


LmTag: functional-enrichment and imputation-aware tag SNP selection for population-specific genotyping arrays

Dat Thanh Nguyen , Quan Hoang Nguyen, Nguyen Thuy Duong and Nam S. Vo

Corresponding author. Dat Thanh Nguyen, Center for Biomedical Informatics, Vingroup Big Data Institute, Hanoi, Vietnam.

E-mail: n.dat@outlook.com; Nam S. Vo, Center for Biomedical Informatics, Vingroup Big Data Institute, Hanoi, Vietnam; College of Engineering and Computer Science, VinUniversity, Vinhome Ocean Park, Hanoi, Vietnam. E-mail: v.namvs@vinbigdata.org

Abstract

Despite the rapid development of sequencing technology, single-nucleotide polymorphism (SNP) arrays are still the most cost-effective genotyping solutions for large-scale genomic research and applications. Recent years have witnessed the rapid development of numerous genotyping platforms of different sizes and designs, but population-specific platforms are still lacking, especially for those in developing countries. SNP arrays designed for these countries should be cost-effective (small size), yet incorporate key information needed to associate genotypes with traits. A key design principle for most current platforms is to improve genome-wide imputation so that more SNPs not included in the array (imputed SNPs) can be predicted. However, current tag SNP selection methods mostly focus on imputation accuracy and coverage, but not the functional content of the array. It is those functional SNPs that are most likely associated with traits. Here, we propose LmTag, a novel method for tag SNP selection that not only improves imputation performance but also prioritizes highly functional SNP markers. We apply LmTag on a wide range of populations using both public and in-house whole-genome sequencing databases. Our results show that LmTag improved both functional marker prioritization and genome-wide imputation accuracy compared to existing methods. This novel approach could contribute to the next generation genotyping arrays that provide excellent imputation capability as well as facilitate array-based functional genetic studies. Such arrays are particularly suitable for under-represented populations in developing countries or non-model species, where little genomics data are available while investment in genome sequencing or high-density SNP arrays is limited. LmTag is available at: <https://github.com/datngu/LmTag>.

Keywords: tag SNP selection, SNP array design, linkage disequilibrium, beam search

Introduction

Single-nucleotide polymorphism (SNP) arrays and recent technology whole-genome sequencing (WGS) have been widely used in genomic research and applications. Although WGS is attractive due to its ability to capture all genetic variation in the genome, SNP arrays have been the most widely used strategy due to several advantages such as cost-effectiveness, reliability of the technology and light computational requirement [1]. SNP arrays still play important roles in Genome-wide association studies (GWAS), which have facilitated the detection of DNA variants associated with human complex traits, including disease traits, leading to numerous proven and potential translational applications towards new diagnoses and therapeutics over the last decade [2].

However, due to the small number of SNPs that can be included, array-based genomic studies often require imputation to increase the number of variants for association tests by predicting the genotypes at the SNPs that are not directly genotyped in the study samples. The performance of imputation is affected by three main factors, including imputation algorithms [3], imputation reference panels [4, 5] and the design of SNP arrays [6].

Available genomic studies have focused mainly on European descent, accounting for approximately 79% of all GWAS participants, while the overall European population comprises about 16% of the total global population [7, 8]. Given that the majority of human functional genetic variants are population-specific and rare [9, 10], the imbalance in current population genetic

Dat Thanh Nguyen completed his MSc at the Tel Aviv University, Israel; and bioinformatics training at the Karolinska Institute, Sweden in 2020. His research interests include bioinformatics and machine learning with an emphasis on gene regulation and multi-omics data analytics.

Dr Quan Hoang Nguyen is a Senior Research Fellow and the head of Genomics and Machine Learning Lab at the Institute for Molecular Bioscience (IMB), The University of Queensland (UQ). He completed a PhD in Bioengineering at UQ in 2013, a postdoc at RIKEN in Japan in 2015, a CSIRO OCE Fellowship in 2016, an ARC DECRA fellow in 2021, and is currently an NHMRC leadership fellow. With spatial technology approaches, he investigates cell types, their spatial organisation and cell-cell interactions underlying differential responses to treatment and risks of cancer metastasis, revealing potentially new gene targets to modulate cancer-immune cell interactions.

Nguyen Thuy Duong has been working on human diseases and identified many pathogenic mutations which were partly used as genetic markers for prenatal diagnosis over the last 15 years. In the past five years, NTD is interested in studying human population genetics of Vietnamese populations.

