scientific reports

Check for updates

OPEN

A machine learning-based SNP-set analysis approach for identifying disease-associated susceptibility loci

Princess P. Silva[1,2], Joverlyn D. Gaudillo[1,2,3]✉, Julianne A. Vilela[4], Ranzivelle Marianne L. Roxas-Villanueva[1,2], Beatrice J. Tiangco[5,6], Mario R. Domingo[3] & Jason R. Albia[1,3,7]

Identifying disease-associated susceptibility loci is one of the most pressing and crucial challenges in modeling complex diseases. Existing approaches to biomarker discovery are subject to several limitations including underpowered detection, neglect for variant interactions, and restrictive dependence on prior biological knowledge. Addressing these challenges necessitates more ingenious ways of approaching the "missing heritability" problem. This study aims to discover disease-associated susceptibility loci by augmenting previous genome-wide association study (GWAS) using the integration of random forest and cluster analysis. The proposed integrated framework is applied to a hepatitis B virus surface antigen (HBsAg) seroclearance GWAS data. Multiple cluster analyses were performed on (1) single nucleotide polymorphisms (SNPs) considered significant by GWAS and (2) SNPs with the highest feature importance scores obtained using random forest. The resulting SNP-sets from the cluster analyses were subsequently tested for trait-association. Three susceptibility loci possibly associated with HBsAg seroclearance were identified: (1) SNP rs2399971, (2) gene LINC00578, and (3) locus 11p15. SNP rs2399971 is a biomarker reported in the literature to be significantly associated with HBsAg seroclearance in patients who had received antiviral treatment. The latter two loci are linked with diseases influenced by the presence of hepatitis B virus infection. These findings demonstrate the potential of the proposed integrated framework in identifying disease-associated susceptibility loci. With further validation, results herein could aid in better understanding complex disease etiologies and provide inputs for a more advanced disease risk assessment for patients.

Understanding the emergence and progression of complex diseases incessantly pose challenges to researchers due to its intricate and multifactorial nature. These diseases are caused by interplays between genetics and environmental factors leading to a plethora of combinations that need to be considered in modeling. From the genetics' aspect, understanding the etiology of complex diseases necessitates an extensive localization of significant genomic variations due to its polygenic nature[1–3]. Identifying these biomarkers, albeit elucidating only a portion of the entire underpinnings of complex diseases, could nevertheless aid in increasing patients' chances of survival by allowing a more personalized and advanced disease risk assessment[4].

A genome-wide association study (GWAS) is the traditional approach employed to discover genetic biomarkers, i.e. single nucleotide polymorphisms (SNPs), associated with various traits and diseases[5]. GWAS has been successful in identifying several risk loci for a wide array of illnesses including cancer[6], Type 2 diabetes mellitus[7], Crohn's disease[8], and coronary artery disease[9], among others. However, despite these achievements,

[1]Data-Driven Research Laboratory (DARELab), Institute of Mathematical Sciences and Physics, University of the Philippines Los Baños, 4031 Los Baños, Laguna, Philippines. [2]Computational Interdisciplinary Research Laboratory (CINTERLabs), University of the Philippines Los Baños, 4031 Los Baños, Laguna, Philippines. [3]Domingo AI Research Center (DARC Labs), 1606 Pasig City, Philippines. [4]Philippine Genome Center Program for Agriculture, Office of the Vice Chancellor for Research and Extension, University of the Philippines Los Baños, 4031 Los Baños, Laguna, Philippines. [5]National Institute of Health, UP College of Medicine, Taft Avenue, 1000 Manila, Philippines. [6]Division of Medicine, The Medical City, 1605 Pasig, Philippines. [7]Present address: Venn Biosciences Corporation Dba InterVenn Biosciences, Metro Manila, Philippines. ✉email: jdgaudillo@up.edu.ph

GWAS faces limitations due to its individual-SNP analysis approach exacerbated by the high dimensionality of genomic datasets. As multitudinous individual association tests are performed, stringent thresholds must be adopted to account for error rates leading to underpowered detection[10]. This increases the probability of not detecting SNPs with small effects that are truly associated with a trait and could significantly contribute to phenotypic variability[11]. The traditional GWAS approach also fails to capture SNP-SNP interactions as it only tests for the marginal effects of SNPs and disregards the variants' joint contributions to phenotypic expression. These interactions require explicit analysis since they are vital in addressing the "missing heritability" problem[12] which states that single genetic variations are insufficient in explaining the entire heritability of a trait.

Under the "polygenic paradigm", refining statistical models, such as increasing sample sizes[13] and reducing the number of tests employed[14], is crucial in increasing the chances of discovering true associations. Empirical evidence[15,16] has shown that as sample size increases, GWAS continues to yield more novel trait-associated loci. However, this approach is not always feasible[14] especially for studies involving small populations and diseases with low prevalence. For this reason, it is more viable to reduce the number of tests employed to relax the stringent conditions used to consider genomic variants as significant. Existing approaches to this latter strategy include haplotype-based association analysis and SNP-set analysis, both of which also address the inability of GWAS to capture SNP-SNP interactions[17,18]. Haplotype-based analysis[19] accounts for linkage disequilibrium between SNPs; while SNP-set analysis, e.g. gene-based[20] and pathway-based analyses[21], considers the joint effects of variants on phenotypic expression. Aside from addressing the aforementioned GWAS' limitations, SNP-set analysis further permits hypothesis testing on associations possibly existing between wider loci and traits[18]. However, when this type of analysis groups SNPs based on prior biological knowledge, a study's success may be hampered when information on genetic variations and competitive pathways related to the trait are insufficient. To allow a less restricted analysis, it is necessary to explore other methods of forming SNP-sets using information independent of a priori biological knowledge.

