International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2022; 7(5): 114-124

www.mathsjournal.com
Received: 20-06-2022
Accepted: 24-07-2022

Makambi Abuga Dennis
Department of Mathematics and Actuarial Sciences, Kisii University, Kisii, Kenya

Fred Monari1
Department of Mathematics and Actuarial Sciences, Kisii University, Kisii, Kenya

Robert Nyamao Nyabwanga
A) Department of Mathematics and Actuarial Sciences, Kisii University, Kisii, Kenya
B) School of Pure and Applied Sciences, Catholic University of Eastern Africa.

Lameck Ondieki Agasa
Department of Mathematics and Actuarial Sciences, Kisii University, Kisii, Kenya

Corresponding Author:
Makambi Abuga Dennis
Department of Mathematics and Actuarial Sciences, Kisii University, P.O Box 408-40200, Kisii, Kenya

Imputation methods on Hardy Weinberg equilibrium for missing genome wide expression data

Makambi Abuga Dennis, Fred Monari, Robert Nyamao Nyabwanga and Lameck Ondieki Agasa

Abstract
Genomic data is a common data source containing missing expression values. Deletion of missing values can lead to a serious bias of the results. One of the common methods of addressing missing values in genomics is through implementing imputation methods. This research sought to address imputation methods which are commonly used for filling missing values. The imputation methods include; RF and KNN methods. This research obtained dataset with no missing values from GENICA study. The missingness structure was created and the imputed values generated using the above-mentioned methods. Missing genotypes were removed and markers evaluated and determined it relation to Hardy−Weinberg equilibrium. Assessment for Hardy−Weinberg equilibrium under the occurrence of missingness by using exact *p*-values and inbreeding coefficients were developed. Data analysis was carried out using R-program version 4.1.2 and the results presented using relevant charts and tables. The results from the findings indicated that, dataset which were missing at random and missing completely at random were taken into consideration. Based on the findings, the study found that, random forest was the most appropriate method in imputing missing values which are missing completely at random. In addition, the method was suitable for estimation of HW proportions with the aim of maintaining HWE. K NN method was the most effective method in imputing values which were missing at random since it gave out small disparities and it was appropriate method in giving close approximations to Hardy−Weinberg equilibrium.

Keywords: Genome data, missing data, random forest, k nn and hardy Weinberg equilibrium

Introduction
The Hardy-Weinberg Equilibrium is a crucial element of population genetics. It states that the genotypic rates in a population setting becomes static between cohorts in absence of disruption by external forces."(Edwards, 2008) [12]. In relation to Hardy-Weinberg equilibrium, for a given locus with alleles *A* and *a* with equivalent frequencies *p* and *q* three genotypes are likely to occur *AA*, *Aa*, and *aa* with estimated values p^2, 2pq, q^2, respectively whereby *p* represents the allelic frequency of *A* and *q=1-P* (Graffelman, 2017) [16]. HW proportions is met in one area of random mating. In case the external forces are absent then the genotypic and allelic frequencies have to change. This phenomenon is termed as Hardy Weinberg equilibrium (HWE). Though, certain aspects like natural selection, on-random mating, gene flow, drift and mutation can cause deviations from Hardy Weinberg equilibrium (Graffelman, 2017) [16]. Genomic data may result in deviation from Hardy Weinberg equilibrium (HWE), however, the latter is not common in humans. Non-random mating from certain niche might be one of the causes in shifting of HWE. This is due to heterozygous shortage in great population at various origins (Graffelman, 2017) [16]. A common cause in shifting Hardy Weinberg equilibrium is occasionally detected in population studies is sequencing errors (Chen, 2017) [10]. It is unusual to have 15 or more missing genotype data in a genomic setup. Though, the rate of missing information may highly vary from one marker to another. For a specific marker, 0 through 100% of the data can be missing because of either corrupt information, inaccurate extraction of information or lack of proper uploading of data correctly.

At certain situations when markers are tested for Hardy Weinberg equilibrium, the missing genotype data are sometimes deleted. Removing missing data can result to larger standard errors due to reduction of population size or immerse loss of data Most studies have compared different imputation methods using normalize root square error or by use of complete case study with the aim of establishing which method is better than the other in imputing the missing genomic data (Graffelman *et al.* 2013) [14]. Little studies have addressed the effects of imputation methods on HWE. This study sought to find out the effectiveness of RF and kNN methods on HWE and come up with a decisive conclusion on which method is suitable in estimating HW equilibrium and apply it in real dataset. Addressing the issue of missingness is important and one ought to effectively understand the mechanism and arrangements of missing information. Different statistical simulations for dealing with missing data purely depends on assumptions such as missing completely at random or missing at random. Such conventions that govern missingness of data ought to be reviewed before applying imputation processes. To address the problem of missingness is not such an easy task. The assignment may be sparingly impractical and missing information may originate from unidentified responses with no way of communicating with the survey respondents. For such circumstances, uncertainty in the concept of missing data will continuously occur. The assumption of basic tests on Hardy Weinberg equilibrium can be incorrect if the missingness is ignored (Graffelman *et al.* 2013) [14]. The study sought to show that HW proportions regions are close to nominal in big datasets if single-nucleotide polymorphisms (SNPs) with missingness are not take into consideration. Knowledge obtained from imputing missing data was useful for example in treatment of certain diseases which have been difficult to handle in genomics.

Review of Related Literature

According to (Enders, 2010) [13], there exists three categories of missing mechanisms this research takes into consideration. Missing at random, missing completely at random and missing not at random. Missing data occurs when information is missing independently for both unobserved and observed data. Evaluation of different approaches for missing data on feature associated to genomics found that RF and KNN algorithms are the most appropriate methods for evaluated dataset (Ben Omega *et al.*, 2021) [43]

This research introduces effective methods for imputation; k-nearest neighbor and random forest machine learning algorithms for imputing missing data as well as regression (Srishti & Verma, 2019). In KNN method, the cost implication was very higher as compared to RF since it involved in Euclidean distance calculation for each training sample although k- nearest neighbor algorithm was time consuming, is was a simple with accuracy that proved effective in several cases.

Random forests (RF) algorithm has become the most popular method due to its applicability in missing data especially in biomedical research as compared to method methods. The advantage of RF in addressing the issue of missingness is that, it does not assume normality or need specific statistical parametric models in its applicability. Though, it is still indecisive on how it perform in non-normality distributed data or at circumstances when non-linear relationship or interactions are considered (Van Buuren, 2018) [39]. Random forest method is described an algorithm made for mixed continuous or categorical data in the issue of compound relations and non-linearity without needful distribution of the variables (Stekhoven DJ, 2012) [35]. Most studies have proposed Random forest algorithm as the most appropriate method for imputing missing data as compared to other standard imputation methods and has been used as a comparison for non-parametric statistical imputation methods (Tang F, 2017) [37].

