

A Two-step Feature Selection Framework for Computationally Efficient Machine Learning on Genome-wide SNP Data

Jinhyun Kim

The University of Suwon

Seungjun Lee

The University of Suwon

Daewon Lee

Hanyang University

Seo Woo Song

Korea Institute of Science and Technology (KIST)

Cheolheon Park

Seoul National University

Huiran Yeom

huiran.yeom@gmail.com

The University of Suwon

Research Article

Keywords: Big data, feature selection, geographic SNP, machine learning, computational efficiency

Posted Date: August 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10747589/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Recent advances in next-generation sequencing (NGS) technologies have substantially increased the scale of genomic datasets, creating major computational challenges for genome-wide analyses involving tens of millions of single nucleotide polymorphisms (SNPs). In this study, we present a two-step feature selection framework designed to improve the computational efficiency of machine learning-based genome-wide SNP analysis.

Results

Using data from the 1000 Genomes Project, we first applied variance-based hard filtering to reduce the number of SNPs from approximately 61 million to 1 million. This preliminary reduction enabled the practical application of more computationally intensive machine learning-based feature selection methods. Among the evaluated secondary selection approaches, XGBoost was selected based on its overall performance, and the resulting SNP subsets were subsequently used to train a support vector machine classifier for country-level geographic origin prediction. The proposed framework reduced the estimated computational requirements of XGBoost-based feature selection by approximately 25-fold while preserving the utility of the selected SNP subsets for downstream classification. We further evaluated prediction performance across a wide range of SNP subset sizes and identified differences among populations in the number of informative SNPs required for classification.

Conclusions

By transforming a 61-million-SNP dataset into a computationally tractable feature space, the proposed framework enables machine learning-based feature selection at a scale that would otherwise require prohibitive computing resources. This two-step strategy therefore provides a practical solution for applying computationally intensive machine learning methods to population-scale genome-wide SNP datasets under realistic hardware constraints.

Background

The advancement of Next-Generation Sequencing (NGS) technologies has dramatically increased the production of genomic data, enabling high-dimensional analyses of millions of single nucleotide polymorphisms (SNPs) across large populations. For example, the 1000 Genomes Project provides an extensive dataset capturing the genetic diversity of global populations, widely applied in disease prediction, ancestry inference, and population structure analysis. However, whole-genome analyses of such large-scale SNP data demand substantial computational resources. In particular, applying machine learning methods to identify informative predictors for these tasks is challenged by the extreme

dimensionality of genome-wide data—where the number of features (p , often millions of SNPs) vastly exceeds the number of samples (n , typically limited by cohort size)—leading to issues such as overfitting to the training data, commonly referred to as the “ $p \gg n$ ” problem.

To address these challenges, various approaches have been proposed, including dimensionality reduction and feature selection. Among these, feature selection can be an effective strategy for improving predictive accuracy and computational efficiency, especially when the selected features need to remain interpretable in the context of the underlying biology. It works by identifying a subset of SNPs that are statistically relevant to the target phenotype or population origin, thereby effectively reducing the high dimensionality of the dataset. When combined with cross-validation, care must be taken to avoid feature selection bias, which can inflate performance estimates if selection is not confined to the training data.

Genetic algorithms (GA) are commonly used for their ability to effectively search for feature subsets that optimize the predictive performance of a given model, such as maximizing classification accuracy or minimizing error rates. However, GAs require substantial computational resources and lengthy processing times, making them less practical for large-scale SNP datasets. Similarly, while XGBoost is highly effective for feature selection due to its capability to handle complex interactions between features, applying it directly to an initial set of over 10 million SNPs is impractical due to memory constraints and excessive computational demands. This highlights the necessity for feature selection methods that balance computational efficiency with high predictive accuracy, ensuring the practical feasibility of genomic data analysis.

Here, we propose an integrated feature selection method that combines multiple techniques to enhance computational efficiency and prediction accuracy in genomic big data analysis. Our approach involves a two-step feature selection process, which includes variance-based hard filtering followed by refined selection methods such as XGBoost. By employing hard filtering to remove low-variance SNPs ($> 60M$) and reduce the dataset to a manageable size ($< 1M$), we enable the effective application of refined feature selection methods, including machine learning techniques. The selected SNP features were then used to train a Support Vector Machine (SVM) model, chosen for its robustness in managing high-dimensional genomic datasets. This phase of our research involved a detailed analysis of the model's performance and resource consumption, providing valuable insights into the trade-offs inherent in computational genomics. Furthermore, we identified the minimal number of SNPs required for accurate differentiation between national populations, enhancing the practical application of our findings in fields such as genetic testing and personalized medicine. By optimizing the feature selection process through this two-step method, our approach ensures high accuracy in geographic origin prediction while efficiently managing computational resources.

Methods

Data acquisition

The data utilized in this study are provided by the International Genome Sample Resource (IGSR)²⁶. These data can be accessed through a specific website dedicated to this collection: <https://www.internationalgenome.org/data-portal/data-collection/30x-grch38>. Within the IGSR data portal, phased VCF files comprising 3202 samples were downloaded.

