Web Based Application for Controlling Data Quality in Phenotype Prediction of Indonesian Rice Genomes

E B Nababan1*, R Nurhasanah1* and A S Huzaifah1*

1 Department of Informatics Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia, 20155

*Email: ernabrn@usu.ac.id, rossynurhasanah@usu.ac.id, adesarah@usu.ac.id

Abstract. Single Nucleotide Polymorphism (SNP) is a form of Deoxyribonucleic Acid (DNA) variation that can be used in predicting phenotypes. Data quality control is a crucial stage in the process of detecting phenotypes using SNP data. In this study, we built a web-based application to carry out the SNP data quality control function. Raw SNP data in string type are filtered by calculating the missing rate, minor allele frequency, and Hardy-Weinberg Equilibrium values. The result is SNP data that has been filtered in numeric form, namely the value 1 represents dominant homozygous, 2 represents heterozygous and 3 represents homozygous recessive. SNP encoding in numerical form aims to make SNP data can be processed into machine learning for the further phenotype prediction step.

1. Introduction

The need for rice as the main staple food in Indonesia is increasing due to a very large population (more than 200 million people) and continues to grow [1]. This condition causes dependence on imported rice and causes vulnerable food security. An effort that can be made to address these food problems is to intensify plant breeding activities, which is a method that exploits the genetic information of plants through the assembly of high yielding and high-quality superior cultivars, as well as resistance to biotic and abiotic threats [2]. Genetic diversity at the DNA level is a source of information and powerful tools for doing plant improvement. One alternative in developing superior varieties is by utilizing DNA markers in the form of Single Nucleotide Polymorphisms (SNP) with the machine learning approach. This technology is required to overcome the limitations of applying conventional breeding techniques and accelerate the achievement of the goals in a breeding program [3].

After the discovery of Next Generation Sequencing (NGS) technology, genetic data from various agricultural commodities are available and stored online by various institutions in the world. In Indonesia, Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development (ICABIOGRAD) is a government institution that provides the agricultural gen bank which is freely accessible. The abundant amount of those agricultural genome data can be used to predict the phenotypic characteristics of plants that are economically valuable.

Machine learning approaches and bioinformatics infrastructure are needed in processing and analyzing SNP data. Various studies on the association of SNP with phenotypes using machine learning techniques have been conducted. Kusuma et al [4] proposed a machine learning approach in predicting the relationship between SNP and phenotype using the Gravitational Search Algorithm (GSA) algorithm combined with exhaustive search. As a result, the proposed method is robust in filtering out SNPs that are not relevant and able to identify the main SNPs associated with the tested phenotype.
Setiawan [5] made a comparison between the Correlation-Adjusted Marginal Correlation Score (CAR Score) with the Influential Score (I-Score) method in case of filtering SNP. The output of the SNP ranking process above is then evaluated using the Random Forest algorithm. Some irrelevant SNP were removed from further steps. The search for SNP combinations that have an association with the observed phenotype was carried out using the Sequential Forward Floating Selection (SFFS) technique with the Support Vector Regression (SVR) algorithm. The proposed method is able to identify the SNPs associated with the phenotype tested and is able to eliminate false positives in the association mapping process.

In association mapping, there needs to be a framework to guarantee the validity of the variable’s sources involved in each process. Guzzetta et al [6] proposed a series of procedures in an association with genotype data with a quantitative phenotype using a regression machine learning approach. The pipeline aims to ensure reproducibility and avoid bias in selection. The conclusion obtained is that machine learning techniques can support the prediction of quantitative properties from genotype data more effectively compared with conventional methods.

Another research proposed a protocol for the steps in controlling data quality in the study of SNP associations with phenotypes [7]. The goal is to clean DNA data by removing DNA samples that can cause bias to the study. This process is a very urgent step in ensuring the quality of the data before carrying out the next process. This protocol explains the stages in quality control for each marker, including the removal of SNPs with low quality, removing SNPs that show significant deviations from Hardy-Weinberg equilibrium (HWE) and eliminating SNPs that have low minor allele frequency values.

