Analyzing the Potential Occurrence of Osteoporosis and its Correlation to Cardiovascular Disease Using Predictive Analytic

Kae Sawada
Jet Propulsion Laboratory
Pasadena, USA
Kae.Sawada@jpl.nasa.gov

Michael Wayne Clark
Biology Department
Pasadena City College
Pasadena, USA
mclark7@pasadena.edu

Zilong Ye
Computer Science Department
California State University, Los Angeles
Los Angeles, USA
zye5@calstatela.ed

Nabil Alshurafa
Medicine and of Computer Science Department
Northwestern University
Evanston, USA
nabil@northwestern.edu

Mohammad Pourhomayoun
Computer Science Department
California State University, Los Angeles
Los Angeles, USA
mpourho@calstatela.edu

Abstract— **In this paper, Big Data Processing was utilized to develop and validate a Predictive Analytics Model with the goal of determining the risk for an individual manifesting osteoporosis in later life. The analyzed dataset consists of the genomic information from over 2,500 individuals from all around the world. This model development leverages the novel genetic pleiotropic information, the two or more apparently unrelated phenotypes caused by a single gene. The dataset was examined for the mutations associated with osteoporosis and cardiovascular disease from the population genetics perspectives. The study also proposes the automatic histogram clustering as an effective and intuitive visualization method for high dimensional dataset. The data visualization and clustering results revealed a significant correlation between a person's regional background and the frequency of occurrence of the 35 single nucleotide polymorphisms (SNPs). These 35 SNPs are associated with osteoporosis and/or cardiovascular disease (CVD). Distinct SNP frequency of occurrence profiles were observed for specific geographic regions. Machine learning algorithms were then applied to predict the occurrence of 7 osteoporosis-related-SNPs based on the existing CVD-related-SNPs input as an experiment. The model's validity was confirmed through a separate experimental result, utilizing a set of data obtained through Affymetrix microarray mRNA expression signal values for the specific SNP(s) in individuals with and without osteoporosis. Furthermore, these results confirmed the genetic linkage between osteoporosis and Cardiovascular related parameters such as High Density Lipoprotein (HDL) and Systolic Blood Pressure (SBP). A useful Predictive Analytics Model for determining these genetic predispositions have been produced.**

Keywords- osteoporosis; 1000 Genome Project; Machine Learning; Predictive Model; Genome Wide Association Study (GWAS); Data Visualization; Clustering; Classifiers; Supervised Learning.

I. INTRODUCTION

Early prediction and detection of chronic diseases such as osteoporosis gives patients an opportunity to make lifestyle changes, which can ameliorate the severity of the disorder. It will also enable patients and doctors to prevent the medically adverse events associated with the chronic disease. Here, machine learning based classifiers were developed, trained, and tested to predict Osteoporosis at an early stage.

This study, originally presented in Biocomputational Systems and Biotechnologies, 2018. [1], investigates the occurrence of 7 selected osteoporosis-related Single-Nucleotide Polymorphism (SNP) [2], and their correlation to 28 CVD related SNPs [3]. In addition, the sample's geographic background, along with the physical proximities between each of the SNPs were examined.

Furthermore, this paper proposes a novel approach to automatic data-driven clustering of histogram presented data for verification and validation of disease related expression in different human populations. As explained in later sections, high dimensional datasets can be effectively and intuitively visualized by the algorithm generating the histogram clustering. This automated process could aid in understanding existing correlations among various types of large datasets.

A. Osteoporosis and Space Exploration

Fractures, as a result of osteoporosis, will have a significant negative impact on an individual's health, quality of life, and work performance. Some modern occupations inevitably expose workers to a significantly increased risk of developing Osteoporosis. An obvious example is space flight. Within a few days of Zero Gravity (zero G), astronauts begin to lose both muscle and bone mass. There is also a zero G suppression effect on the immune system, as well as an increase in the aging process of particular cells. These effects occurred in all individuals who exposed themselves to a zero G environment, which can last for months after returning to Earth [4]. It is reasonable to assume that a predisposition to Osteoporosis might increase the occurrence of the condition. Having a reasonably reliable predictive model to reveal the predisposition for osteoporosis can allow us to apply preventive approaches or medical intervention to prevent the occurrence of the condition.

The 1000 Genome Project [5] sequenced over 7 TB of genetic data. The resulting datasets are collected from over 2,500 individuals from all over the world [6]. One effective way of utilizing such data is to understand the correlations between the observed variations in the DNA sequences in specific locations on the chromosome. Understanding the variations that influence specific diseases and conditions allows the prediction of risk for developing specific diseases/conditions in an individual's life time. Genome-Wide Association Studies (GWAS) have shown more than 107 genes and 129 SNPs to be associated with osteoporosis [7]. Such large genomic data sets allow for the study of complex disease states such as osteoporosis and cardiovascular disease. These pathologies involve not only multiple gene mutation interactions, but also nutritional and age-related components [8] [2]. Unfortunately, many of these gene mutations still maintain some level of phenotypic ambiguity [9] [10]. Modern Machine Learning analysis can potentially be employed to overcome the difficulty imposed by the multi-dimensional nature of the osteoporosis and CVD risk factors in analyzing the sequences. Such applications allow for early forecast of a patient-specific frequency of occurrence for a particular genetic disorder in later life. It allows for appropriate preventative measures to be followed.

The available sequenced genomes from the 1000 Genome Project enables direct examination of mutation variants with known relations to specified medical conditions such as osteoporosis. Some Machine Learning algorithms have been successfully utilized to decompose and understand the complex nature of genomic sequences in recent studies. Hidden Markov model utilization succeeded in identifying protein coding genes [11]. Support Vector Machine and Artificial Neural Network models identified certain genes' functional RNA encoding [12]. Random forest algorithms were able to predict the phenotypic effects of non-synonymous single nucleotide polymorphisms (SNPs) [13].

Several classification algorithms were utilized in this paper to predict genetic markers of osteoporosis, given genetic markers of CVD. The results of the predictor were used to analyze predisposition to osteoporosis in relation to demographic background. The first machine learning model developed in this study confirms a genetic link between osteoporosis and CVD, which has been observed in patients. These results provide clues to understand genetic linkages in the context of population generics. Particularly, when analyzing larger population data sets such as the 100,000 genomes in the UK, a result set obtained here will provide essential insights into future analysis.

B. Mutations: SNPs

This paper analyzes 35 SNPs, that are commonly observed as genetic disease related mutations. It is a mutation of one base for another, which occurs in more than one percent of the general population [14]. 7 of these SNPs have direct indications in the expression of osteoporosis [9], while the other 28 SNPs have implications in both CVD and osteoporosis [4].