Preliminary feature selection

Prior to preliminary feature selection, we filtered out biallelic SNPs (i.e., genomic variant features) that were present in fewer than 3 individuals or in more than $n - 3$ individuals (where n is the total number of samples). In this context, a SNP was considered "present" if it had at least one non-reference allele in a sample. SNPs with extremely low or high allele counts are generally uninformative for distinguishing geographic origin and were excluded to reduce noise and computational burden during downstream analysis. After this initial filtering, we employed variance-based preliminary feature selection. Specifically, the variance of each feature was calculated, and the 1 million features with the highest variances were selected. This process reduced the dataset from approximately 61.6 million SNPs to about 43 million SNPs based on sample support counts, and further to 1 million SNPs through variance-based selection. Consequently, the variance-based preliminary feature selection retained approximately 2.3% of the variants. As a baseline for the preliminary feature selection, we substitute variance-based feature selection into random feature selection to compare the effect of the preliminary feature selection method. We did not explore other well-known univariate feature selection methods such as "chi2" and "f_classif" implemented in scikit-learn, due to their prohibitive memory requirements.

Machine learning-based secondary feature selection

For the secondary feature selection, we utilized machine learning algorithms to further refine the features identified during the preliminary stage. We employed the XGBoost model, implemented using the XGBoost Python package²⁴. To optimize the model, we conducted systematic hyperparameter tuning by evaluating combinations from the following grid: `n_estimators = [100, 500, 1000]`, `max_depth = [2, 3, 6]`, `gamma = [0, 1]`, `subsample = [0.5, 0.8, 1]`, `reg_lambda = [1, 2]`, and `reg_alpha = [0, 0.5]`, with `colsample_bytree` fixed at 0.8. Early stopping with `early_stopping_rounds = 10` was applied based on validation loss to prevent overfitting in the high-dimensional feature space. The final hyperparameter configuration—`learning_rate: 0.1`, `n_estimators: 1000`, `max_depth: 3`, `gamma: 0`, `subsample: 0.8`, `colsample_bytree: 0.8`, `reg_lambda: 1`, and `reg_alpha: 0`—was selected based on the highest validation accuracy. Feature importance scores from the trained XGBoost model were then used to rank and select the most relevant features.

Similarly, we applied the RandomForestClassifier from the scikit-learn library and performed hyperparameter tuning over the following grid: `n_estimators = [100, 500, 1000]`, `max_features = ['sqrt', 'log2']`, `max_depth = [3, None]`, `min_samples_split = [2, 3, 5]`, and `min_samples_leaf = [1, 2]`. The selected

configuration—`n_estimators`: 500, `max_features`: 'sqrt', `max_depth`: None, `min_samples_split`: 2, and `min_samples_leaf`: 1—yielded the best validation performance.

We evaluated the effectiveness of these models by selecting various numbers of top-ranked features for downstream classification: 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768, 65536, and 131072. These features were selected from the 1 million features identified during the preliminary feature selection stage.

Univariate secondary feature selection

The `chi2` and `f_classif` functions from the scikit-learn library were utilized to perform chi-square statistical tests and obtain ANOVA F-values, respectively. For variance-based feature selection, variances were calculated across samples for each feature. The top *k* highest scoring features were selected from the 1 million features identified during the preliminary feature selection stage.

Prediction of geographic origin using selected features

Using the features selected during the secondary feature selection process, we applied a Support Vector Machine (SVM) model to predict geographic origin. The hyperparameters for SVM training were optimized using the `GridSearchCV` function, exploring a range of *C* values: 0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, and 10. We opted for the linear kernel, as it consistently outperformed the `rbf` and `sigmoid` kernels in high-dimensional feature spaces (Supplementary Fig. 2A). To ensure the reproducibility of our method, all experiments, including feature selection and machine learning model training, were conducted in five replicates with different random seeds. Although we also tested other classification algorithms, including XGBoost, SNP-BLUP, and Random Forest, these models demonstrated inferior performance compared to the linear SVM (Supplementary Fig. 2B).

Identifying minimum SNPs for accurate geographic origin differentiation

To determine the minimum number of features required for classifying each geographic origin, we first reclassified the data by designating a specific geographic origin as the "target" class, while all remaining classes were grouped into a single class named "others." We then applied the XGBoost feature selection method, as previously described, with various target feature numbers ranging from 1 to 8192. An SVM classifier was trained on the selected features under the target vs. others classification scheme. The minimum number of SNPs was determined by identifying the point at which the test set *f1* macro score differed by no more than 2 percentage points from the *f1* macro score before feature selection. The *f1* macro score was chosen due to the inherent label imbalance in this classification task. These procedures were iteratively performed for each geographic origin to ascertain the class-specific minimum SNPs. The overlap between the identified minimum SNPs across different geographic origins

was then calculated. The minimum SNP identification process began with the 8192 SNPs selected by XGBoost, which demonstrated the highest geographic origin prediction accuracy (Fig. 2B).

Results

Comprehensive feature selection using 1000 Genomes data

The study leverages data from the 1000 Genomes Project, offering an extensive and diverse dataset encompassing multiple continents and populations (Fig. 1A). The dataset includes 3,202 samples representing 26 country-level geographic origins, with each country containing between 74 to 179 samples. These samples are partitioned into 60% training, 20% validation, and 20% test sets. Notably, even the country with the fewest samples had 15 test samples, ensuring sufficient data for performance evaluation. This diverse dataset is essential for ensuring that feature selection and subsequent model training are robust and generalizable across different genetic backgrounds.