Zeng et al [6] provide an overview of the various strategies and approaches that are often used in analyzing data in the genome-wide association study. One of them is about data quality control, which is the most critical stage in the association mapping process. At this stage, the validity of the data will be ensured using a series of statistical procedures. First of all, SNPs with high missing rates (e.g. > 5%) will be eliminated. Then SNPs that have a low minor allele frequency (e.g. <1%) will be deleted. Next, check for dependency between individuals, interrelated individuals will be excluded from the dataset. Finally, checking the Hardy-Weinberg equilibrium (HWE) value, SNPs with P-values less than $10^{-5}$ - $10^{-6}$ will be deleted from the dataset.

In this research, a quality control process will be carried out on the SNP data of Indonesian rice varieties. We built web-based applications that function to run data quality control processes. The end result is SNP data in numeric form, which has been filtered out. The filter process is based on missing rate, minor allele frequency, and Hardy-Weinberg Equilibrium values.

2. Methods
This research will follow the procedure that was carried out in [6–9]. Assessment will be conducted for each SNP. For each item that does not meet the requirements will be removed from the dataset. The SNP quality control procedure can be seen in Figure 1.
2.1. Missing Rate

The missing rate refers to the percentage of SNPs whose genotypes are not called. In each SNP, if the missing rate is high, it means that the DNA quality is low due to problems during the genotyping process. So that the SNPs are removed from the dataset because they can produce bias in further association process [6]. There are no definite provisions in determining the missing rate threshold. In general, SNPs that have a missing rate of more than 5% will be filtered.

2.2. Minor Allele Frequency

Simply, the calculation of allele frequency is done by counting the number of each allele occurrences in a population. Calculation of minor allele frequency is very important to do in association mapping. Minor alleles are the lowest allele frequencies in the population [7]. In association mapping, SNPs with very low MAF values are removed from the dataset. In general, the MAF limit used in research is 1-2% [5]. SNPs with very low MAF values indicate that the SNPs' ability to detect associations is very weak and can lead to wrong associations. This is due to the genotypes coming from individuals who are very few in number, thus causing statistical power is also very low.

2.3. Evaluation of Hardy-Weinberg Equilibrium (HWE)

Deviations from HWE can indicate problems during the genotyping process, or even a signal of true genetic association. Every SNP will be examined using the exact tests of Hardy-Weinberg Equilibrium. The threshold in determining whether SNP is in the HWE varies between several previous studies (P-value in the range of $10^{-3}$ and $5.7 \times 10^{-7}$) [7]. In general, SNPs with P-values of less than $10^{-5}$-$10^{-6}$ are safely eliminated from the next study [6].

2.4. SNP Encoding

SNP data needs to be encoded in the numerical form so that it can be processed using various machine learning algorithms. There are several ways to represent the genotype data in numerical form. Each encoding method contains different biological meanings and will affect the results of the machine learning algorithm used [4]. Kim et al [7] compared three SNP encoding models, namely additive models, recessive/dominant models and genotypic models. In the additive model, the genotype data is encoded to 0 for dominant homozygous, 1 for heterozygous and 2 for recessive homozygous. Guzzetta et al [4] made comparisons of other numerical values, namely {-1, 0, 1} and {100, 010, 001}. In Support
Vector Regression (SVR), no significant difference was found between encoding \{0, 1, 2\} and \{-1, 0, 1\}, while encoding with \{100, 010, 001\} resulted in poor SVR predictive ability. In the recessive/dominant model, the number of occurrences of each allele will be calculated. Each allele is set to 0 if the allele does not appear, and set to 1 if the allele appears. In the genotypic model, each SNP consists of 3 features that represent the appearance of each genotype. If one genotype appears then it is given a value of 1 and a value of 0 for the other 2 features. The three models are compared by calculating the Area Under ROC Curve (AUC) value. As a result, additive models produce the highest maximum AUC values compared to other models. So this research will use an additive model in conducting SNP encoding. An overview of these three encoding models can be seen in Table 1.