Machine learning (ML) is an innovative and powerful approach used in solving complex problems in various fields and disciplines due to its capability to handle and analyze high-dimensional datasets[22–24]. Several studies have already demonstrated the usability of ML in genomic datasets[25–27]; however, to our knowledge, there is only a handful of existing literature discussing its application to SNP-set formation[28–31]. These studies employed cluster analysis to form SNP-sets in a data-driven manner. This approach could subsequently lead to the identification of novel risk loci associated with a trait[31], albeit there may be problems related to computational complexity and cost. As genomic datasets are usually of high dimension, it is susceptible to the "curse of dimensionality"[32,33], a problem that could be addressed by solely clustering the SNPs found in certain genomic regions that are known to play a role in trait development[29,30]. However, this approach defeats the purpose of performing an inclusive analysis as the search for significant biomarkers is restricted by relatively narrow regions. For a more varied selection of SNPs to analyze, dimensionality reduction techniques based on random forest (RF) could be used to reduce dataset dimensions before conducting cluster analysis. RF has been widely incorporated in SNP research[25,34–36] due to its significant properties: (1) a nonparametric nature that allows the establishment of predictive models without the need for preliminary statistical assumptions, and (2) the capability to provide an importance score, i.e. variable importance measure (VIM) for each SNP, which increases the probability of detecting highly relevant biomarkers.

Cluster analysis and random forest have already been proven applicable and effective in genomic data analysis, specifically in identifying predictive and presumably disease-associated SNPs[31,37]. However, based on the literature review, the integration of these approaches has not been explored on SNP data. This study aims to incorporate these two techniques to augment previous GWAS findings and allow the discovery of novel trait-associated susceptibility loci. The study implements the proposed integrated framework using the following three-step algorithm:

1. Dimensionality reduction through RF;
2. SNP-set formation through cluster analysis involving top-ranking SNPs from Step 1 and SNPs considered by GWAS to be significantly associated with the trait of interest (termed in this study as 'GWAS-identified SNPs'); and
3. Association testing on the resulting SNP-sets from Step 2.

In Step 1, dimension reduction is implemented using random forest feature selection to circumvent the "curse of dimensionality" problem associated with analyzing high-dimensional SNP datasets[35]. In Step 2, top-ranking SNPs determined from the results of Step 1 and GWAS-identified SNPs are subjected to cluster analysis to evaluate shared similarities among the variants and form SNP-sets. Finally, Step 3 involves testing the SNP-sets derived from Step 2 for trait-association. The proposed methodology was applied to the GWAS data by[38] wherein the phenotype of interest is hepatitis B virus surface antigen (HBsAg) seroclearance, a marker for clearance of chronic hepatitis B virus (HBV) infection.

Methodology

This study aims to discover novel trait-associated susceptibility loci through a machine learning-based SNP-set analysis approach built on the integration of RF, cluster analysis, and previous GWAS findings. The entire analysis is divided into three main parts: dimension reduction, SNP-set formation, and association testing. Figure 1 shows the architecture of the proposed integrated framework.

Data description and preprocessing. The data used in this study was adopted from the GWAS conducted by[38] which aimed to identify susceptibility loci associated with HBsAg seroclearance among patients with

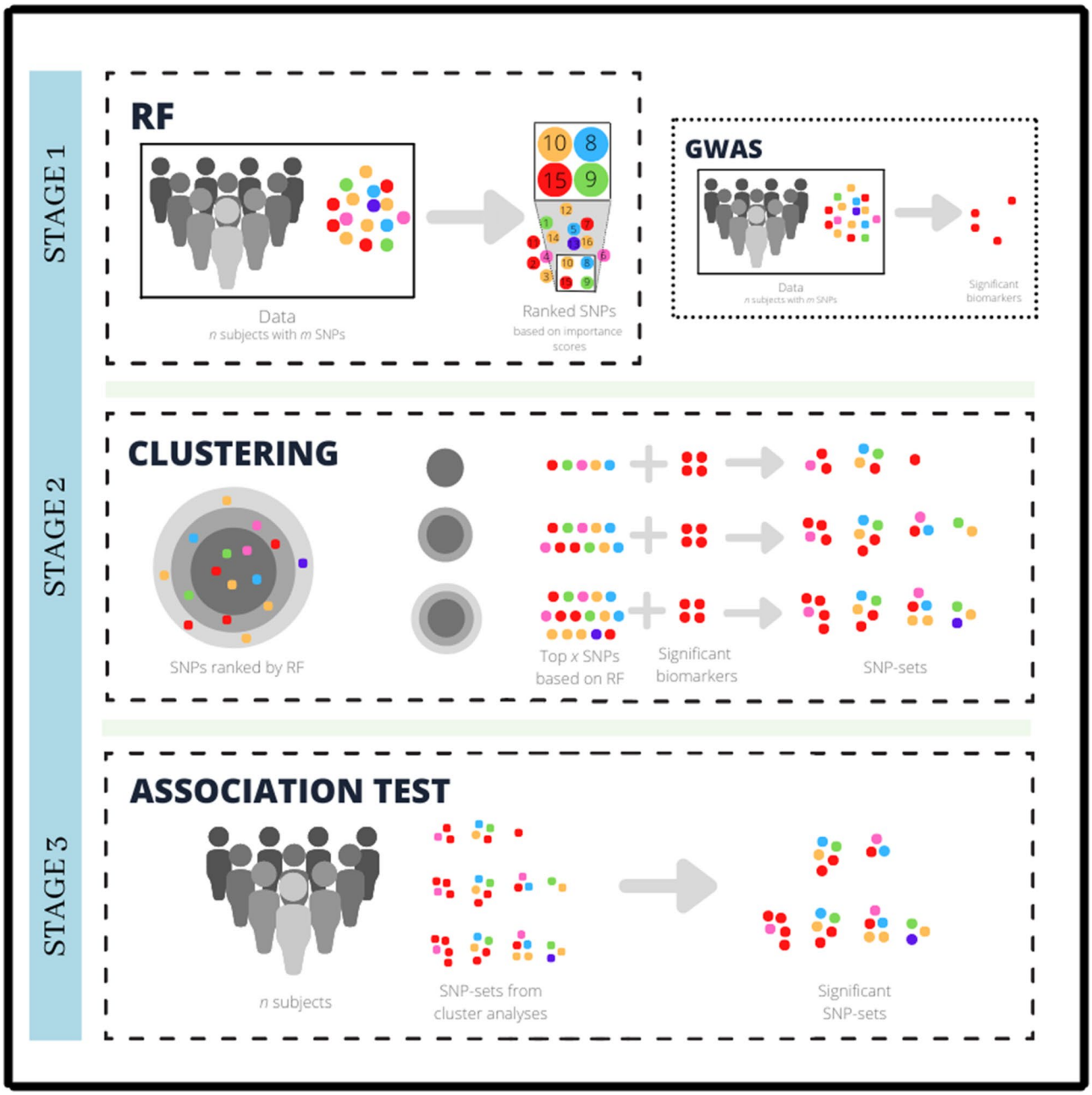

Figure 1. The architecture of the proposed integrated framework. In Stage 2, SNPs in concentric circles in darker shades of gray represent higher-ranking SNPs based on RF. (Image generated using Canva[40]).