The dataset comprises approximately 61 million SNPs, far exceeding the number of samples, making it impractical for direct use in machine learning-based geographic prediction tasks. Our analysis revealed that random selection of up to 100,000 SNPs does not result in clear population separation, as demonstrated by the principal component analysis (PCA) plots for 100, 1,000, 10,000, and 100,000 randomly selected SNPs (Fig. 1B). These findings underscore the need for selecting an adequate number of informative SNPs to achieve reliable ancestry inference.

To address this challenge, we propose an optimal feature selection methodology designed to identify the minimal number of significant SNPs necessary for accurate geographic origin prediction. Our comprehensive comparison of several feature selection methods begins with phased SNP data from the 1000 Genomes Project. The workflow includes the following key steps: preprocessing the raw SNP data to convert variant call files (VCF) into a numeric matrix, applying various feature selection methods to determine the optimal feature selection strategy, and training machine learning models to predict geographic origin (Fig. 1C).

A critical aspect of our approach is the implementation of a two-step feature selection process, which balances computational efficiency and high predictive accuracy. Given the resource constraints associated with processing over 61 million SNPs, most feature selection methods were not directly applicable. Therefore, we divided the feature selection workflow into two distinct stages (Fig. 1C). Initially, we employed a preliminary feature selection step using methods based on variance, F_{st} , and minor allele frequency (MAF), which are computationally less demanding, to reduce the dataset to approximately 1 million SNPs. Subsequently, we applied refined feature selection methods, including machine learning-based methods such as XGBoost and random forest, and univariate feature selection methods such as ANOVA F-value, chi-squared statistics, variance, and random selection, to further narrow down the dataset to a few hundred or thousand SNPs. After feature selection, we utilized a

Support Vector Machine (SVM) model to predict geographic origin. This systematic, two-step feature selection process ensures the identification and utilization of the most informative SNPs, thereby enhancing the accuracy and efficiency of genomic predictions while optimizing resource usage.

Integrated feature selection for efficient and optimal SNP selection

We began by reducing the dataset from 61 million SNPs to approximately 43 million SNPs by discarding those supported by fewer than three samples, as these are likely non-informative for machine learning training. For preliminary feature selection, we compared random selection and variance-based hard filtering to select 1 million SNPs. Following this, we evaluated five different secondary feature selection methods, selecting various numbers of features ranging from 128 to approximately 100,000.

Without preliminary feature selection (using random selection), all other secondary feature selection methods performed better than the baseline random selection (Fig. 2A, Supplementary Fig. 1A). Overall, XGBoost achieved the highest test set accuracy ($78.2 \pm 2.4\%$ s.e.m.; $n = 35$), followed by random forest ($76.7 \pm 2.4\%$ s.e.m.; $n = 35$), variance ($74.7 \pm 2.6\%$ s.e.m.; $n = 35$), ANOVA F-value ($72.0 \pm 2.8\%$ s.e.m.; $n = 35$), chi-squared ($70.8 \pm 2.8\%$ s.e.m.; $n = 35$), and random ($57.6 \pm 3.3\%$ s.e.m.; $n = 35$). For this overall accuracy calculation, we considered cases where the number of selected features ranged from 128 to 8,192, which is the required number of features for training with 3,202 samples.

After preliminary feature selection using variance hard filtering, the results showed a remarkably different pattern (Fig. 2B, Supplementary Fig. 1B). XGBoost continued to achieve the highest test set accuracy ($81.7 \pm 2.1\%$ s.e.m.; $n = 35$), with a peak accuracy of $95.2 \pm 0.2\%$ s.e.m. ($n = 5$) when 8,192 features were selected. However, all other secondary feature selection methods performed worse compared to the baseline random selection. This suggests that preliminary feature selection effectively pre-screens informative SNPs, allowing XGBoost to select the most informative SNPs. Specifically, random forest ($72.3 \pm 2.6\%$ s.e.m.; $n = 35$), variance ($57.7 \pm 2.5\%$ s.e.m.; $n = 35$), ANOVA F-value ($52.7 \pm 2.7\%$ s.e.m.; $n = 35$), and chi-squared ($51.9 \pm 3.1\%$ s.e.m.; $n = 35$) all showed lower overall accuracy compared to random selection ($77.9 \pm 2.6\%$ s.e.m.; $n = 35$).

Importantly, the combination of variance hard filtering and XGBoost feature selection yielded the best geographic prediction accuracy (81.7%) among all other combinations of feature selection methods (Fig. 2C, Supplementary Fig. 1). Using variance filtering alone (57.7%) did not outperform random selection alone (57.6%). Direct application of XGBoost to the initial feature data (61 million SNPs) is computationally infeasible, requiring an estimated CPU time of approximately 1,250 hours and peak memory of approximately 2.5 TB (Fig. 2D). Therefore, our two-step feature selection approach facilitates computationally efficient preliminary selection, making more sophisticated machine learning-based feature selection feasible and enabling the identification of the most informative features with the highest accuracy.