There exists different methods for addressing missingness of values. The most common methods widely used in addressing the same is imputation and complete case analysis criterion. The applicability of complete case analysis is applied when dataset information with any missing data are completely deleted hence only cases with full complete data are considered. Complete case analysis helps in streamlining the issue by assuming data, but this method is infective since it has serious negative implications on the result. Some of the drawbacks in using complete case analysis include; it yields large standard errors for decreased sample sizes of estimates. The attained results will be biased when considering data which is missing at random, it means that, the cases with missingness are systematically different from complete case analysis method. This can lead to biasness of the expected results. Important loss of information is another key to address. When their exist a lot of missing values, the corresponding number of complete cases will become small and large dataset will be deleted. In general, selection of an imputation method should be taken into account since any imputation method has its benefits and disadvantages (Janssen, 2010) [20].

The introduction of of Mendel's law of segregation and independent assortment has contributed to foundation of population genetics. In 1908, British mathematician Godfrey H. Hardy and German physician Wilhelm Weinberg independently discovered the existence of mutual relationship between genes and genotypic frequencies called Hardy-Weinberg (HW) equilibrium. The principle gained familiarity in statistical research in both applied research and theoretical research in qualitative genetics (Chen, 2010) [9].

Hardy Weinberg Principle states that," When the organisms are diploid in a genetic population, allele frequencies are equal in the sexes, no migration, mutation, or selection, generations do not overlap, population size is significantly large and sexual reproduction then the genotypic frequencies in a certain population is given as the product of allele frequencies " In certain conditions where there exists one locus with two allelic frequencies such that [(p, q)] or (A, a)] give the genotypic frequencies (AA, aa, Aa) $p^2 + q^2 + 2pq = 1$ (Bosco F, 2012) [5]. Deviation from HWE highly imply that at least one of hypothetical is not taken in consideration.(Chen, 2010) [9]. An investigation conducted by (Graffelmam *et al.*, 2020), on chi-square exact test for HW proportions with missing genotypes showed that, dataset with missing values are deleted when markers are tested for HWE which can give inaccuracy results in statistical inference about equilibrium.

Random forest and k-nearest neighbor can recuperate inferences on equilibrium. The findings showed that, analysis of a group of markers with large amount of missing values indicated exact inferences on equilibrium and altered significantly when the level of missing is taken into consideration. For the case with high missing ness (>5%), the study found that, single imputation methods weaken the results for HWE.

In summary, (Olive, 2008) [45]; suggests that, HW principle together with Newton's law of motion states that" a body remains constant or continue to propel at a steady speed unless an external pull or push acts upon it" if that conditions remains, then it gives the foundation for detection and estimation of effects on population through considering the missing mechanism of data and sought effective imputation methods to maintain the balance.

Material and Methods 522123 70098k

Research Method

The research sought to address two methods of imputing missing values using exact-test method based on statistical inference for Hardy Weinberg equilibrium in regard to missing genomic data. Missing values can be deleted or a complete case analysis is applied when SNPs are tested for HWE. This can results to immerse loss information and give bias results. Performing single imputation methods can improve inferences on equilibrium. The study developed tests for HWE by using exact *p*-values and inbreeding confidents (equivalently, χ^2 statistics). This study restricted to biallelic genotypic markers with alleles A and B such that P_A and P_B are allele frequencies in such way that, n_{AA} n_{AB}, n_{BB}, and n_A, n_B as the genotype and allele counts, and n represents total sample size.

Imputation methods

Single imputation method is a common tool for addressing missingness in Genome wide data (Little, 2002) [26]. It consists of test statistics which will be calculated for each single imputed data. In case Hardy Weinberg equilibrium is tested by exact chi-square statistics, then single imputation can be applied together with inbreeding coefficients (Graffelman J. a., 2013) [14]. Incase exact chi-square exact test is used for approximation of HWE, then SI method can be applied using
p-values.

Inbreeding coefficients

Hardy Weinberg equilibrium can be parameterized by inbreeding coefficient *f*. this can be formulated by;

$$P_{AA}=P_A^2+P_AP_Bf,$$

$$P_{AB}=2P_AP_B(1-f), \quad \text{(I)}$$

$$P_{BB}=P_B^2+P_AP_Bf,$$

$$\text{With}\frac{-P_m}{1-P_m} \leq f \leq 1$$

whereby P_m represents minor allele frequency which is $min(P_A,P_B)$.
If$f=0$, then the genotype frequencies resembles to the HE proportions. For the case $f>0$ it indicates deficiency of heterozygotes, and for $f<0$, there exist excess heterozygotes. f can be estimated by maximum likelihood (ML) through multinomial distribution for genotype frequencies, and its variance can be represented as

$$f=\frac{4n_{AA}n_{BB}-n_{AB}^2}{n_An_B}$$

and

$$var(f)=\frac{(1-f)^2(1-2f)}{n}+\frac{f(1-f)(2-f)}{2nP_A(1-P_A)} \quad \text{(II)}$$

From the above equations we note that, MLE is associated to the classical χ^2 test statistic for HWE by $\chi^2=nf^2$. The study take into confide ration a multi normal regression model that customs allele frequency and uses markers as predictors. Single imputation methods yield a group of *m* data matrices of genotype information. To carry out statistical estimation for Hardy Weinberg equilibrium, the f and its corresponding variances are calculated for all imputed genomic data, also the estimates are put together. If f_i represents estimate off from the i^{th} of *m* imputations, then

$$f^- = \frac{1}{m}\sum_{i=1}^{m} f_i$$

$$w = \frac{1}{m}\sum_{i=1}^{m} var(f) \quad (3)$$

w represents normal within-imputation variance, (B) represents between-imputation variance and (T) indicates the total variance. This phenomenon is combined and illustrated as

$$B = \frac{1}{(m-1)}\sum_{i=1}^{m}(f_i - f^-)^2$$

$$T = W + (1+1/M)\text{B} \quad (4)$$

Exact p-values for HWE

Hardy Weinberg equilibrium is estimated through a consideration of discrete distribution of heterozygotes given by the allele count n_A

$$P(N_{AB}=n_{AB}|N_A=n_A)=\frac{n!n_A!n_B!2^{n_AB}}{(2n!n_{AB}!n_{AA}!n_{BB}!)} \quad (5)$$

The *p-value* for exact chi-square test statistic is calculated through summing up all the probabilities for all expected samples that are likely or unlikely under Hardy weinberg equilibrium than the observed sample. In this Thesis, mid *p-value* is taken into consideration, as put forward for use in exact Hardy Weinberg's Equilibrium testing by (Graffelman and Moreno, 2013) [15]. The *p-value* is described as the probability of observed sample and the probability of all samples at extreme than the observed one (Agresti, 2002). The study applies Rubin's findings for summing up *p-values* from various imputed data sets (Liublinska and Rubin, 2014) [27].

Let P_i denote *p-value* of the i^{th} imputed data, then we get

$$z_i = \varphi^{-1}(1-P_i) \quad (6)$$

Whereby φ^{-1} represents the inverse proportionality distribution function of Z- random variable. Incase P_i has a uniform distribution, z_i is *N (0, 1)*.

