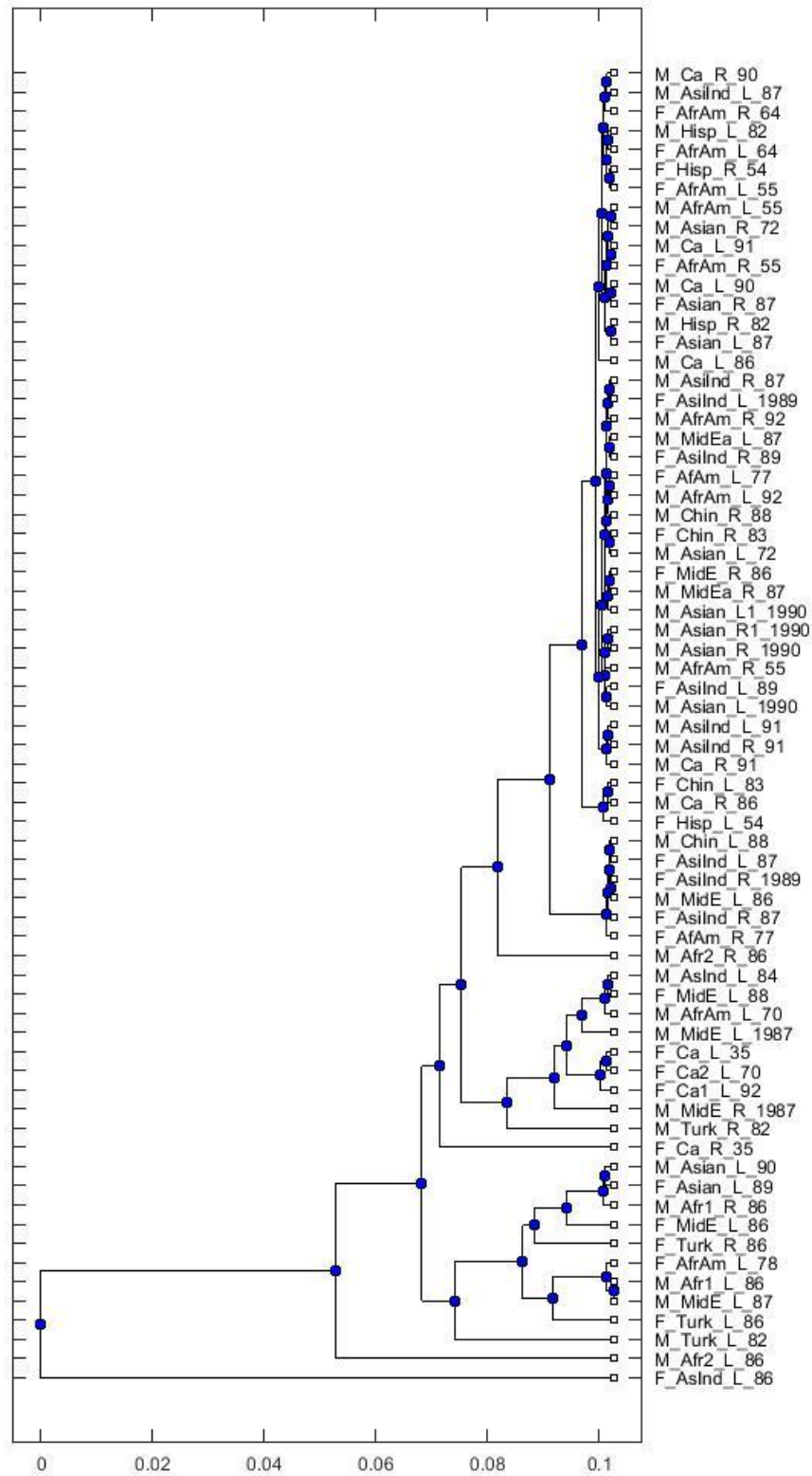


Fig 4.23 Phylogenetic tree based on KLD analysis of unweighted k-mer (k-2) frequencies from mapped sequences

4.3.5 KLD analysis over weighted k-mer frequencies from mapped sequences:

Figures 4.25 and 4.26 are the phylogenetic trees from KLD on weighted k-mer frequencies from mapped OTU sequences. 0.3787 and 0.3315 were the average Euclidean pair-wise distances between resultant and reference distance matrices for k-1 and k-2 respectively. Among weighted and unweighted frequencies from both unmapped and mapped sequences, weighted k-2 frequencies from mapped OTU sequences produced relatively effective results with least dissimilarity from reference distance matrix. Phylogenetic tree from k-2 frequencies (Fig 4.26) has longer branches suggesting that k-mer frequencies with k-2 are more distinguishing than k-1 (Fig 4.25). Fig 4.24 shows that samples in case of k-2 frequencies are more widely spread than in case of k-1, indicating that k-2 frequencies are more distinguishing. PCoA plots on other population groups are listed in the Appendix E section of this thesis.

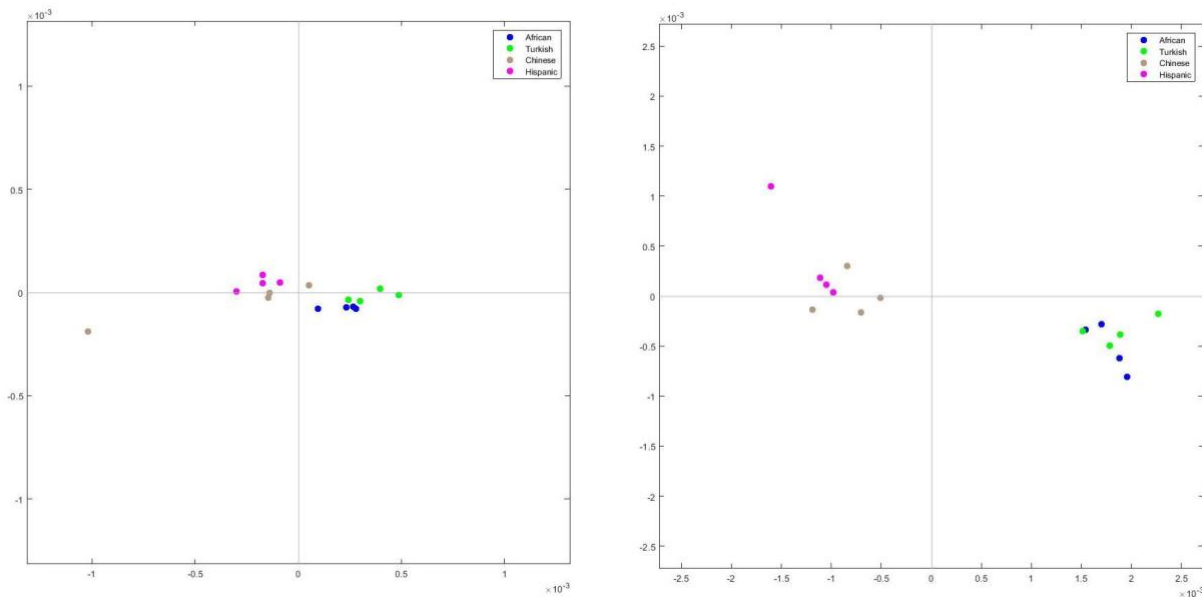


Fig 4.24 PCoA plot from KLD analysis of weighted k-mer (k-1 on the left; k-2 on the right) frequencies of 16 samples from 4 population groups

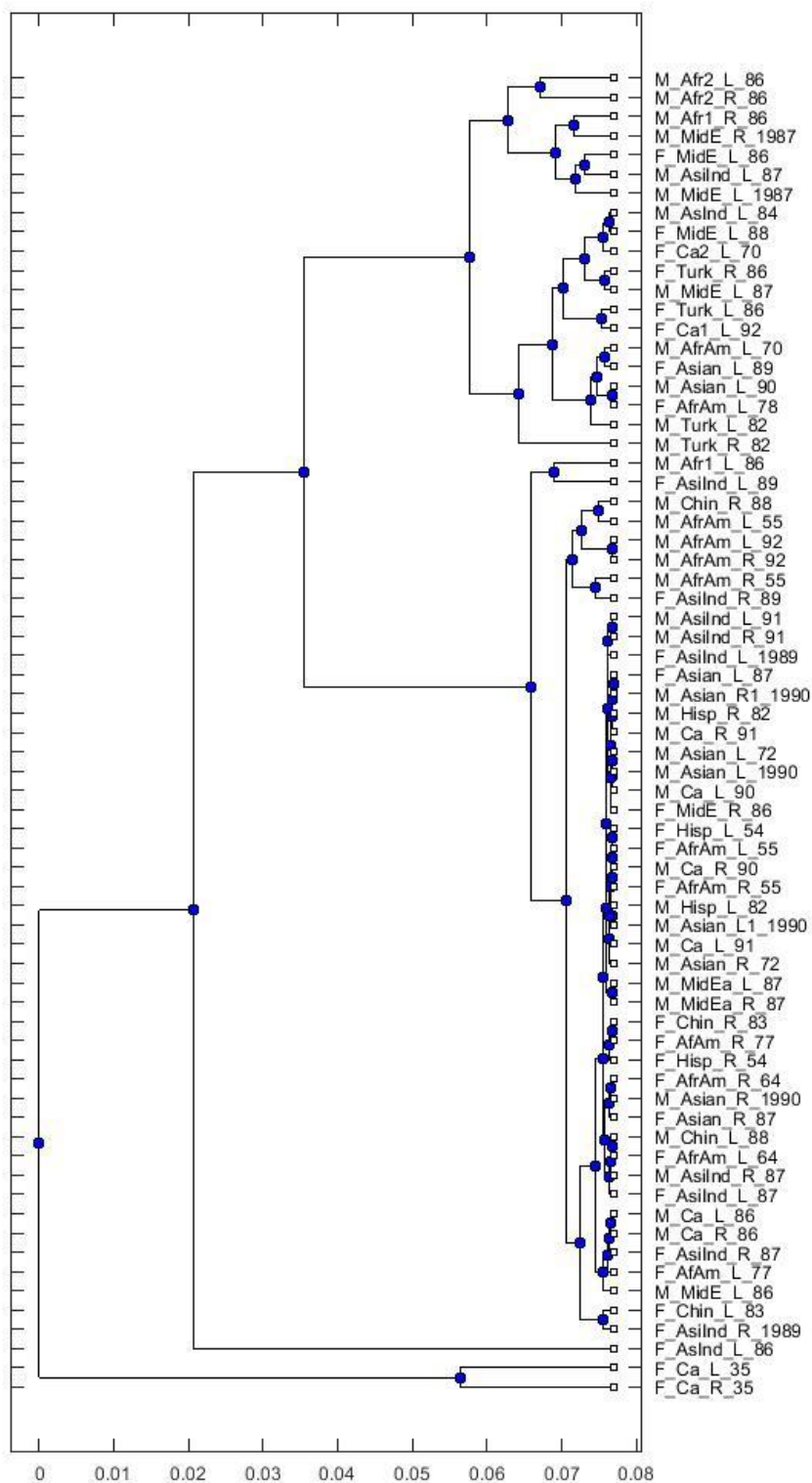


Fig 4.25 Phylogenetic tree based on KLD analysis of weighted k-mer (k-1) frequencies from mapped sequences

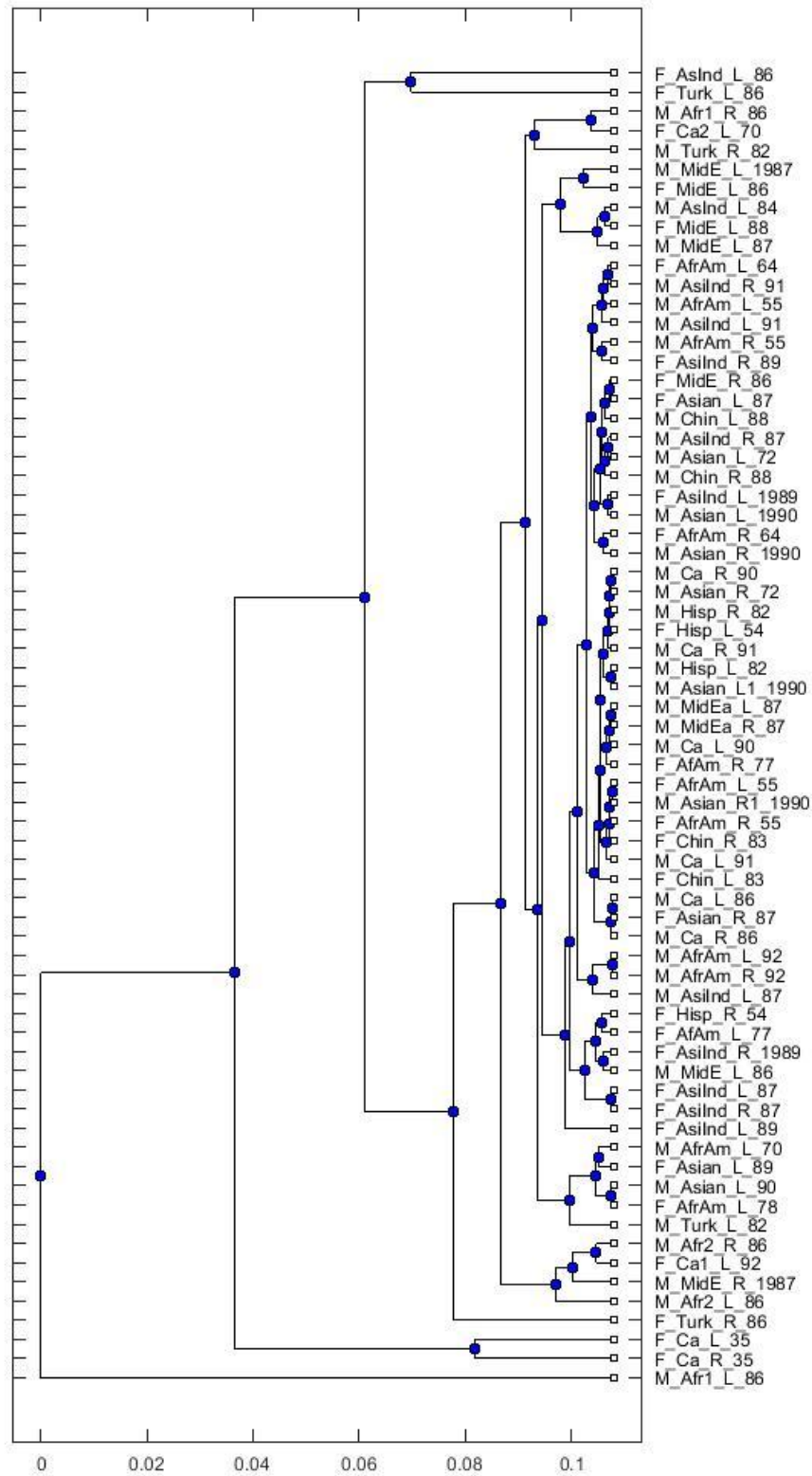


Fig 4.26 Phylogenetic tree based on KLD analysis of weighted k-mer (k=2) frequencies from mapped sequences

4.4 Unique signatures of k-mer frequencies

In previously mentioned methods, frequencies and k-mers of identified and classified OTUs were used where as the sequences which could not be identified as an OTU by the GreenGenes reference database were ignored. In an effort to include the unclassified sequences, a novel technique is used to classify the sequences depending on their k-mer profiles. FASTA files of representative sequences from QIIME were aligned and used to extract 65 nucleotides of V3 hypervariable region. K-mer frequencies, both k-1 and k-2 were calculated using Bioinformatics toolbox in MATLAB. Unique k-mer frequencies from the entire data set were found and their frequency in each sample was recorded. Additionally, sequences were mapped as explained in section 2.8, and unique k-mer frequencies (k-1 and k-2) and their frequencies in each sample were calculated. Number of unique k-mer frequencies found in the entire data for 4 different cases are put in table 4.1.

Table 4.1 Number of unique k-mer frequencies found in different cases of k

k-mer	Unmapped	Mapped
k-1	3459	32731
k-2	35647	37286
k-3	26529	

K-mer frequencies were then clustered using K-means clustering to compress the data. To decide the number of clusters (value of K), average silhouette and number of negative silhouette values for different cluster sizes are calculated and compared. Value of K that resulted in higher average silhouette value and less negative silhouette values was chosen for different cases of k-mer from figures 4.27 to 4.31. The chosen value of K for each case of k-mer is put in table 4.2

Table 4.2 Number of clusters (K) chosen for different cases of k

k-mer	Unmapped	Mapped
k-1	225	100
k-2	150	275
k-3	450	

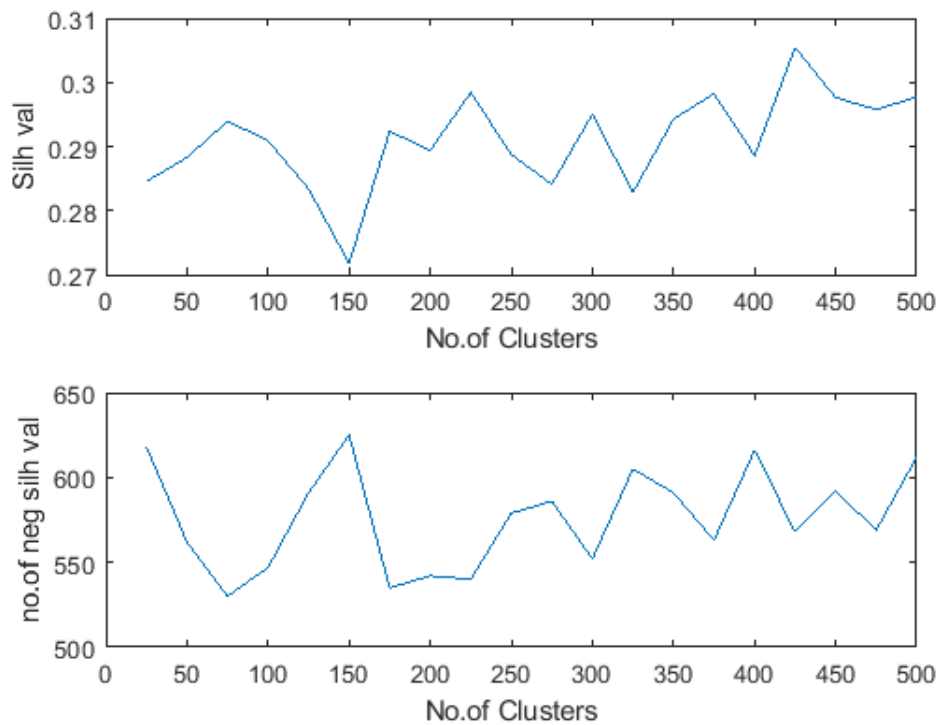


Fig 4.27 Average Silhouette values and number of negative silhouette values for different number of clusters for k-1

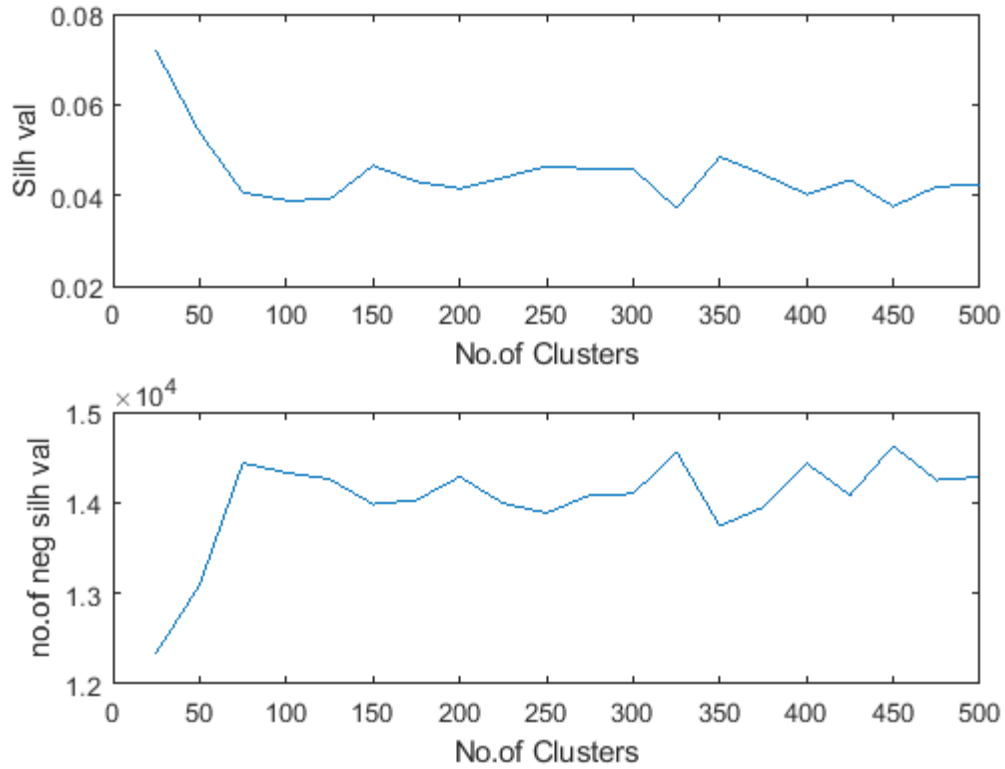


Fig 4.28 Average Silhouette values and number of negative silhouette values for different number of clusters for k-2