Identifying minimum SNPs required for accurate geographic origin differentiation

By applying our feature selection process, we sought to identify the minimum number of SNPs required for the differentiation of each geographic origin (Fig. 3A, Supplementary Table 1). The Puerto Rican population (PUR) required the highest number of SNPs, with 8,192 needed for accurate differentiation. Conversely, the Esan population (ESN) required the fewest SNPs, with only 128 necessary for differentiation, followed by the Japanese (JPT), Colombian (CLM), and Kinh Vietnamese (KHV) populations, each requiring 256 SNPs. Additionally, we assessed the overlap of these minimal SNP sets across different geographic origins. Using Finland as a reference, the highest overlap was observed with the Peruvian population (PEL), where 68.3% of the SNPs overlapped with those of Finland among the 1,024 minimum SNPs identified in PEL. The Bengali population (BEB) showed the second-highest overlap, with 67.8% of its 1,024 minimum SNPs overlapping with those of Finland. Conversely, the lowest overlap was observed in the Kinh Vietnamese population (KHV), with only 9.8% of its 256 minimum SNPs overlapping with those of Finland, likely reflecting the geographical and genetic separation between Finland and Vietnam.

We also measured the class-level accuracy for each geographic origin (Fig. 3B). Populations such as the Iberian (IBS), Bengali (BEB), Peruvian (PEL), CEPH (CEU), Mende (MSL), Luhya (LWK), Gambian Mandinka (GWD), Finnish (FIN), Yoruba (YRI), and Kinh Vietnamese (KHV) all exhibited perfect prediction accuracy ($100 \pm 0.0\%$ s.e.m.; $n = 3$), indicating that their genetic characteristics are distinct and easily identifiable. In contrast, the British population (GBR) showed significantly lower prediction accuracy (3.7%), reflecting the greater genetic diversity within this population.

The confusion matrix (Fig. 3C) provides further insights into prediction accuracy and misclassification rates among different population groups. Notably, 94% (17/18) of British (GBR) samples were misclassified as CEPH (CEU), suggesting a high degree of genetic similarity between these groups. Additionally, 52% (12/23) of African Caribbean (ACB) samples were misclassified as Yoruba (YRI), and 43% (9/21) of Han Chinese (CHB) samples were misclassified as Southern Han Chinese (CHS). The Punjabi population (PJT) frequently showed misclassifications with other South Asian groups, such as Gujarati (GIH), Telugu (ITU), and Tamil (STU), likely due to the genetic similarity within the South Asian region. These results underscore the complexity and challenges of accurately predicting genetic ancestry, particularly in populations with high genetic diversity.

Discussion

In this study, we introduce an integrated feature selection pipeline aimed at improving computational efficiency while preserving predictive performance in the analysis of genome-wide SNP data, which often contains tens of millions of variants. Our two-step feature selection strategy—consisting of initial variance-based hard filtering followed by XGBoost-based refinement—allowed us to effectively reduce

the dataset from approximately 61 million SNPs to a much smaller subset of highly informative features, while preserving the model's classification performance.

In the first step, a variance-based hard filter reduced the dimensionality from 61 million SNPs to approximately 1 million. This reduction was essential for making it computationally feasible to apply more advanced feature selection algorithms. In the second step, we applied XGBoost, a tree-based ensemble learning method known for its efficiency and high performance in high-dimensional settings. Our results showed that this refined selection process significantly reduced computational burden: the estimated runtime for XGBoost feature selection decreased from over 1,250 hours for the full dataset to just 50 hours after initial filtering—a 25-fold reduction. Despite the reduction in SNP count, the model maintained competitive prediction accuracy, comparable to previous studies.

In addition to evaluating different feature selection strategies, we systematically analyzed prediction accuracy across a wide range of SNP counts (128 to 131,072) to assess the trade-off between the number of features and model performance. The results showed that for certain populations (e.g., Esan), accuracies above 95% could be achieved with as few as 128 SNPs, while others such as Japan, Colombia, and Kinh Vietnamese required 256 SNPs to reach similar accuracy (Fig. 3A). Many populations also maintained stable performance with only a few thousand SNPs (e.g., 2,048 or 8,192). This variability highlights that our contribution is not to propose a single universal SNP set, but to provide empirical evidence that enables researchers to tailor SNP numbers to specific populations and resource constraints.

However, a notable limitation of our study is the lack of a detailed analysis of population admixture, which can substantially affect geographic origin predictions. As global human migration increases, individuals with mixed ancestry are becoming more common, and traditional population labels may no longer fully reflect genetic structure. Recent studies have introduced methods such as ADMIXTURE, STRUCTURE, RFMix, and Local Ancestry Inference to account for admixture patterns in genomic data. Although our focus was on optimizing feature selection workflows, integrating admixture-aware algorithms in future studies could potentially improve the resolution and accuracy of geographic origin inference.

To mitigate potential confounding effects from populations with substantial historical admixture in the present study, we excluded entire geographic origin classes that consistently showed low prediction accuracy (< 90%) across multiple independent runs during preliminary experiments. This decision was guided by known challenges reported in previous population genomics studies, where certain groups exhibited unclear genetic boundaries or high label noise. Importantly, we removed entire classes rather than individual samples, and we re-ran the full analysis pipeline—including both preliminary and secondary feature selection as well as hyperparameter tuning—after this cleaning step to avoid any risk of information leakage. While this procedure improved interpretability for the remaining populations, we acknowledge that it also reduced the representativeness of the dataset.

Another limitation lies in our inability to evaluate model performance on the full 61 million SNP dataset due to server resource constraints. This prevented a direct comparison of alternative methods across the complete genome. Furthermore, we emphasize the importance of rigorous validation, ensuring that feature selection is performed exclusively within the training set in each fold of cross-validation to avoid optimistic bias.