Working of kNN

KNN algorithm method is an imputation method suitable in handling both categorical and continuous data sets (Jerez, 2010) [21]. The method uses the Euclidean distance calculation between the k-nearest neighbor to impute missing values. For continuous variables, the missing values is imputed with mean of the k-nearest neighbors. For the case that the missing

values is a class variable, then the imputed values are district by majority vote. Before the process of imputing starts, the metric distance should be well defined. To calculate the Euclidean metric distance, the Euclidean distance between g_x and g_y can be calculated as;

g_x=$\delta_{x1,}\delta_{x2},\delta_{x3,}$ … … … … . $\delta_{xn,}$

and

g_y=$\delta_{y1,}\delta_{y2},\delta_{y3,}$ … … … … . $\delta_{yn,}$

If Euclidean distance measurement is employed for two genes expression vectors, then

$$dis(g_x, g_y) = \sqrt{\sum_{t=1}^{n}(\delta_{xt}, \delta_{yt})^2} \quad (7)$$

For the missing value estimation, the corresponding entire is calculated as the weighted average values as the selected k expression vectors given by

G_{IJ}=$\sum_{i=1}^{k} w_i \; x \; \delta_{iJ}$

Where w_i is the weighted value which can be calculated by

w_i=$\frac{1}{dis(g^*,g_i)\,\Delta}$ where Δ=$\sum_{i=1}^{k}\left[\frac{1}{dis(g^*,g_i)}\right]$

g^* denotes the set of k -gene closest to g_i

In general

$$dis(g_x, g_y) = \frac{|(\delta_{xt}-\delta_{yt}|}{\max((\delta_x)-\min((\delta_y)} \quad (8)$$

According to KNN impute method, the unmarked data is categorized by defining which classes its neighbors lie to. KNN method applies this model in its determination. A particular value of *k* is fixed thus grouping the unidentified tuple. When a new unidentified tuple is met in the dataset, k-nearest neighbor plays two operations; the first one, it determines the *k* points nearer to the new data points.

The second one is by using neighbors' categories such that k-nearest neighbor finds to which category the new data should be placed. When the new dataset is introduced, it classifies the data consequently. It is advantageous in dataset which is unevenly divided into groups and belongs to a particular area of the data point.

Thus, this method gives more precision in dividing the data points into different categories in a clear manner. k- Nearest neighbor identifies the class having the highest number of points sharing the least Euclidean distance from the data points that are required to be categorized. Hence the distance is supposed to be calculated between the test sample and the identified training sample. After grouping the k-nearest neighbor, most of them are taken for estimation of the training sample.

Some of the hinderances of k-nearest neighbor's workability include; the challenge in determination of the value of k, calculation of Euclidean distance and normalization of the parameters. For easy understanding of the working of the algorithm, the steps are as follows: After keeping the training samples, all set parameters should be normalized in such a way that the calculations become simpler. The estimation outcome is sensitive to the value of "k". The input variable 'k' decides the total numbers of neibours to be included. The final value of 'k' affects the algorithm by building the boundaries of each class.

Miss Forest (Random Forest method)

Random forest method is an imputation method which is a non-parametric for mixed genomic data. It is a common iterative method based on training random forest. Miss Forest is an effective method due to its computation criterion and efficiency with high dimensional data particularly in genomic dataset (Stekhoven & Bühlmann, 2012) [35].

(Stekhoven & Bühlmann, 2012) [35]. Presents random forest method as an algorithm that can deal with input data at the same time have little assumptions as possible. A similar study conducted by Stekhoven *et al.* (2012) [36] illustrate this method as a very effective and competitive algorithm as compared to others regardless of type, composition, data dimensionality and amount of missingness. Dataset with complex interactions with UN linear relations are very complicated to capture parametric procedures. Parametric procedures needs changing of a parameter. The selection of the parameter without prior knowledge is challenging because it leads to great reduction in its performance since Miss Forest does not make any assumptions about the data. Random forest is based on the observed values and continuous prediction of missing values until prevention principle is achieved. However, random forest method averages over various unpruned estimation or regression trees, the algorithm be considered a multiple imputation method. Using out-of-bag (OOB) error rates, imputation errors can be predicted without the need of a test set. Stekhoven *et al.* (2012) [36] show that the difference between OOB error rates and real error rates usually do not differ more than 10-15%.

Multiple Imputation Using Random Forest Method

Assume that $X = x_1, x_2, x_3, \ldots \ldots x_p$ is a nxp-dimensional data matrix. The research sought to establish how random forest method is imputing missing observations. Miss Forest method has been designed to deal with values that are missing by weighing the frequency of values with the nearest training samples initially imputed mean data (Breiman, 2001). This method required a response variable that is wide-ranging and essential for forest training. Instead of finding out the values of all missing variables directly, use of the Miss Forest that is trained on observed data is considered.

Instead of estimation the values of all the missing values directly, use of a random forest that is trained on the observed data set is considered, where *X* represents the matrix vector for the complete data. $X_s$contains all missing values.

$i^s_{miss}\underline{C}\{1, \ldots . n\}$. The data set can be separated into four parts:

$y^{(s)}_{obs}$: the observed values of X_s

$y^{(s)}_{miss}$: the missing values of X_s

$x^{(s)}_{obs}$: the observation,

i^s_{miss}= (1….n)\ i^s_{miss} that belong in the other variables **X**s.

$x^{(s)}_{miss}$: the observation, i^s_{miss} in other variables **X**s

Note that $x^{(s)}_{obs}$ and $x^{(s)}_{miss}$ are not completely observed, as the index i^s_{obs} relates to the observed values in **X**s

3.7. Simulation and data simulation method

This Thesis used data simulation method, the study simulated the data such a way that,

$Y_{it}; \forall\, i = 1,2,3,4,5 \ldots .150; j = 1,2,3,4,5 \ldots \ldots$ for the i^{th} according to multinormal model such that;

$$Y_{it} = \beta_0 + \beta_1 x + \varepsilon \quad (10)$$

where β_0 *is the intercept*

β_1 *is the slope*

The variance at each period varied over a given time limit while r (correlation coefficient)

between x_{is} *and* x_{it} is expected value which is positive as represented as (r) of the 1^{th} order $AR(1)$.

In general, dataset Y with nxm matrix is drawn from multinormal distribution with a zero mean vector and variance-covariance matrix $\sum$ *given as*;

$$\begin{bmatrix} r_{11} \cdots & r_{13} \cdots & r_{15} \\ r_{21} \cdots & r_{23} \cdots & r_{25} \\ r_{51} \ldots & r_{53} \cdots & r_{55} \end{bmatrix} \quad (11)$$

$AR(1)$ autocorrelation model can be summarized as;

$$\begin{bmatrix} 1 \cdots & r \cdots & r^2 \cdots r^4 \\ r^2 \cdots & 1 \cdots & 1 \ldots 1 \\ r^4 \ldots & 1 \cdots & 1 \ldots 1 \end{bmatrix}$$

Where, $r_{it} = \frac{r_{it}}{\sqrt{r_{ii}}\sqrt{r_{tt}}}$ where $i = 1,2, \ldots 150; t = 1,2, \ldots 5$

The r in this study is set to be 0.8 indicating the strong relationship among variables. The study considered a total of 18 $SNPs$ and there existed 10 variables at 10 given time limit for each variable.