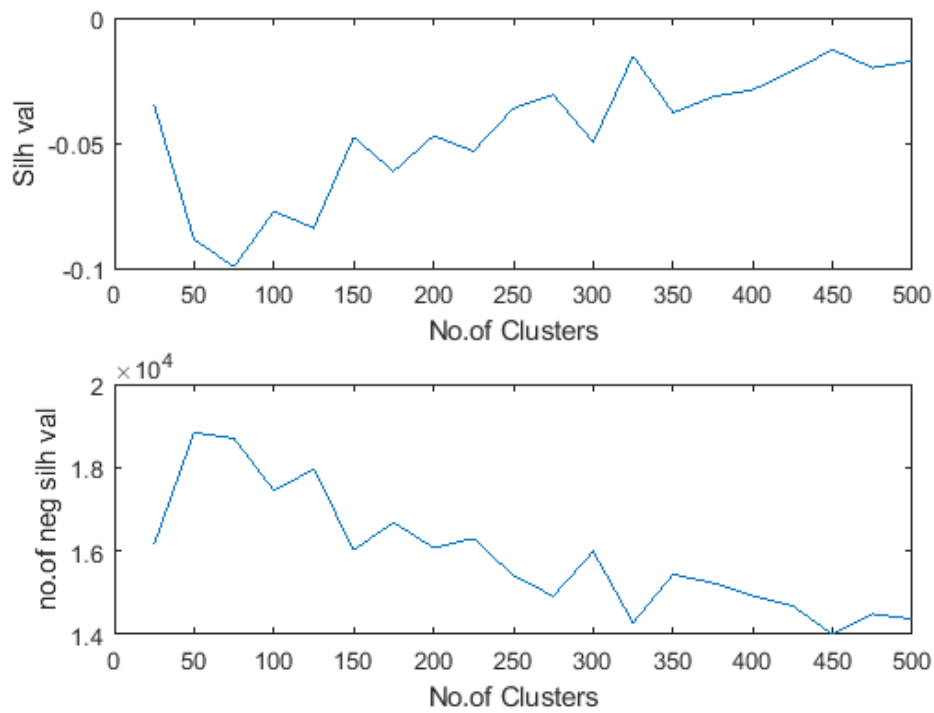


Fig 4.29 Average Silhouette values and number of negative silhouette values for different number of clusters for k-3

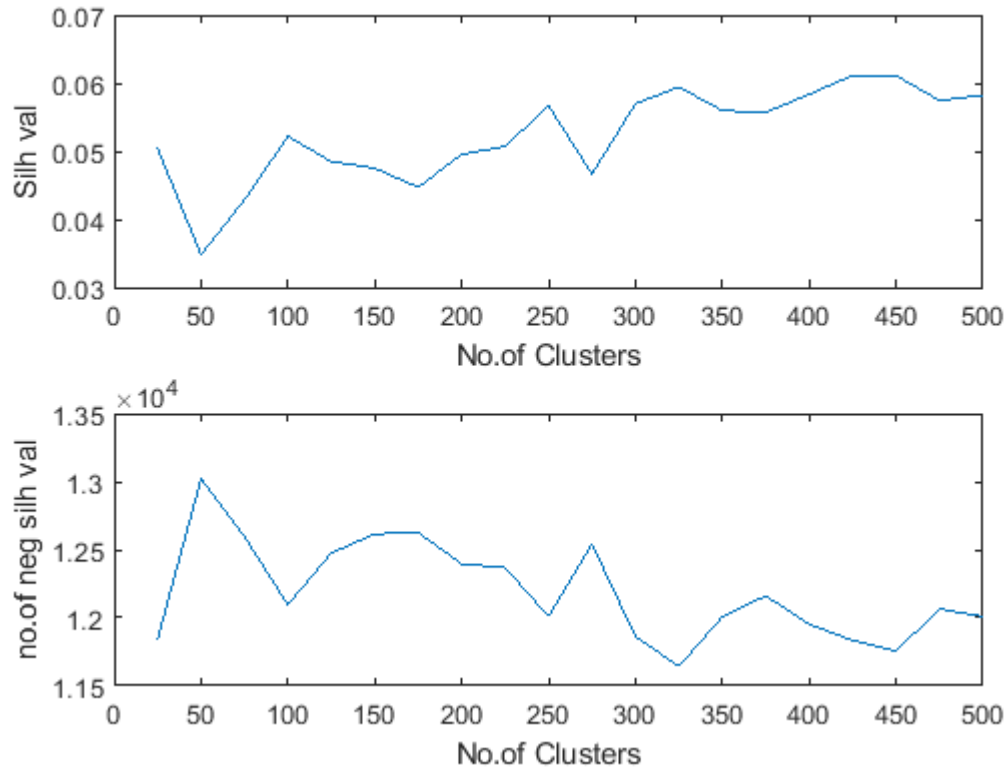


Fig 4.30 Average Silhouette values and number of negative silhouette values for different number of clusters for k-1 from mapped sequences

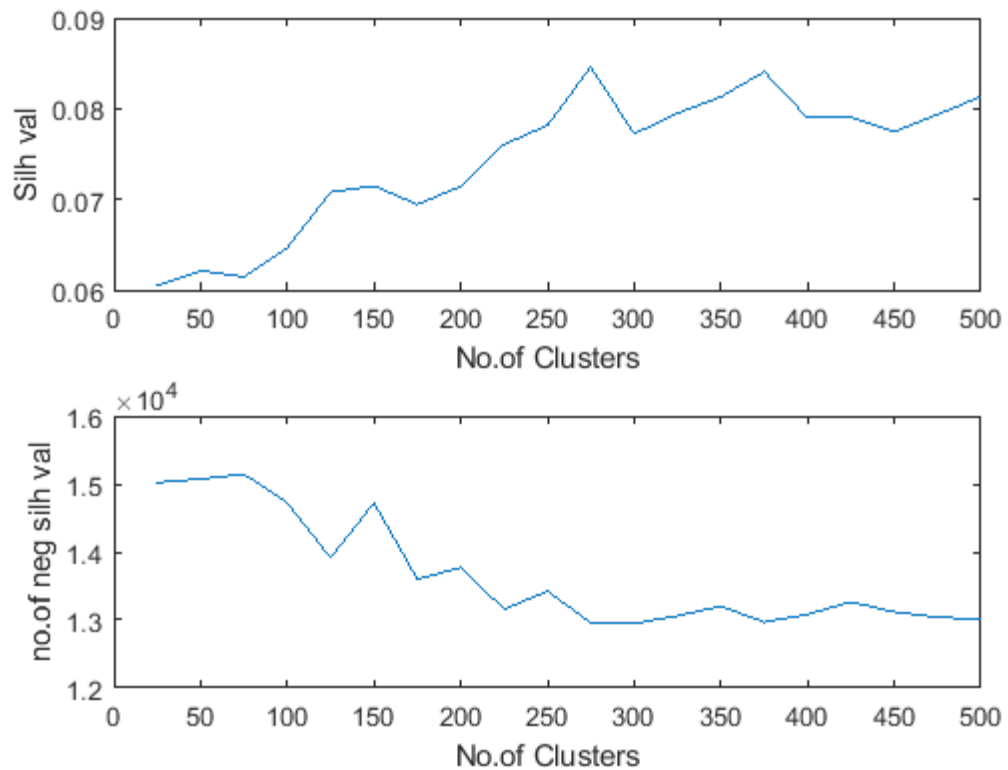


Fig 4.31 Average Silhouette values and number of negative silhouette values for different number of clusters for k-2 from mapped sequences

After clustering, mean of k-mer frequencies in each cluster was considered as the signature of the cluster and the sum of frequencies of k-mer frequencies within a cluster was considered the occurrence of that particular cluster. Occurrence of clusters in each sample were used to calculate the frequencies of clusters in a sample. These frequencies were further used to calculate KLD between samples and build distance matrix using the equations 4.1 and 4.2 where value of t= {225, 150, 450, 100 and 275} for k-1, k-2, k-3 mapped k-1 and mapped k-2 respectively.

$$D_{KLD}^{sign}(S_i||S_j) = \sum_{r=1}^t Q_i^r \log \frac{Q_i^r}{Q_j^r} \quad \text{Equation 4.1}$$

$$d_{i=1-n, j=1-n}^{sign} = \frac{1}{2} [D_{KLD}^{sign}(S_i||S_j) + D_{KLD}^{sign}(S_j||S_i)] \quad \text{Equation 4.2}$$

4.5 Unsupervised machine learning of population groups using unique k-mer frequencies

Frequencies of clusters that were built from unique k-mer frequencies in section 4.4 were used to apply KLD analysis between samples. Figures 4.32, 4.33 and 4.34 are the phylogenetic trees from KLD analysis on signatures from k-mer frequencies with k-1, k-2 and k-3 respectively. Dissimilarity from reference phylogenetic tree that is found by calculating average Euclidean Pair-wise distances between resultant and reference distance matrices for k=1, 2 and 3 were 0.3189, 0.3298 and 0.3232 respectively suggesting that unique k-mer signatures for k=1 are more similar to reference phylogenetic tree than k-2 and k-3; moreover, longer branch lengths in Fig 4.32 suggests that signatures from unique k-1 frequencies are more distinguishing than signatures from unique k-2 and k-3 frequencies. PCoA plots on various population groups are listed in the Appendix F section of this thesis.

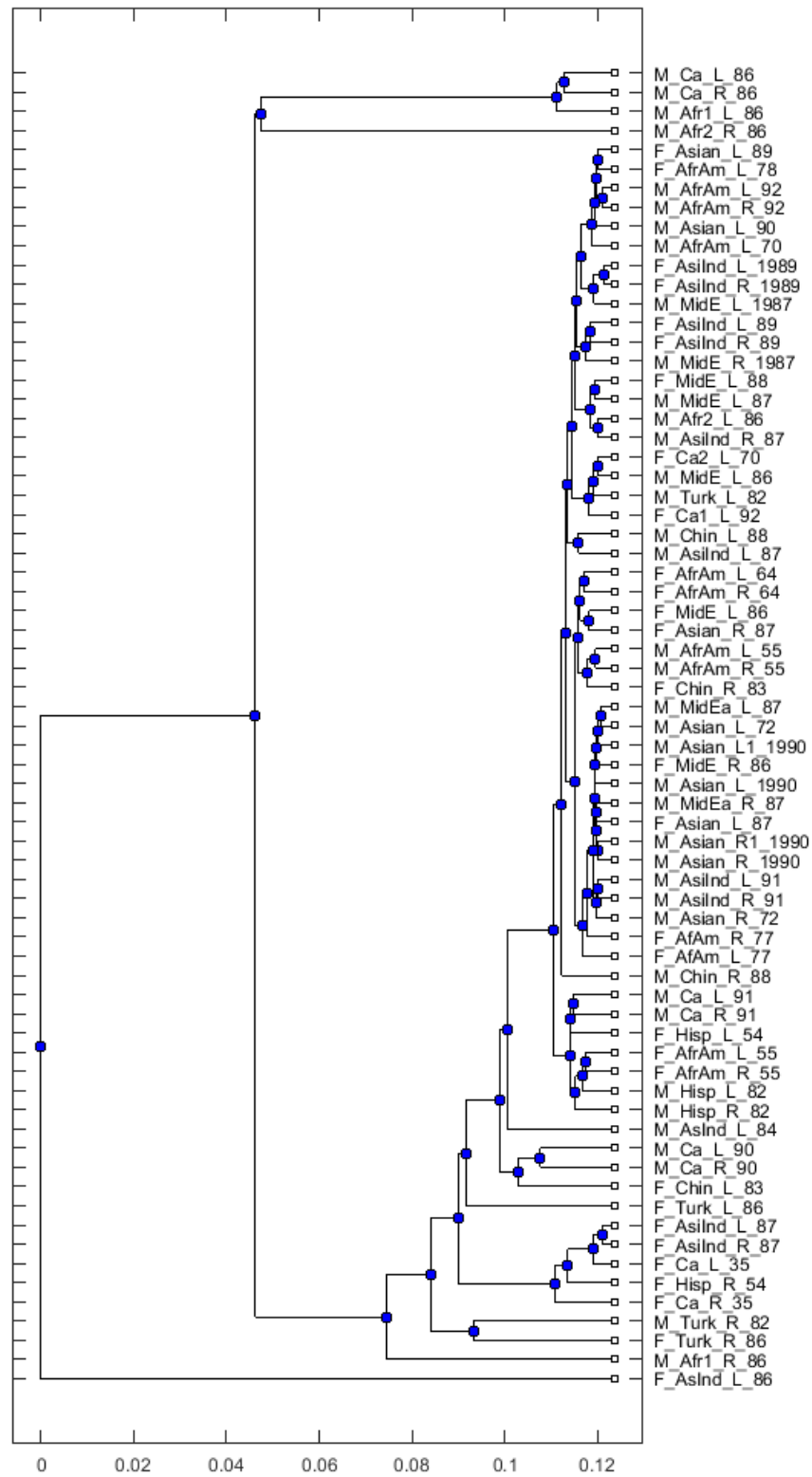


Fig 4.32 Phylogenetic tree based on KLD analysis on unique signatures of k-mer (k-1) frequencies

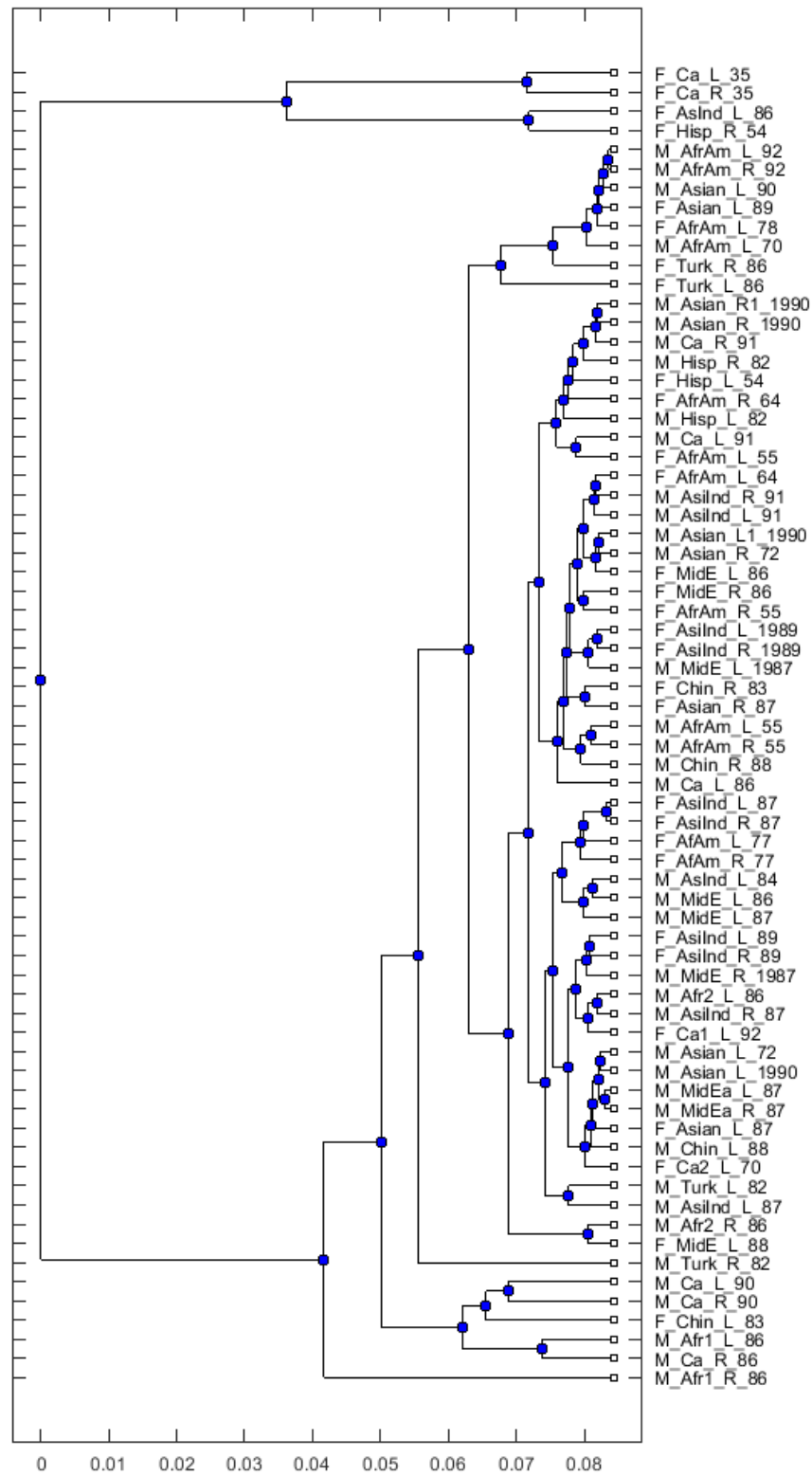


Fig 4.33 Phylogenetic tree based on KLD analysis on unique signatures of k-mer (k-2) frequencies

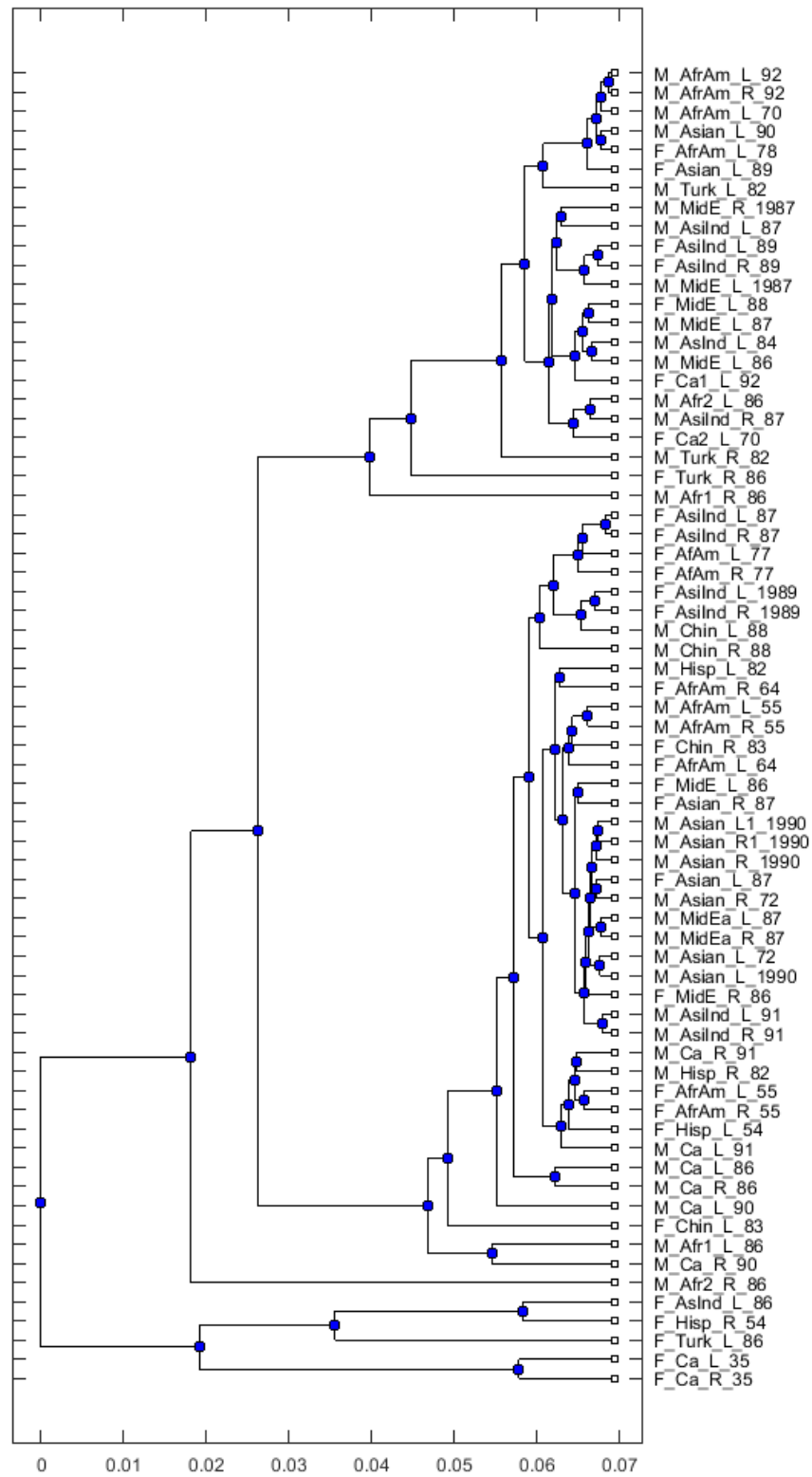


Fig 4.34 Phylogenetic tree based on KLD analysis on unique signatures of k-mer (k-3) frequencies

Furthermore, frequencies of clusters that were built from unique k-mer frequencies, with $k=1$ and 2 , from mapped sequences were used to apply KLD analysis between samples. Figures 4.36 and 4.37 are the respective phylogenetic trees from $k=1$ and $k=2$ frequencies. Average Euclidean Pair-wise distances between resultant and reference distance matrices for $k=1$ and $k=2$ were 0.3409 and 0.3084 respectively. Signatures from unique k-mer frequencies ($k=1$) from mapped sequences yielded the least dissimilarity among on all the cases that were applied in this work; moreover, distinct and longer branch lengths in Fig 4.37 suggest that signatures from unique $k=2$ frequencies distinguish better. PCoA plot on the right of Fig 4.35 is fairly more distinct than that on the left indicating that signatures from unique $k=2$ frequencies from mapped sequences are more distinguishing. PCoA plots on other population groups are listed in the Appendix section of this thesis.