C. Preceding Related Work

i. Osteoporosis-related-SNP selection

The 7 osteoporosis-related-SNPs were chosen based on the study led by Hsu et al. [2], published in 2010. The phenotypic association of these SNPs to osteoporosis was demonstrated by the GWAS study [7].

ii. Genetic Pleiotropy

Using False Discovery Rate (FDR) statistical methods, Reppe et al. [3] revealed a potential genetic link between Cardio-Vascular Disease (CVD) and Osteoporotic conditions. In this paper, the potential mutant gene interactions between the osteoporosis-related SNPs and CVD SNPs are analyzed with: big data processing and analytics, predictive analytics based on machine learning algorithms, data visualization, and clustering.

This paper is organized as follows: Section I is the introduction. In Section II, the methods and materials employed during the analysis and experimental developments are introduced. Included topics are the datasets and methods used, feature selections and dataset label, and evaluation methods. In Section III, the results and observations are discussed. Finally, Section IV will discuss conclusions.

II. METHODS AND MATERIALS

This section describes the datasets and methods employed in this paper.

A. Datasets

i. The first dataset includes the genotypes of 35 SNPs (See Table I and Table II) from the 1000 Genome Project collected from 2504 human subjects, both male and female, from 26 regions worldwide [5]. Given the 28 SNPs related to CVD, the objective was to predict occurrence of the other 7 Osteoporosis-related SNPs (Table I),. In other words, feed corresponding features that represents SNPs listed in Table II, and predict the occurrence of SNPs listed in Table I. After preprocessing the data, a total number of 112 features (See

section *II-B-ii*) were extracted for the 28 SNPs (four possible pairs per SNPs), along with gender and region. The output label is a binary label indicating whether or not the individual has two or more of the high-risk Osteoporosis-related SNP(s) simultaneously. Note that this dataset does not contain the information indicating whether or not an individual developed osteoporosis.

The input dataset used in this study included 35 sets of comma-separated values (CSV) files, corresponding to each of the observed SNPs. Each set consists of 2504 samples, which consists of the following information:

- Sample ID with gender implication (Male/Female/Unknown).
- Genotype (forward strand).
- Population(s): 5 population categories, each divided into 4 - 7 subcategories.

Note: Gene locations of each SNP were appended to the given datasets to observe the proximities of each SNP to another

ii The second dataset is from Reppe et al's study mentioned in Section I-C-ii [16]. The samples of this dataset are collected from the 84 post-menopausal females between the ages of 50 to 86 years old in Lovisenberg Diakonale Hospital located in Sweden. There are two components to this dataset:
(1) The result of Affymetrix Microarray analysis of the patients, the Affymetrix microarray signal values per sample, one ".CELL" file per patient.
(2). A set of biopsy results, sample ID (anonymous), age, gender, and the biopsy results of bone density scores, both T- and Z-scores, consisting of average neck, total hip, and average spine of each subject.
In this paper, all ".CELL" files from (1) were processed and interfaced with the library, pd.hg.u133.plus.2 [17], to obtain the gene symbol per an array cell. Then, it was interfaced with SNP identifier (e.g., #rsxxx..x, with x composing an integer) to acquire the existing SNP(s) data per each sample. In acquisition of SNP existence, the signal threshold value was determined the average score of all samples per column (per array cell). Such determination was made based on the preceding study, which defined signal thresholds in DNA microarray analysis [18]. Precise mapping among gene identifier, SNP identifier, and chromosomal location was achieved via ICSC Genome Browser [19] .

The label per sample was acquired based on the data from (2). Samples with the T-scores of -2.5 or less in one or more of neck, hip, and spine are marked as osteoporotic. The threshold was determined, following the World Health Organization (WHO) international reference standard for osteoporosis diagnosis [20]. T-scores were used instead of Z-scores, based on the sample's age and the guidelines provided by WHO.

B. Method Design

i. Problem Definition

This study investigates the occurrence of 7 selected osteoporosis-related Single-Nucleotide Polymorphism (SNP) [2] and their correlation to 28 CVD related SNPs [3] (See I-C-i for osteoporosis-related SNPs and I-C-ii for CVD-related SNPs). In addition, the sample's geographic background, along with the physical proximities between each of the SNPs were examined. The CVD related SNPs were divided into six subcategories: High Density Lipids proteins cholesterol (HDL), Low Density Lips proteins cholesterol (LDL), Systolic Blood Pressure (SBP), Diatonic Blood Pressure (DBP), Type 1 Diabetes (T1D), and Triglycerides (TG). The table of osteoporosis-related SNPs (See Appendix A) describes the SNP identification number, normal (ancestral) base, high-risk (mutated) base, and the homozygous base pairs that are associated with a high risk of Bone Mineral Density (BMD) loss, and consequently the development of osteoporosis.

ii. Features

As mentioned in the section II-B-i, 28 CVD-related SNPs, gender, and regional background of each sample were fed to the predictor to determine the presence of the 7 osteoporosis-related SNPs. The label for this predictor is a boolean value, whether the individual has one or more high-risk osteoporosis-related SNP(s).

iii. Predictive Analytics Algorithms: Classifiers

Various predictor algorithms were applied and compared: KNN, Logistic Regression, Decision Tree, Naive Bayes, Adaboost, Random Forest, and Support Vector Machine. A systematic aggregation of these classifier results is future work, for example using ensemble learning algorithm.

iv. Affymetrix Microarray

The second dataset introduced in II-A-ii-2) contains both the sample's genetic profile and BMD profile. However, it must be noted the dataset does not represent a wide scope of populations.
Running the mentioned predictor on this dataset will not only allow for validation of the proposed predictor design, but also provide an idea of how likely the mutant genes are to express their phenotype.

DNA microarray analysis allows for an evaluation of gene expressions profiles in a living individual. Such analysis provides data about the actual utilization of a specific gene by a specific organism. Thus, a SNP can be shown to be actively used by the individual increasing the likelihood of the SNP being related to the phenotype, the manifestation of a disease state.

Affymetrix microarray consists of a grid of oligonucleotide probes produced to have a known DNA sequence. The grid Microarray thus holds a specific SNP

mutation at a specific locus on the grid. Preparations of labelled mRNA (cDNA / cRNA) taken from the individual patients can then be exposed to the entire grid containing the variety of SNP mutations. Identification of a specific SNP in the patient is determined by the measured level of hybridization with the corresponding target grid position and the labelled cDNA/cRNA.