Conclusions

The two-step SNP selection approach proposed in this study was primarily designed to improve computational efficiency in the analysis of genome-wide SNP data. By filtering out SNPs with minimal variability that limits their discriminatory power, our method enables the more efficient application of machine learning algorithms to high-dimensional genomic datasets containing millions of SNPs. However, the prediction accuracy observed in this study remained modest for certain countries, which may be attributed to factors such as limited sample sizes, complex population structures, and mismatches between political boundaries and genetic variation. Therefore, the main contribution of our method lies in enhancing computational tractability rather than achieving the highest possible predictive accuracy. We believe that with access to more refined reference populations or ancestry-informed labels in future studies, the predictive performance could be substantially improved.

Abbreviations

ANOVA

analysis of variance

CPU

central processing unit

FST

fixation index

GA

genetic algorithm

IGSR

International Genome Sample Resource

MAF

minor allele frequency

NGS

next-generation sequencing

PCA

principal component analysis

SNP

single nucleotide polymorphism

SVM

support vector machine

VCF

variant call format

s.e.m.

standard error of the mean

Declarations

Ethics approval and consent to participate

Not applicable. This study used publicly available, de-identified data from the International Genome Sample Resource/1000 Genomes Project and did not involve new participant recruitment, specimen collection, or identifiable individual data.

Consent for publication

Not applicable.

Availability of data and materials

The data utilized in this study are provided by the International Genome Sample Resource (IGSR) and are available at <https://www.internationalgenome.org/data-portal/data-collection/30x-grch38>. Phased VCF files comprising 3,202 samples were downloaded from the IGSR data portal. Analysis scripts are available on GitHub at https://github.com/Jinhyun-Kim/geography_prediction.git.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Bio&Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00440370, RS-2026-25486095, RS-2025-24803203). This work was supported by the research grant of The University of Suwon, 2023.

Authors' contributions

H.Y. and J.K. analyzed data, wrote the main manuscript text, and S.L. prepared figures. D.L., S.W.S., and C.P. reviewed the manuscript.

Acknowledgements

Not applicable.

Authors' information

Not applicable.

References

1. Altshuler DM, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56–65.
2. Auton A, et al. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
3. Abdellaoui A, Dolan CV, Verweij KJH, Nivard MG. Gene–environment correlations across geographic regions affect genome-wide association studies. *Nat Genet*. 2022;54:1345–54.
4. Byrska-Bishop M, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*. 2022;185:3426–e344019.
5. Carrot-Zhang J, et al. Genetic ancestry contributes to somatic mutations in lung cancers from admixed latin american populations. *Cancer Discov*. 2021;11:591–8.
6. Mahajan A, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet*. 2014;46:234–44.
7. Wang Y, et al. Ancestry inference using reference labeled clusters of haplotypes. *BMC Bioinformatics*. 2021;22:1–14.
8. Carlson CS et al. Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLoS Biol* 11, (2013).
9. Esmaeili F, Narimani Z, Vasighi M. Discovering SNP-disease relationships in genome-wide SNP data using an improved harmony search based on SNP locus and genetic inheritance patterns. *PLoS ONE*. 2023;18:1–18.
10. Waters KM et al. Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. *PLoS Genet* 6, (2010).
11. Kidd KK, et al. Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int Genet*. 2014;10:23–32.
12. Bermingham ML, et al. Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Sci Rep*. 2015;5:1–12.
13. Kusmec A, Schnable PS, FarmCPUpp. Efficient large-scale genomewide association studies. *Plant Direct* 2, (2018).
14. Yelmen B, et al. Deep convolutional and conditional neural networks for large-scale genomic data generation. *PLoS Comput Biol*. 2023;19:1–21.
15. Heinrich F, et al. Exploring the potential of incremental feature selection to improve genomic prediction accuracy. *Genet Selection Evol*. 2023;55:1–10.
16. Hong SC, Muyas F, Cortés-Ciriano I, Hormoz S. scAI-SNP: a method for inferring ancestry from single-cell data. *BMC Methods* 2, (2025).
17. Wang C, et al. A biogeographical ancestry inference pipeline using PCA-XGBoost model and its application in Asian populations. *Forensic Sci Int Genet*. 2025;77:103239.

18. Mota LFM, et al. Variable selection strategies for genomic prediction of growth and carcass related traits in experimental Nellore cattle herds under different selection criteria. *Sci Rep.* 2025;15:1–15.
19. Kavakiotis I, Samaras P, Triantafyllidis A, Vlahavas IFIFS. A data mining method for informative marker selection in high dimensional population genomic data. *Comput Biol Med.* 2017;90:146–54.
20. Hemphill E, Lindsay J, Lee C, Măndoiu II, Nelson CE. Feature selection and classifier performance on diverse bio- logical datasets. *BMC Bioinformatics.* 2014;15:1–14.
21. Neunhoeffler M, Sternberg S. How cross-validation can go wrong and what to do about it. *Political Anal.* 2019;27:101–6.
22. Goldberg DE, Holland JH. Genetic Algorithms and Machine Learning. *Mach Learn.* 1988;3:95–9.
23. Kloska A, et al. A Machine-Learning-Based Approach to Prediction of Biogeographic Ancestry within Europe. *Int J Mol Sci.* 2023;24:1–12.
24. Chen T, Guestrin C, XGBoost. A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17-Aug, 785–794* (2016).
25. Altinigneli C, Konten B, Rujescir D, Bohm C, Plant C. Identification of SNP interactions using data-parallel primitives on GPUs. *Proceedings – 2014 IEEE International Conference on Big Data, IEEE Big Data 2014 539–548* (2014) 10.1109/BigData.2014.7004271
26. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res.* 2020;48:D941–7.
27. Holtgrewe M, Beule D. VCFPy: a Python 3 library with good support for both reading and writing VCF. *J Open Source Softw.* 2016;1:85.
28. Pedregosa F. Scikit-learn: Machine Learning in Python. *J ofMachine Learn Res.* 2011;12:2825–30.
29. Elhaik E et al. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat Commun* 5, (2014).
30. Cheung EYY, Gahan ME, McNevin D. Prediction of biogeographical ancestry in admixed individuals. *Forensic Sci Int Genet.* 2018;36:104–11.
31. Nievergelt CM et al. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig Genet* 4, (2013).
32. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009;19:1655–64.
33. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000;155:945–59.
34. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet.* 2013;93:278–88.
35. Jin Y, Schaffer AA, Feolo M, Holmes JB, Kattman BL. GRAF-pop: A fast distance-based method to infer subject ancestry from multiple genotype datasets without principal components analysis. *G3: Genes, Genomes, Genetics* 9, 2447–2461 (2019).