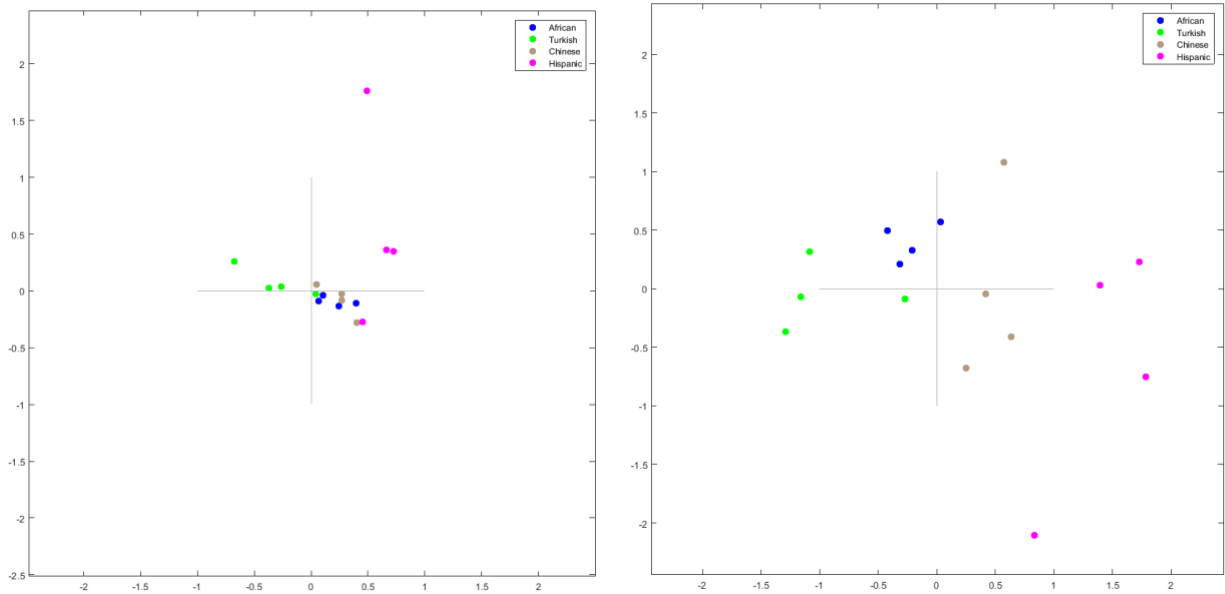


Fig 4.35 PCoA plot from KLD analysis of signatures from unique k-mer ($k=1$ on the left; $k=2$ on the right) frequencies from mapped sequences of 16 samples from 4 population groups

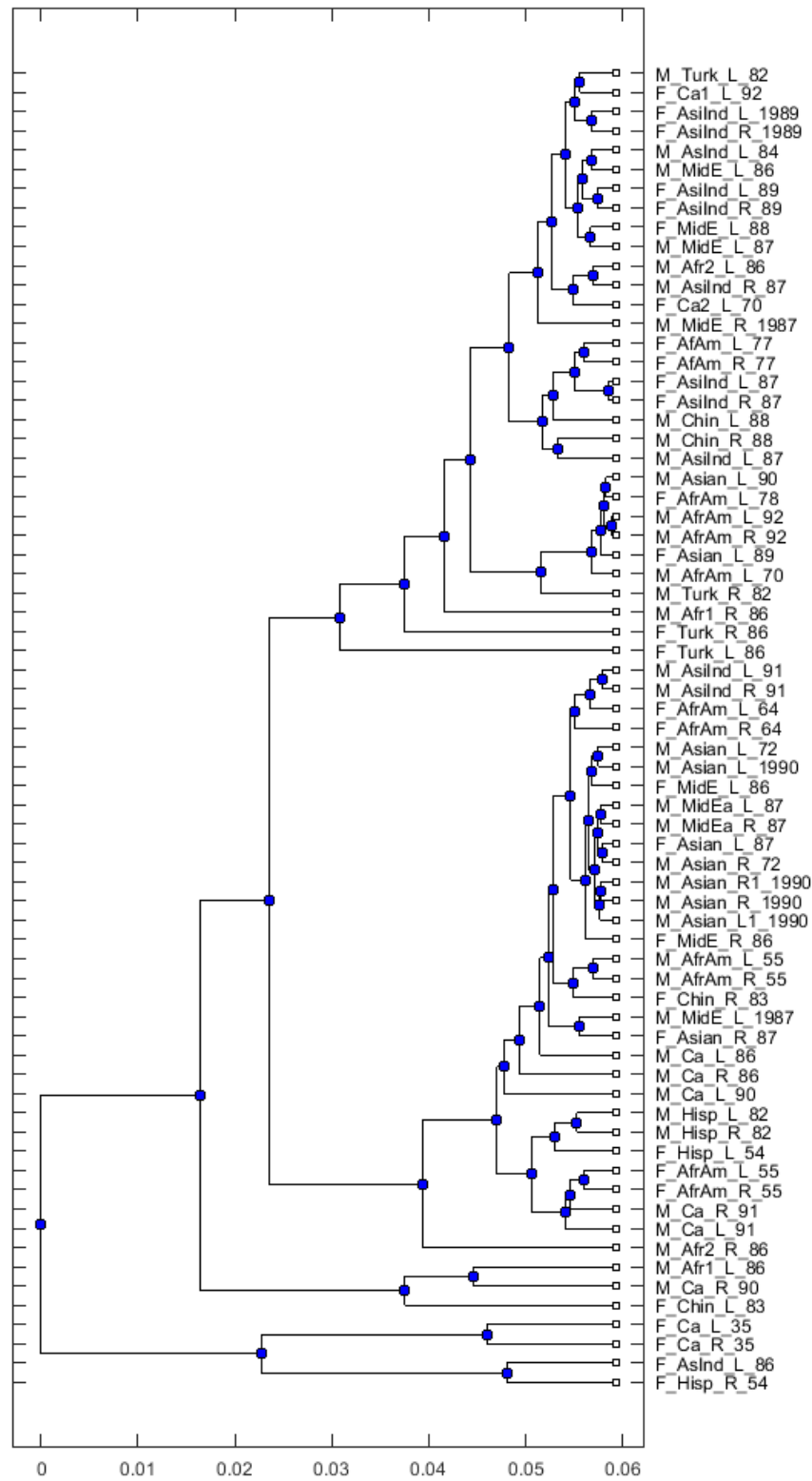


Fig 4.36 Phylogenetic tree based on KLD analysis on unique signatures of k-mer (k-1) frequencies from mapped sequences

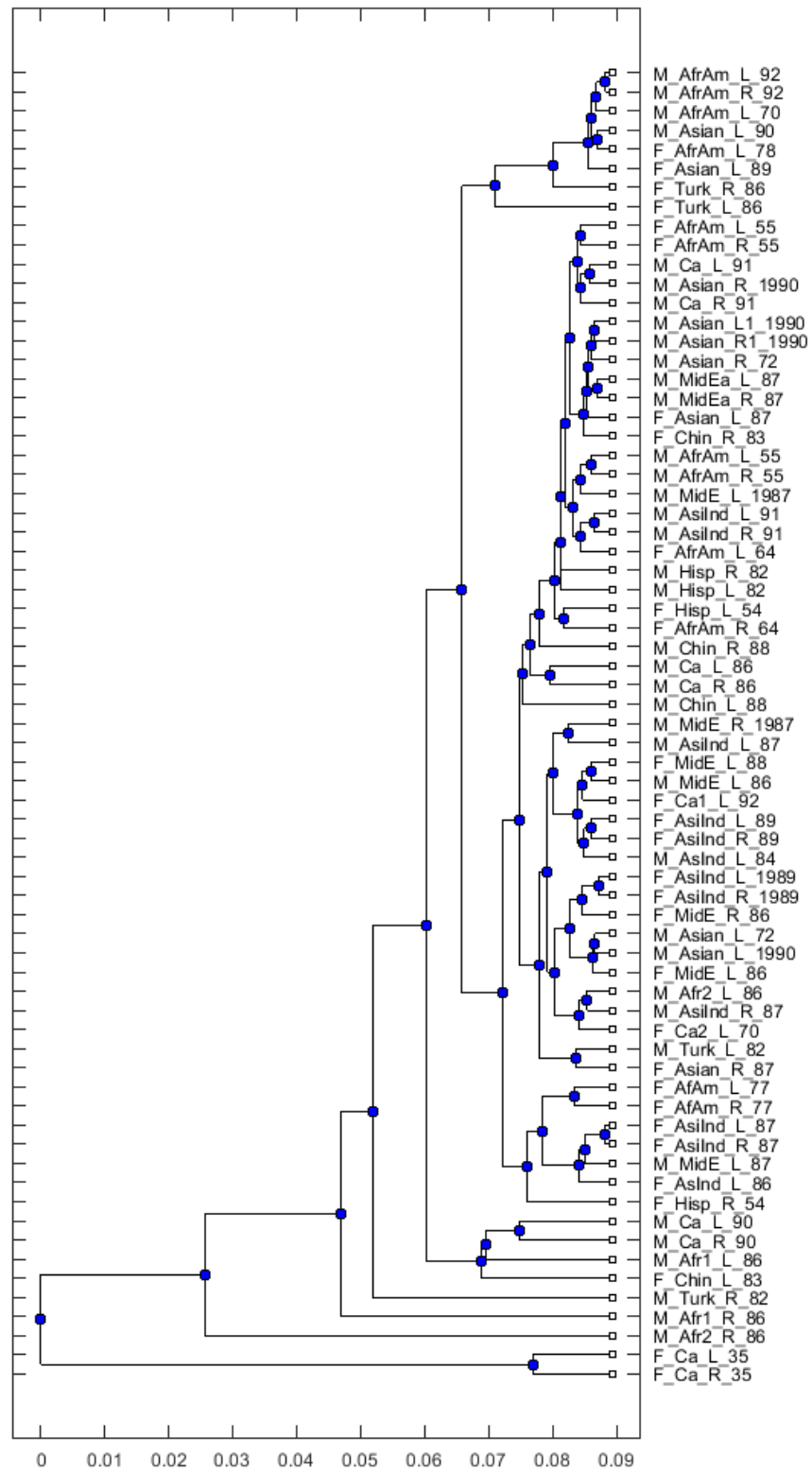


Fig 4.37 Phylogenetic tree based on KLD analysis on unique signatures of k-mer (k-2) frequencies from mapped sequences

4.6 Ensemble learning of samples using OTU and signature frequencies

Frequencies of OTU sequences and k-mer signatures were further used to identify the population group of samples using Ensemble learning classifier. Due to small number of samples, Leave One Out Cross Validation (Loo-CV) technique was adopted where the classifier was tested on one sample while the rest of the data was used to train the classifier. In an effort to avoid having biased data, equal number of samples from two different population groups were used while training the classifier. For example, to classify African and Asian population groups, 4 African samples along with every possible combination of 4 from 10 Asian samples were classified and their average accuracy rate is recorded. Number of classification iterations executed for every two population groups are listed below from tables 4.3 to 4.10.

Table 4.3 Number of iterations to classify African with rest of the population groups

	African (4 samples)
Turkish (4 samples)	1
Chinese (4 samples)	1
Hispanic (4 sample)	1
Middle Eastern (9 samples)	$C(9,4) = 126$
Caucasian (10 samples)	$C(10,4) = 210$
Asian (10 samples)	$C(10,4) = 210$
Asian Indian (12 samples)	$C(12,4) = 495$
African American (12 samples)	$C(12,4) = 495$
	Total=1539

Table 4.4 Number of iterations to classify Turkish with rest of the population groups

	Turkish (4 samples)
Chinese (4 samples)	1
Hispanic (4 sample)	1
Middle Eastern (9 samples)	$C(9,4) = 126$
Caucasian (10 samples)	$C(10,4) = 210$
Asian (10 samples)	$C(10,4) = 210$
Asian Indian (12 samples)	$C(12,4) = 495$
African American (12 samples)	$C(12,4) = 495$
	Total=1538

Table 4.5 Number of iterations to classify Chinese with rest of the population groups

	Chinese (4 samples)
Hispanic (4 sample)	1
Middle Eastern (9 samples)	$C(9,4) = 126$
Caucasian (10 samples)	$C(10,4) = 210$
Asian (10 samples)	$C(10,4) = 210$
Asian Indian (12 samples)	$C(12,4) = 495$
African American (12 samples)	$C(12,4) = 495$
	Total=1537

Table 4.6 Number of iterations to classify Hispanic with rest of the population groups

	Hispanic (4 samples)
Middle Eastern (9 samples)	$C(9,4) = 126$
Caucasian (10 samples)	$C(10,4) = 210$
Asian (10 samples)	$C(10,4) = 210$
Asian Indian (12 samples)	$C(12,4) = 495$
African American (12 samples)	$C(12,4) = 495$
	Total=1536

Table 4.7 Number of iterations to classify Middle Eastern with rest of the population groups

	Middle Eastern (9 samples)
Caucasian (10 samples)	$C(10,9) = 10$
Asian (10 samples)	$C(10,9) = 10$
Asian Indian (12 samples)	$C(12,9) = 220$
African American (12 samples)	$C(12,9) = 220$
	Total=460

Table 4.8 Number of iterations to classify Caucasian with rest of the population groups

	Caucasian (10 samples)
Asian (10 samples)	1
Asian Indian (12 samples)	$C(12,10) = 66$
African American (12 samples)	$C(12,10) = 66$
	Total=133

Table 4.9 Number of iterations to classify Asian with rest of the population groups

	Asian (10 samples)
Asian Indian (12 samples)	$C(12,10) = 66$
African American (12 samples)	$C(12,10) = 66$
	Total=132

Table 4.10 Number of iterations to classify Asian Indian with African American

	Asian Indian (12 samples)
African American (12 samples)	1
	Total=1

4.6.1 Ensemble learning using OTU frequencies:

After applying Ensemble Bagging algorithm with 500 Tree learning cycles for each iteration, accuracy rates for each population group are calculated from the resulting confusion matrix in Fig 4.38.

OTU (702 genus otus)	African	Turkish	Chinese	Hispanic	Middle Eastern	Caucasian	Asian	Asian Indian	African American	Sum of samples
African	24.11	1	1	1	1.06	0.98	0.99	0.87	1	32
Turkish	2	27.38	0	0	1.01	0.73	0.18	0.64	0.07	32
Chinese	1	0	18.39	2	2.85	2.41	1.93	1.4	2.02	32
Hispanic	0	0	0	28.6	0	1.74	0.65	0.03	0.97	32
Middle Eastern	1.64	0.79	2.35	1.18	39.82	0.7	1.3	2.95	1.25	52
Caucasian	1.5	1.08	2.43	2.74	2.4	39.58	0	2.89	2.38	55
Asian	0.38	0.03	0.99	2.8	0.4	1	44.81	1.38	3.21	55
Asian Indian	1.52	1.01	2.35	1.84	4.35	2.74	2	40.19	1	57
African_American	0.75	0.37	2.09	3.18	1.07	5.5	3.89	1	39.14	57

Fig 4.38 Confusion matrix from classifying samples using OTU frequencies

For each row that represents different population group in Fig 4.38, last column in green color represents the sum of samples in that particular population group. The values that are marked darker blue in shape of a diagonal represent the number of samples that were predicted accurately. From confusion matrix Fig 4.38, accuracy rates of population groups and average accuracy of the classification model is calculated as explained from section 2.10.

Table 4.11 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using OTU Frequencies

Population group	Accuracy %
African	75.34
Turkish	85.56
Chinese	57.46
Hispanic	89.37
Middle Eastern	76.58
Caucasian	71.96
Asian	81.47
Asian Indian	70.51
African American	68.67
Average accuracy	75.21

Table 4.11 display average accuracy of Loo-CV Ensemble classification of population groups using OTU frequencies as 75.21% with highest accuracy recognized in Hispanic and lowest in Chinese.

4.6.2 Ensemble learning using signatures from k-mer (k-1) frequencies:

Figure 4.39 is the resultant confusion matrix from ensemble learning of population groups using frequencies of signatures from k-mer frequencies (k-1). The values that are marked darker blue in shape of a diagonal in Fig 4.39 represent the number of samples that were predicted accurately out of total number of samples marked green in the last column. Table 4.12 shows that highest accuracy was found in Turkish and lowest in Caucasian.

k1 (225 clusters)	African	Turkish	Chinese	Hispanic	Middle Eastern	Caucasian	Asian	Asian Indian	African American	Sum of samples
African	27.44	1	0	0	2.02	0.41	0.17	0.94	0.02	32
Turkish	0	29.62	0	0	1.06	0.52	0.05	0.75	0	32
Chinese	0	1	21.27	1	2.06	1.81	1.9	1.13	1.83	32
Hispanic	1	0	2	18.06	1.27	2.52	2.91	1.29	2.95	32
Middle Eastern	1.66	1.15	1.96	0.75	39.02	0.6	2.1	3.23	1.53	52
Caucasian	1.75	1.52	2.97	3.07	3.3	31.76	4	3.94	2.7	55
Asian	0.12	0.36	1.4	2.31	1.8	2	40.15	3.27	3.59	55
Asian Indian	1.55	1.44	1.49	1.31	3.61	3.23	2.61	38.77	3	57
African American	0.31	0.24	1.75	2.76	1.43	1.85	4.52	2	42.16	57

Fig 4.39 Confusion matrix from classifying samples using signature frequencies from unique k-mer (k-1) frequencies

Table 4.12 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using signatures of k-mer (k-1) frequencies

Population group	Accuracy %
African	85.74
Turkish	92.55
Chinese	66.47
Hispanic	56.43
Middle Eastern	75.04
Caucasian	57.75
Asian	73
Asian Indian	68.02
African American	73.96
Average accuracy	72.11

4.6.3 Ensemble learning using signatures from k-mer (k-2) frequencies:

Figure 4.40 is the resultant confusion matrix from ensemble learning of population groups using frequencies of signatures from k-mer frequencies (k-2). Values marked in blues depict the number of samples that were accurately predicted out of total number of samples marked in green.