chronic hepatitis B. The dataset is composed of 1,365,088 SNPs collected from 200 subjects of Korean ethnicity. The subjects were further divided into two groups: the cases (n = 100), which consist of patients who had experienced HBsAg seroclearance before the age of 60, and the controls (n = 100) comprising of patients who exhibited high levels (> 1000 IU/mL) of HBsAg at ≥ 60 years of age. An additive genetic model was utilized to transform the SNP dataset. A SNP marker is encoded as 0, 1, or 2 depending on the number of minor alleles it carries.

Dimension reduction. Dimension reduction is commonly a prerequisite in analyzing SNP datasets as large amounts of features impedes the capability of analytical approaches in performing fast and effective analyses. In this study, features are only selected for cluster analysis if they were considered by a previous GWAS to be statistically significant or if they are one of the top-ranking SNPs as per RF. RF ranks SNPs based on their feature importance scores which is a measure of the usefulness of a marker in predicting a target variable, in this case, trait occurrence. RF has been widely utilized in analyzing SNP data primarily due to its capacity to build a predictive model without making any assumptions about the underlying relationship between genotype and phenotype[39]. In RF, the predictive abilities of multiple decision trees, which are trained on bootstrap samples of the data, are consolidated to generate the final output prediction. In addition, randomization is not only induced

Cluster analysis experiment ID	Number of SNPs included
1	1047
2	2044
3	3041
4	4038
5	5036

Table 1. Number of SNPs subjected to cluster analysis.

by bootstrapping but also introduced at the node level when growing a tree. It selects a random subset of SNPs at each node of the tree as candidates to find the best split for the node. In estimating the importance of SNPs, RF calculates the Gini importance which quantifies the difference between a node's impurity and the weighted sum of the impurities of the two descendent nodes.

Mathematically, the importance of SNP_j is determined by summing the decrease in impurity (ΔI) for all the nodes t, where SNP_j is split. The decreases in impurity are weighted by fractions of samples in the nodes $p(t)$ and averaged over all trees in the forest. The Gini variable importance is then given by,

$$VI_{gini}^{(k)}(SNP_j) = \sum_{t \in T_k : v(s_t)} p(t)\Delta I(s_t, t)$$

where T_k is the number of nodes in the kth tree, $p(t) = \frac{n_t}{n}$ is the fraction of the samples reaching node t; and $v(s_t)$ is the variable used in the split s_t.

Since one round of calculations is not enough to ensure robustness of scores, the Leave-One-Out cross-validation (LOOCV) strategy was adopted. For each fold in the LOOCV, RF is trained on the (N-1) dataset, where N is the number of observations, and a corresponding score (the Gini variable importance) for each SNP is calculated. The scores obtained by a SNP for all folds are then averaged and the result would be the final feature importance score of that variant. In symbols,

$$SNP_i = \frac{\sum_{j=0}^{N} VI_{gini}^{k}(SNP_j)}{N}$$

where $VI_{gini}^{k}(SNP_j)$ is the SNP importance for the jth fold, N is the number of subjects, and SNP_i is the final SNP score of the ith SNP. The final scores are then used to rank the SNPs. The number of top biomarkers included in the clustering process is determined by the researchers as it is already outside the scope of RF. A detailed description on how SNPs are ranked is provided in the 'Appendix' section.

SNP-set formation. This study exploited the similarities shared among SNPs to identify novel susceptibility loci associated with HBsAg seroclearance. The analysis utilized the unsupervised machine learning method known as cluster analysis which aims to separate data points into distinct groups such that more similarities are shared among objects within the same group than objects belonging to different groups. Similarities between SNPs can be quantified in terms of *agreement,* i.e. based on the occurrence of sequence alterations computed via matching coefficients and measures of correlation, or *dependence*, i.e. based on the presence or absence of dependence quantified via measures based on the $\chi 2$-statistic[41]. This study adopts an *agreement*-based similarity measure by employing the method proposed in[30]. This method modified an agglomerative hierarchical clustering algorithm with average linkage for continuous data to develop a Hamming distance-based algorithm for determining SNP-sets. Hamming distance is a similarity measure used to calculate the number of dissimilar components between two categorical data points of the same size[42]. Applied to SNP data, the Hamming Distance d^{HAD} between SNPs i and j would be,

$$d^{HAD}(i,j) = \sum_{k=0}^{n-1} \left[y_{i.k} \neq y_{j.k} \right]$$

where n is the total number of subjects and y_k is the genotype of the kth subject. The similarity measure was adapted on SNP datasets based on the premise that the more individuals carrying the same genotype concerning two given SNPs or two SNP-sets (signified by a relatively small Hamming distance), the more similar the variants are and more likely to cluster[30].

Multiple cluster analyses were performed exclusively on GWAS-identified and top-ranking SNPs obtained by random forest. As shown in Table 1 Column 2, the number of SNPs analyzed was gradually increased to achieve a higher likelihood of discovering novel susceptibility loci. The set of SNPs included in each cluster analysis is the union of the 52 significant SNPs from Kim et. al.'s GWAS[38] and the top biomarkers identified by random forest (starting from top 1000 to top 5000 SNPs in increments of 1000). Each implementation resulted in candidate SNP-sets identified using the following parameters: *percentile cut* which specifies the height wherein a dendrogram will be cut and *minimum cluster size* which dictates the minimum number of SNPs for all clusters.