The corresponding SNP IDs were mapped through the affy ID and a gene symbol that are assigned to each cell, as well as the manual mappings of the target SNPs through a capability of the genome browser provided by University of California, Santa Cruz [19].

v. Prediction Accuracy Measure

In all cases of this study, the accuracy of the predictor was measured using 15 to 35-fold cross-validation [21].

vi. Phenotype Expression Measurement

The first dataset (see II-A-i), acquired from 1000 Genome Project lacked information on whether each individual developed osteoporosis or not. Therefore, another set of samples was sought to evaluate the geno-pheno-transfer rate as well as the validity of the predictor developed in this study. Two pieces of information had to be present in the data:

1) Each sample's SNP profiles
2) Presence of the condition, osteoporosis, in each sample

vii. Visualization

For most humans, visualizing the dataset with a dimension greater than 3 is difficult, if not impossible. This study visualizes the dataset of over 30 essential features by generating various histograms and applying a K-mean clustering algorithm implementation to cluster the resulting plots into groups. As demonstrated in the result section, the existing correlations in the dataset are clearly displayed through this visualization method. Such method of visualization can aid the observers in developing an intuitive understanding of the dataset, effectively mirroring the datasets' characteristics and patterns in them. In addition, this visualization method will aid in the automated identification of patterns in large datasets.

This section describes the result of the predictors ran against the two datasets, the data visualization and observed patterns.

A. Predictor results – Dataset from 1000 Genome Project

The predictions were performed by the classifiers in 32 different scenarios, each ran against an element of the powerset. For example, only the SNPs associated with T1D was input for the first iteration. TG-related SNPs were input for the second iteration, then DBP for the next iteration. Next, T1D-related and DBP-related SNPs are fed for another iteration, and so forth (see the 1st column of Table III for the combination of SNP inputs). As the Feature Length (the second column in Table III) increased, specific correlation stood out. The results obtained from various combinations of SNP inputs clearly showed a strong correlation between the 7 osteoporosis-related SNPs and the HDL[2] SNPs. Similarly, another strong correlation was found between the 7 osteoporosis-related SNPs and SBP[2] SNPs. Our best classifier achieved the accuracy score of 0.7769.

TABLE I: SNP ASSOCIATED WITH OSTEOPOROSIS AND CVD

SNP (rs ID)	Ancestral allele	Mutated allel	Possible pair	High Risk Genotype	Phenotype (associated condition)
rs2278729	G	A	AA, GG, AG, GA	AA	Osteoporosis
rs12808199	A	G	AA, GG, AG, GA	GG	Osteoporosis
rs7227401	G	T	GG, TT, GT, TG	TT	Osteoporosis
rs494453	T	C	TT, CC, TC, CT	CC	Osteoporosis
rs12151790	G	A	AA, GG, AG, GA	AA	Osteoporosis
rs2062375	C	G	CC, GG, CG, GC	GG	Osteoporosis
rs17184557	T	A	TT, AA, TA, AT	AA	Osteoporosis

TABLE II: SNP ASSOCIATED WITH OSTEOPOROSIS AND CVD

SNP (rs ID)	Ancestral allele	Mutated allel	Possible pair	High Risk Genotype	Phenotype (associated condition)
rs4957742	A	G	AA, GG, AG, GA	GG	DBP
rs665556	C	T	TT, CC, TC, CT	TT	DBP
rs10779702	A	G	AA, GG, AG, GA	GG	HDL
rs12137389	T	C	TT, CC, TC, CT	CC	HDL
rs9309664	G	A	AA, GG, AG, GA	AA	HDL
rs7594560	T	C	TT, CC, TC, CT	CC	HDL
rs10953178	C	T	TT, CC, TC, CT	TT	HDL
rs980299	T	C	TT, CC, TC, CT	CC	HDL
rs10746070	T	C	TT, CC, TC, CT	CC	HDL
rs7175531	C	T	TT, CC, TC, CT	TT	HDL
rs3198697	C	T	TT, CC, TC, CT	TT	HDL
rs756632	C	T	TT, CC, TC, CT	TT	HDL
rs4820539	G	A	AA, GG, AG, GA	AA	HDL
rs6583337	G	A	AA, GG, AG, GA	AA	LDL
rs11809524	C	T	TT, CC, TC, CT	TT	SBP
rs11675051	G	A	AA, GG, AG, GA	AA	SBP
rs13005335	A	G	AA, GG, AG, GA	GG	SBP
rs12995369	A	G	AA, GG, AG, GA	GG	SBP
rs10464592	G	A	AA, GG, AG, GA	AA	SBP
rs1670346	A	G	AA, GG, AG, GA	GG	SBP
rs13272568	A	C	AA, CC, AC, CA	CC	SBP
rs600231	G	A	AA, GG, AG, GA	AA	SBP
rs258415	C	A	AA, CC, AC, CA	AA	SBP
rs11614913	C	T	TT, CC, TC, CT	TT	SBP
rs199529	C	A	AA, CC, AC, CA	AA	SBP
rs8090312	G	A	AA, GG, AG, GA	AA	T1D
rs2282930	G	A	AA, GG, AG, GA	AA	TG
rs10851498	T	C	TT, CC, TC, CT	CC	TG

III. RESULTS

First dataset did not contain label information, indicating an individual's BMD score. The strength of this dataset was the abundance and diversity of SNP data across all available samples. Leveraging on the well-formatted and well-standardized dataset, the predictive analytics model was trained to predict the presence of osteoporosis-related SNPs. The prediction was a binary label - whether or not 2 or more homozygous osteoporosis-related SNPs were present in each

individual. The predictor predicts 'true/1' if an individual were to have 2 or more osteoporosis related SNPs simultaneously.