Table 4.13 display the accuracy rates of population groups calculated from confusion matrix Fig 4.40 with highest accuracy in African samples and least in Caucasian.

k2 (150 clusters)	African	Turkish	Chinese	Hispanic	Middle Eastern	Caucasian	Asian	Asian Indian	African American	Sum of samples
African	31.33	0	0	0	0.4	0.18	0	0.05	0.04	32
Turkish	2	27.87	0	0	0.47	0.86	0.01	0.78	0.01	32
Chinese	0	0	24.1	1	1.56	1.52	1.99	0.44	1.37	32
Hispanic	0	0	0	26.7	0.41	1.6	1.25	0.07	1.97	32
Middle Eastern	1.68	1.17	2.59	0.66	36.51	2	2	3.98	1.41	52
Caucasian	1.58	1.71	3.24	3.11	3.3	28.39	5	2.98	5.68	55
Asian	0.39	0.75	2.31	1.57	1.5	2	42.21	1.42	2.85	55
Asian Indian	1.56	1.64	2.2	0.93	4.46	2.02	3.35	38.83	2	57
African American	0.85	0.64	2.62	2.67	2.33	4.5	4.86	1	37.53	57

Fig 4.40 Confusion matrix from classifying samples using signature frequencies from unique k-mer (k-2) frequencies

Table 4.13 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using signatures of k-mer (k-2) frequencies

Population group	Accuracy %
African	97.9
Turkish	87.1
Chinese	75.33
Hispanic	83.44
Middle Eastern	70.2
Caucasian	51.61
Asian	76.75
Asian Indian	68.13
African American	65.84
Average accuracy	75.14

4.6.4 Ensemble learning using signatures from k-mer (k-3) frequencies:

Figure 4.41 is the resultant confusion matrix from ensemble learning of population groups using frequencies of signatures from k-mer frequencies (k-3). Values marked in blues depict the number of samples that were accurately predicted out of total number of samples marked in green.

Table 4.14 display the accuracy rates of population groups calculated from confusion matrix Fig 4.41 with highest accuracy in African samples and least in Caucasian.

k3 (150 clusters)	African	Turkish	Chinese	Hispanic	Middle Eastern	Caucasian	Asian	Asian Indian	African American	Sum of samples
African	27.96	1	0	0	1	0.85	0.24	0.5	0.44	32
Turkish	1	27.7	1	0	0.66	0.55	0.1	0.68	0.31	32
Chinese	0	0	24.53	0	2.12	1.11	1.71	0.71	1.82	32
Hispanic	0	0	1	22.13	0.68	2.44	2.34	0.57	2.84	32
Middle Eastern	1.44	1.09	2.21	0.87	40.17	0.9	1.9	2.32	1.1	52
Caucasian	1.68	1.5	2.95	3.1	2.7	30.57	4	3.06	5.44	55
Asian	0.41	0.57	1.74	1.92	0.8	1	43.9	0.52	4.15	55
Asian Indian	1.36	1.65	1.9	1.18	3.37	1.15	2.12	41.27	3	57
African American	0.87	0.77	2.42	2.74	1.88	4.42	5.67	2	36.22	57

Fig 4.41 Confusion matrix from classifying samples using signature frequencies from unique k-mer (k=3) frequencies

Table 4.14 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using signatures of k-mer (k=3) frequencies

Population group	Accuracy %
African	87.39
Turkish	86.56
Chinese	76.66
Hispanic	69.15
Middle Eastern	77.25
Caucasian	55.58
Asian	79.82
Asian Indian	72.4
African American	63.55
Average accuracy	74.27

4.6.5 Ensemble learning using signatures from k-mer (k-1) frequencies of mapped sequences:

Figure 4.42 is the resultant confusion matrix from ensemble learning of population groups using frequencies of signatures from k-mer frequencies (k-1) from mapped sequences. Values marked in blues depict the number of samples that were accurately predicted out of total number of samples marked in green. Table 4.15 display the accuracy rates of population groups calculated from confusion matrix Fig 4.42 with highest accuracy in African samples and least in Caucasian, similar to the results in case of k-2 frequencies.

k1m (100 clusters)	African	Turkish	Chinese	Hispanic	Middle Eastern	Caucasian	Asian	Asian Indian	African American	Sum of samples
African	31.61	0	0	0	0.25	0	0.01	0.07	0.06	32
Turkish	0	29.63	0	0	1.06	0.64	0.07	0.59	0.01	32
Chinese	0	1	22.03	2	2.51	1.86	0.71	0.78	1.11	32
Hispanic	0	0	1	23.42	1.24	2	1.84	0.26	2.25	32
Middle Eastern	1.53	1.61	3.01	0.8	36.78	1.8	2.1	3.46	0.9	52
Caucasian	1.8	1.87	3.12	3.04	2.8	30.72	5	2.83	3.82	55
Asian	0.86	0.8	1.67	2.03	2	2	41.15	0.42	4.08	55
Asian Indian	1.76	1.83	2.05	0.67	3.89	2.82	1.62	37.37	5	57
African American	1.08	0.93	2.65	2.66	1.62	2.91	3.74	2	39.39	57

Fig 4.42 Confusion matrix from classifying samples using signature frequencies from unique k-mer (k-1) frequencies from mapped sequences

Table 4.15 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using signatures of k-mer (k-1) frequencies from mapped sequences

Population group	Accuracy %
African	98.78
Turkish	92.6
Chinese	68.85
Hispanic	73.17
Middle Eastern	70.74
Caucasian	55.85
Asian	74.82
Asian Indian	65.56
African American	69.11
Average accuracy	74.39

4.6.6 Ensemble learning using signatures from k-mer (k-2) frequencies of mapped sequences:

Fig 4.43 is the confusion matrix from ensemble learning of population groups using frequencies of signatures from k-mer (k-2) frequencies from mapped sequences. Values marked in blues depict the number of samples that were accurately predicted out of total number of samples marked in green. Table 4.16 display accuracy rates of population groups calculated from confusion matrix Fig 4.43. Signatures from k-mer frequencies (k-2) of mapped sequences resulted in the most efficient result with 75.71% average accuracy with highest in African, Turkish and lowest in Caucasian.

k2m (275 clusters)	African	Turkish	Chinese	Hispanic	Middle Eastern	Caucasian	Asian	Asian Indian	African American	Sum of samples
African	27.42	2	0	0	0.37	0.83	0.45	0.6	0.33	32
Turkish	2	27.43	0	0	0.81	0.7	0.23	0.58	0.26	32
Chinese	0	0	25.18	0	2.67	0.93	1.21	0.86	1.15	32
Hispanic	0	0	1	24.22	0.81	1.55	1.81	0.85	1.76	32
Middle Eastern	1.56	1.37	2.4	0.85	39.74	0.7	1.2	3.15	1.04	52
Caucasian	1.73	1.44	2.34	3	2.8	33.45	3	2.45	4.79	55
Asian	0.63	0.75	1.2	1.76	1.6	0	43.43	2.12	3.52	55
Asian Indian	1.37	1.42	1.87	1.06	4.38	2.3	2.79	40.81	1	57
African American	0.74	0.74	1.65	2.73	1.44	5.06	4.98	1	38.65	57

Fig 4.43 Confusion matrix from classifying samples using signature frequencies from unique k-mer (k-2) frequencies from mapped sequences

Table 4.16 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using signatures of k-mer (k-2) frequencies from mapped sequences

Population group	Accuracy %
African	85.7
Turkish	85.72
Chinese	78.68
Hispanic	75.69
Middle Eastern	76.43
Caucasian	60.82
Asian	78.96
Asian Indian	71.6
African American	67.8
Average accuracy	75.71

4.6.7 Ensemble learning using relatively more accurate signatures from k-mer frequencies for each population group:

From tables 4.12 to 4.16, signature k-mer frequencies that caused the highest accuracy for each population group was noted and combined to build a new confusion matrix Fig 4.44. Table 4.17 presents the highest accuracy achieved for each population group and their average accuracy. An average of 79.65% was achieved, which is 4% more than the accuracy achieved by using OTU frequencies.

Better performing case of k-mer	African	Turkish	Chinese	Hispanic	Middle Eastern	Caucasian	Asian	Asian Indian	African American	Sum of samples
African (k1m)	31.61	0	0	0	0.25	0	0.01	0.07	0.06	32
Turkish (k1m)	0	29.63	0	0	1.06	0.64	0.07	0.59	0.01	32
Chinese (k2m)	0	0	25.18	0	2.67	0.93	1.21	0.86	1.15	32
Hispanic (k2)	0	0	0	26.7	0.41	1.6	1.25	0.07	1.97	32
Middle Eastern (k2m)	1.56	1.37	2.4	0.85	39.74	0.7	1.2	3.15	1.04	52
Caucasian (k2m)	1.73	1.44	2.34	3	2.8	33.45	3	2.45	4.79	55
Asian (k3)	0.41	0.57	1.74	1.92	0.8	1	43.9	0.52	4.15	55
Asian Indian (k3)	1.36	1.65	1.9	1.18	3.37	1.15	2.12	41.27	3	57
African_American (k1)	0.31	0.24	1.75	2.76	1.43	1.85	4.52	2	42.16	57

Fig 4.44 Confusion matrix from application of relatively better performing signature k-mer frequencies for each population group

Table 4.17 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups using better performing signatures of k-mer frequencies for each population group

Population group	k-mer signature	Accuracy %
African	Mapped K1	98.78
Turkish	Mapped K1	92.6
Chinese	Mapped K2	78.68
Hispanic	K2	83.44
Middle Eastern	Mapped K2	76.43
Caucasian	Mapped K2	60.82
Asian	K3	79.82
Asian Indian	K3	72.4
African American	K1	73.96
Overall accuracy		79.65

Summarization of accuracy levels for different population groups for different cases of k-mer is given in table 5.3.

Chapter 5 : Conclusions

Chapter 5 restates the problem statement and goal of this thesis. This chapter summarizes the achieved results and the limitations that were encountered in the process. Prospective work from this research is also addressed in this chapter.

5.1 Summary

The main objective of this research is to distinguish and identify the population group of individuals by studying the skin bacterial communities on the palm regions. For this particular work, 69 hand-swab samples collected from 39 people of 9 different population groups were used to analyze hypervariable region (V3) of the bacterial 16S ribosomal RNA (rRNA) gene. Representative consensus sequences of genus level OTUs extracted from V3 hypervariable region which is 65 nucleotide long are the prime focus of this thesis. In addition to OTUs, frequencies of nucleotide k-mers with k=1, 2 and 3 i.e. frequencies of {A, C, G,T}, {AA, AC, AG, AT, ... TT} and {AAA, and AAC, AAG, AAT, ... TTT} respectively were considered to determine clustering of multiple population groups. The strategy of using k-mer frequencies as a feature to distinguish individuals was verified by comparing k-mer frequencies of sequences representing same OTU, but from 4 different samples in Fig 5.1. After application of KLD analysis, dissimilarity of the resultant phylogenetic tree from a reference phylogenetic tree is considered as the measure of performance for different applications of k-mer frequencies. Furthermore, structure of the hypervariable region V3 shown in Fig 5.2 was taken into consideration to perform KLD analysis of k-mer frequencies. 65 positions of hypervariable region V3 of 16S rRNA were mapped into 47 elements based on the nucleotide links in the V3 hypervariable region as shown in Fig 5.3.

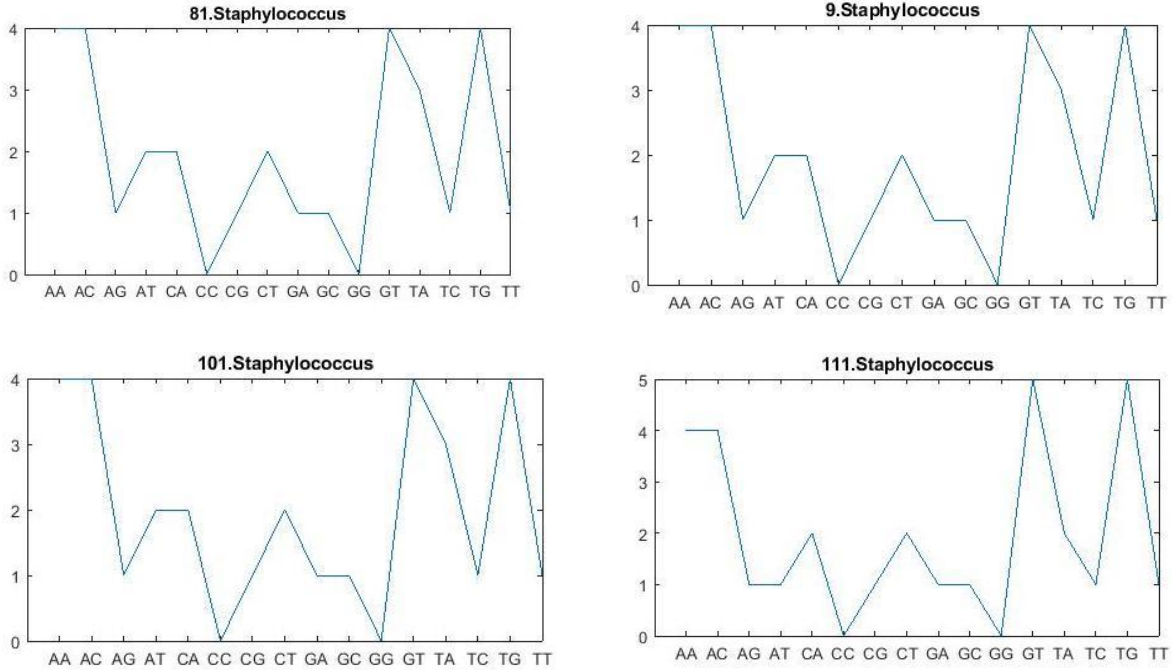


Fig 5.1 k-mer (k=2) frequencies in genus level OTU Staphylococcus from 4 different samples

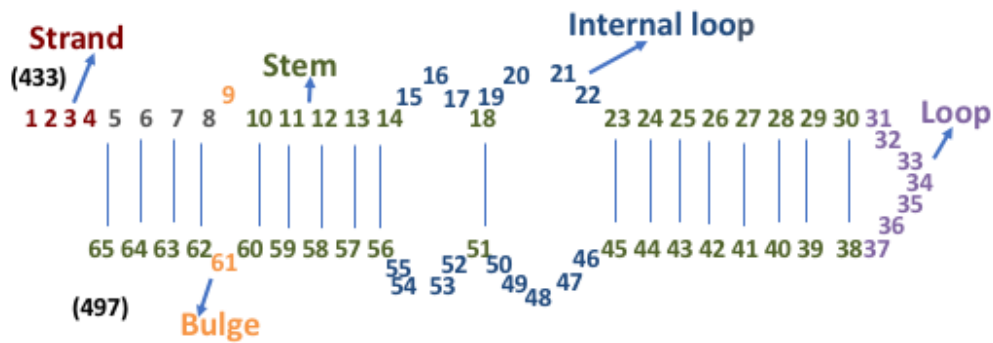


Fig 5.2 Nucleotide bond structure in 16S rRNA V3 region

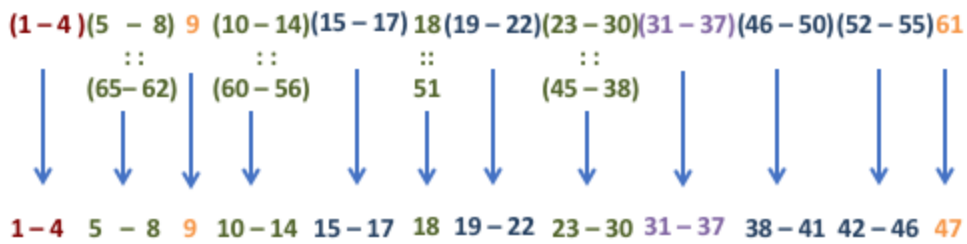


Fig 5.3 Mapping of 65nt of 16S rRNA into 47 elements

K-mer frequencies of the mapped sequences were further used to apply KLD analysis. In case of mapped sequences, k-mer analysis is stopped with k=2 because k=3 resulted in 8000 k-mers where the majority of the k-mers could not be found in the mapped sequences. Table 5.1 shows that mapped OTU frequencies resulted in the least dissimilarity when compared to using different cases of k-mer frequencies. Table 5.1 shows that OTU frequencies resulted in the least dissimilarity when compared to using different cases of k-mer frequencies. It is also apparent that mapped sequences based on their structure perform better than unchanged sequences with an exception of k1.

Table 5.1 Dissimilarity of resultant phylogenetic tree for different applications of KLD

Dissimilarity of resultant phylogenetic tree (Average Euclidean Pair-wise Distance)	Unmapped Sequences	Mapped Sequences
Features adopted to calculate KLD between samples		
OTU frequencies	0.3276	
Unweighted k-mer frequencies for k-1	0.3806	0.3797
Weighted k-mer frequencies for k-1	0.3718	0.3787
Unweighted k-mer frequencies for k-2	0.3479	0.3363
Weighted k-mer frequencies for k-2	0.3425	0.3315
Unweighted k-mer frequencies for k-3	0.3602	-
Weighted k-mer frequencies for k-3	0.3630	-

In an attempt to include the ignored unclassified sequences in OTU methods, k-mer frequencies of all the representative consensus sequences, both unclassified and classified, from

the data set were calculated. K-mer frequencies that were unique in the data set, and their frequencies in each sample were noted. Unique k-mer frequencies were further clustered to compress the data, and sum of the frequencies of all the sequences in a single cluster was considered as the occurrence of that particular cluster. Frequencies of these k-mer clusters in each sample were then used to apply KLD between samples. Table 5.2 presents the dissimilarity between resultant and reference phylogenetic trees in different cases of k-mer signatures. K-mer signatures with k-2, from mapped sequences resulted in comparatively less dissimilarity than the rest; moreover, signatures from k-1 frequencies, and k-2 frequencies of mapped sequences both resulted in comparatively lesser dissimilarity than 0.3276 of OTU frequencies

Table 5.2 Dissimilarity of resultant phylogenetic tree for different signatures of k-mer

Set of K-mer frequency signatures	Dissimilarity of resultant phylogenetic tree (Average Euclidean Pair-wise Distance)
k-1	0.3189
k-2	0.3298
k-3	0.3248
k-1 from mapped sequences	0.3409
k-2 from mapped sequences	0.3084

Frequencies from OTUs and signature k-mers were further used to identify the population group using Leave-one-out Cross Validation of Ensemble Bag Tree learning. Accuracy rates for each population group in different cases of k-mer signatures are displayed in table 5.3. It can be observed that signatures of k-2 from mapped sequences resulted in comparatively better results

than rest of the signature k-mers, including OTU frequencies. African and Turkish achieved highest accuracy rates, while Caucasian showed the least accuracy.

Table 5.3 Accuracy rates of Loo-CV of Ensemble bag tree learning of population groups in different cases

Frequencies of... Population group	OTU	Signatures of k-1	Signatures of k-2	Signatures of k-3	Signatures of k-1 from mapped sequences	Signatures of k-2 from mapped sequences
African	75.34 %	85.74 %	97.9 %	87.39	98.78 %	85.7 %
Turkish	85.56 %	92.55 %	87.1 %	86.56	92.6 %	85.72 %
Chinese	57.46 %	66.47 %	75.33 %	76.66	68.85 %	78.68 %
Hispanic	89.37 %	56.43 %	83.44 %	69.15	73.17 %	75.69 %
Middle Eastern	76.58 %	75.04 %	70.2 %	77.25	70.74 %	76.43 %
Caucasian	71.96 %	57.75 %	51.6 %	55.58	55.85 %	60.82 %
Asian	81.47 %	73 %	76.75 %	79.82	74.82 %	78.96 %
Asian Indian	70.51 %	68.02 %	68.13 %	72.4	65.56 %	71.6 %
African American	68.67 %	73.96 %	65.84 %	63.55	69.11 %	67.8 %
Average Accuracy	75.21 %	72.11 %	75.14 %	74.27	74.39 %	75.71 %

Table 5.4 presents the highest achieved accuracy in each population group and their average value, which is considered as the overall accuracy of the approach. Overall accuracy of ensemble classification of population groups using k-mer frequencies is 79.65 % which is ~2% greater than the accuracy acquired by applying Support Vector Machine classification on 5 most occurring OTUs [22].