SNP-set	SNPs	Gene[a]	Chromosome
1	rs1809862, rs10769023, rs10838245, rs2017434, rs2047456, rs7945342, rs872751	UBQLNL; rs7945342 - OLFM5P	11
2	rs2399971, rs10508462, rs2153442, rs4748035	BEND7	10
3	rs2215905, rs2192611, rs199869387, rs887941, rs12464531, rs13018470	–	2
4	rs2119977, rs6826277, rs11931577	–	4
5	rs6749972, rs1558599, rs11891860, rs17584600	–	2
6	rs35689347, rs2173091, rs8037510	AGBL1	15
7	rs6462008, rs6947275, rs6462003	rs6462008 - EVX1, HOXA13; rs6947275 - HOTTIP, EVX1; rs6462003 - HOXA13	7
8	rs1505687, rs12620748, rs13382813	rs12620748 and rs13382813 - LINC01246	2
9	rs741229, rs12151705, rs6737829	–	2

Table 2. Cluster memberships of the SNPs that obtained a p-value less than 10^{-4} in Kim et al.'s GWAS[38].

Association test. Hamming distance-based association tests (HDAT)[30] were employed to identify the candidate SNP-sets significantly associated with HBsAg seroclearance. The presence of association depends on the amount of difference in the biomarkers found in cases and controls. Minor alleles were incorporated in the equations as it reveals more similarities in the genomes of two individuals than common alleles[43]. A comprehensive discussion of the equations used in HDAT can be found in[30]. Permutation test, a non-parametric test used to evaluate the statistical significance of a model through randomization, is used to compute the *p*-value of each SNP-set. The test calculates the *p-value* by permuting the dataset and constructing a test-statistic distribution and evaluating the probability that a test-statistic would be equal to or more extreme than the initial computed value.

Ethics approval and consent to participate. This study used the data provided in[38] which was a project approved by the ethics committees at Korea University Anam Hospital (ED13220) and conducted in agreement with the ethical principles of the Declaration of Helsinki. According to the project's ethical declaration, all patients provided written informed consent for participation and use of their data for research purposes.

Results

Top-ranking SNPs from dimension reduction. This study used random forest feature selection to reduce dataset dimensions prior to conducting cluster analysis. Specifically, random forest was employed to rank SNPs based on their feature importance score, a measure which determines a variant's relevance in making accurate phenotype predictions. SNPs are assigned a feature importance score based on the average scores for every fold in LOOCV to eliminate bias and ensure robustness. Investigation into the functional significance of three of the top five biomarkers ranked by RF led to possible connections between the variants and HBsAg seroclearance. SNPs rs28588178 (top-ranking SNP), rs1994209 (3rd-ranking SNP), and rs7958186 (5th-ranking SNP) are linked with Cadherin 4 (CDH4), PIG11, and PCED1B, respectively—genes reported to be associated with hepatocellular carcinoma (HCC)[44–46], a disease that can develop due to the presence of the hepatitis B virus.

Generated SNP-sets. Upon performing multiple cluster analyses, a total of 108 candidate SNP-sets were identified at a percentile cut of 0.9 and a minimum cluster size of 3. SNP-sets with the maximum number of SNPs were chosen in cases where there were overlaps to maximize the information obtained from the analyses.

SNP-sets containing SNPs which were considered significant in a previous GWAS were investigated as the variants sharing high degrees of similarity with GWAS-identified SNPs may also provide insights into trait etiology. As shown in Table 2, SNPs rs2399971, rs2119977, rs6826277, rs35689347, rs1505687, and rs741229 were grouped with at least one of the variants reported to be significantly associated with HBsAg seroclearance (Note: SNPs in boldface are those that obtained a p-value less than 10^{-4} in Kim et. al.'s GWAS[38]). Genes were retrieved from dbSNP[47] and[38].) No information regarding possible association existing between the latter five SNPs and the phenotype of interest was found; meanwhile, the opposite was true for rs2399971. Notably, albeit rs2399971 had not reached the cut-off value used in the GWAS performed by Kim et al.[38] on the whole study population, it was nevertheless found to be significantly associated with HBsAg seroclearance in the subjects who had received antiviral treatment[38]. Figure 2 shows the dendrogram of the GWAS-identified SNPs together with the aforementioned six variants and as presented, the SNPs belonging to the SNP-set which contains rs2399971 shows the least height differences, indicating that the SNPs in the set are more similar to each other than the variants found in other clusters.

Significant SNP-sets. Hamming distance-based association test (HDAT) was performed on the candidate SNP-sets to further identify SNPs possibly associated with HBsAg seroclearance. After performing a Bonferroni

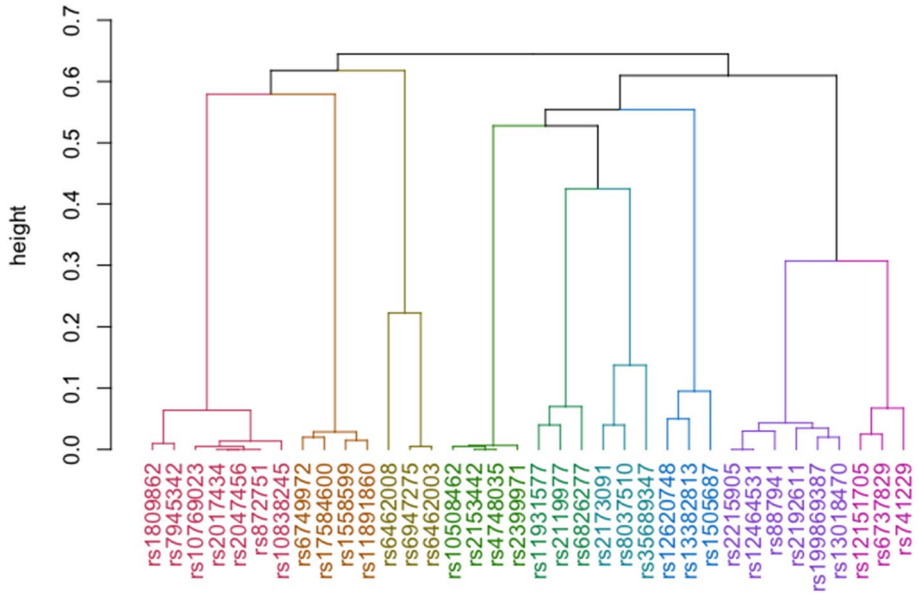

Figure 2. Dendrogram of the SNPs listed in Table 2.