TABLE III: PREDICTION SCORES

name	Feature Length	Adaboost	Decision Tree	KNN	Logistic Regression	Naive Bayes	Random Forest
NONE	2	0.7331	0.7331	0.49	0.7331	0.7331	0.7331
T1D	6	0.7331	0.7331	0.6932	0.7331	0.7331	0.7331
TG	10	0.7331	0.7331	0.6773	0.7331	0.7331	0.7331
DBP	10	0.7331	0.7331	0.7052	0.7331	0.7331	0.7331
T1D_DBP	14	0.7331	0.7251	0.6892	0.7331	0.7331	0.7251
TG_T1D	14	0.7331	0.7291	0.6813	0.7331	0.7331	0.7211
TG_DBP	18	0.7331	0.6932	0.6693	0.7331	0.7331	0.6892
TG_T1D_DBP	22	0.7331	0.6853	0.6972	0.7331	0.7331	0.6773
HDL	45	0.7331	0.6096	0.6972	0.7331	0.7211	0.6972
SBP	46	0.757	0.6494	0.7052	0.753	0.7331	0.7171
HDL_T1D	49	0.7331	0.6255	0.7131	0.7331	0.7012	0.6972
T1D_SBP	50	0.757	0.6693	0.741	0.761	0.7331	0.7251
HDL_TG	53	0.7331	0.6016	0.6574	0.7331	0.7211	0.6693
HDL_DBP	53	0.7331	0.5498	0.6853	0.7331	0.7171	0.7052
TG_SBP	54	0.761	0.5777	0.6932	0.757	0.753	0.7171
SBP_DBP	54	0.7649	0.6892	0.7052	0.7689	0.7291	0.741
HDL_TG_T1D	57	0.7331	0.6295	0.6813	0.7331	0.7052	0.6773
HDL_T1D_DBP	57	0.7331	0.5976	0.7012	0.7331	0.7012	0.6773
TG_T1D_SBP	58	0.7649	0.6693	0.7211	0.7729	0.741	0.7291
T1D_SBP_DBP	58	0.761	0.7012	0.7052	0.7769	0.7251	0.745
HDL_TG_DBP	61	0.7331	0.6135	0.6653	0.7331	0.7211	0.7012
TG_SBP_DBP	62	0.757	0.6494	0.6892	0.749	0.7331	0.7171
HDL_TG_T1D_DBP	65	0.7331	0.5777	0.6972	0.7331	0.6932	0.6932
TG_T1D_SBP_DBP	66	0.7729	0.6175	0.7012	0.7729	0.7291	0.7291
HDL_SBP	89	0.757	0.6096	0.7092	0.757	0.6972	0.7131
HDL_T1D_SBP	93	0.7729	0.6614	0.7291	0.7729	0.6693	0.7729
HDL_TG_SBP	97	0.753	0.6096	0.6972	0.753	0.7012	0.7251
HDL_SBP_DBP	97	0.757	0.6653	0.6972	0.761	0.7012	0.7331
HDL_T1D_SBP_DBP	101	0.7649	0.6614	0.7291	0.7729	0.6813	0.7331
HDL_TG_T1D_SBP	101	0.7649	0.6733	0.7331	0.7689	0.6853	0.753
HDL_TG_SBP_DBP	105	0.761	0.6215	0.7012	0.757	0.7052	0.757
HDL_TG_T1D_SBP_DBP	109	0.7649	0.6693	0.7092	0.761	0.6813	0.7211

Table III shows the prediction results. The column headers point to the classifier algorithms used, and the row headers, name", point to the varied combination of CVD-related SNP feature sets. First, classical algorithms, KNN, Naïve Bayyes, Decision Tree, and Logistic Regression were applied to the given dataset. Most algorithms yielded a better result as the number of features increased, while the results of Decision Tree worsened, demonstrating overfitting. Stacking (ensemble learning) algorithm was implemented with the 4 classic algorithms noted above, logistic regression, and Support Vector Machine as its base classifiers. The performance was measured with the 15-fold cross-validation. Predictor's performance improved with this classifier orchestration. The more robust and optimized ensemble learning methods, Adaboost and Random Forest implementations with decision stumps as their base classifiers were applied, and results were compared. A standard ANN implementation did not perform well with this dataset due to the small size of the dataset. The highest result of 0.7769 accuracy score was captured with Logistic Regression when HDL, TG, SBP, and DBP subcategories of the CVD-SNPs were used to train and test the model. For more information on parameter values applied, please contact the authors.

The prediction results demonstrate the relationships between CVD-related SNPs and osteoporosis disease. These results show calculable relationships between CVD and Osteoporosis. The Predictor could now be used for early stage warning for future disorders.This implies that the CVD patients may have a higher chance of developing osteoporosis. It also indicates the possibility of predicting Osteoporosis at an early stage based on CVD-related factors such as Cholesterol, Blood Pressure, and Triglycerides levels.

B. Predictor results – Affy dataset

The performances of the predictive model for 3 different scenarios are listed in the following. We calculated ROC AUC (Receiver Operating Characteristic – Area Under Curve) to evaluate the performance of the system. Unlike the previous experiment, this experiment setting did not distinguish high-risk heterozygous pair vs. high-risk homozygous pair. Here are the results for 3 different

Scenarios:

scenario 1 - Osteoporosis SNPs only – AUC = 0.7285
scenario 2 - CVD SNPs only – AUC = 0.7569
scenario 3 - Both CVD and Osteo SNPs - AUC = 0.8571

The predictor strategy developed here has been confirmed with the available disease related database of Reppe et al. The predictor was shown to be correct in the majority of the patients. These results also confirm the correlation of osteoporosis SNPs with CVD SNPs reported in the previous section.

In most cases, the chance of a mutant genotype expressing its aberrant phenotype is much less than 100%. As a well-studied example, the two mutations related to Breast Cancer, BRCA – 1 and BRAC – 2, show such occurance. The chances these two mutations expressing their phenotypes by the time a woman is 70 years old are 55% for BRCA-1, and 45% for BRCA-2, according to a report by the Susan G. Komen Foundation [22].

The accuracy of the predictor developed with this paper of close to 70% demonstrated an effective prediction. Such a score would thus be reasonably effective to caution individuals about having a higher risk of BMD loss. With such knowledge, individuals can take the necessary preventative measures to prevent the development of undesirable conditions and disease, starting at an early stage of their lives.

C. Statistical Analytics and Data Visualization

To effectively visualize the existing correlations among this moderately high dimensional dataset, various histograms were drawn, using python Matplotlib library [23]. The histograms shown in the following figures demonstrate the number of occurrences of each SNP that corresponds to the indices on the x-axis. For the numerical values of all visualized data, please contact the authors.

The histograms drawn based on gender, male and female, did not display significant variation in their osteoporosis-related SNP profiles. However, distinct patterns were observed in the histograms drawn based on various geographical regions. Figure 1 demonstrates two completely different SNP profiles that come from two

regions that are located far apart. Figure 2 demonstrates two similar SNP profiles that come from two regions that are closely located.