Table 5.4 Highest accuracy rates of each population group and overall accuracy from k-mer signatures

Population group	k-mer signature	Accuracy %
African	Mapped K1	98.78
Turkish	Mapped K1	92.6
Chinese	Mapped K2	78.68
Hispanic	K2	83.44
Middle Eastern	Mapped K2	76.43
Caucasian	Mapped K2	60.82
Asian	K3	79.82
Asian Indian	K3	72.4
African American	K1	73.96
Overall accuracy		79.65

5.2 Conclusions and Future work

The set goal of classifying population groups by studying the hypervariable region V3 of 16S rRNA genome of hand bacterial samples was achieved with 79.6 % accuracy. The emphasis of this thesis was using nucleotide sequences of hypervariable region V3 and their k-mers to study hand bacterial samples, in addition to the conventional OTU approach. Additionally, information that resides in the structure of 16S rRNA hypervariable region V3 was included through mapping technique. The highpoint of this thesis was, unclassified sequences that are generally ignored in other OTU methods were ensured to be included through a novel k-mer classification model. Among 9 different population groups from all over the world, 6 population groups namely African, Turkish, Chinese, Hispanic, Middle Eastern and Asian achieved accuracy greater than 75%, with African and Turkish achieving more than 90% accuracy. Table 5.3 shows that k-mer (k-2) frequencies from mapped sequences resulted .2% greater accuracy than the conventional OTU method, encouraging the idea of k-mer usage.

However, hand bacterial samples that were included in this study does not specify other factors that could influence skin microbiome, for example, there is no record of when the individual last washed their hands before sample collection, or if the individual has lived in a country different from that of the origin of their population group. Also, among 9 hypervariable regions, only a single region was considered for the extraction and analysis of nucleotide sequences. Therefore, there is a scope for improving the methodology by considering other hypervariable regions of 16S rRNA. Considering multiple hypervariable regions could be beneficial while studying the structure of 16S rRNA and would allow using longer k-mers for classification, which was limited to k-3 in this study due to only 65 nucleotides. In addition to 16S rRNA, other parts of 70S ribosome i.e. 23S RNA could offer better understanding of bacteria with the application of k-mer classification. Furthermore, increasing the number of samples from different population groups would improve the performance of classification model by studying more and better patterns. More number of samples would allow to extend the study in terms of age and gender.

References

- [1] E. A. Grice and J. A. Segre, “The skin microbiome,” *Nat. Rev. Microbiol.*, vol. 9, no. 4, pp. 244–253, Apr. 2011.
- [2] Z. Gao, C. Tseng, Z. Pei, and M. J. Blaser, “Molecular analysis of human forearm superficial skin bacterial biota,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 8, pp. 2927–2932, Feb. 2007.
- [3] J. Si, S. Lee, J. M. Park, J. Sung, and G. Ko, “Genetic associations and shared environmental effects on the skin microbiome of Korean twins,” *BMC Genomics*, vol. 16, p. 992, Nov. 2015.
- [4] N. Fierer, C. L. Lauber, N. Zhou, D. McDonald, E. K. Costello, and R. Knight, “Forensic identification using skin bacterial communities,” *Proc. Natl. Acad. Sci.*, vol. 107, no. 14, pp. 6477–6481, Apr. 2010.
- [5] C. L. Lauber, N. Zhou, J. I. Gordon, R. Knight, and N. Fierer, “Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples,” *FEMS Microbiol. Lett.*, vol. 307, no. 1, pp. 80–86, Jun. 2010.
- [6] J. Oh, A. L. Byrd, M. Park, H. H. Kong, and J. A. Segre, “Temporal Stability of the Human Skin Microbiome,” *Cell*, vol. 165, no. 4, pp. 854–866, May 2016.
- [7] E. A. Grice *et al.*, “Topographical and temporal diversity of the human skin microbiome,” *Science*, vol. 324, no. 5931, pp. 1190–1192, May 2009.
- [8] J. Peterson *et al.*, “The NIH Human Microbiome Project,” *Genome Res.*, vol. 19, no. 12, pp. 2317–2323, Dec. 2009.
- [9] N. Fierer, M. Hamady, C. L. Lauber, and R. Knight, “The influence of sex, handedness, and washing on the diversity of hand surface bacteria,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 46, pp. 17994–17999, Nov. 2008.
- [10] M. G. Dominguez-Bello *et al.*, “Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns,” *Proc. Natl. Acad. Sci.*, vol. 107, no. 26, pp. 11971–11975, Jun. 2010.
- [11] M. J. Blaser *et al.*, “Distinct cutaneous bacterial assemblages in a sampling of South American Amerindians and US residents,” *ISME J.*, vol. 7, no. 1, pp. 85–95, Jan. 2013.
- [12] D. N. Fredricks, “Microbial ecology of human skin in health and disease,” *J. Investig. Dermatol. Symp. Proc.*, vol. 6, no. 3, pp. 167–169, Dec. 2001.

- [13] S. Ying *et al.*, “The influence of age and gender on skin-associated microbial communities in urban and rural human populations,” *PloS One*, vol. 10, no. 10, p. e0141842, 2015.
- [14] J. Urban *et al.*, “The effect of habitual and experimental antiperspirant and deodorant product use on the armpit microbiome,” *PeerJ*, vol. 4, p. e1605, Feb. 2016.
- [15] Giovannoni SJ, Britschgi TB, Moyer CL, and Field KG, “Genetic diversity in Sargasso Sea bacterioplankton,” *Nature*, vol. 345, no. 6270, pp. 60–3, 1990.
- [16] W. Li, L. Han, P. Yu, C. Ma, X. Wu, and J. Xu, “Nested PCR-denaturing gradient gel electrophoresis analysis of human skin microbial diversity with age,” *Microbiol. Res.*, vol. 169, no. 9, pp. 686–692, Sep. 2014.
- [17] P. H. Janssen, “Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes,” *Appl. Environ. Microbiol.*, vol. 72, no. 3, pp. 1719–1728, Mar. 2006.
- [18] E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight, “Bacterial community variation in human body habitats across space and time,” *Science*, vol. 326, no. 5960, pp. 1694–1697, Dec. 2009.
- [19] D. Hospodsky *et al.*, “Hand bacterial communities vary across two different human populations,” *Microbiology*, vol. 160, no. 6, pp. 1144–1152, 2014.
- [20] C. A. Lozupone, M. Hamady, S. T. Kelley, and R. Knight, “Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities,” *Appl. Environ. Microbiol.*, vol. 73, no. 5, pp. 1576–1585, Mar. 2007.
- [21] M. H. Y. Leung, D. Wilkins, and P. K. H. Lee, “Insights into the pan-microbiome: skin microbial communities of Chinese individuals differ from other racial groups,” *Sci. Rep.*, vol. 5, p. 11845, Jul. 2015.
- [22] A. B. Holbert, H. P. Whitelam, L. J. Sooter, and J. M. Dawson, “Evaluation of Hand Bacteria as a Human Biometric Identifier,” in *IEEE International Conference on Bioinformatics and Bioengineering*, 2014, pp. 83–89.
- [23] N. N. Schommer and R. L. Gallo, “Structure and function of the human skin microbiome,” *Trends Microbiol.*, vol. 21, no. 12, pp. 660–668, Dec. 2013.
- [24] A. SanMiguel and E. A. Grice, “Interactions between host factors and the skin microbiome,” *CMLS*, vol. 72, no. 8, pp. 1499–1515, Apr. 2015.
- [25] G. I. Perez Perez *et al.*, “Body site is a more determinant factor than human population diversity in the healthy skin microbiome,” *PLoS ONE*, vol. 11, no. 4, p. e0151990, Apr. 2016.

- [26] S. L. Edmonds-Wilson, N. I. Nurinova, C. A. Zapka, N. Fierer, and M. Wilson, "Review of human hand microbiome research," *J. Dermatol. Sci.*, vol. 80, no. 1, pp. 3–12, Oct. 2015.
- [27] A. Benohanian, "Antiperspirants and deodorants," *Clin. Dermatol.*, vol. 19, no. 4, pp. 398–405, Jul. 2001.
- [28] A. Camarinha-Silva *et al.*, "Comparing the anterior nare bacterial community of two discrete human populations using Illumina amplicon sequencing," *Environ. Microbiol.*, vol. 16, no. 9, pp. 2939–2952, Sep. 2014.
- [29] D. Yadav, T. S. Ghosh, and S. S. Mande, "Global investigation of composition and interaction networks in gut microbiomes of individuals belonging to diverse geographies and age-groups," *Gut Pathog.*, vol. 8, p. 17, 2016.
- [30] J. Zhang *et al.*, "A phylo-functional core of gut microbiota in healthy young Chinese cohorts across lifestyles, geography and ethnicities," *ISME J.*, vol. 9, no. 9, pp. 1979–1990, Sep. 2015.
- [31] S.-H. Park, K.-A. Kim, Y.-T. Ahn, J.-J. Jeong, C.-S. Huh, and D.-H. Kim, "Comparative analysis of gut microbiota in elderly people of urbanized towns and longevity villages," *BMC Microbiol.*, vol. 15, p. 49, Feb. 2015.
- [32] C. Dominianni *et al.*, "Sex, Body Mass Index, and Dietary Fiber Intake Influence the Human Gut Microbiome," *PLoS ONE*, vol. 10, no. 4, p. e0124599, Apr. 2015.
- [33] A. Andoh *et al.*, "Comparison of the gut microbial community between obese and lean peoples using 16S gene sequencing in a Japanese population," *J. Clin. Biochem. Nutr.*, vol. 59, no. 1, pp. 65–70, Jul. 2016.
- [34] R. J. Newton *et al.*, "Sewage reflects the microbiomes of human populations," *mBio*, vol. 6, no. 2, pp. e02574-14, May 2015.
- [35] Servier, "Oral cavity", Creative Commons Attribution-3.0 Unported License, https://smart.servier.com/smart_image/oral-cavity/
- [36] W. G. Wade, "The oral microbiome in health and disease," *Pharmacol. Res.*, vol. 69, no. 1, pp. 137–143, Mar. 2013.
- [37] M. Kilian *et al.*, "The oral microbiome – an update for oral healthcare professionals," *Br. Dent. J.*, vol. 221, no. 10, pp. 657–666, Nov. 2016.
- [38] J. Li *et al.*, "Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa," *BMC Microbiol.*, vol. 14, p. 316, Dec. 2014.

- [39] J. Wu *et al.*, “Cigarette smoking and the oral microbiome in a large study of American adults,” *ISME J.*, vol. 10, no. 10, pp. 2435–2446, Oct. 2016.
- [40] R. Krajmalnik-Brown, Z.-E. Ilhan, D.-W. Kang, and J. K. DiBaise, “Effects of gut microbes on nutrient absorption and energy regulation,” *Nutr. Clin. Pract. Off. Publ. Am. Soc. Parenter. Enter. Nutr.*, vol. 27, no. 2, pp. 201–214, Apr. 2012.
- [41] R. T. Coupe, “Evaluating the effects of resources and solvability on burglary detection,” *Polic. Soc.*, vol. 26, no. 5, pp. 563–587, Jul. 2016.
- [42] J. T. Hampton-Marcell, J. V. Lopez, and J. A. Gilbert, “The human microbiome: an emerging tool in forensics,” *Microb. Biotechnol.*, vol. 10, no. 2, pp. 228–230.
- [43] S. Lax *et al.*, “Forensic analysis of the microbiome of phones and shoes,” *Microbiome*, vol. 3, p. 21, May 2015.
- [44] I. Klymiuk, I. Bambach, V. Patra, S. Trajanoski, and P. Wolf, “16S based microbiome analysis from healthy subjects’ skin swabs stored for different storage periods reveal phylum to genus level changes,” *Front. Microbiol.*, vol. 7, 2016.
- [45] C. E. Kandel, A. E. Simor, and D. A. Redelmeier, “Elevator buttons as unrecognized sources of bacterial colonization in hospitals,” *Open Med.*, vol. 8, no. 3, pp. e81–e86, Jul. 2014.
- [46] S.-Y. Lee, S.-K. Woo, G.-W. Choi, Y.-J. Hong, and Y.-B. Eom, “Microbial forensic analysis of bacterial fingerprint by sequence comparison of 16S rRNA gene,” *J. Forensic Res.*, vol. 6, no. 5, pp. 1–4, Jun. 2015.
- [47] S.-Y. Lee, S.-K. Woo, S.-M. Lee, and Y.-B. Eom, “Forensic analysis using microbial community between skin bacteria and fabrics,” *Toxicol. Environ. Health Sci.*, vol. 8, no. 3, pp. 263–270, Sep. 2016.
- [48] E. Nishi, Y. Tashiro, and K. Sakai, “Discrimination among individuals using terminal restriction fragment length polymorphism profiling of bacteria derived from forensic evidence,” *Int. J. Legal Med.*, vol. 129, no. 3, pp. 425–433, May 2015.
- [49] H. R. Johnson *et al.*, “A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval,” *PLoS ONE*, vol. 11, no. 12, p. e0167370, Dec. 2016.
- [50] E. M. Bik, “The hoops, hopes, and hypes of human microbiome research,” *Yale J. Biol. Med.*, vol. 89, no. 3, pp. 363–373, Sep. 2016.
- [51] “Human Microbiome Project,” *Wikipedia*.
- [52] “Nucleotide”, *Wikipedia*.

- [53] A. M. Helmenstine, "Differences between DNA and RNA", Spunk / Wikimedia Commons / CC BY-SA 3.016s .
- [54] P. A. Pevzner, H. Tang, and M. S. Waterman, "An Eulerian path approach to DNA fragment assembly," *Proc. Natl. Acad. Sci.*, vol. 98, no. 17, pp. 9748–9753, Aug. 2001.
- [55] P. Melsted and J. K. Pritchard, "Efficient counting of k-mers in DNA sequences using a bloom filter," *BMC Bioinformatics*, vol. 12, p. 333, Aug. 2011.
- [56] R. S. Roy, D. Bhattacharya, and A. Schliep, "Turtle: Identifying frequent k -mers with cache-efficient algorithms," *Bioinformatics*, vol. 30, no. 14, pp. 1950–1957, Jul. 2014.
- [57] A. Sievers *et al.*, "K-mer content, correlation, and position analysis of genome DNA sequences for the identification of function and evolutionary features," *Genes*, vol. 8, no. 4, Apr. 2017.
- [58] V. B. Dubinkina, D. S. Ischenko, V. I. Ulyantsev, A. V. Tyakht, and D. G. Alexeev, "Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis," *BMC Bioinformatics*, vol. 17, p. 38, Jan. 2016.
- [59] "16S ribosomal RNA," *Wikipedia*.
- [60] S. Chakravorty, D. Helb, M. Burday, N. Connell, and D. Alland, "A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria.," *J. Microbiol. Methods*, vol. 69, no. 2, pp. 330–339, May 2007.
- [61] M. M. Kattar *et al.*, "Application of 16S rRNA gene sequencing to identify *Bordetella hinzii* as the causative agent of fatal septicemia," *J. Clin. Microbiol.*, vol. 38, no. 2, pp. 789–794, Feb. 2000.
- [62] J. E. Clarridge, "Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases," *Clin. Microbiol. Rev.*, vol. 17, no. 4, pp. 840–862, Oct. 2004.
- [63] "Molecular biology - What causes the variable/conserved structure in the 16S rRNA gene?," <https://biology.stackexchange.com/questions/54823/what-causes-the-variable-conserved-structure-in-the-16s-rna-gene> .
- [64] S. McGinn and I. G. Gut, "DNA sequencing – spanning the generations," *New Biotechnol.*, vol. 30, no. 4, pp. 366–372, May 2013.
- [65] M. Morey, A. Fernández-Marmiesse, D. Castiñeiras, J. M. Fraga, M. L. Couce, and J. A. Cocho, "A glimpse into past, present, and future DNA sequencing," *Mol. Genet. Metab.*, vol. 110, no. 1, pp. 3–24, Sep. 2013.

- [66] “MiSeq System | Focused power for targeted gene and small genome sequencing.”
<https://www.illumina.com/systems/sequencing-platforms/miseq.html>.
- [67] B. Canard and R. S. Sarfati, “DNA polymerase fluorescent substrates with reversible 3’-tags,” *Gene*, vol. 148, no. 1, pp. 1–6, Oct. 1994.
- [68] Illumina, *Illumina Sequencing by Synthesis*.
<https://www.youtube.com/watch?v=fCd6B5HRaZ8>.
- [69] “Getting Started.”
<https://developer.basespace.illumina.com/docs/content/documentation/getting-started/overview>. [Accessed: 03-Nov-2018].
- [70] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, “Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy,” *Appl. Environ. Microbiol.*, vol. 73, no. 16, pp. 5261–5267, Aug. 2007.
- [71] J. G. Caporaso *et al.*, “QIIME allows analysis of high-throughput community sequencing data,” *Nat. Methods*, vol. 7, no. 5, pp. 335–336, May 2010.
- [72] S. Kullback and R. A. Leibler, “On Information and Sufficiency,” *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.
- [73] David J.C MacKay, *Information theory, Inference, and Learning Algorithms*, 1st ed. Cambridge University Press.
- [74] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.
- [75] A. Ramette, “Multivariate analyses in microbial ecology,” *Fems Microbiol. Ecol.*, vol. 62, no. 2, pp. 142–160, Nov. 2007.
- [76] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.
- [77] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
- [78] L. C. Paulino, C.-H. Tseng, B. E. Strober, and M. J. Blaser, “Molecular analysis of fungal microbiota in samples from healthy human skin and psoriatic lesions,” *J. Clin. Microbiol.*, vol. 44, no. 8, pp. 2933–2941, Aug. 2006.
- [79] G. . Baker, J. . Smith, and D. . Cowan, “Review and re-analysis of domain-specific 16S primers,” *MIMET J. Microbiol. Methods*, vol. 55, no. 3, pp. 541–555, 2003.
- [80] G. Muyzer, E. C. de Waal, and A. G. Uitterlinden, “Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain

- reaction-amplified genes coding for 16S rRNA.,” *Appl. Environ. Microbiol.*, vol. 59, no. 3, pp. 695–700, Mar. 1993.
- [81] A. K. Bartram, M. D. J. Lynch, J. C. Stearns, G. Moreno-Hagelsieb, and J. D. Neufeld, “Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads,” *Appl. Environ. Microbiol.*, vol. 77, no. 11, pp. 3846–3852, Jun. 2011.
- [82] “BaseSpace User Guide,” p. 98.