SNP-set	List of SNPs	*p*-value
1	rs6731235, rs199703414, rs16829541, rs1485096, rs2341849	0.0002
2	rs28365850, rs62625038, rs17102970	0.0004
3	rs59659073, rs10754962, rs2380525	0.0004
4	rs200957040, rs1499880, rs4857702	0.0004
5	rs12644266, rs13130260, rs6815422	0.0001

Table 3. SNP-sets obtaining the lowest *p*-values (excluding those that harbor variants reported by Kim et al.[38] to be significantly associated with HBsAg seroclearance).

correction for multiple tests, 11 SNP-sets significantly associated with HBsAg seroclearance ($p\text{-}value < 0.0005$) were identified, the majority of which (7 out of 11) were found to harbor at least one of the GWAS-identified SNPs. Among the SNP-sets obtaining the lowest *p*-values, the set which obtained the highest test statistic is the one composed of rs1809862, rs10769023, rs10838245, rs2017434, rs2047456, rs7945342, and rs872751—all GWAS-identified SNPs[38]. All these variants reside in 11p15.4, a region that shows a possible correlation with HBsAg seroclearance. In a study by[48], it was observed that among hepatocellular carcinoma cases, more than 20 percent loss of heterozygosity (LOH) was shown for locus 11p, wherein region 11p15 was commonly affected. Moreover, a significant correlation was found to exist between LOH on 11p and HBsAg positivity. Specifically, results showed that there is a significantly higher frequency of LOH on 11p among hepatitis B virus carriers[48].

Table 3 shows the five significant SNP-sets which do not hold any of the GWAS-identified SNPs (*p*-values were obtained from 10000 permutations). No supporting evidence was found regarding possible associations between the individual variants belonging to the five SNP-sets and HBsAg seroclearance. Nonetheless, interesting findings were discovered when SNPs were analyzed collectively. Results showed that three out of the five SNP-sets in Table 3 harbor SNPs residing in similar genes, i.e. there is a corresponding gene for each distinct set. These are the following: (1) LOC105373438 for SNP-set 3, (2) LINC00578 for SNP-set 4, and (3) STOX2 for SNP-set 5. In[49], LINC00578 was reported to be a prognostic marker for pancreatic cancer (PC), a disease for which hepatitis B has been suggested to be a risk factor[50–52], increasing the likelihood of PC by 24%[53].

Discussion

This study aims to discover novel trait-associated susceptibility loci by augmenting previous GWAS findings using a machine learning-based SNP-set analysis approach built on the integration of RF and cluster analysis. By analyzing SNP-sets instead of individual variants, we increase the chances of discovering other existing true associations in two ways. First, by exploiting the similarities shared among the variants, SNPs that are truly associated with the trait of interest but which did not pass the threshold of significance can still be detected when grouped with statistically significant SNPs. Second, by reducing the unit of analysis into groups, a substantial decrease in the number of tests ensues which eliminates the necessity to adopt stringent thresholds used in considering a SNP significant. Investigation into the functional relevance of variants found in the same SNP-set

containing GWAS-identified SNPs and SNP-sets obtaining significant p-values led to the discovery of loci that may also contribute to phenotypic expression yet overlooked by GWAS as a consequence of its individual SNP analysis approach. The novelty in our proposed method lies in the GWAS-based and data-driven approach in feature selection prior to cluster analyses. This study did not restrict the discovery of susceptibility loci to a certain genomic region alone as the criteria for selecting SNPs depend on statistical significance and predictive powers. As a result, the resulting SNP-sets implicated a varied selection of genes and cytobands.

The proposed method was applied on an HBsAg seroclearance GWAS data[38] and was able to enhance the GWAS findings in two ways. First is through the discovery of SNPs highly similar with GWAS-identified variants. As shown in Table 2, statistically significant SNPs tend to cluster together. This acts as justification for further investigation of all variants belonging to the sets which contain GWAS-identified SNPs. It is possible that they may be false negatives or linked with the phenotype in some way. For example, rs2399971, a variant detected in the cluster analyses stage, was not considered significant in the GWAS conducted on the whole study population as it did not reach the threshold that was used (obtaining a p-value of 1.05×10^{-4} wherein the cut-off p-value used was 1.00×10^{-4}). Nevertheless, it was found to be significantly associated with HBsAg seroclearance in patients who had received antiviral treatment[38]. The other way in which the proposed approach has successfully enhanced the previous GWAS findings is through the identification of SNP-sets significantly associated with the trait of interest. Variants in Table 3 were not considered as statistically significant by the previous GWAS. However, since the SNP-sets where they belong showed association with the phenotype upon testing, we could say that some, if not all of them, could still be susceptible SNPs. This assumption is based on how HDAT results are interpreted as defined by[30]. Identifying these significant SNP-sets also allows hypothesis generation not only on SNPs but also on other larger biological units such as genes or cytobands[18,29]. For instance, gene LINC00578 and locus 11p15, regions implicated by two of the SNP-sets with the lowest p-values, have shown potential in understanding HBsAg seroclearance as both are linked with diseases associated with the presence of hepatitis B virus infection. By mapping out these implicated regions and identifying shared susceptibility loci with a well-researched phenotype, a better understanding of the intricate underpinnings of the trait of interest could be achieved. For instance, some of the SNPs associated with height may be considered in understanding the etiology of HBsAg seroclearance as 11p15 has been reported to harbor genes responsible for growth and development[54]. Furthermore, elevations in alanine transaminase (ALT) level, a consideration in declaring HBsAg seroclearance, was found to be an important factor for growth impairment in children[55].