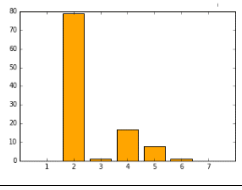
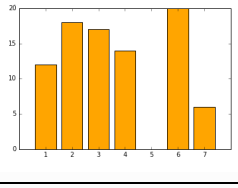
SNP Frequency-African Caribbean in Barbados	**SNP frequency-Northern and Western Europe region**
	
rs2278729_A|A 0.0 rs12808199_G|G 79.0 rs7227401_T|T 1.0 rs494453_C|C 17.0 rs12151790_A|A 8.0 rs2062375_G|G 1.0 rs17184557_A|A 0.0	rs2278729_A|A 12.0 rs12808199_G|G 18.0 rs7227401_T|T 17.0 rs494453_C|C 14.0 rs12151790_A|A 0.0 rs2062375_G|G 20.0 rs17184557_A|A 6.0

Figure 1: Varied SNP Profiles

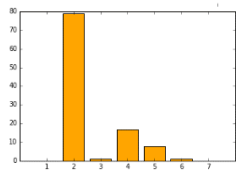
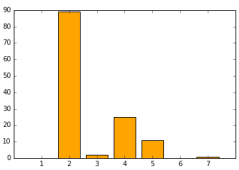
SNP frequency: African Caribbean in Barbados	**SNP frequency: Esan in Nigeria**
	
rs2278729_A|A 0.0 rs12808199_G|G 79.0 rs7227401_T|T 1.0 rs494453_C|C 17.0 rs12151790_A|A 8.0 rs2062375_G|G 1.0 rs17184557_A|A 0.0	rs2278729_A|A 0.0 rs12808199_G|G 89.0 rs7227401_T|T 2.0 rs494453_C|C 25.0 rs12151790_A|A 11.0 rs2062375_G|G 0.0 rs17184557_A|A 1.0

Figure 2: Similar SNP Profiles

To further investigate the histogram patterns, all histograms were manually grouped based on profile similarity. The results of the division are five larger categories and twelve subcategories.

The aim of this section is to group the histograms into several categories solely based on the shape of the plots, thus SNP profile. To eliminate any bias in the histogram grouping, we removed all labels from the plots. The results demonstrated that the grouping solely based on plot shapes perfectly matched the grouping based on the geographical regions. The groups also demonstrated that regional backgrounds known to have higher rates of Osteoporosis also have more SNPs present (see Fig 3). Thus, two important conclusions can be drawn here:

1) The 7 SNP profiles are dependent on an individual's regional background.

2) The more of these 7 SNPs the individuals have, the more they are predisposed to developing osteoporosis.

1. Regional Divide

The grouping results demonstrated the divides in people's SNP characteristics. This result is consistent with the idea that a person's genetic construct is dependent on the region, in which his/her genes have evolved.

2. Number of SNPs in an Individual

It has been statistically shown that Europeans tend to develop osteoporosis more frequently and they are more prone to bone mass density (BMD) loss, thus leading to major bone fractures, compared to Africans. It has also been shown that compared to Swedish Europeans, East Asians have a lesser chance of developing osteoporosis, particularly in women [4]. Even though we concluded that there were no significant differences in osteoporosis-related SNP distributions among the two genders, the differences in hormonal systems and the pregnancy events make females more vulnerable to a development of osteoporosis [24]. The results obtained here are consistent with such claims.

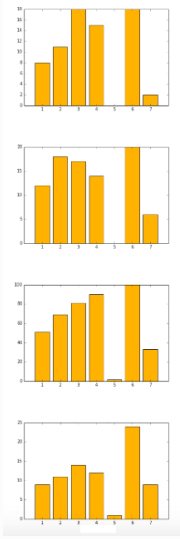
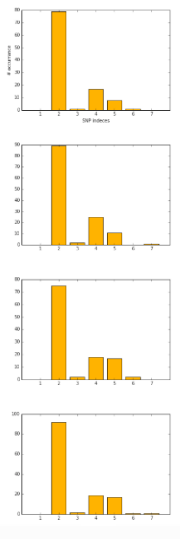
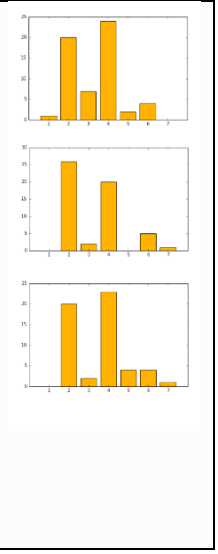
European Regional Profiles	African Regional Profiles	Asian Regional Profiles
		

Figure 3: SNP Regional SNP Profile Comparison

A high frequency of Osteoporosis-related SNPs is observed in Europe while a low frequency of them are observed in Africa region. European Regional Profiles: [British in England and Scotland (1st graph, top to bottom), Northern and Western European Ancestry (2nd graph), Europa (3rd graph), Iberian Population in Spain (4th graph)], African Regional Profiles [African Caribbean in Barbados (1st graph), Esan in Nigeria (2nd graph), Mende in Sierra Leone (3rd graph), Yoruba in Ibadan Nigeria (4th graph)], Asian Regional Profiles: [Kinh in Ho Chi Minh City – Vietnam (1st graph), Han Chinese in Beijing – China (2nd graph), Southern Han Chinese (3rd graph)]

As shown in Figure 3, the 7 osteoporosis-related SNPs are found much more frequently in European region, compared to Africa or Asia. This observation is consistent with the fact that European women tend to have a higher incident of osteoporosis and hip fractures. Individuals of European heritage have a much greater possibility of the presence of osteoporotic mutations than African individuals [25].

Such a result leads to the following question: Do these regional groupings change with the inclusion of another SNP related to osteoporosis? What would happen to the histogram profile if we added more SNPs that are associated with osteoporosis/BMD to the histogram plots?

According to the GWAS Catalogue, there are over 60 to 107 identified SNPs that have correlations to osteoporosis/BMD loss, in addition to the 7 osteoporosis-related SNPs that are strongly associated by GWAS [7].

F. Plot Grouping (8 SNPs) to Test Correlation

In this section, another SNP (randomly chosen from the osteoporosis-related SNPs) was added, histograms were drawn, and grouping was done in the same way as on the original 7 SNP regional profiling. The 8th SNP, rs2569031, was the identified high-risk base pair added here. The overview of the grouping results is shown in Figure 4.

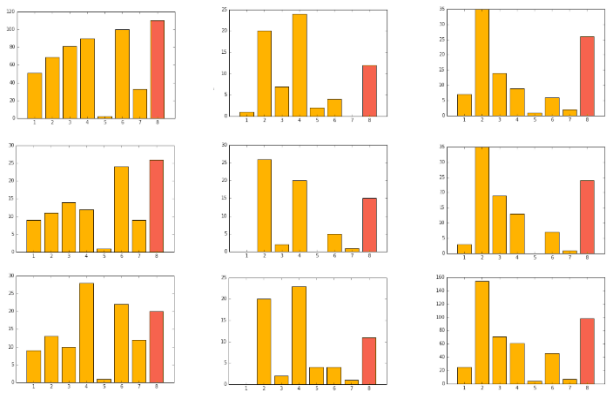


Figure 4: Randomly Chosen Additional SNP Profile Injection

The new SNP profile clustering (Figure 4) clearly follows the trend derived from the 7 SNPs grouping. The 8th SNP was found frequently in the leftmost group (European region), whereas it was found less frequently in the rightmost group (African region).