Appendix A: KLD analysis of OTU frequencies

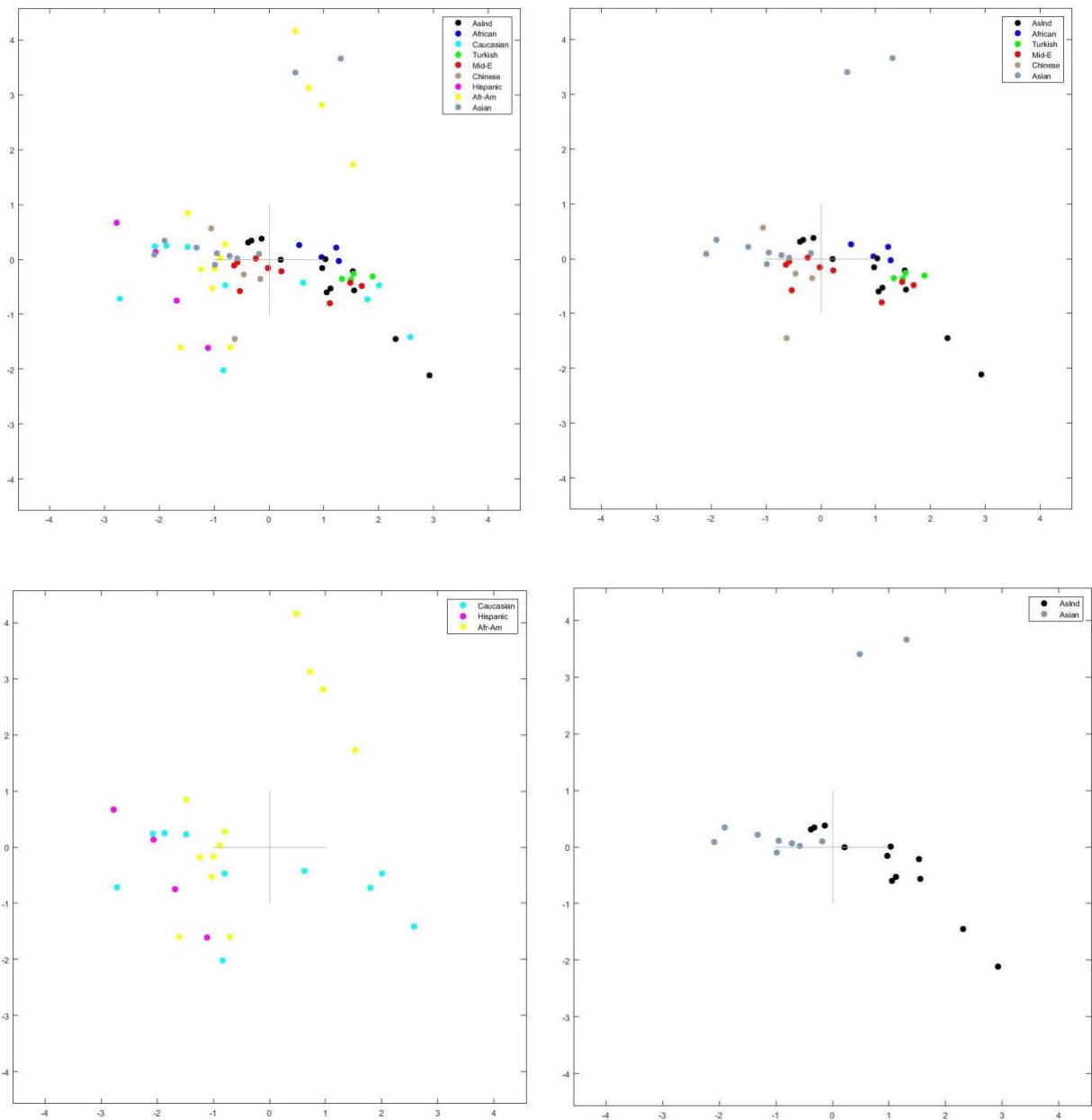


Fig A.1 PCoA plot from KLD analysis of OTU frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

Appendix B: KLD analysis of unweighted k-mer frequencies

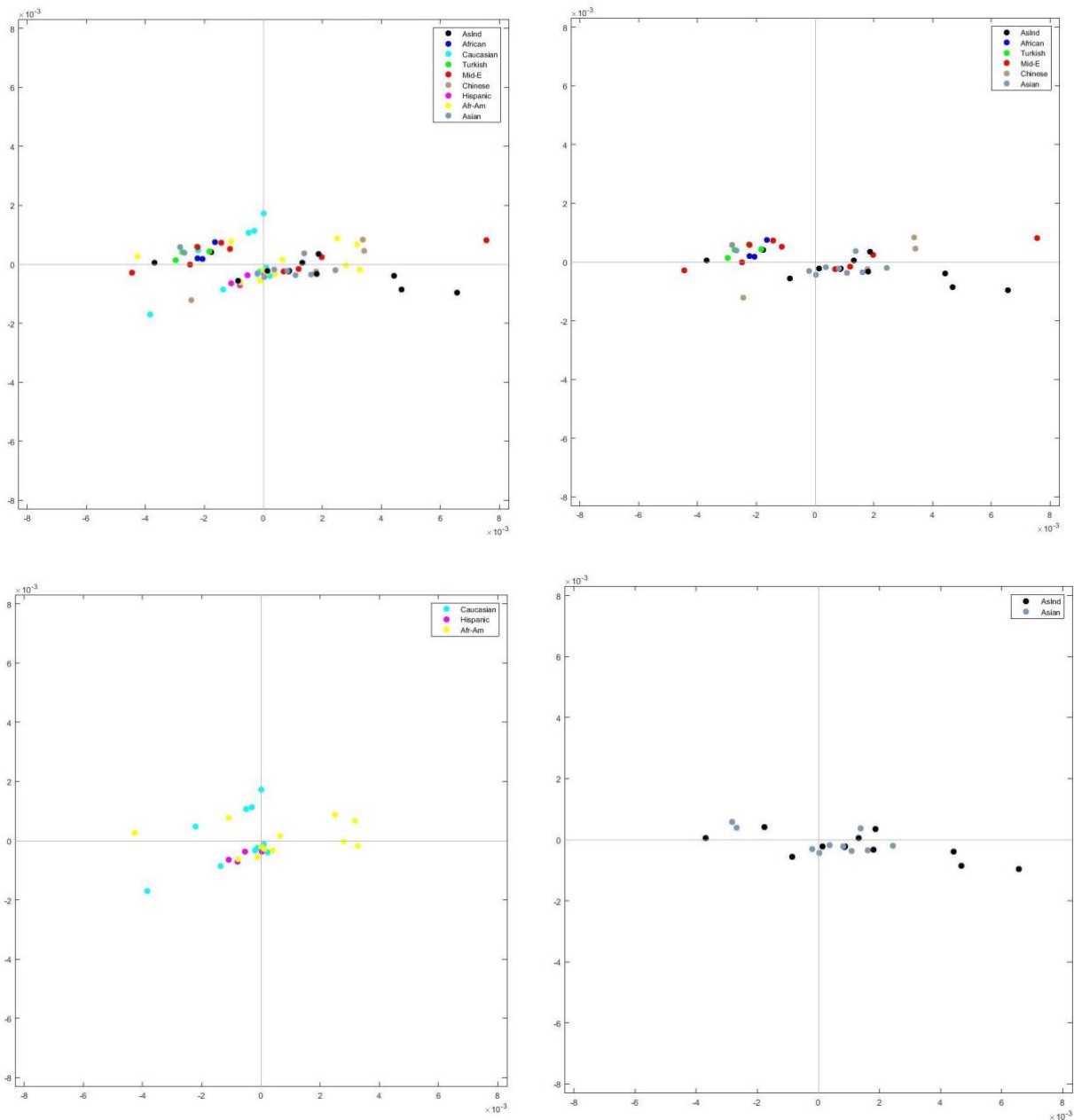


Fig B.1 PCoA plot from KLD analysis of unweighted k-mer ($k=1$) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

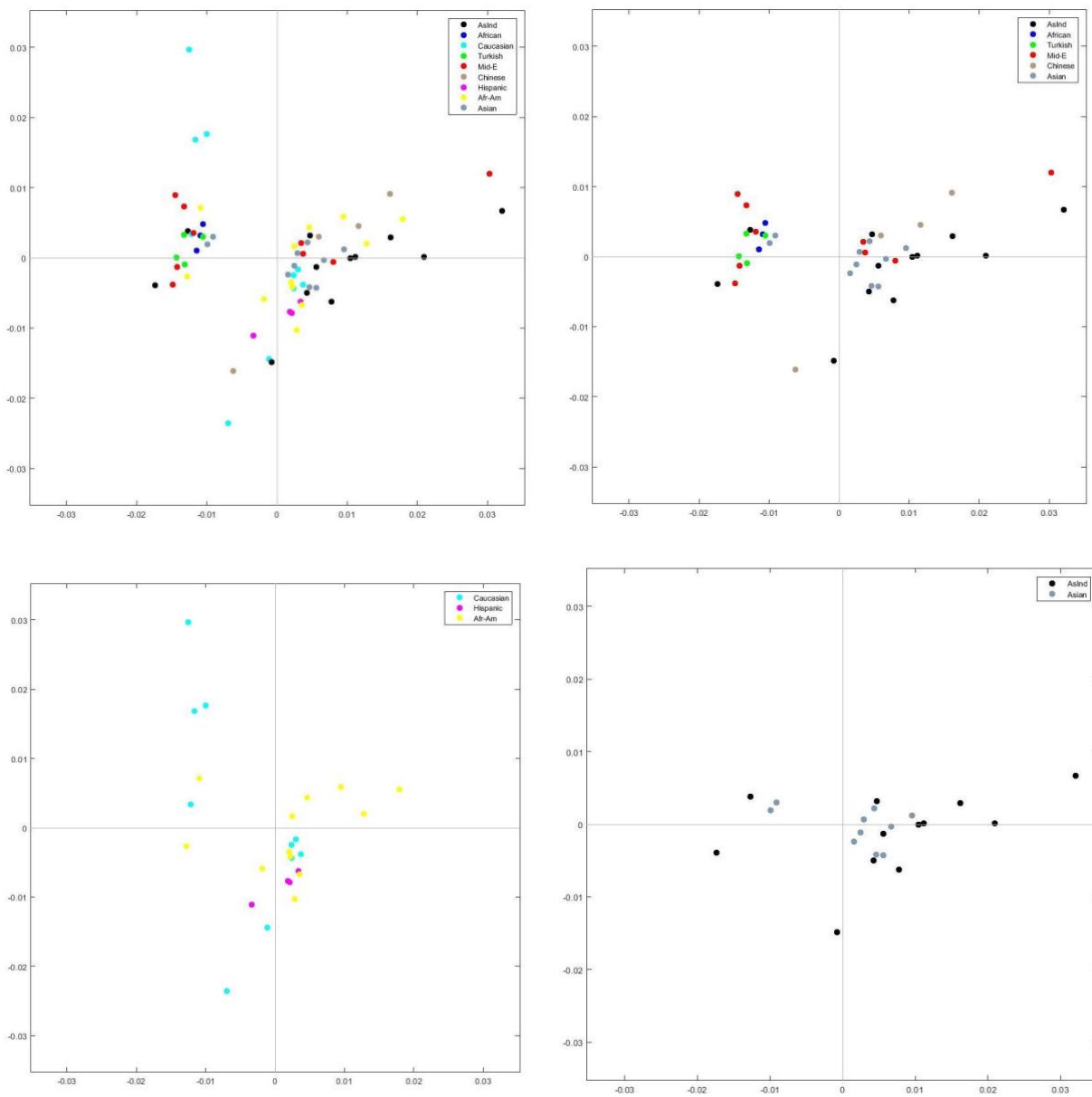


Fig B.2 PCoA plot from KLD analysis of unweighted k-mer ($k=2$) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

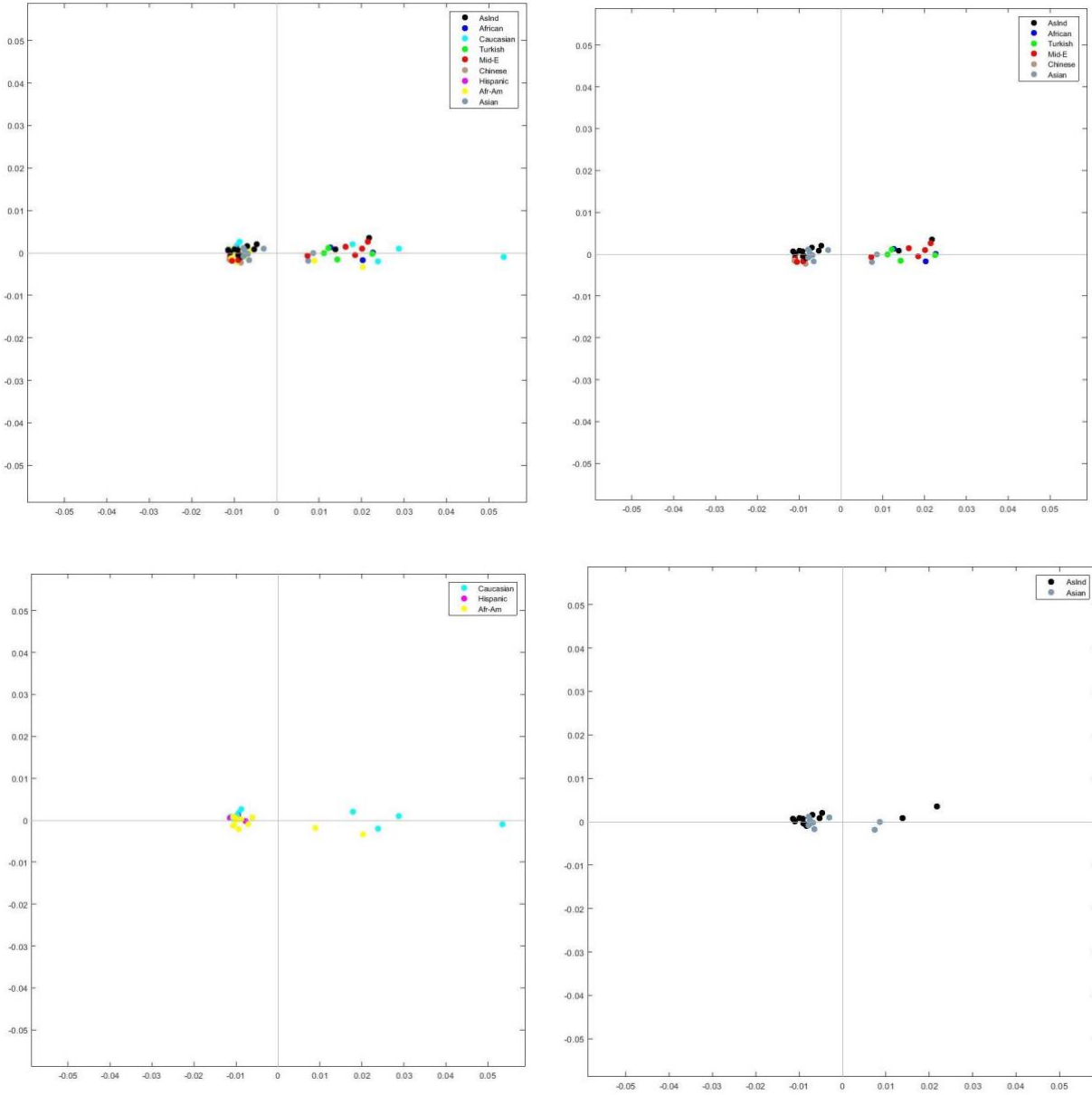


Fig B.3 PCoA plot from KLD analysis of unweighted k-mer (k=3) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

Appendix C: KLD analysis of weighted k-mer frequencies

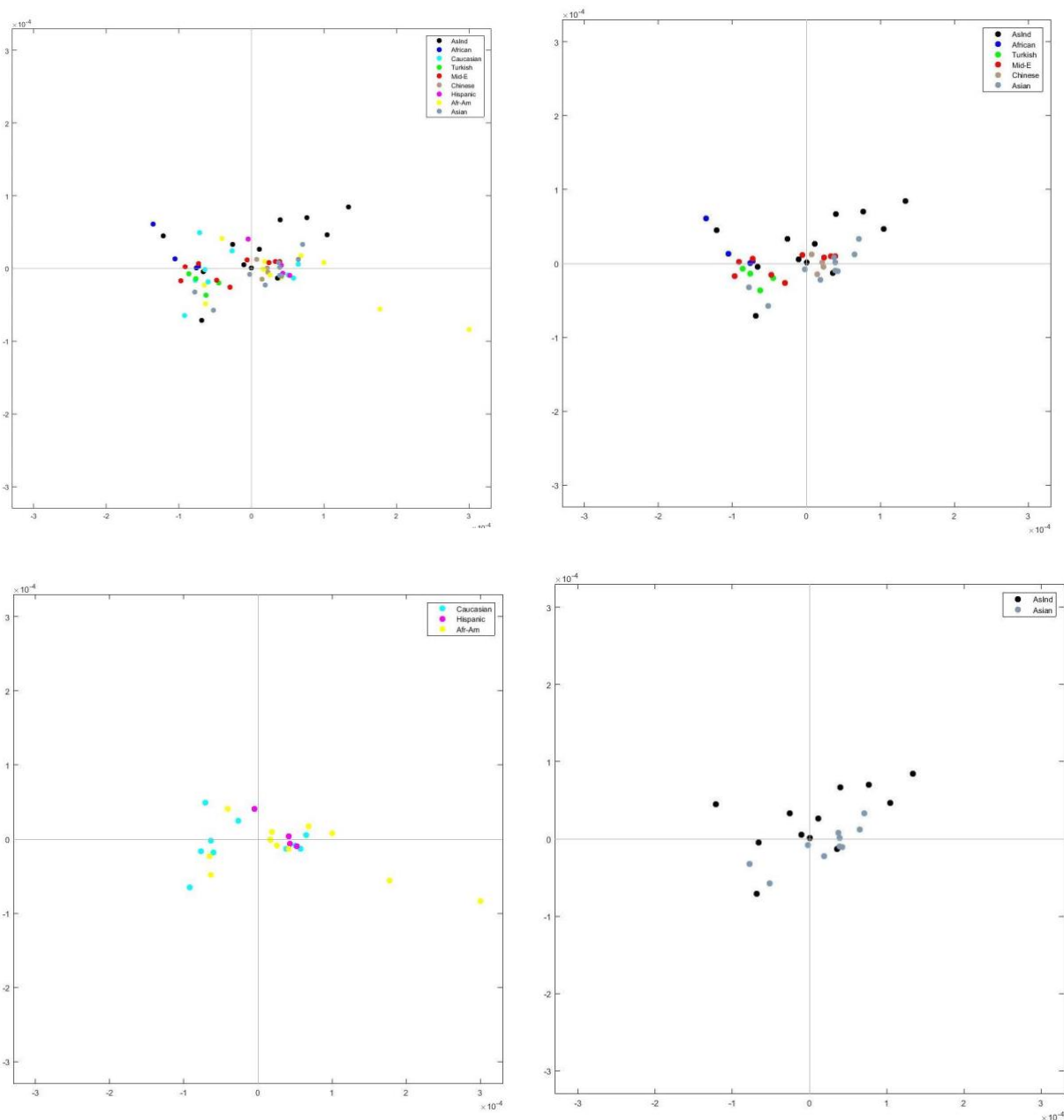


Fig C.1 PCoA plot from KLD analysis of weighted k-mer (k=1) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

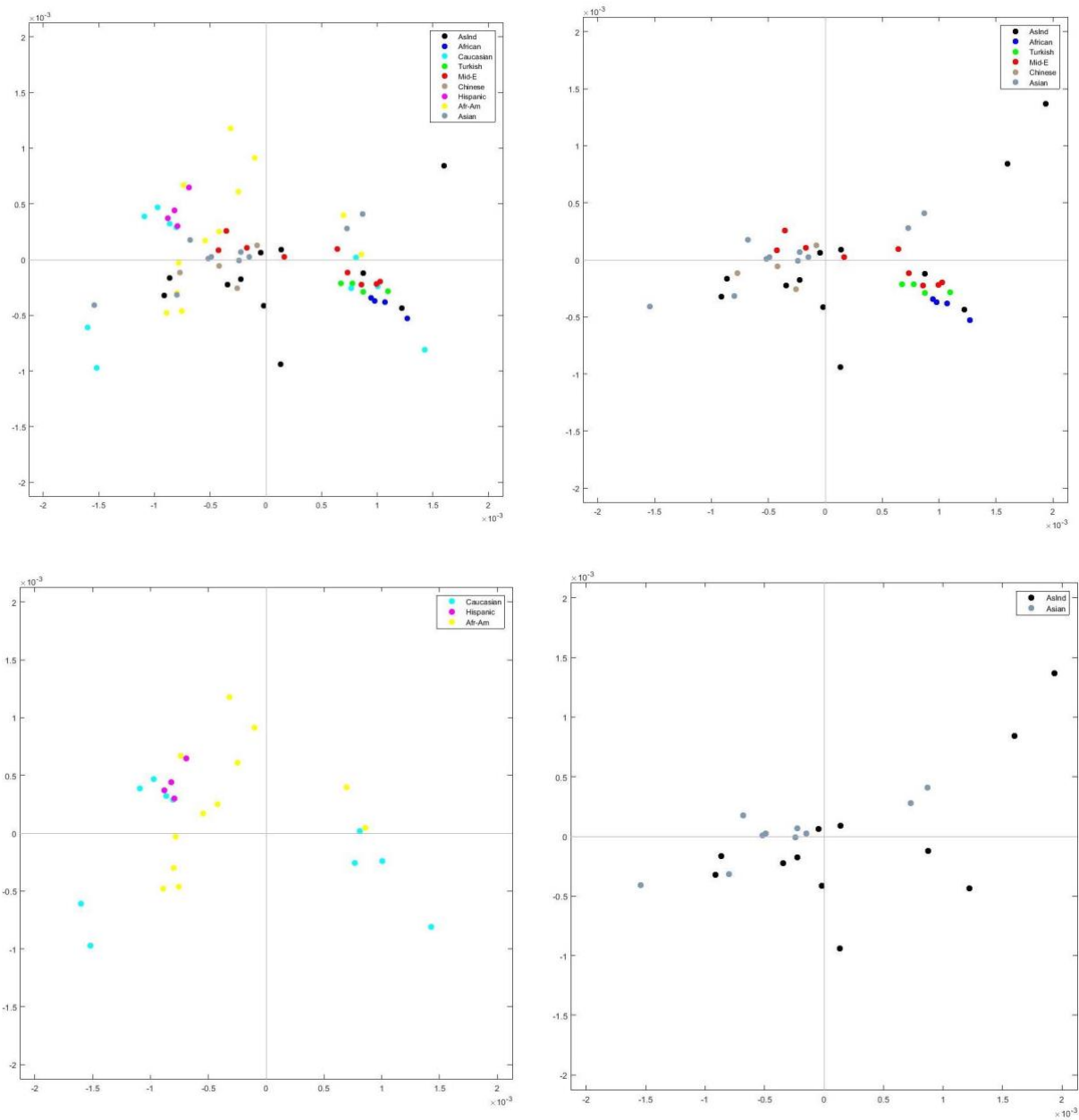


Fig C.2 PCoA plot from KLD analysis of weighted k-mer (k=2) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