36. Gaspar HA, Breen G. Probabilistic ancestry maps: A method to assess and visualize population substructures in genetics. BMC Bioinformatics. 2019;20:1–11.

37. Das R, Upadhyai P. Application of geographic population structure (GPS) algorithm for biogeographical analyses of populations with complex ancestries: A case study of South Asians from 1000 genomes project. BMC Genet. 18, (2017).

38. Garcia-Calleja J, Biagini SA, de Cid R, Calafell F, Bosch E. Inferring past demography and genetic adaptation in Spain using the GCAT cohort. Sci Rep. 2025;15:1–13.

Figures

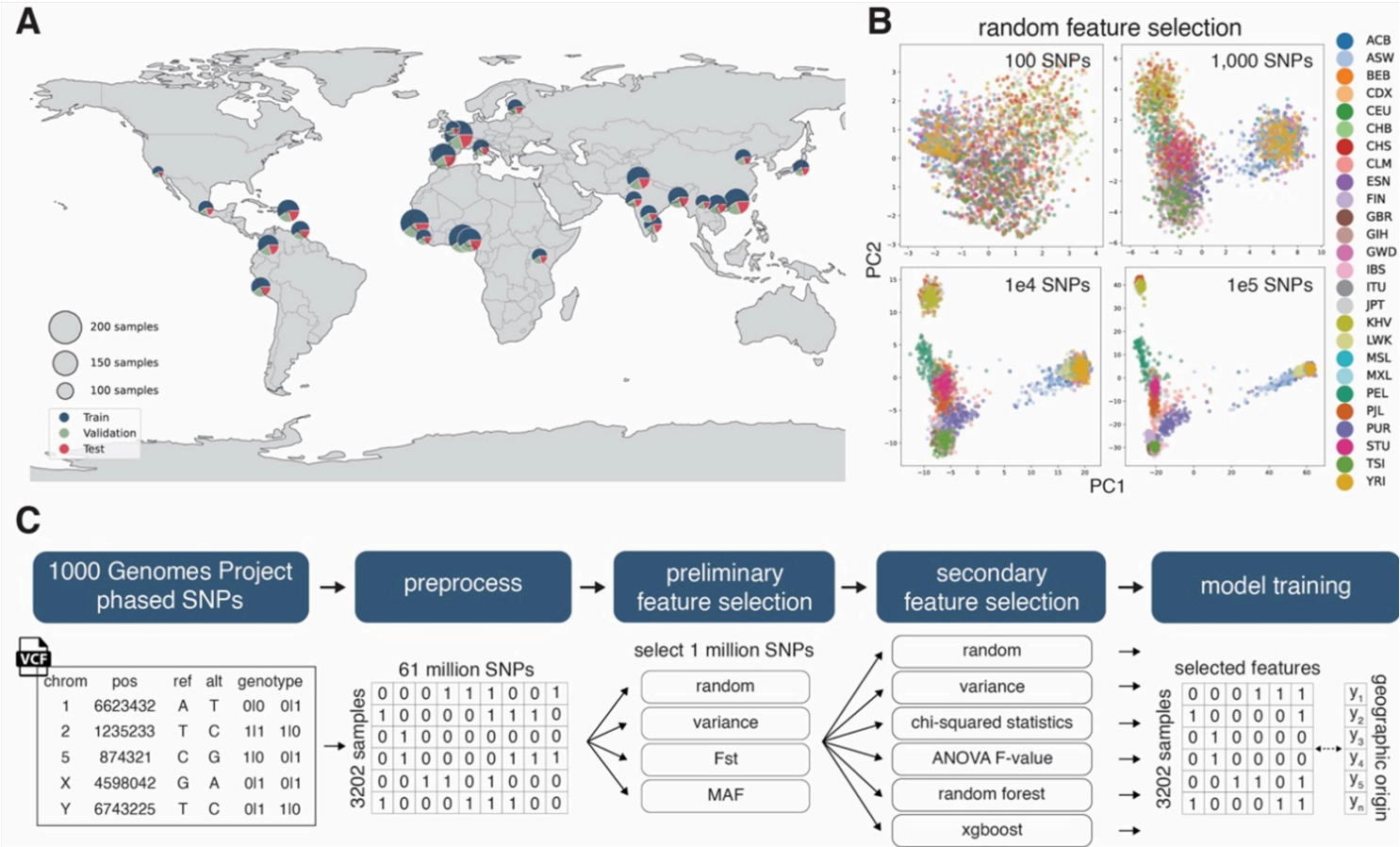


Figure 1

Efficient geographic origin prediction from genome-wide SNP data. (A) The 1000 Genomes Project data spans 3,202 individuals across 26 diverse geographic origins, with each origin represented proportionally by the size of the circles and the ratios of training, validation, and test sets shown in different colors. (B) Principal component analysis (PCA) of 100, 1,000, 10,000, and 100,000 randomly selected SNPs illustrates that feature selection of up to 100,000 SNPs does not achieve clear population separation, highlighting the necessity for more effective feature