The identical procedures were applied to the CVD-related SNPs. The results of the clustering confirmed high correlations between region and the SNP profile, particularly for SBP and HDL related SNPs (See Figure 5).

Figure 6 shows the comparison with the HDL SNP profiles of people of Europe and Africa. It clearly reaffirms to the points made above.

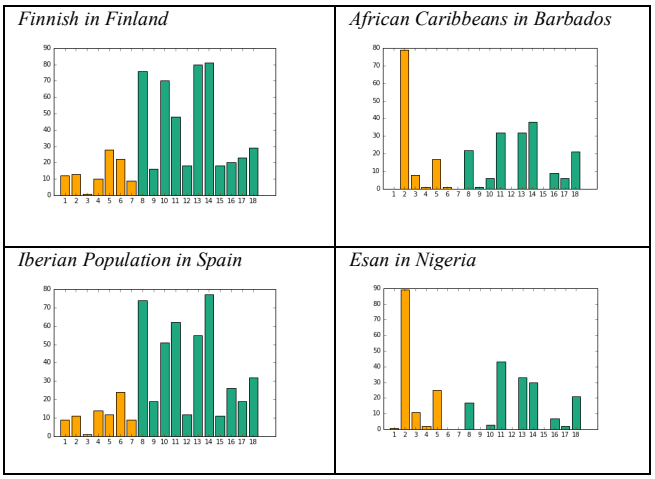


Figure 5: SNP Profiles - SBP

European Regions vs. African Regions

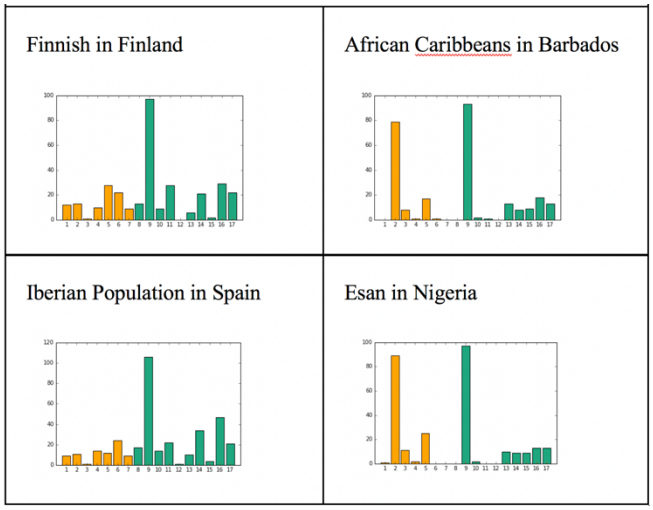


Figure 6: SNP Profiles - HDL

G. Clustering Automation

To verify the manual regional grouping of the histogram profiles, an automated clustering process was developed, utilizing KMeans clustering algorithm [26].

KMeans clustering is accomplished by minimizing the sum of the distances between the chosen centroids and each data point within the group, to which each centroid belongs to (See Equation 1). This algorithm was determined to be suitable for this experiment for two reasons: 1) The dimensionality of the dataset is not extremely large. 2) The number of clusters that the algorithm was approximated during the manual grouping.

$$t = \sum_{K=1}^{K} \sum_{i \in C_K} \|x_i - \mu_k\|. \quad (1)$$

Figure 7 shows a sample result of this automated clustering.

Sample groups obtained in automated histogram clustering result. The results verify the regional divide of osteoporosis- and CVD-related SNP profiles. Left Most Column: [Europa, Iberian Population in Spain, Utah Residents (CEPH) with Northern and Western European Ancestry, British in England and Scotland, Finnish in Finland, Toscani in Italia], Middle Column: [Bengali from Bangladesh, Sri Lankan Tamil from the UK, Indian Telugu from the UK, Gujarati Indian from Houston, Texas, Punjabi from Lahore, Pakistan], Right Most Column: [Peruvians from Lima, Peru, Puerto Ricans from Puerto Rico, Mexican Ancestry from Los Angeles USA, Colombians from Medellin, Colombia]. The result of clustering confirms the regional divide of the histogram profiles.

To demonstrate the value of this clustering automation, the final number of clusters to be formed by the algorithm was varied. As shown in Figure 8, when the number of clusters is set to 12, the algorithm groups South Asian SNP profiles and American SNP profiles into two separate groups, perfectly distinguishing the two separate regions. However, when the number of clusters is set to 11, the South Asian profiles and American profiles are bundled up into a single group. Such a result shows the possibility of further the effects of osteoporosis on humans are even important to NASA. Their studies have found that within a few days at low gravity, astronauts show significant bone density loss. NASA and other space agencies are actively investigating the exact effects of the outer space specific environments on human biological systems. Understanding such effects on skeletal system, cardiovascular system, nervous system, reproductive system, and our genome is crucial when attempting to adapt to an unknown environment. Being able to use an astronaut's genetic predisposition to predict how that individual's body might respond to zero or low Gravity could only improve the outcome of commercial space flight and the expansion of humanity into space. The Chinese state that they will have a manned Moon base by the end of 2019. What will be the effect of one sixth Earth gravity of those individuals? The data has yet to be collected. Such a result shows the possibility of further dissection of the regionalism seen here by the automated process. Automation of the histogram profiles also allowed increased number of SNPs to be analyzed all at once.

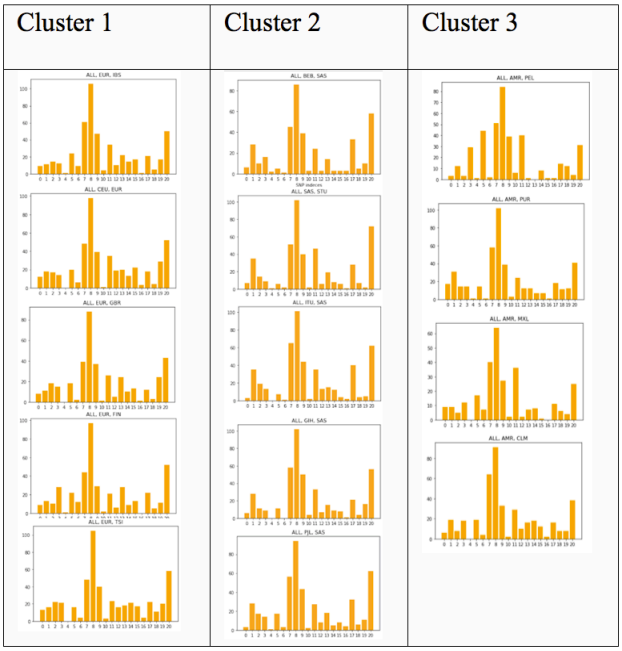


Figure 7: A Sample Histograms Demonstrating Regional Divides

H. The Number of SNPs per Individual

An unexpected finding in this study was that there was no individual who had all of the 7 osteoporosis-related SNPs simultaneously. The maximum number of SNPs that occurred in one individual was 4 out of the 7 SNPs, and 5 out of the 8 osteoporosis-related SNPs. Although four of the 7 osteoporosis-related SNPs are more likely to be found in osteoporotic females, and some are more likely to be found in males [2], all 7 SNPs were seen in at least 26 counts among males, 35 counts among females
Such cases, in which individuals have all 7 mutations were perhaps eliminated over the course of evolution.