Generating each dataset was based upon the following assumption; given time limit for each variable $(t = 0)$ the data is completely observed. Secondly, in this research, data was considered to be MAR and $MCAR$ missing mechanisms. Lastly, the missing pattern was monotone.

In this study, the simulation process took three steps to check whether the dataset is $MAR\ or\ MCAR$. The steps was as follows; the first step included, generating the nine-time point dimension for each variable by a random number from multi normal distribution and repeats the steps 18 $genes$ given;

$Y_{it}; for\ all\ i = 1,2,3,4 \ldots 5\ 18; j = 1,2,3 \ldots 5.$

The next step was to generate the $MAR\ or\ MCAR$ dataset. $Y_{it(miss)}$ by the missing rates $i.e\ 5\%, 10\%, 20\%$ for each test variable as indicated in $table$ 1 below. The last step, was to evaluate $MAR\ or\ MCAR$ mechanisms using little's $MCAR$ criterion to find out whether the generated result are either$MAR\ or\ MCAR$.

Table 1: Missing rate at Time t

Missing Mechanism	$SNPs$	β_0	β_1	σ^2	r	1	2	3
MAR	18	0.2 2.0	20	1	0.8	0%	5%	10%
MCAR	18	0.2 2.0	20	1	0.8	0%	5%	10%

Data simulation methodology

The findings from table 1 above indicates the different types of parameters used in data simulation process. (β_0) values were set to be 20; while (β_1) the value were estimated to be0.2. The variance and covariance were set to be1 and r was estimated to be0.8, whereby $r = 0.8$a correlation coefficient value is set to indicate moderately to high correlation. (Nakai, 2014) [44] Approximated both $r = 0.01\ and\ r = 0.7$ in his study.

Results and Discussion

Genic Data Set

GENICA study is an experimental study conducted by a study group on Gene Environment Interaction and breast Cancer in Germany (http://www.genica.de), a combined initiative of German scientists aimed to detect genetic and environmental aspects leading to development of sporadic breast cancer. The cases and controls of this age-matched and population–based study have been employed in the dataset in Germany. In this thesis, the aim is to determine the subset of the genotypic data from GENICA study. More specifically, data of 50,198 women (cases were 22225 and 27973 controls) and 18 SNPs containing estrogen hormone, all the 18 SNPs with more than 8 missing values were deleted from analysis giving a total of 50190 women (cases were 22225 and 27973 controls). This research, however, preferred to delete such values – as long as their number was small – since observations with many replaced values may add large uncertainties to further analyses with, e.g., discrimination methods such as deletion method. This research aimed to investigate how HW proportions behaves with different levels of missingness and how HW equilibrium is maintained. After the new data were generated, the targeted variables at different time limits were set to missing completely at random or missing at random mechanism. However, the baseline values for each value was expected to be always observed. In the case of missing completely at random, the missingness of the variables were produced randomly at visits through 9 based on the missing probabilities as indicated in table 1. In general, the missingness does not be contingent either unobserved or observed data.

Simulation of Results on Hardy Weinberg Proportions for MAR Mechanisms on Genomic Missing Data

Table 2: HW proportions for MAR

Missing Mechanisms	Hardy Weinberg Iteration History						
	ML	p	q	r	p-value	Inbreeding coef (f)	λ
	0%	0.33333	0.33333	0.3333	0.0034	0.2861	0.273
	1%	0.30845	0.37845	0.3130	0.0473	0.1866	0.222
	2%	0.30778	0.3897	0.3024	0.1837	0.1181	0.133
	3%	0.30928	0.3937	0.2969	0.4605	0.0626	0.048
MAR	4%	0.31036	0.39550	0.2941	0.4782	0.0481	0.031
	5%	0.31098	0.3963	0.2926	0.4791	0.0444	0.029
	6%	0.31131	0.39684	0.2918	0.4821	0.0393	0.028
	7%	0.31149	0.39709	0.2914	0.4832	0.0383	0.022
	8%	0.31159	0.39722	0.2911	0.4842	0.0282	0.021
	9%	0.31164	0.39729	0.2910	0.4923	0.0279	0.018

The results indicates that, HW proportions with MAR mechanism varies in regard to the level of missingness. Missing values in genomic data varies from one experimental design to another. The obtained results from table 4.3.1 shows that, the level of missingness varies independently with the HW proportions at each time t. The analysis also indicates that, datasets with MAR mechanism changes proportionately until they attain the equilibrium state which is termed as HWE.The level of missing (ML) increases, i.e. from 1% to 9%, the HW proportion (p) also increases. More specific, when ML increases from 1%-5%, a percentage increase of 19.95% is achieved also, when ML increases downward from 5%-9%, an increase of 80.04% is obtained. This indicates a direct influence of ML on p and vice versa

Consequently, there is a general increase in q HW proportions as shown from table 1. as the missing levels increases, from 1% to 9%, there exists also a direct increase in q proportions across the proportions. This direct increase in q proportions with ML indicates a direct correlation of HW proportions q with respect to level of missingness. In regard to r HW proportions, the study findings portrays a general decrease in r proportions with an increase in ML. This indicates an inverse change in r with respect to the level of missingness. In another words, as the level of missing ness (ML) increases, the r HW proportions decreases.

Inbreeding coefficients (f) has a statistical significance in determination of HWE and at the same time, shows how HW proportions changes with time (t). In general, where the inbreeding coefficient is greater than zero there exist a deficiency of heterozygotes and for the case $f < 0$, there exists excess of the same. The study findings also indicates that, the inbreeding coefficients (f) decreases gradually with respect to the change in level of missing ness. As the f coefficients increases, the level of missingness decreases. This shows an inverse proportional change in inbreeding coefficients with respect to ML. In particular, when the level of missingness changes from 3%-9%, there exists 23.1% change in the f coefficients; but at the same time, when the ML increases from 1%-3%, there exists 15.4% change in inbreeding coefficients.

The parameter λ in this research shows a general decrease but with a minimal variance. In comparison with the level of missingness, λ decreases at a rate of 0.6% with an increase with ML. when the level of missingness from 1% to 2% the λ values shows 23.6%, but when the level of missingness changes from 2% to 9%, the λ indicates 24.7%. In general, the results of λ portrays a drastic change w.r.t MLThe $p - value$ shows an increase w.r.t ML. as the ML increases from 1%-9%, the $p - value$ also increases. This indicates a direct proportional change in $p - value$ with regard to ML.

HW proportions attains the level of equilibrium when the level of missingness is set at 0% ($p = 0.333333, q = 0.33333, r = 0.33333$). This shows that the alleles and genes frequencies remain static within a given period of time (t). The first point is crucial since it indicates that genetic composition of actual population can be defined in terms of frequencies of moderately limited alleles rather than abundant great arrays of all likely genotypes. This study finding are similar to Graffelmam (2020), who established that, missing alleles are frequently deleted in missing at random mechanism when markers are estimated for HWE, which can lead to biasness and precision in statistical inference about equilibrium. The fidings are further illustrated in figure 1 below.