**Table 1.** Overview of three genotype encoding models.

| Genotype (SNP) | Additive Model | Recessive/Dominant Model | Genotypic Model |
|----------------|----------------|--------------------------|-----------------|
|                |                | A | G | AA | AG | GG |
| AA             | 0              | 1 | 0 | 1  | 0  | 0  |
| AG             | 1              | 1 | 1 | 0  | 1  | 0  |
| GG             | 2              | 0 | 1 | 0  | 0  | 1  |

3. Result and Discussion

To carry out data quality control functions, we build computer programs in Python. The program accepts a file as an input, which is SNP data in the form of a comma-separated value (CSV) format. In the SNP screening process, the first step undertaken is to calculate the missing rate of each SNP. In some previous studies, it was mentioned that SNPs with a missing rate of > 5% will be removed from the dataset. Because the quality of the SNP is low and can cause bias in the phenotype prediction process. In the application that we built, the user can input the desired threshold value (in float type) into the text box provided.

The second step is to calculate the minor allele frequency (MAF). In several related studies, SNPs with a MAF value of less than 5% were eliminated from the dataset because the SNP was indicated to be irrelevant. In this application, we also provide a text box that accepts input threshold values from the user. At a later stage, the Hardy-Weinberg Equilibrium (HWE) value for each SNP is calculated. This application provides a text box that can be filled by the user to set the HWE threshold value. The application display that we built is as shown in Figure 2 below.
Figure 2. The interface of SNP Quality Control Application.

The raw SNP data will be filtered using this application based on the three criteria above. Then the SNP data is encoded into a numerical form using the additive model so that it can be processed using machine learning in the next process. We made a web-based interface to make it easier for users to use this application.

4. Conclusion

We build this SNP quality control application to help researchers filter data on the association mapping process. This application interface is web-based, making it easier for users to use it. The process of screening raw SNP data is based on three criteria, namely the calculation of the missing rate, the calculation of minor allele frequency (MAF) and the calculation of the Hardy-Weinberg Equilibrium (HWE) values. In order to be processed into a machine learning algorithm, the final SNP results are encoded into numerical form based on additive models with 0 representing dominant homozygous, 1 representing heterozygous and 2 representing homozygous recessive. This quality control process produces SNP data that has been filtered in a comma separated value (CSV) file format so that it can be easily used in further computing processes.

References

[1] BPS Proyeksi Penduduk Indonesia Indonesia Population Projection (Jakarta).
[2] Jannink J, Lorenz A J and Iwata H 2010 Genomic selection in plant breeding : from theory to practice 9 166–77.
[3] Halldórsson B V, Istrail S and De La Vega F M 2004 Optimal Selection of SNP Markers for Disease Association Studies *Hum Hered* 58 190–202.
[4] Kusuma W A, Hasibuan L S and Istiadi M A 2016 SNPs Selection using Gravitational Search Algorithm and Exhaustive Search for Association Mapping *IOP Conference Series: Earth and Environmental Science* vol 31 (Institute of Physics Publishing).
[5] Setiawan D, Soroako A T, Kusuma W A and Wigena A 2018 J estr.
[6] Guzzetta G, Jurman G and Furlanello C 2010 A machine learning pipeline for quantitative phenotype prediction from genotype data 11.
[7] Anderson C A, Pettersson F H, Clarke G M, Cardon L R, Morris P and Zondervan K T 2011 Europe PMC Funders Group Data quality control in genetic case-control association studies 5
1564–73.

[8] Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J and Liu J 2015 Statistical analysis for genome-wide association study 29 285–97.

[9] Mardianto, M F F, Tjahjono, E and Rifada M 2019 Statistical Modelling for Prediction of Rice Production in Indonesia Using Semiparametric Regression Based on Three Forms of Fourier Series Estimator 4 2763 - 2770.