Potential of BMD loss in individuals. Such a technique can provide an effective aid to medical professionals in the diagnosis of future disease expression based on a person's genetic profiles, biological signs, and family histories. An appropriate set of treatments/ remedies can thus be generated, utilizing the predictor.

IV. Conclusions and Discussion

The mutation frequency profiles related to Osteoporosis displayed geographical regionalism. These data are consistent with the occurrence of the clinical observation for the development of osteoporosis. This observation lead to the development of a predictive model to measure the

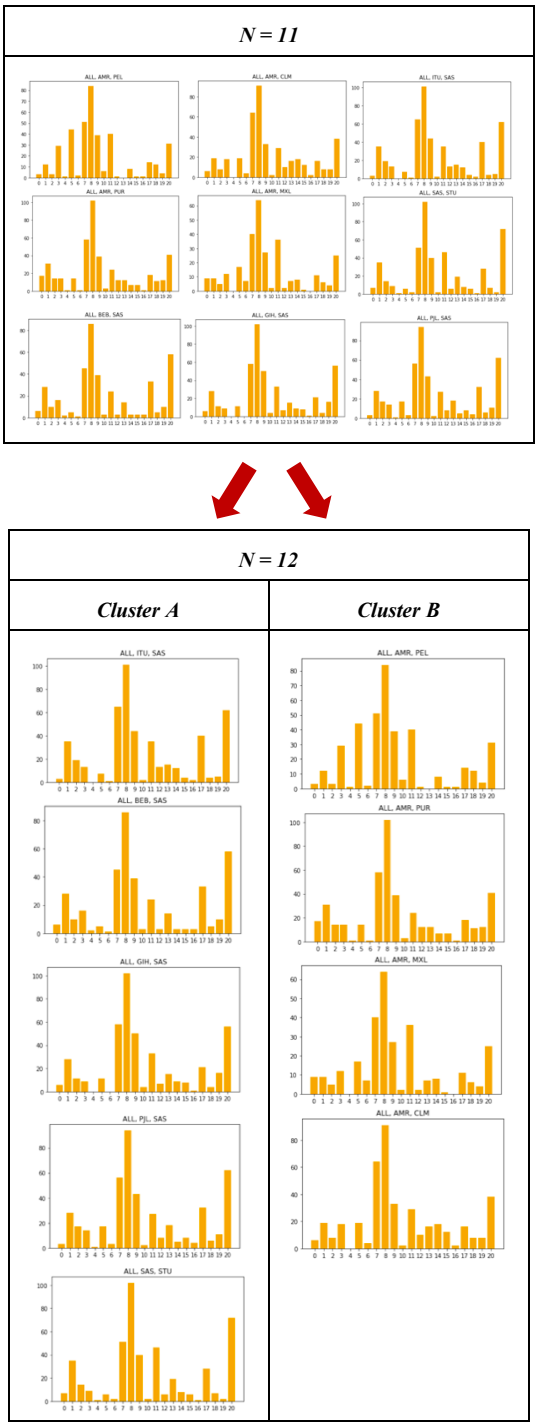


Figure 8: A Sample Auto-Clustering Result

Within the past few years, over 1200 ancient genome have been sequenced [27]. If these osteoporosis SNPs display regionalities, it would be of interest to determine how far back in human history these SNPs might go. With the publication of many ancient human genomes, we now have the opportunity to examine the fuller history of mutations among humans. We have taken a brief look at the Neanderthals genome [28] and found the presence of these SNPs. Further analysis would allow us to construct a phylogenetic tree of osteoporosis mutations. Closer look into our past may help us better educated look into the future.

Summary of the Results:

i. Our results demonstrated that there was a high correlation between a person's regional background and the occurrence of the selected 35 SNPs associated with osteoporosis and/or CVD. This finding conformed to the claims in some of the preceding Osteoporosis population/demographic studies: highest fracture rates are found in white women of European descent. African Americans tend to have higher BMD [25]. Swedish elderly women tend to suffer from bone fracture more often than Asian elderly women [5].

ii. In all 2504 individuals examined, the minimum and maximum occurrence of the selected 7 Osteoporosis-related SNPs in an individual was between 0 and 4 respectively, and 1 and 16 in all 35 SNPs. No individual with all 7 Osteoporosis-related SNPs was present among the 2504 individuals.

iii. Our results demonstrated that there were distinct correlations between the 7 Osteoporosis-related SNPs and CVD related SNPs (8 SNPs). This finding strongly indicated the CVD patients might have a higher chance of developing Osteoporosis. Such correlation data can be utilized to predict Osteoporosis at an early stage based on CVD-related factors such as Cholesterol level, Blood Pressure, and Triglycerides level.

iv. Our algorithm demonstrated that the set of CVD-/ Osteoporosis -related SNPs from Reppe et al's dataset can be used to predict a person's likelihood of developing osteoporosis on their own without factoring in any non-genetic factors such as food intake, exercise habits, and other drug intake and medication.

v. We have shown that Histogram analysis and auto-clustering of such histograms is a highly effective visualization technique, capturing existing correlations among a high dimensional dataset.

The United Kingdom is now half way through its 100,000 Genome Project. The goal of obtaining this massive human genome dataset was to determine genetic predisposition to genetic related diseases such as cancer, CVD and osteoporosis. We hope that the automated systems developed here will be of significant assistance in achieving these goals.

REFERENCES

[1] K. Sawada, M. W. Clark, Z. Ye, N. Alshurafa, and M. Pourhomayoun, "Predictive Analytics to Determine the Potential Occurrence of Genetic Disease and their Correlation: Osteoporosis and Cardiovascular Disease," the Proceeding of The Tenth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies, 2018.

[2] H. Yi-Hsiang, et al. "An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits." PLoS genetics 6, no. 6, e1000977, 2010.

[3] S. Reppe, Y. Wang, W.K. Thompson, L. K. McEvoy, A. J. Schork, V. Zuber et al. GEFOS Consortium. "Genetic sharing with cardiovascular disease risk factors and diabetes reveals novel bone mineral density loci." PloS one 10, no. 12, e0144531, 2015.