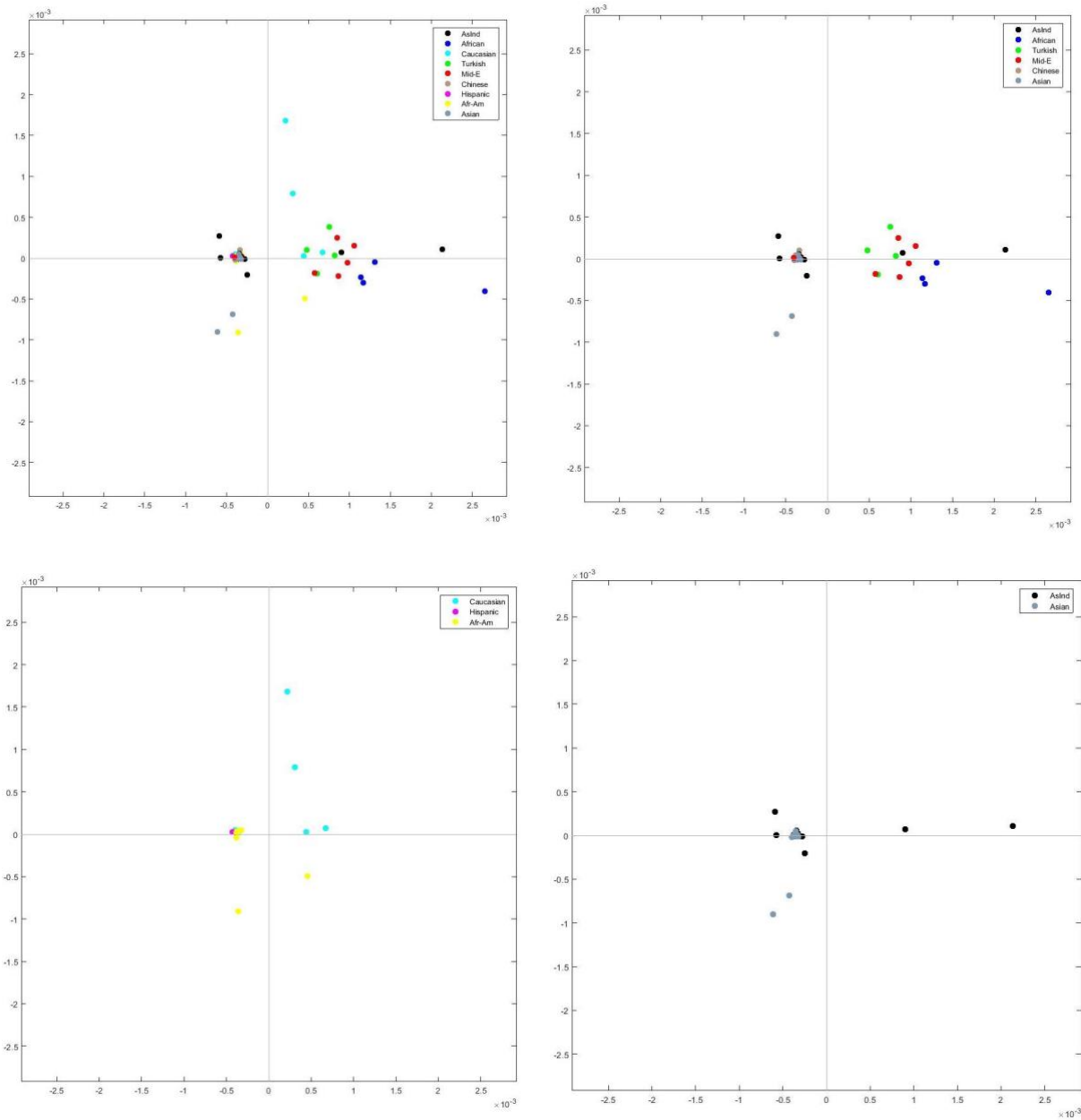


Fig C.3 PCoA plot from KLD analysis of weighted k-mer (k=3) frequencies of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

Appendix D: KLD analysis of unweighted k-mer frequencies from mapped sequences

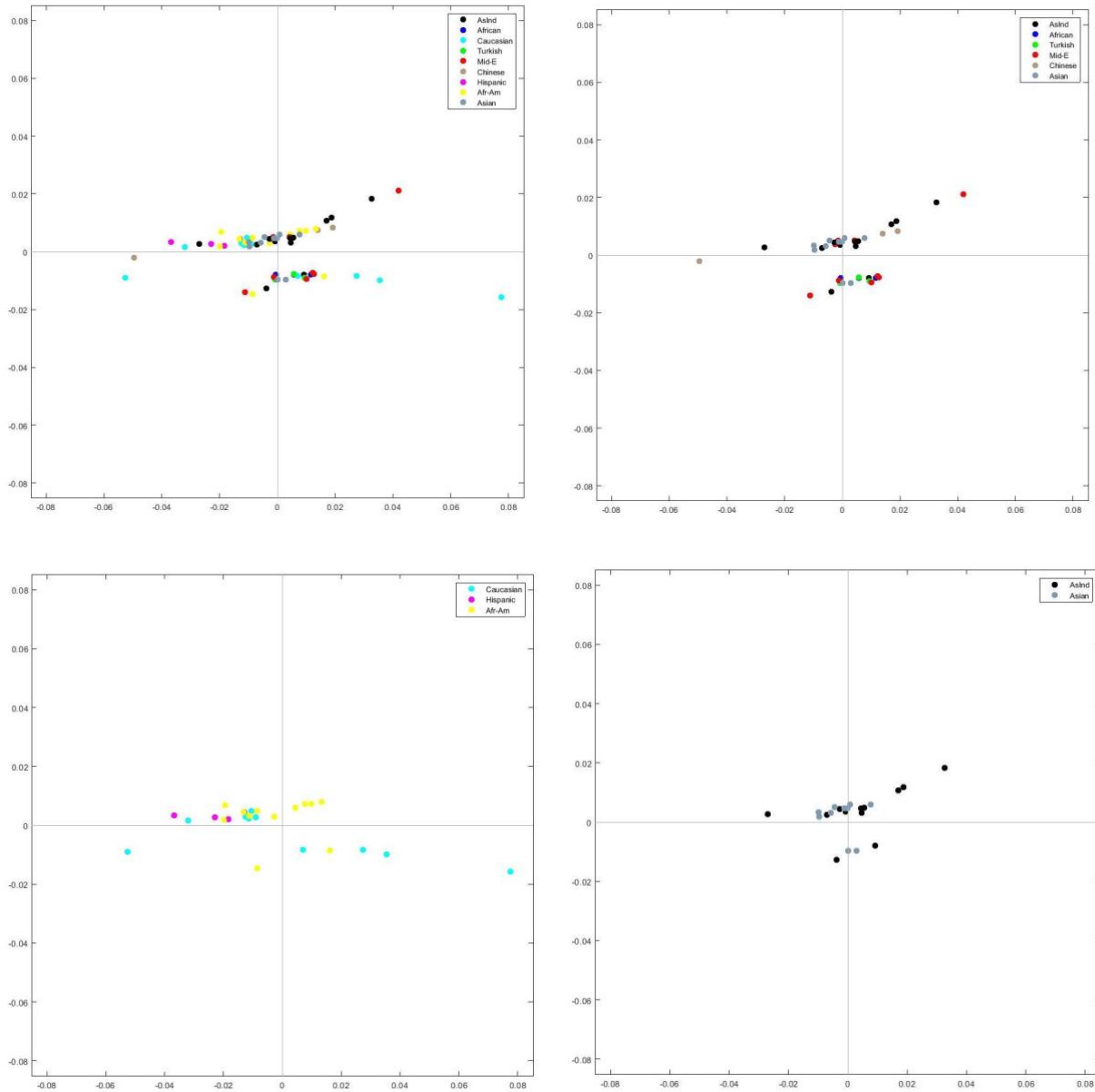


Fig D.1 PCoA plot from KLD analysis of unweighted k-mer (k=1) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

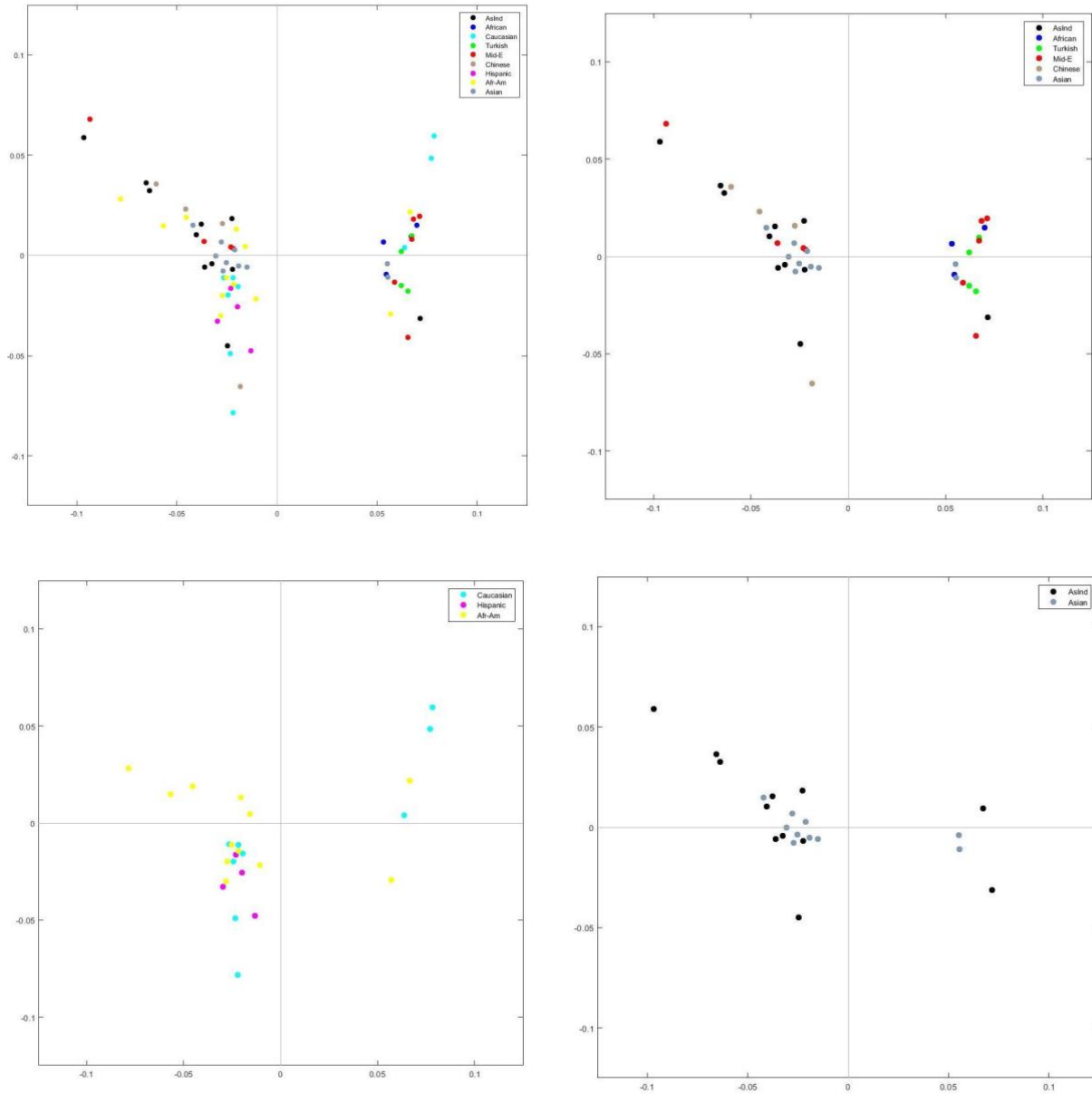


Fig D.2 PCoA plot from KLD analysis of unweighted k-mer (k=2) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

Appendix E: KLD analysis of weighted k-mer frequencies from mapped sequences

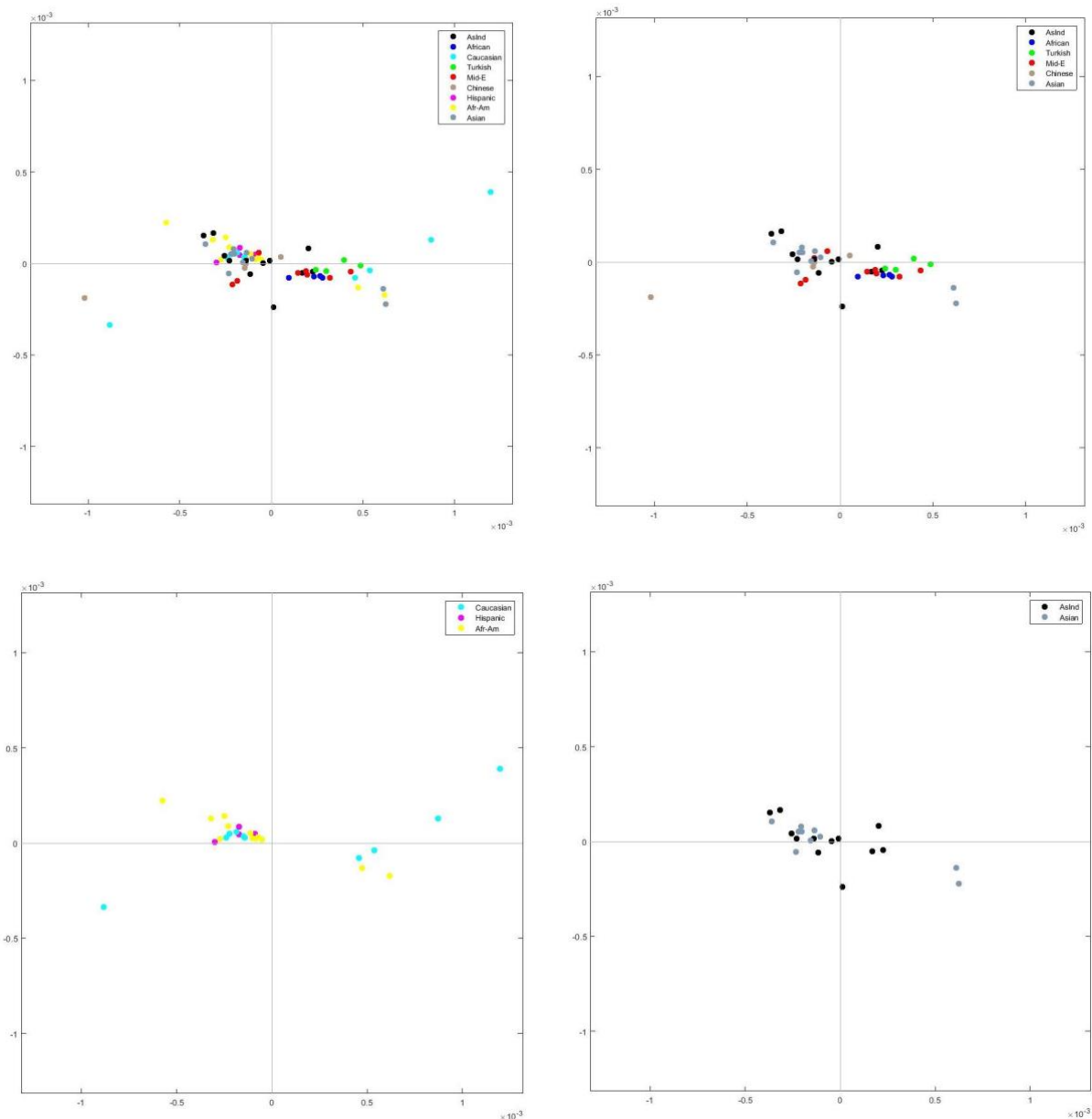


Fig E.1 PCoA plot from KLD analysis of weighted k-mer ($k=1$) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

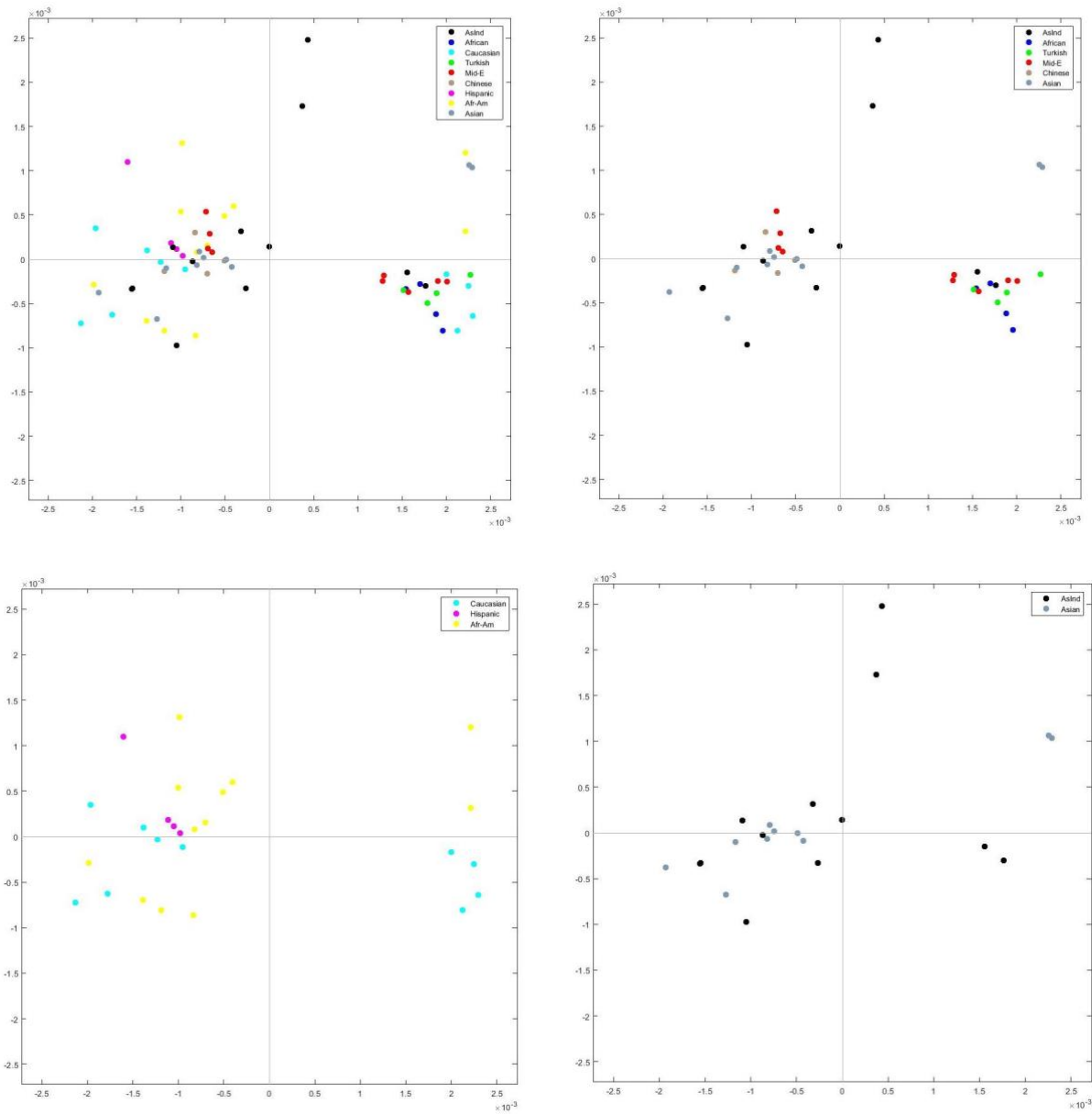


Fig E.2 PCoA plot from KLD analysis of weighted k-mer (k=2) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