Despite the advantages of the proposed method, several issues remain to be resolved. First, the total number of SNPs to consider in the clustering process should be optimized in future implementations of the approach so that variants possibly associated with the trait but obtaining low feature importance scores could still have higher chances of being discovered. Secondly, parameter values would still have to be tuned by utilizing specific measures such as gap statistics[56,57] to ensure an optimal number and a more cohesive composition of SNP-sets. Thirdly, the type of clustering procedure and association test employed on the SNP-sets should be modified depending on the goals of a study. HDAT, the association test used herein, only evaluates whether a SNP-set could distinguish different disease phenotypes. It does not determine if there is a presence of interaction in the set and even more so if that interaction is significant. Accounting for complex SNP interactions and nonlinear effects would require employing a different type of test on the SNP-sets such as the logistic kernel-machine-based test by[18].

Aside from optimizing the settings directly involved in the actual data analysis, another fundamental issue that needs to be addressed in this study is the possible presence of population structure in the dataset which could negatively affect the clustering results and could lead to spurious associations. Some approaches that could be used to correct for this include principal component analysis (PCA) and clustering techniques which utilize similarity measures such as the allele-sharing distance (ASD)[58]. However, the results of these techniques could be distorted when there is a large number of correlated markers due to linkage disequilibrium (LD)[58]. To address this problem, one could employ a clustering-based strategy on the SNPs initially in order to minimize the number of markers to only the most informative ones[59]. Following these premises, a proposed approach that could be utilized to avoid unreliable results is to perform clustering on the SNPs first to select representative markers, then on the patients to identify subpopulations, and lastly on the variants once again to identify the final list of SNP-sets that would then be subjected to the association tests. Although correcting for population structure is a prerequisite for any genetic data analysis, it should be proceeded with caution when the dataset being analyzed contains only a few observations (small sample size). In these settings, handling population stratification could be more complicated than usual especially if statistical power is at stake[30]. Mentioned that even though it is intuitive to address population stratification first before conducting association tests on the case and control groups of the same population, if the stratified populations are only of smaller sizes, then it could just lead to unstable findings.

Conclusion

This study aims to identify disease-associated susceptibility loci by augmenting previous GWAS findings using the integration of RF and cluster analysis. The proposed approach was applied to a hepatitis B virus surface antigen (HBsAg) seroclearance GWAS data[38]. Thereafter, the researchers were able to detect rs2399971, a variant that was not considered to be significantly associated with the phenotype in the main GWAS, but which obtained a significantly low p-value in a subgroup analysis[38]. Results of the association tests conducted on the generated SNP-sets led to the implication of gene LINC00578 and locus 11p15. The former was linked with pancreatic cancer[49] and the latter with hepatocellular carcinoma[48], diseases associated with hepatitis B virus infection. There are three ways in which readers could reinforce their findings using the proposed approach. The first one could be done during the dimension reduction phase wherein random forest is employed to identify SNPs which are highly predictive of the trait of interest. If a researcher found that the variant they discovered to be associated with a phenotype is also one of the top-ranking SNPs as per RF, then this could provide strong evidence

for follow-up investigations on the said variant. Predictive importance and association do not always coincide and so if they do, it could show important promise for clinical translation. The second way that readers could reinforce their findings is by looking at the SNP-sets which are significantly associated with the trait of interest. By identifying the genomic regions implicated by these sets, further evidence is provided to studies reporting on the significance of the said regions on a given phenotype. And lastly, one could check if the variant discovered to be associated with a phenotype belongs to a set which contains GWAS-identified SNPs. For instance, SNP rs2399971 was discovered in this study as it shared a high degree of similarity with variants significantly associated with HBsAg seroclearance. This somehow supports[38]'s finding on the association existing between rs2399971 and HBsAg seroclearance on patients who had received antiviral treatment[38]. Researchers who aim to extend this study could experiment on different supervised learning techniques for feature selection and utilize other similarity measures for clustering SNPs. With further investigation and validation, insights gleaned using the proposed framework could also be integrated into prediction models to aid in quantifying patients' risks for trait or disease development.

Data availability

The dataset used in this study can be accessed through this link: https://figshare.com/articles/dataset/gtReport_txt/6614975. The Python code used for implementing random forest can be found in https://github.com/jdgaudillo/SNP-ML.git while the R codes for clustering and association tests are available at http://homepage.ntu.edu.tw/~ckhsiao/HammingDistance/HD.htm.

Received: 23 May 2022; Accepted: 2 September 2022
Published online: 22 September 2022