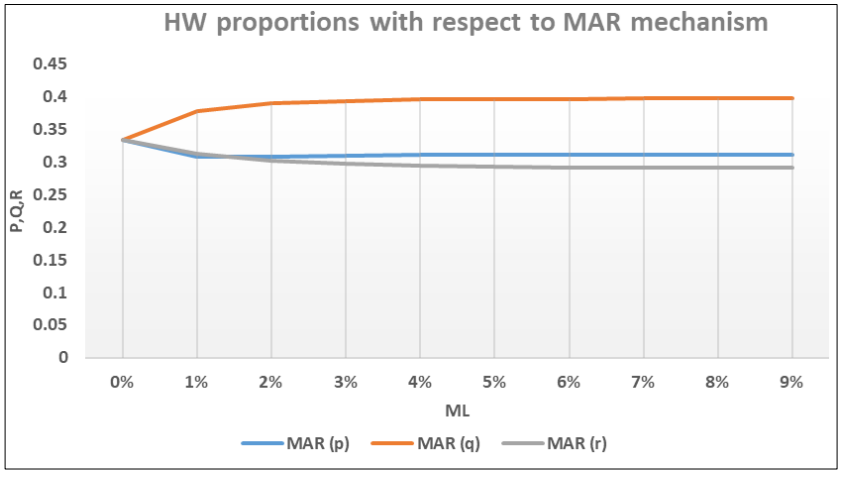

Fig 1: HW proportions at MAR and MCAR.

The figure above indicates HW proportions and their behaviors with respect to level of missingness. From the figure 1, it is noted that, the HW proportions have a common origin that i s zero but shows various disparities until they become constant. It has been noted that, the HW proportions $(p, q \ and \ r)$ behaves differently at MAR mechanism until they become constant. It is evident that, when q proportions increases with the level of messiness, both $p \ and \ r$ decreases w.r.t ML respectively. Generally, when q HW proportion increases from 0% to 1%, both $p \ and \ r$ decreases. Both p,q and r attains their normality at a level of 3% up to 9%.

Simulation of Results on Hardy Weinberg Proportions for MCAR Mechanisms on Genomic Missing Data

Table 3: HWE results for MCAR

Missing Mechanisms	Hardy Weinberg Iteration History						
	ML	p	q	r	p-value	Inbreeding coef (f)	λ
	0%	0.33333	0.33333	0.3333	0.0231	0.2562	0.271
	1%	0.31002	0.38815	0.3019	0.6234	0.1764	0.222
	2%	0.32114	0.3264	0.3524	0.6376	0.1285	0.133
	3%	0.33226	0.3008	0.3669	0.6437	0.0424	0.049
	4%	0.34338	0.28252	0.3741	0.6693	0.0385	0.032
	5	0.35441	0.2629	0.3826	0.6834	0.0343	0.023
	6	0.36552	0.2426	0.3918	0.6933	0.0293	0.021
MCAR	7%	0.37662	0.21198	0.4114	0.6955	0.0284	0.020
	8%	0.38773	0.1971	0.4151	0.6966	0.0185	0.029
	9%	0.31164	0.39729	0.2910	0.4923	0.0279	0.018

The study findings indicates that, as the HW proportions (p) increases, the ML also increases. This represents 0.56% increase in p w.r.t ML. as the level of missingness increases from 1% to 9%, there exists a gradual increase in the HW proportion (p). This indicates a direct influence of ML with respect to HW proportion (p).Unlike with MAR, the HW proportion (q) indicates a gradual decrease which represents 2.5% w.r.t ML. In general, an increase in level of missingness, results to a decrease in q and vice versa.

The study findings also indicates that, the HW proportion (r) increases with an increase in levels of missingness. This shows a direct influence of ML on r. It has been noted that, as the level of missingness increases from 1%-3% the HW proportions (r) values decreases. This research study findings are similar to (Enders, 2010) [13], who established that, two attributes, represented by *m* and *n*, missing value *n* neither depends on *m* nor *n* in a such way that, when one genotype value increases the corresponding value decreases. The $p-values$ in this research indicates a general increase as the level of missingness increases. The increase in $p-values$ indicates a direct relationship that exist when the ML.

In comparison with ML, and inbreeding coefficients (f), the level of missingness increases from 1% to 9% as the inbreeding coefficients (f) increases too. This is a clear indication of a direct correlation of ML and inbreeding coefficients (f). The values of λ decreases with an increase in levels of missingness. This phenomenon is termed as an inverse proportional change of λ w.r.t ML. In comparison with $p-values$ with λ values, as the $p-values$ increases with levels of missingness, the λ values decreases with ML but inbreeding coefficients decreases with decrease in λ values indicating an inverse proportionality in the values.Missing dataset completely at random has got a higher impact in HW proportions as compared with missing dataset missing at random. The findings are further supported by figure 2

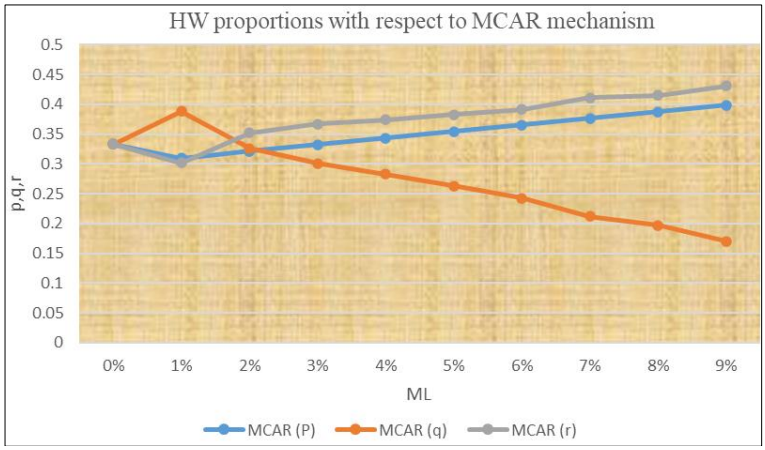

Fig 2: HW Proportions at MCAR

The study findings from table 2 above indicates that, as the level of missingness increases to 1% both p and r HW proportions increases. q starts dropping gradually, as the level of missingness increases from 1% to 9%. Also, the study findings indicates that, there exists a great disparities in $p, r \ and \ q$ between 3% and 9% at different missing levels. At 9% missing level, there exists a small disparity between $p \ and \ r$.It has been noted that, both $p \ and \ r$ coincides at 2% missing level but starts showing a gradual disparity as the level of missingness starts to increase to 9%. At 2% ML, p starts to increase and q starts to decrease. This behavior can be termed as inverse proportionality of $p \ and \ r$ w.r.t ML.