Appendix F: KLD analysis of signatures of unique k-mer frequencies

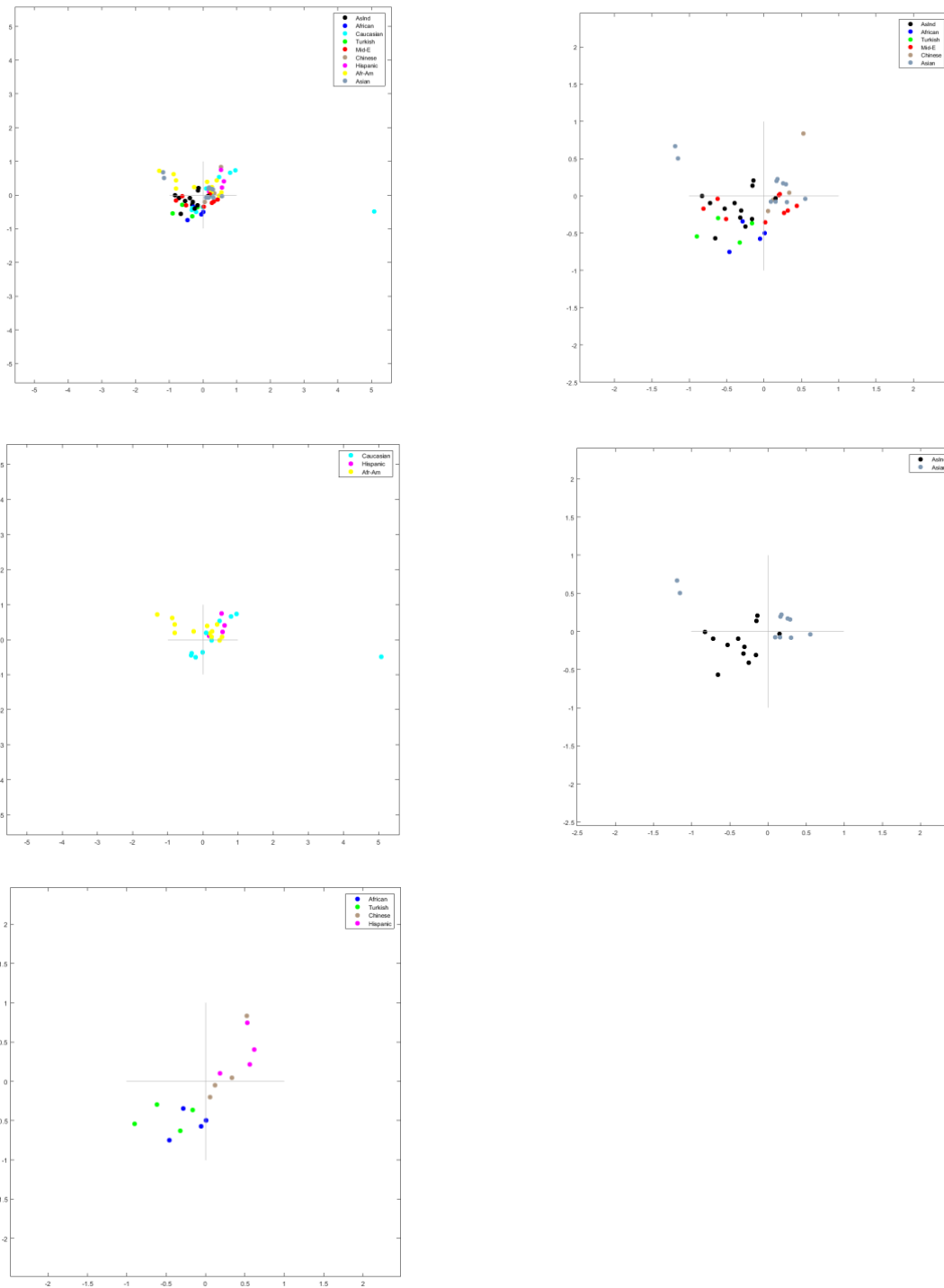


Fig F.1 PCoA plot from KLD analysis of signatures from unique k-mer (k=1) frequencies from of all 69 samples (top left), 39 samples (top right), 26 samples (middle left), 22 samples (middle right) and 16 samples (bottom left)

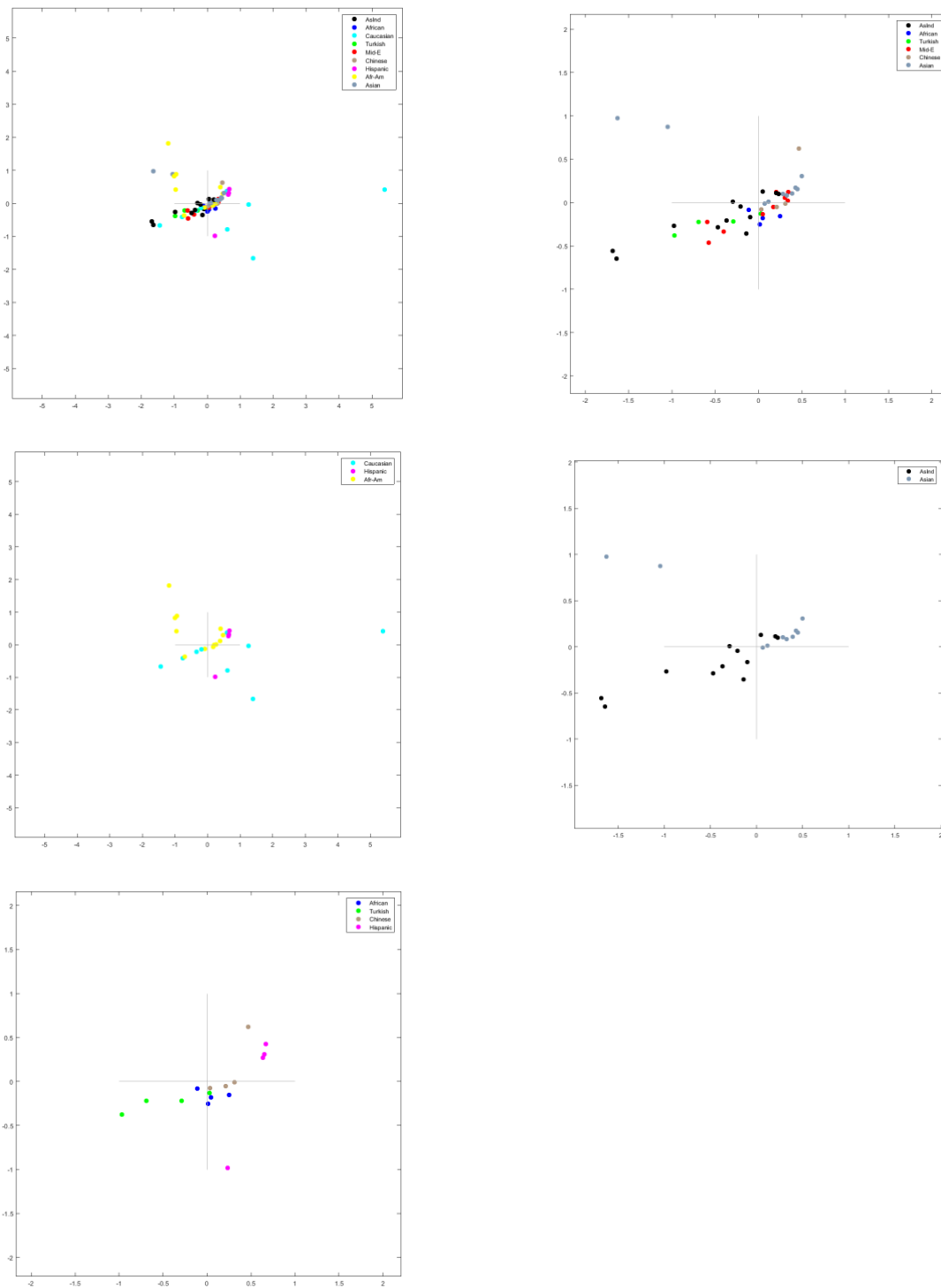


Fig F.2 PCoA plot from KLD analysis of signatures from unique k-mer ($k=2$) frequencies from of all 69 samples (top left), 39 samples (top right), 26 samples (middle left), 22 samples (middle right) and 16 samples (bottom left)

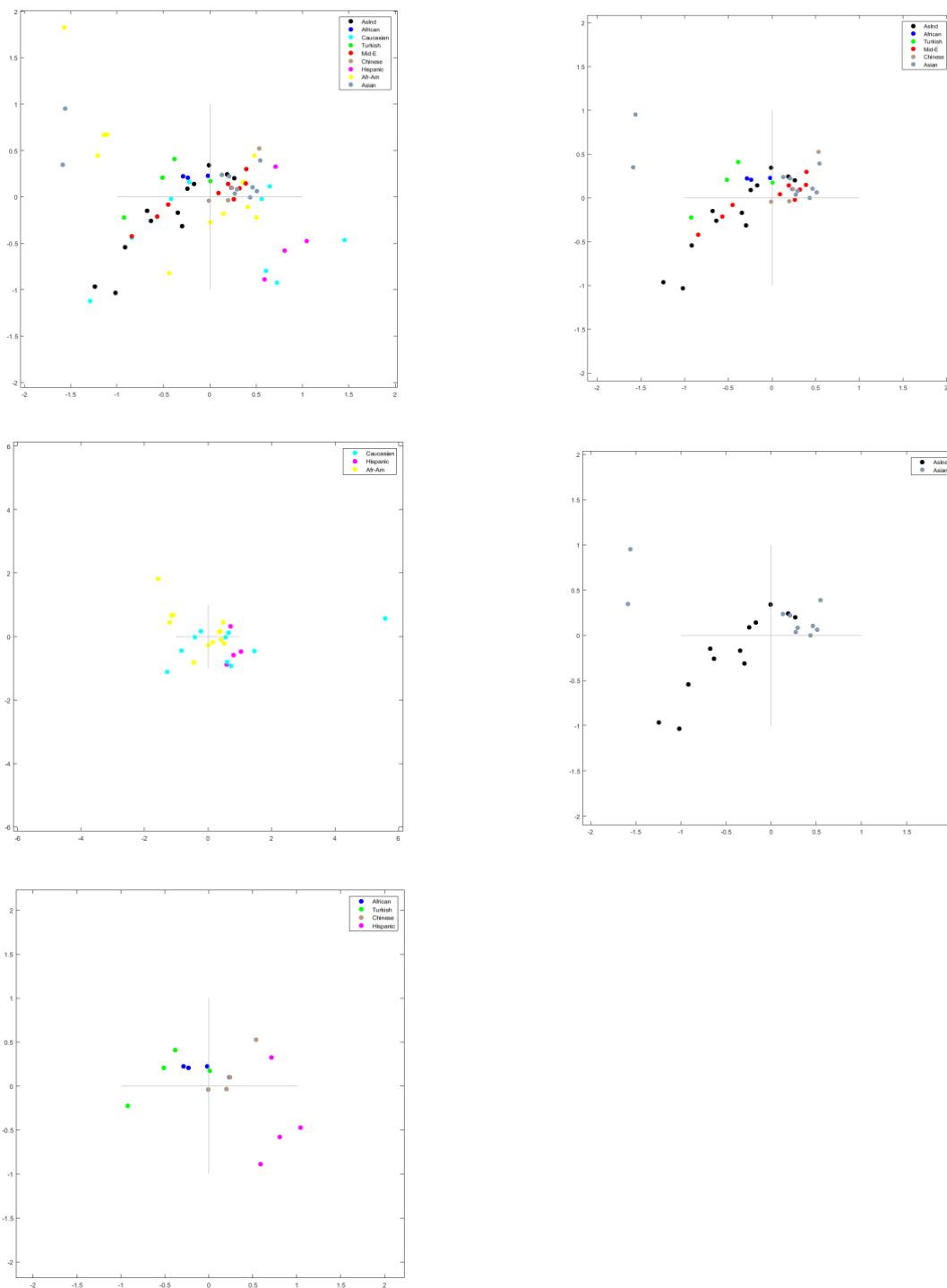


Fig F.3 PCoA plot from KLD analysis of signatures from unique k-mer ($k=3$) frequencies from of all 69 samples (top left), 39 samples (top right), 26 samples (middle left), 22 samples (middle right) and 16 samples (bottom left)

Appendix G: KLD analysis of signatures of unique k-mer frequencies from mapped sequences

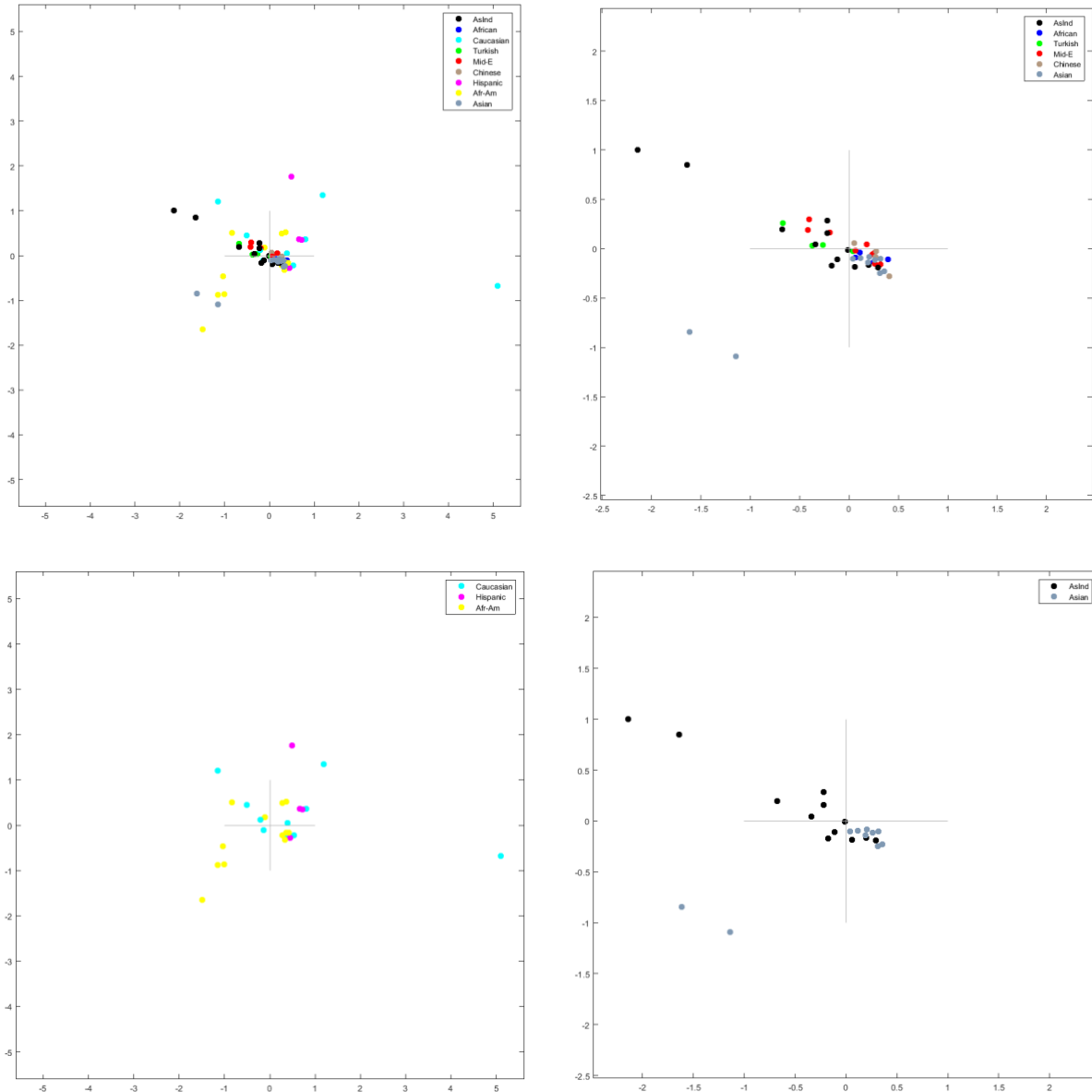


Fig G.1 PCoA plot from KLD analysis of signatures of unique k-mer ($k=1$) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups

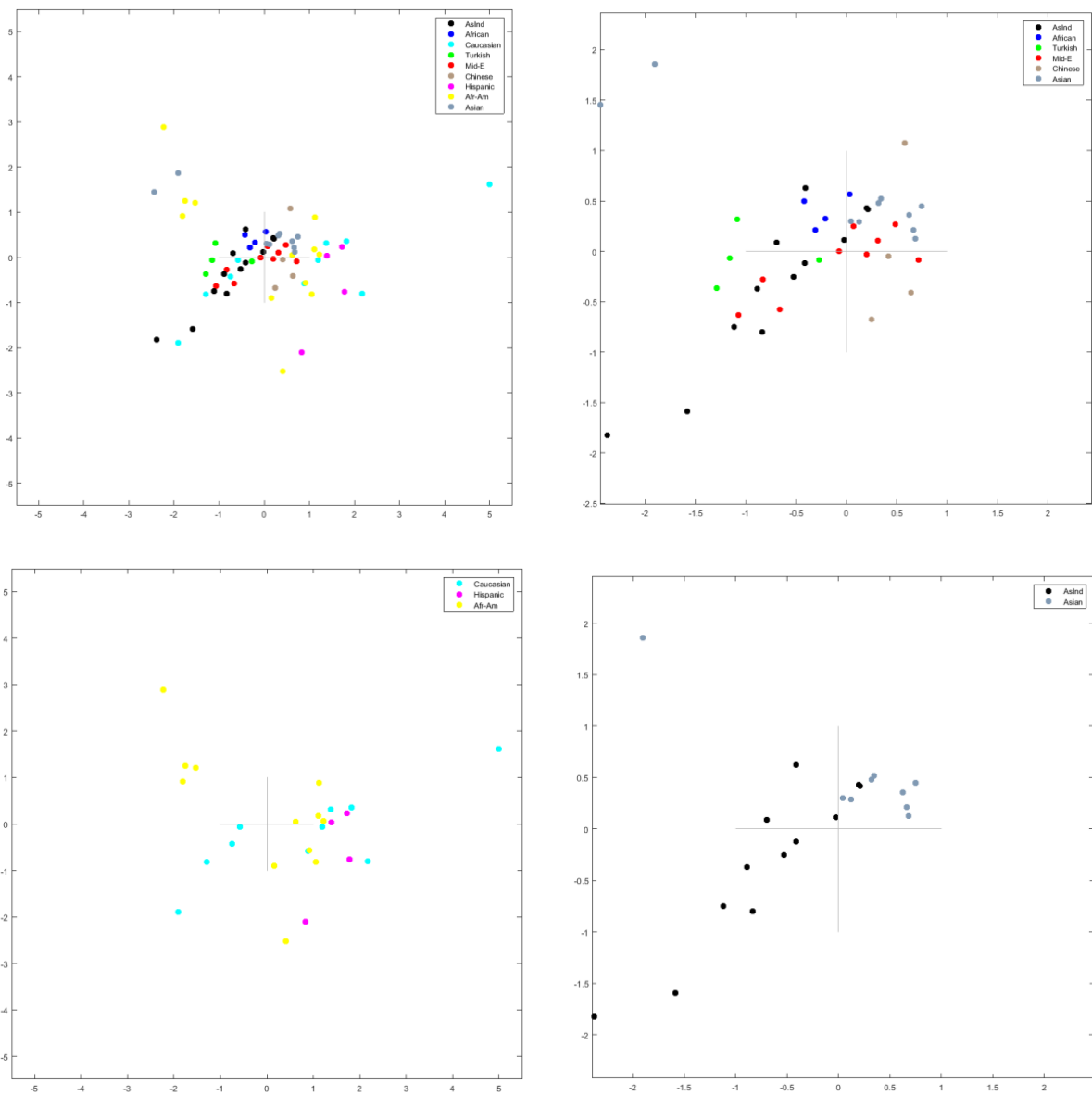


Fig G.2 PCoA plot from KLD analysis of signatures of k-mer ($k=2$) frequencies from mapped sequences of all 69 samples (top left) from 9 population groups, 39 samples (top right) from 5 population groups, 26 samples (bottom left) from 3 population groups and 22 samples (bottom right) from 2 population groups