[4] E. Roberge. "The Gravity of It All: From osteoporosis to immunosuppression, exploring disease in a microgravity environment holds promise for better treatments on Earth." IEEE pulse 5, 2014.

[5] IGSR and the 1000 Genomes Project. [Online] Available at: http://www.internationalgenome.org/ [retrieved: Nov. 2018].

[6] Which populations are part of your study? [Online] Available at: http://www.internationalgenome.org/faq/which-populations-are-part-your-study [retrieved: Nov. 2018].

[7] GWAS Catalog. [Online] Available at: https://www.ebi.ac.uk/gwas/ [retrieved: Nov. 2018].

[8] O. Hitomi, S. Sasaki, H. Horiguchi, E. Oguma, K. Miyamoto, Y. Hosoi, M. K. Kim, and F. Kayama. "Dietary patterns associated with bone mineral density in premenopausal Japanese farmwomen–." The American journal of clinical nutrition 83, no. 5, pp. 1185-1192, 2006.

[9] D. Karasik, H. Yi-Hsiang, Y. Zhou, L. A. Cupples, D. P. Kiel, and S. Demissie. "Genome-wide pleiotropy of osteoporosis-related phenotypes: the Framingham study." Journal of Bone and Mineral Research 25, no. 7, pp. 1555-1563, 2010.

[10] L. Langsetmo, et al. "Using the same bone density reference database for men and women provides a simpler estimation of fracture risk." Journal of Bone and Mineral Research, no.10, pp. 2108-2114, 2010.

[11] I. M. Meyer, and R. Durbin. "Gene structure conservation aids similarity based gene prediction." Nucleic acids research 32, no. 2, pp. 776-783, 2004.

[12] R. J. Carter I. Dubchak, and S. R. Holbrook. "A computational approach to identify genes for functional RNAs in genomic sequences." Nucleic acids research 29, no. 19, pp. 3928-3938, 2001.

[13] T. Chen, M. Y. Kao, M. Tepel, J. Rush, and G. M. Church. "A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry." Journal of Computational Biology 8, no. 3, pp. 325-337, 2001.

[14] U.S. National Library of Medicine | What are Single Nucleotide Polymorphisms (SNPs)? [Online] Available at: https://ghr.nlm.nih.gov/primer/genomicresearch/snp [retrieved: Nov. 2018].

[15] Ensemble Genome Browser [Online] Available at: https://uswest.ensembl.org/index.html [retrieved: Nov. 2018].

[16] E-MEXP-1618 - Transcription profiling of bone biopsies from postmenopausal females identifies 8 genes highly associated with bone mineral density [Online] Available at: https://www.ebi.ac.uk/arrayexpress/experiments/E-MEXP-1618/files/ [retrieved: Nov. 2018].

[17] Bioconductor 3.6 Annotation package, pd.hg.u133.plus.2 [Online] Available at: http://bioconductor.org/packages/release/data/annotation/html/pd.hg.u133.plus.2.html [retrieved: Nov. 2018].

[18] M. Bilban, L. K. Buehler, S. Head, G. Desoye, and V. Quaranta. "Defining signal thresholds in DNA microarrays: exemplary application for invasive cancer." BMC genomics 3, no. 1, 19, 2002.
2015 ISCD Official Positions – Adult [Online] Available at: https://www.iscd.org/official-positions/2015-iscd-official-positions-adult/ [retrieved: Nov. 2018].

[19] UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly [Online] Available at: https://genome.ucsc.edu/ [retrieved: Nov. 2018].

[20] World Health Organization – WHO Criteria for Diagnosis of Osteoporosis [Online] Available at: http://www.4bonehealth.org/education/world-health-organization-criteria-diagnosis-osteoporosis/ [retrieved: Nov. 2018]

[21] The documentation for *sklearn.model_selection.cross_val_score* [Online] Available at: http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html [retrieved: Jan. 2018].

[22] Suzan G. Koman Foundation [Online] Available at: https://ww5.komen.org/ [retrieved: Nov. 2018]

[23] The Matplotlib API [Online] Available at: https://matplotlib.org/2.0.2/api/index.html [retrieved: Nov. 2018].

[24] P. Salari, and M. Abdollahi. "The influence of pregnancy and lactation on maternal bone health: a systematic review." Journal of family & reproductive health 8, no. 4, p. 135, 2014.

[25] N. Hae-Sung, S. S. Kweon, J. S. Choi, J. M. Zmuda, P. C. Leung, L. Y. Lui, et al. "Racial/ethnic differences in bone mineral density among older women." Journal of bone and mineral metabolism 31, 2013.

[26] klearn.cluster.KMeans [Online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html [retrieved: Nov. 2018].

[27] Harvard Medical School. "Ancient-DNA researchers surpass the 1,000-genome milestone." ScienceDaily. ScienceDaily, 21 February 2018. [online] Aailable at: www.sciencedaily.com/releases/2018/02/180221131848.htm retrieved: Nov. 2018].

[28] Ensemble Neandertal Genome browser [online] Available at: http://neandertal.ensemblgenomes.org/Homo_sapiens/Location/View?r=13:31787617-31871809 [retrieved: Nov. 2018].

[29] K. Sawada, M. W. Clark, N. Alshurafa, and M. Pourhomayoun, "Analyzing the Mutation Frequencies and Correlation of Genetic Diseases in Worldwide Populations Using Big Data Processing, Clustering, and Predictive Analytics," International Conference on Computational Science and Computational Intelligence, pp. 1459 – 1464, 2017.

[30] Epidemiology by International Osteoporosis Foundation [Online] Available at: https://www.iofbonehealth.org [retrieved: Nov. 2018].

[31] J. B. Richards, H. F. Zheng, and T. D. Spector, "Genetics of osteoporosis from genome-wide association studies: advances and challenges," Nature Reviews Genetics, 2012.

[32] US Department of Health and Human Services. "Bone health and osteoporosis: a report of the Surgeon General." Rockville, MD: US Department of Health and Human Services, Office of the Surgeon General 87, 2004.

[33] J. Greenbaum, K. Wu, L. Zhang, H. Shen, J. Zhang, and H. W. Deng. "Increased detection of genetic loci associated with risk predictors of osteoporotic fracture using a pleiotropic cFDR method." Bone 99, pp. 62-68, 2017.

[34] D. Karasik, and M. Cohen-Zinder. "Osteoporosis genetics: year 2011 in review." BoneKEy reports 1, no. 8, 2012.

[35] L. Qin, Y. Liu, et al. "Computational characterization of osteoporosis associated SNPs and genes identified by genome-wide association studies." PloS one 11, no. 3, e0150070, 2016.