selection methods. (C) Schematic overview of the integrated feature selection process. The workflow includes data preprocessing, preliminary feature selection, secondary feature selection, and machine learning model training.

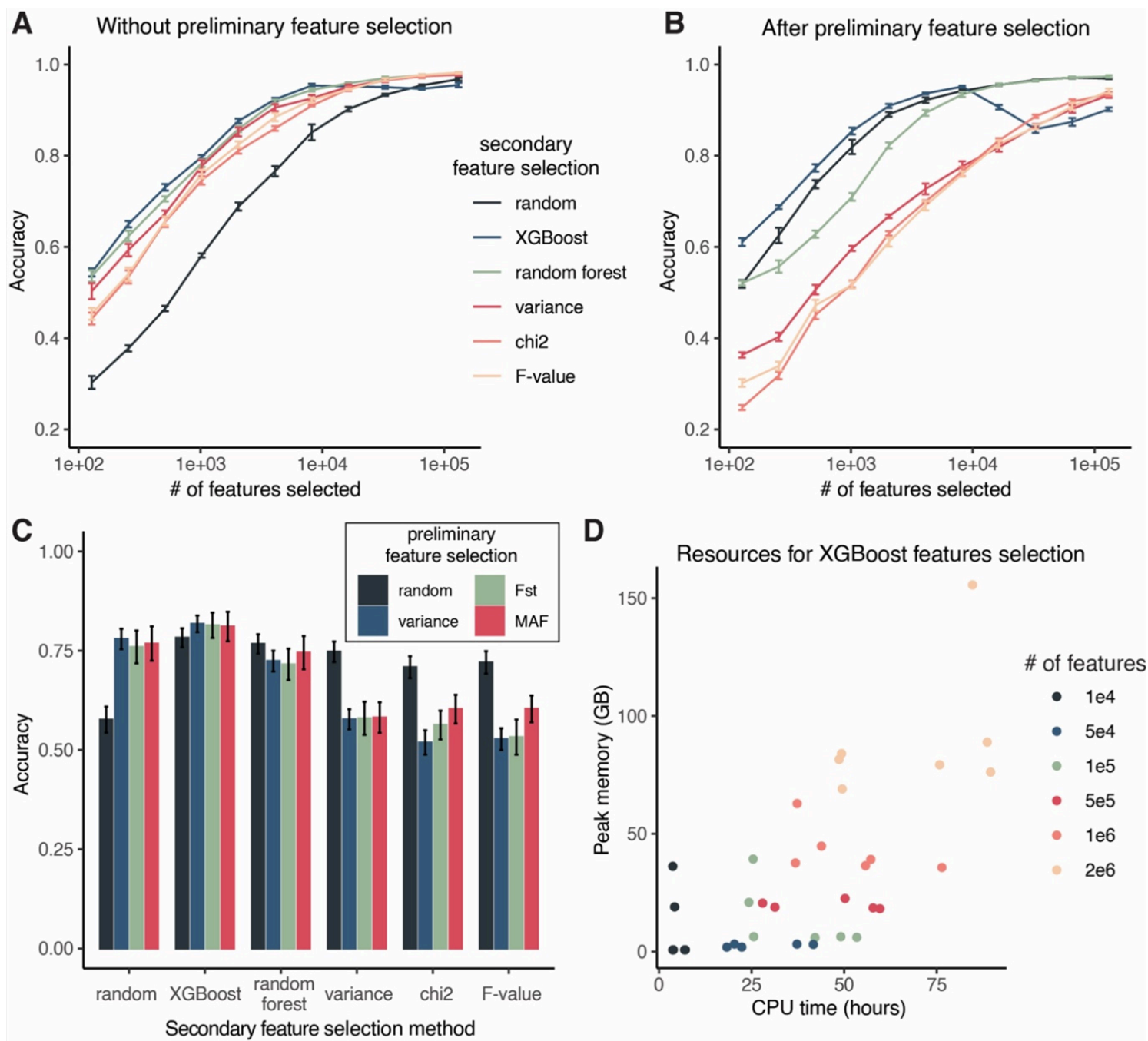


Figure 2

Integrated feature selection optimizes SNP selection efficiency. (A, B) Test set accuracy for different numbers of SNPs selected by six secondary feature selection methods: random, XGBoost, random forest, variance, chi-squared, and F-value. Results are shown for preliminary feature selection using random selection (A) and variance-based selection (B). Lines and error bars represent the mean $\pm$ s.e.m. ($n = 5$). (C) Overall test set accuracy for various secondary feature selection methods following preliminary feature selection based on random selection, variance, minor allele frequency (MAF) or FST. The results of selecting 128 to 8,192 numbers of SNPs in secondary feature selection were aggregated into a single plot. Bars and error bars represent the mean $\pm$ s.e.m. ($n = 5$). (D) CPU time and peak memory required for XGBoost-based feature selection for different numbers of input features