References

1. Lvovs, D., Favorova, O. O. & Favorov, A. V. A polygenic approach to the study of polygenic diseases. *Acta Naturae.* **4**(3), 59–71 (2012).
2. Schork, N. J. Genetics of complex disease: Approaches, problems, and solutions. *Am. J. Respir. Care Med.* **156**(4), S103–S109. https://doi.org/10.1164/ajrccm.156.4.12-tac-5 (1997).
3. Visscher, P. M. *et al.* 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101**(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005 (2017).
4. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**(9), 581–590. https://doi.org/10.1038/s41576-018-0018-x (2018).
5. Norrgard K. Genetic variation and disease: GWAS. In: Nat Educ. https://www.nature.com/scitable/topicpage/genetic-variation-and-disease-gwas-682/#. Accessed 8 Mar 2022.
6. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**(7678), 92–94. https://doi.org/10.1038/nature24284 (2017).
7. Zhao, W. *et al.* Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat. Genet.* **49**(10), 1450–1457. https://doi.org/10.1038/ng.3943 (2017).
8. Kakuta, Y. *et al.* A genome-wide association study identifying RAP1A as a novel susceptibility gene for Crohn's disease in Japanese individuals. *J. Crohns Colitis.* **13**(5), 648–658. https://doi.org/10.1093/ecco-jcc/jjy197 (2019).
9. Antikainen, A. A. V. *et al.* Genome-wide association study on coronary artery disease in type 1 diabetes suggests beta-defensin 127 as a risk locus. *Cardiovasc Res.* **117**(2), 600–612. https://doi.org/10.1093/cvr/cvaa045 (2021).
10. Chen, Z., Boehnke, M., Wen, X. & Mukherjee, B. Revisiting the genome-wide significance threshold for common variant GWAS. *G3* **11**(2), jkaa056 (2021).
11. Génin, E. Missing heritability of complex diseases: Case solved?. *Hum Genet.* **139**(1), 103–113. https://doi.org/10.1007/s00439-019-02034-4 (2020).
12. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**(6), 446–450. https://doi.org/10.1038/nrg2809 (2010).
13. Klein, R. J. Power analysis for genome-wide association studies. *BMC Genet.* **8**(1), 1–8. https://doi.org/10.1186/1471-2156-8-58 (2007).
14. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1 (2019).
15. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**(12), 1452–1458. https://doi.org/10.1038/ng.2802 (2013).
16. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. *Hum. Mol. Genet.* **27**(20), 3641–3649. https://doi.org/10.1093/hmg/ddy271 (2018).
17. Ken-Dror, G., Humphries, S. E. & Drenos, F. The use of haplotypes in the identification of interaction between SNPs. *Hum. Hered.* **71**(1), 44–51. https://doi.org/10.1159/000350964 (2013).
18. Wu, M. C. *et al.* Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.* **86**(6), 929–942. https://doi.org/10.1016/j.ajhg.2010.05.002 (2010).
19. Howard, D. M. *et al.* Genome-wide haplotype-based association analysis of major depressive disorder in Generation Scotland and UK Biobank. *Transl. Psychiatry.* **7**(11), 1–9. https://doi.org/10.1038/s41398-017-0010-9 (2017).
20. Alonso-Gonzalez, A., Calaza, M., Rodriguez-Fontenla, C. & Carracedo, A. Gene-based analysis of ADHD using PASCAL: A biological insight into the novel associated genes. *BMC Med. Genet.* **12**(1), 1–2. https://doi.org/10.1186/s12920-019-0593-5 (2019).
21. Jin, L. *et al.* Pathway-based analysis tools for complex diseases: A review. *GPB.* **12**(5), 210–220. https://doi.org/10.1016/j.gpb.2014.10.002 (2014).
22. McCarthy, J. F. *et al.* Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. *Ann. NY Acad. Sci.* **1020**(1), 239–262. https://doi.org/10.1196/annals.1310.020 (2004).
23. Roy, A. A classification algorithm for high-dimensional data. *Procedia Comput. Sci.* **53**, 345–355. https://doi.org/10.1016/j.procs.2015.07.311 (2015).
24. Thottakkara, P. *et al.* Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS ONE* **11**(5), e0155705. https://doi.org/10.1371/journal.pone.0155705 (2016).
25. Gaudillo, J. *et al.* Machine learning approach to single nucleotide polymorphism-based asthma prediction. *PLoS ONE* **14**(12), e0225574. https://doi.org/10.1371/journal.pone.0225574 (2019).