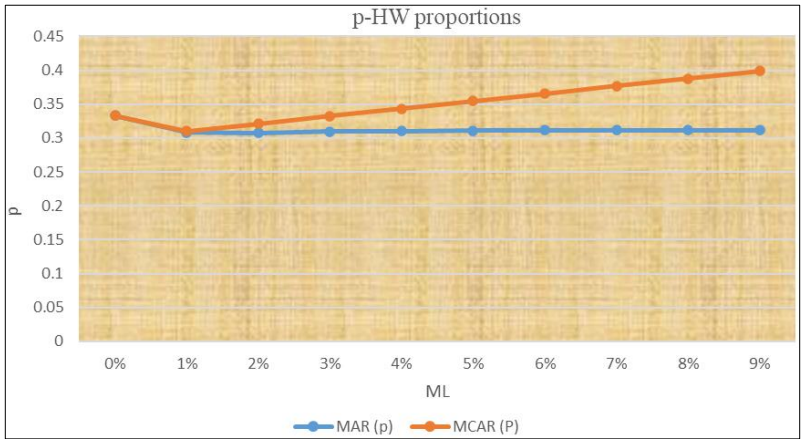

Fig 3: p - HW proportions for both MAR and MCAR

It is evident that, both p-HW proportions of MAR and MCAR have got a common origin that is to say, they originate from HWE (0.333).As the ML increases from 1% to 9%, the p-HW proportions of MCAR increases drastically as the p-HW proportions of MAR decreases. The disparity between the two mechanisms widens as the ML increases. This behavior can be termed as direct influence in proportionality of the missing mechanisms.

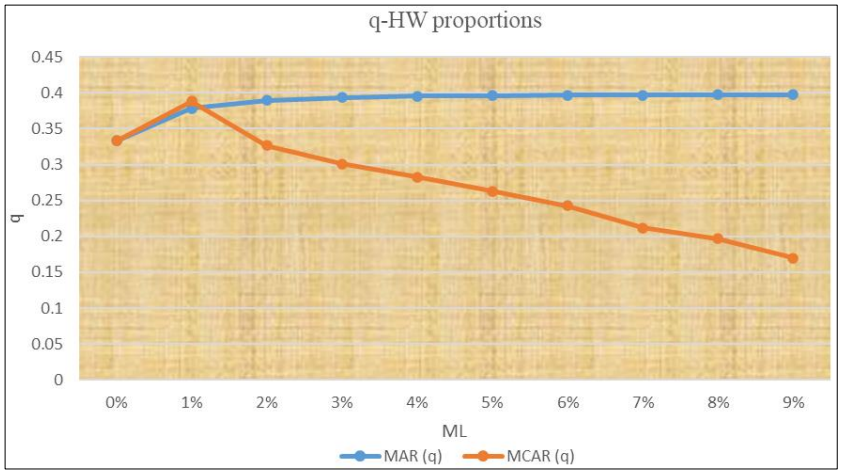

Fig 4: q - HW proportions for both MAR and MCAR

Figure 4 above indicates q-HW proportions of both MAR and MCAR mechanisms.

The study findings indicate that, q-HW proportions of MAR and MCAR originates from HWE (0.333) and starts to increase upwards as the level of missingness increases up to 4%. As the ML increases to 5%, the q-HW proportions becomes constant until it reaches 9% ML.

As the level of missingness increases from 1%, q-HW proportions of MCAR starts to decrease slowly until it reaches when ML is 9%, but at the same time q-HW proportions of MAR increases drastically to 3% ML and becomes constant until it reaches 9% ML.

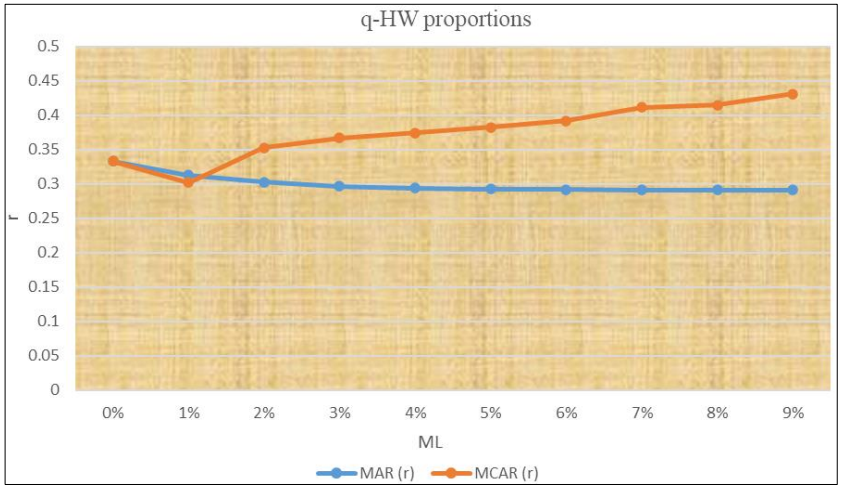

Fig 5: r - HW proportions for both MAR and MCAR

It is clearly that q-HW proportions originates at 0% which is HWE (0.333) and proportionately starts to decreases as the ML increase upwards to 9%.

After 1% ML, r-proportions for MCAR increases upwards until they reach 9% ML, but at the same time, R-proportions for MAR decreases downwards until it reaches 5% ML but becomes constant when it reaches 9%ML. Their levels of disparities widen as the ML of the respective mechanisms increases.

Table 3: Multinomial Coefficients, average of 50198 samples of n = 18 SNPs, k = 5, 5% or 10% of missing data

Missing Mechanism	HWE	Imputation Method	β0 =20					β1 =(2.0)		
			5%		10%		5%		10%	
				Variance	NMSE	Variance	NMSE	Variance	NMSE	Variance	NMSE
MAR	0.3333	KNN	0.63	0.0186	0.324	0.0139	0.319	0.0119	0.253	0.0126	0.263
		RF		0.0125	0.331	0.0136	0.321	0.0111	0.274	0.0124	0.282
MCAR	0.3333	KNN		0.0147	0.317	0.0298	0.341	0.0116	0.331	0.0127	0.311
		RF		0.0142	0.319	0.0213	0.349	0.0118	0.302	0.0125	0.330

Effectiveness of KNN and RF in estimation of HWE

The study findings show that, KNN and RF methods with 5% ML depicts a close range of positive deviations w.r.t HWE. KNN method, shows a positive deviation of 0.0093 and RF of 0.0023 as the *y* intercept is set to be 20.In MCAR mechanism, KNN and RF method had a close range of 0.0163 by KNN and 0.0143 by RF w.r.t NMSE. This indicates that, RF moves close to HWE with a positive deviation of 0.0143 in MCAR w.r.t NMSE. This is in contradict with KNN method. This study finding are similar to (Marietta *et al.*, 2019), who established that, RF method recorded the lowest NRMSE in terms of estimation of missing values with data MAR and MCAR.

In regard to 10% ML, RF and KNN in MAR mechanism, depicts a positive deviation of (KNN=0.0143, RF=0.0123) w.r.t HWE proportions. This shows that, as the level of missingness increases in MAR mechanism, the deviations from HWE also increases. Also in MCAR mechanism, as the level of missingness increases to 10% KNN shows a negative deviation of -0.0077 and RF of -0.0157 w.r.t HWE. This indicates that, in MCAR mechanism, both KNN and RF methods shows negative deviations but with a higher deviation of KNN as compared of RF. When the *y* intercept is set to be 20, the gradient function increases with the increase in ML and with more deviations of KNN and RF from HWE.