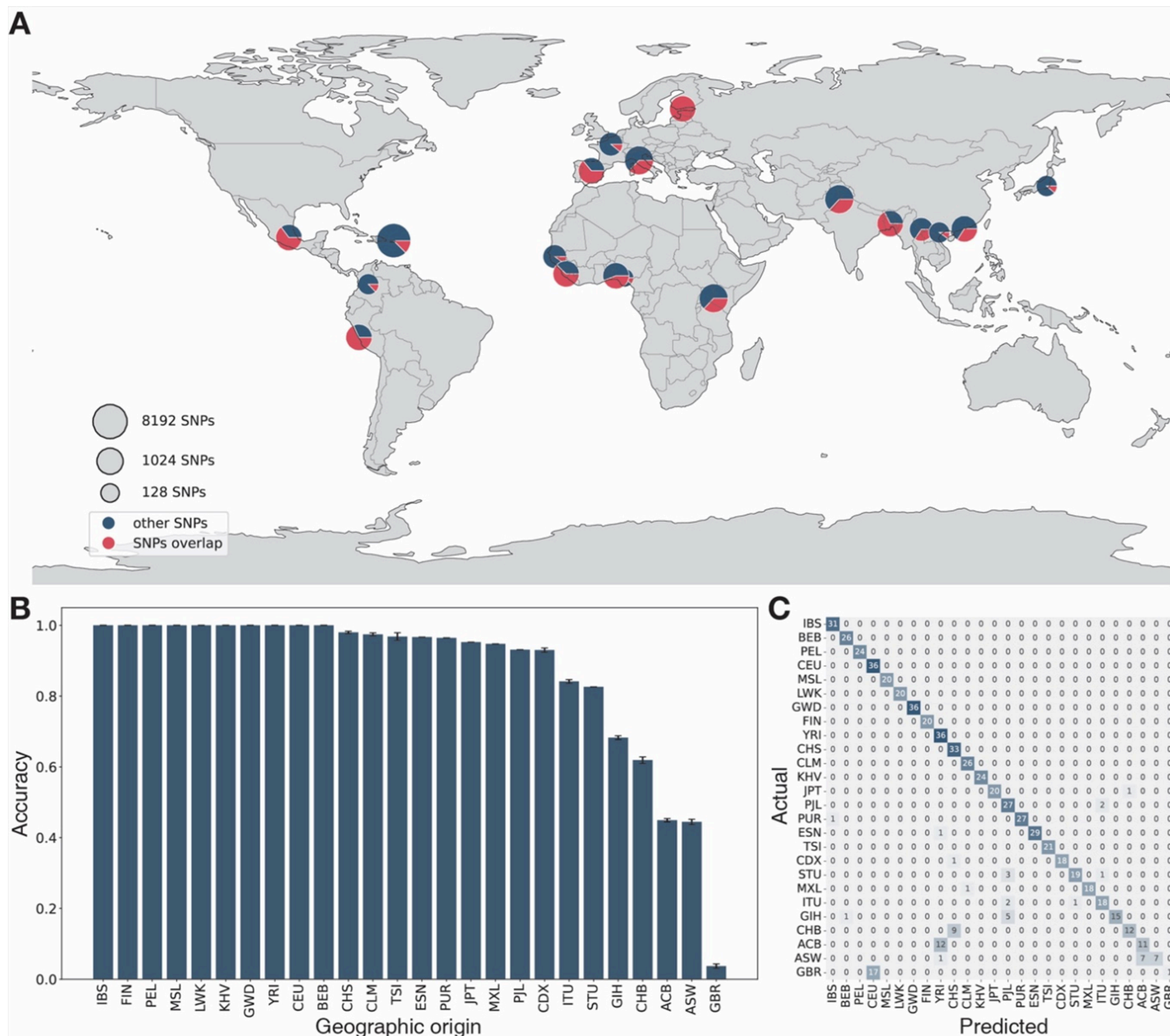


Figure 3

Geographic origin prediction using minimal SNPs. (A) The minimum number of SNPs required to differentiate each geographic origin is shown. The overlap of these SNPs between geographic origins is indicated in red, with Finland as the reference point, and the circle size represents the number of SNPs. (B) Prediction accuracy for 26 different geographic origins: African Caribbean (ACB), African Ancestry SW (ASW), Bengali (BEB), Dai Chinese (CDX), CEPH (CEU), Han Chinese (CHB), Southern Han Chinese (CHS), Colombian (CLM), Esan (ESN), Finnish (FIN), British (GBR), Gujarati (GIH), Gambian Mandinka (GWD), Iberian (IBS), Telugu (ITU), Japanese (JPT), Kinh Vietnamese (KHV), Luhya (LWK), Mende (MSL), Mexican Ancestry (MXL), Peruvian (PEL), Punjabi (PJI), Puerto Rican (PUR), Tamil (STU), Toscani (TSI), and Yoruba (YRI). Bars and error bars represent the mean $\pm$ s.e.m. ($n = 3$). (C) Confusion matrix of geographic origin prediction