26. Ramezani, M. *et al.* Investigating the relationship between the SNCA gene and cognitive abilities in idiopathic Parkinson's disease using machine learning. *Sci Rep.* **11**(1), 1–10. https://doi.org/10.1038/s41598-021-84316-4 (2021).
27. Zhang, Z. & Liu, Z. P. Robust biomarker discovery for hepatocellular carcinoma from high-throughput data by multiple feature selection methods. *BMC Med. Genet.* **14**(1), 1–12. https://doi.org/10.1186/s12920-021-00957-4 (2021).
28. Ickstadt, K., Mueller, T. & Schwender, H. Analyzing SNPs: Are there needles in the haystack?. *Chance mag.* **19**(3), 21–26. https://doi.org/10.1080/09332480.2006.10722798 (2006).
29. Ng, M.K., Li, M.J., Ao, S.I., Sham, P.C., Cheung, Y.M., Huang, J.Z. Clustering of SNP data with application to genomics, Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06). 2006:158–162. https://doi.org/10.1109/ICDMW.2006.43.
30. Wang, C., Kao, W. H. & Hsiao, C. K. Using Hamming distance as information for SNP-sets clustering and testing in disease association studies. *PLoS ONE* **10**(8), e0135918. https://doi.org/10.1371/journal.pone.0135918 (2015).
31. Xu, Y., Xing, L., Su, J., Zhang, X. & Qiu, W. Model-based clustering for identifying disease-associated SNPs in case-control genome-wide association studies. *Sci. Rep.* **9**(1), 1–10. https://doi.org/10.1038/s41598-019-50229-6 (2019).
32. Venkat, N. The curse of dimensionality: inside out, Pilani (IN): Birla Institute of Technology and Science, Pilani, Department of Computer Science and Information Systems (2018). https://doi.org/10.13140/RG.2.2.29631.36006.
33. Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nat. Methods.* **15**(6), 399–400. https://doi.org/10.1038/s41592-018-0019-x (2018).
34. Nguyen, T. T., Huang, J. Z., Wu, Q., Nguyen, T. T. & Li, M. J. Genome-wide association data classification and SNPs selection using two-stage quality-based random forests. *BMC Genom.* **16**(2), 1–11. https://doi.org/10.1186/1471-2164-16-S2-S5 (2015).
35. Roshan, U., Chikkagoudar, S., Wei, Z., Wang, K. & Hakonarson, H. Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. *Nucleic Acids Res.* **39**(9), e62. https://doi.org/10.1093/nar/gkr064 (2011).
36. Zhou, W., Bellis, E.S., Stubblefield, J., Causey, J., Qualls, J., Walker, K., *et al.* Minor QTLs mining through the combination of GWAS and machine learning feature selection. BioRxiv [Preprint] (2019). https://doi.org/10.1101/702761.
37. Bureau, A. *et al.* Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.* **28**(2), 171–182. https://doi.org/10.1002/gepi.20041 (2005).
38. Kim, T. H. *et al.* Identification of novel susceptibility loci associated with hepatitis B surface antigen seroclearance in chronic hepatitis B. *PLoS ONE* **13**(7), e0199094. https://doi.org/10.1371/journal.pone.0199094 (2018).
39. Botta, V., Louppe, G., Geurts, P. & Wehenkel, L. Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS ONE* **9**(4), e93379. https://doi.org/10.1371/journal.pone.0093379 (2014).
40. Free design tool: Presentations, video, social media | CANVA. Available from: https://www.canva.com/
41. Selinski, S. Similarity measures for clustering SNP and epidemiological data. Technical Report, No. 2006,25, Dortmund (DE): University of Dortmund, Collaborative Research Center 'Reduction of Complexity in Multivariate Data Structures' (SFB 475). 2006. http://hdl.handle.net/10419/22668.
42. Hamming, R. W. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**(2), 147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x (1950).
43. Wessel, J. & Schork, N. J. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am. J. Hum. Genet.* **79**(5), 792–806. https://doi.org/10.1086/508346 (2006).
44. Gao, Y. *et al.* Long non-coding RNA linc-cdh4-2 inhibits the migration and invasion of HCC cells by targeting R-cadherin pathway. *Biochem. Biophys. Res. Commun.* **480**(3), 348–354. https://doi.org/10.1016/j.bbrc.2016.10.048 (2016).
45. Wu, Y. *et al.* PIG11 is involved in hepatocellular carcinogenesis and its over-expression promotes Hepg2 cell apoptosis. *Pathol. Oncol. Res.* **15**(3), 411–416. https://doi.org/10.1007/s12253-008-9138-5 (2009).
46. Ding, H., He, J., Xiao, W., Ren, Z., Gao, W. LncRNA PCED1B-AS1 is overexpressed in hepatocellular carcinoma and regulates miR-10a/BCL6 axis to promote cell proliferation. Res Sq. (2020). https://doi.org/10.21203/rs.3.rs-79374/v1.
47. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677–679 (1999).
48. Sheu, J. C. *et al.* Loss of heterozygosity and microsatellite instability in hepatocellular carcinoma in Taiwan. *Br. J. Cancer.* **80**(3), 468–476. https://doi.org/10.1038/sj.bjc.6690380 (1999).
49. Zhang, B., Li, C. & Sun, Z. Long non-coding RNA LINC00346, LINC00578, LINC00673, LINC00671, LINC00261, and SNHG9 are novel prognostic markers for pancreatic cancer. *Am. J. Transl. Res.* **10**(8), 2648 (2018).
50. Ben, Q. *et al.* Hepatitis B virus status and risk of pancreatic ductal adenocarcinoma: A case-control study from China. *Pancreas* **41**(3), 435–440. https://doi.org/10.1097/MPA.0b013e31822ca176 (2012).
51. Iloeje, U. H. *et al.* Risk of pancreatic cancer in chronic hepatitis B virus infection: data from the REVEAL-HBV cohort study. *Liver Int.* **30**(3), 423–429 (2010).
52. Wang, Y. *et al.* Hepatitis B virus status and the risk of pancreatic cancer: A meta-analysis. *Eur. J. Cancer Prev.* **22**(4), 328–334 (2013).
53. Desai, R. *et al.* Association between hepatitis B infection and pancreatic cancer: a population-based analysis in the United States. *Pancreas* **47**(7), 849–855. https://doi.org/10.1097/MPA.0000000000001095 (2018).
54. Weksberg, R., Smith, A. C., Squire, J. & Sadowski, P. Beckwith-Wiedemann syndrome demonstrates a role for epigenetic control of normal development. *Hum. Mol. Genet.* **12**(1), R61–R68. https://doi.org/10.1093/hmg/ddg067 (2003).
55. Gerner, P., Hörning, A., Kathemann, S., Willuweit, K. & Wirth, S. Growth abnormalities in children with chronic hepatitis B or C. *Adv. Virol.* https://doi.org/10.1155/2012/670316 (2012).
56. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc. B.* **63**(2), 411–423. https://doi.org/10.1111/1467-9868.00293 (2001).
57. Yan, M. & Ye, K. Determining the number of clusters using the weighted gap statistic. *Biometrics* **63**(4), 1031–1037. https://doi.org/10.1111/j.1541-0420.2007.00784.x (2007).
58. Alhusain, L. & Hafez, A. M. Nonparametric approaches for population structure analysis. *Hum. Genomics* **12**(1), 1–2. https://doi.org/10.1186/s40246-018-0156-4 (2018).
59. Paschou, P., Lewis, J., Javed, A. & Drineas, P. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *J. Med. Genet.* **47**(12), 835–847. https://doi.org/10.1136/jmg.2010.078212 (2010).

Acknowledgements

The authors would like to thank the Department of Science and Technology—Philippine Council for Health Research and Development (DOST-PCHRD) for providing the necessary funding and assistance that made this study possible. The analysis conducted herein acts as a preliminary study for the project sponsored by DOST-PCHRD entitled "AI-driven Integration of Genomic, Ultrasound, Serum Biomarkers, and Clinical data for Early diagnosis of Liver Cancer" under the program "Early CANcer Detection in the LivEr of Filipinos with Chronic Hepatitis B Using AI-Driven Integration of Clinical and Genomic Biomarkers (CANDLE Study)".

Author contributions

P.P.S.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing—Original Draft, Writing—Review and Editing, Visualization; J.D.G.: Conceptualization, Methodology, Validation, Investigation, Writing Original Draft, Writing—Review and Editing, Supervision, Data Curation; J.A.V.: Conceptualization, Writing—Review and Editing; R.M.L.R.-V.: Resources, Writing—Review and Editing, Project Administration, Funding Acquisition; B.J.T.: Resources, Project Administration, Funding Acquisition; M.R.D.: Resources, Project Administration, Funding Acquisition; J.R.A.: Resources, Writing—Review and Editing, Project Administration, Funding Acquisition.

Funding

Funding for this work was provided by the Department of Science and Technology—Philippine. Council for Health Research and Development (DOST-PCHRD).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-19708-1.

Correspondence and requests for materials should be addressed to J.D.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.