At $\beta_{0\ =20}$ and at 5% ML, KNN decreases at MAR mechanism with 0.7% deviation but at the same time, RF decreases with 1.5% deviation w.r.t NMSE. In MCAR, and $\beta_{0\ =20}$ and at 5% ML, KNN method indicated-0.024 while RF method indicated -0.03. This indicates that, as the level of missingness increases from 5% to 10% in MCAR, RF performs better as compared to KNN w.r.t HWE. The study findings indicate that, change in slope has got a significant impact w.r.t ML.

The results in MAR indicates that, at 5% ML and at $\beta_{1=2.0,}$ KNN and RF methods depicts a higher deviation from HWE with a deviation of (KNN=0.08, RF=0.059). This is a higher positive deviation in 5% ML as compared when the y intercept is set to be 20.

In MCAR, the study findings show a smaller deviation of the two methods from HWE as compared in MAR mechanism i.e. (KNN=0.002, RF=0.031).This indicates that, when the gradient is increased, both KNN and RF decreases their Levels of deviations from HWE w.r.t NMSE regardless of the ML.

When comparing the two methods at 5% ML with MAR mechanism when $\beta_{0\ =20}$ and β_1 is set to be 2.0, the values decreases with a deviation of 0.071 in KNN while RF decreases with a deviation of 0.057. This tell us that, as the gradient function increases to a given set level both KNN and RF values decreases w.r.t NMSE. This phenomenon is termed as inverse proportionality of gradients w.r.t imputation methods.

When comparing KNN and RF at 5% ML at MCAR mechanism when $\beta_{0\ is}$ set to be 20 and $\beta_{1\ =2.0,}$ the values shows a smaller deviations w.r.t NMSE such that (KNN=0.014, RF=0.016). This indicates that, regardless of ML, the levels of deviations from HWE differs significantly between the two methods.

When the ML is being increased, to 10%, at MAR mechanism, both KNN and RF shows negative deviations such that KNN=-0.056, RF=-0.03. This indicates that, when the level of missingness is increased for both methods, there exists a significant deviation from HWE regardless of missing mechanisms. In MCAR, the deviation of both methods shows a negligible deviation (KNN=-0.03, RF=-0.019) when the slope is increased, indicating that, regardless of ML the missing mechanism determines the gradual deviation of the two methods from HWE.

For large slope (slope = 2.0) and high missing rate the values of Root Mean Squared Error, KNN achieved large differences as opposed to the Random Forest method. This is not unexpected since the Random Forest method produces unbiased predictable variables under Missing Completely At Random with only 5% missing mechanism.

For the case of Missing At Random mechanism, the obtained results showed that Random Forest is the most appropriate method in estimation of Hardy Weinberg equilibrium despite the missing rates and gradient size based on Root Mean Squared Error. According to Rubin (2010), in the literature the research findings are well documented. For small slope and low missing rate, the obtained results did not differ meaningfully between Random Forest and k-Nearest Neighbor methods. Additionally the two missing mechanism, MAR and MCAR, are considered.

Based on the research findings, the researcher made the following conclusions; the most appropriate method for estimation of HWE and handling data set with MCAR, is Random Forest. Secondly, in all simulated situations particularly under MAR setting, KNN method is the most effective. It has a good (95%) Computational Probability index compared to RF method in estimation of HWE and produces the smallest biases.

Comparison of KNN and RF w.r.t HWE at different missing mechanisms

The two methods can be compared using their variances w.r.t HWE and considering their differences in NRMSE.

In MAR mechanism with 5% ML, RF records a low variance (0.0125) as compared with KNN (0.0186) with reference to HWE. This shows that, with smaller variances of RF the

better the method in estimation of HWE as compared with KNN method.

When the ML is increased to 10%, the variances of both KNN and RF also increases as compared when the level of missingness is 5%. The variances of RF is low (0.0136) as compared with the variances of KNN (0.0139) which tells us that, as the level of missingness is increased, the corresponding variances increases too with significant deviations among the two methods.

In MCAR mechanism with 5% ML, RF method has got a low variance w.r.t HWE (RF=0.0142) as compared when the level of missingness is increased to 10% (RF=0.0213). this shows that, as the level of missingness is increased from 5% to 10%, the level of deviations in RF increases. This show a direct proportionality between the variables of interest.

In the case of KNN, as the level of missingness increases from 5% to 10% in MCAR mechanism, the level of variances increases from 0.0147 to 0.0298. This indicates that, when the y intercept is set to be 20, there exist a direct proportionality of imputation methods.

When the slope is set to be 2.0 at 5% ML at MAR mechanism, RF shows a lower variance as compared to KNN (RF=0.0111, KNN=0.0119) but in MCAR, KNN shows a lower variance (0.0116) while RF (0.0118) records a higher variance. This indicates that, when the slope is increased, the variances of the two corresponding methods also increases implying that, as ML increases with increase in slope, the variances also increases w.r.t HWE

At 10% ML at MAR, the variances of RF is lower (0.0124) as compare to KNN (0.0126). This indicates that, there exists a statistical significant differences in the level of missingness between the two imputation methods.

In MCAR mechanism, at 10% ML, the variances of RF is lower as compared to KNN method. This shows that, regardless to the missing levels, the level of variances of the imputation methods differs with different missing mechanism. This study finding are similar to (Huang and Carriere, 2006), who established that the variances and NMSE are suitable parameters in checking accuracy, precision and performance of each imputation method in relation to estimation of HWE to come up with a conclusion that is decisive.

In general, when the ML is 5% and 10%, RF method records a low variances w.r.t HWE as compared with KNN method hence the most appropriate method in estimation of HWE.

Conclusions and Recommendations:

Even though the simulation findings suggested that unlike KNN method, Random Forest was the exclusive method under MCAR missing method and KNN was superior to RF methods under MAR, the performance of these two methods depended on factors like missing rate and time effect. Generally there exist no imputation method that is appropriate in all situations in estimation of HWE. The values of estimated RMSE for RF method was very close under the assumption of MCAR missing mechanism as the missing rate increased from low to moderate (slope = 2.0).

Regardless of the missing rate and slope size based on RMSE and variance for the MAR and MCAR missing data, the simulation results revealed that RF was the best method.

This study considered a longitudinal study with a total of 18 SNPs and nine visiting time points. Three different slopes and Two V possible missing rates are used to mimic the real-world situations. Additionally, two missing mechanisms (MAR and MCAR), are considered.

On the simulation results the study reached this significant conclusions: for estimation of HW proportion with the aim of maintaining HWE and handling MCAR missing data RF method is the most appropriate method. In simulated situations particularly under MAR setting KNN method is the most effective because it gives only close approximation to HWE and produces small biases.

However for longitudinal data analysis inferior methods such as KNN are still in use. The simulation data provide a good rationale and reference in choosing missing data imputation method and the most effective method in estimation of HWE. According to Kenward and Muhlenberg's (2018), KNN method which is well supported by our simulation results should be not be avoided.

References

1. Perera S, Perrizo W. Gene Function Prediction.,in Research Gate, 2009, 26-31.
2. Agresti A. Categorical Data Analysis. New York.: Wiley; c2002.
3. Aittokallio T. Dealing with missing values in large-scale studies: microarray data imputation and beyond. Briefings in Bioinformatics; 2020. 2010 Mar 1;11(2):253-64.
4. Batista G, Monard M. An analysis of four missing data treatment methods for supervised learning. Applied artificial intelligence; 2018, 2003 May 1;17(5-6):519-33.
5. Bosco FCD. Neutral and stable equilibria of genetic systems and the Hardy–Weinberg principle: limitations of the chi-square test and advantages of auto-correlation functions of allele frequencies. Front Genet. 2012;3:1-10.
6. Breiman L. Random forests. Mach. Learn. 2001;45(1):5-32.
7. Buuren SV. Multivariate imputation by chained equations in R. Journal of Statistical software. 2011;45:1-67.
8. Tsai CF, WEY. Genetic algorithms in feature and instance selection. Knowl.-Based Syst., 2013;39:240-247.
9. Chen. The Hardy-Weinberg principle and its applications in modern population genetics. Front Biol. 2010;5(4):348-353.
10. Chen BCG. Departure from hardy weinberg equilibrium and genotyping error. Front. Genet. 2017;8:167.
11. Di Iulio J, Bartha I, Wong EH, Yu HC, Lavrenko V, Yang D, Jung I, Hicks MA, Shah N, Kirkness EF, Fabani MM. The human noncoding genome defined by genetic diversity. Nature genetics. 2018 Mar;50(3):333-7.
12. Edwards AW. Hardy (1908) and hardy–weinberg equilibrium. Genetics. 2008;179(3):1143-1150.
13. Enders C. Applied Missing Data Analysis. New York, NY, USA: Guildford Press; c2010.
14. Graffelman and Moreno. The mid p-value in exact tests for. Stat. Appl. Genet. Mol. Biol. 2013;12(4):433-448.
15. Graffelman JA. The mid p-value in exact tests for Hardy-Weinberg equilibrium. Stat. Appl. Genet. Mol. Biol. 2013;12(4):433-448.
16. Graffelman JJ. A genome-wide study of Hardy–Weinberg equilibrium with next generation sequence data. Genetics. 2017;136(6):727-741.
17. Grünwald NJ, Everhart SE, Knaus BJ, Kamvar ZN. Best practices for population genetic analyses. Phytopathology. 2017 Sep 16;107(9):1000-10.
18. Ishioka T. Imputation of missing values for unsupervised data using the proximity in random forests, in 'The Fifth International Conference on Mobile, Hybrid, and On-line Learning',The National Center for University Entrance

Examinations,. The National Center for University Entrance Examinations; c2013. p. 30-36.
19. Ishwaran. Random Forest missing data algorithms. tat Analysis Data Mining. 2017;10(6):363-77.
20. Janssen KJ. 'Missing covariate data in medical research: to impute is better than to ignore. Journal of Clinical Epidemiology. 2010;63(7):721-727.
21. Jerez JM. 'Missing data imputation using statistical and machine learning methods in a real. Artificial Intelligence in Medicine, 2010 Oct 1;50(2):105-15.
22. Li Y, Li T. Recent advances in feature selection and its applications. Know l. Inf. Syst, 2017 Dec;53(3):551-577.
23. Lin WC, Tsai CF. Missing value imputation. A review and analysis of the literature (2006-2017). 2020;53(2):1487-1509.
24. Little R, RD. Statistical Analysis with Missing Data; c2014.
25. Little R, RD. International Encyclopedia of the Social & Behavioral Sciences. Missing Data, 2020, 15.
26. Little RJ. Statistical Analysis with Missing Data. New York.: Wiley; c2002.
27. Liublinska and Rubin. Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. Stat. Med. 2014;33(24):4170–4185.
28. Mack C, SZ. Managing Missing Data in Patient Registries for Healthcare Research and Quality. US: Rockville Agency; c2018.
29. Milewski RC. Application of Hardy-Weinberg law in biomedical research. Stud Log Gramm Rhetor. 2011;25(38):7-27.
30. Rentzsch P, Witten D. Predicting the deleteriousness of variants throughout the human genome; 2018, 2019 Jan 8;47(D1):D886-94.
31. Ritchie G, Dunham I. Functional annotation of noncoding sequence variants. Nat Methods. 2014 Mar;11(3):294-6.
32. Rodriguez EA. The Treatment of Missing Values and its Effect on Classifier Accuracy. in Classification, Clustering, and Data Mining Applications, 2010, 639-647.
33. S, VB. Flexible imputation of missing data: chapman and hall/CRC; c2018.
34. Shah AD, BJ. Comparison of random forest and parametric imputation models for imputing missing data using MICE. Am J Epidemiol. 2014;179(6):764–74.
35. Stekhoven and Bühlmann. 'Missforest-Non-parametric missing value imputation for mixed-type data', Bioinformatics. 2012;28(1):112-118.
36. Stekhoven DJ, BP. Miss Forest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-8.
37. Tang F, IH. Random Forest missing data algorithms. Stat Analysis Data Mining. 2017;10(6):363-77.
38. Tang J, Alelyani S. Feature selection for classification. Boca Raton FL. USA: Aggarwal, C.C, Ed.; c2014.
39. Van Buuren S, GOK. Multivariate Imputation by Chained Equations in R. Stat Soft. 2011;45(3):1-67.
40. Zhang M, Browne. A survey on evolutionary computation approaches to feature selection. IEEE Trans. Evol. Comput. 2016;20(4):606-626.
41. Zhang S, Qin S. Missing values in cost-sensitive decision. Missing is useful. 2005;17:1689-1693.
42. Zhang S, Hu R, Zhu. Local and global structure preservation for robust unsupervised spectral feature selection. IEEE Trans. Know l. Data Eng. 2018;30(3):517-529.
43. Ardourel M, Felgerolle C, Pâris A, Acar N, Ramchani Ben Othman K, Ueda N, Rossignol R, Bazinet A, Hébert B, Briault S, Ranchon-Cole I. Dietary supplement enriched in antioxidants and omega-3 promotes glutamine synthesis in Müller cells: a key process against oxidative stress in retina. Nutrients. 2021 Sep 16;13(9):3216.
44. Nakai Y, Honda K, Yanagi K, Kataura H, Kato T, Yamamoto T, Maniwa Y. Giant Seebeck coefficient in semiconducting single-wall carbon nanotube film. Applied Physics Express. 2014 Jan 29;7(2):025103.
45. Cyburt RH, Fields BD, Olive KA. An update on the big bang nucleosynthesis prediction for 7Li: the problem worsens. Journal of Cosmology and Astroparticle Physics. 2008 Nov 17;2008(11):012.