Nam S. Vo is Director of Center for Biomedical Informatics at Vingroup Big Data Institute and an Affiliate Faculty at College of Engineering and Computer Science at VinUniversity, Vietnam. His current research interests focus on analysis and interpretation of large-scale multi-omics data towards understanding disease risk and adverse drug reaction

Received: March 6, 2022. **Revised:** May 2, 2022. **Accepted:** May 31, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

data resources implies a critical problem. Important variants with low frequencies or completely absent in European populations may be missed by GWAS discoveries so far [11]. Consequently, disease risk predictions, which benefit the clinical arena, are currently restricted in the European ancestry population [12]. This is a critical issue, especially for the majority of the world population, who are under-represented in genomic studies. These under-represented populations include both minority ethnic groups in high-income countries, and citizens of low and middle-income countries [13]. This fact leads to an urgent and unmet demand to develop and use customized genotyping platforms for under represented populations [1]. Indeed, population-specific genotyping arrays such as the UK Biobank Axiom Array [14], the Axiom-NL Array [15], the TWB Array [16], the Axiom China Kadoorie Biobank Array [17], the Japonica and Japonica NEO Arrays [18, 19], and the Axiom KoreanChip [20] have been successfully implemented to facilitate genomic studies in these populations.

To develop such arrays, various strategies to select tag SNPs are employed. A tag SNP is a SNP that can represent a group of SNPs called a haplotype due to strong associations between these neighboring alleles (known as linkage disequilibrium, LD). Tag SNP selection methods can be classified into two main categories including block-based [21–23], and LD-based approaches [24–27]. The former approach involves partitioning the whole chromosome into blocks, often relying on a predefined haplotype block structures or simply based on genomic distance. For example, in the early generation of human genotyping SNP array, tag SNPs were selected at intervals of approximately each 5-kilobase with a minor allele frequency of at least 5% [28]. This strategy has also been widely adopted in animal genetics, commonly referred to as the equidistance method [29, 30]. On the other hand, the latter approach utilizes LDs among nearby SNPs to find tag SNPs with a greedy approach to maximize LD coverage [19, 27, 31]. A typical algorithm starts with a set of targeted SNPs, then weighs each SNP candidate by the number of neighbor SNPs (within a specific genomic distance) that have pairwise LD r^2 greater than or equal to a specific threshold, e.g., 0.8. The SNP with the highest score is then selected, and the associated SNPs are removed from the targeted set. These steps are iterated until reaching the desired number of tag SNPs or no more SNP satisfying the LD r^2 threshold [24, 31]. In addition, multi-marker LD approach [25, 32, 33], pairwise LD hybrid tag SNP selection [26], cross-population prioritizing scheme [27] also aim to improve LD coverage and imputation accuracy. Despite the efforts, these strategies still have certain limitations. Firstly, it is unclear that tag SNP selection approaches to maximize LD coverage or genomic distance can provide the best imputation accuracy performance, which is the golden standard of SNP array assessment nowadays [6, 27]. Secondly, SNPs on genotyping arrays are typically not causal variants because they are chosen to be highly LD correlated with neighboring SNPs to cover large genomic regions to allow

for imputing unmeasured SNPs, a common design practice in the greedy paradigm [34].

To address these challenges, we introduce a novel method called LmTag, which facilitates the design of functional-enrichment, imputation-aware, and population-specific SNP arrays. Firstly, LmTag uses a robust statistical modeling to systematically integrate LD information, minor allele frequency (MAF) and physical distance of SNPs into the imputation accuracy score to improve tagging efficiency. Secondly, LmTag adapts the beam search framework [35] to prioritize both imputation scores and functional scores to solve the tag SNP selection problem. We apply LmTag and comprehensively compare it with common approaches of tag SNP selection using a wide range of both public and in-house genomics datasets. Our benchmarking results suggest that LmTag improves both imputation performance and prioritization of functional variants. Furthermore, we show that tagging efficiency of tag SNP sets selected by LmTag are sustainability higher than existing genotyping arrays, indicating the potential improvements for future genotyping platforms.

Materials and methods

Overview of LmTag pipeline

An overview of LmTag is presented in Figure 1. The method includes three key steps: (i) Imputation accuracy modeling, (ii) Functional scoring and (iii) Functional tag SNP selection. In the first step, a theoretical array (set of tag SNPs) is simulated, and imputation accuracy scores of the corresponding tagged SNPs are estimated by leave-one-out cross-validation (details in the next section). A linear model is then employed to assess imputation accuracy scores of tagged SNPs based on pairwise LD r^2 , MAF of tag SNPs (those included in the array), MAF of tagged SNPs (not included in the array), and distances between tag SNPs and tagged SNPs. In the second step, SNPs are functionally scored based on public databases including the GWAS catalog [36], the ClinVar [37] and the Combined Annotation-Dependent Depletion (CADD) [38] to enrich functional variants in the array design. Finally, parameters from the model are used to estimate imputation accuracy score for each SNP. These estimated scores, together with the functional ranking of SNPs, are then used in functional-enrichment tag SNP selection by the beam search algorithm with beam width parameter K [35]. Further details are described in the next sections.

Imputation accuracy modeling

Our aim is to combine systematically information from both pairwise LD r^2 , MAF and genomic distance to improve imputation accuracy of tag SNP selection. To this end, we model imputation accuracy as a linear model:

$$r = \beta_0 + \beta_1 \cdot l + \beta_2 \cdot m_{tag} + \beta_3 \cdot m_{tagged} + \beta_4 \cdot d, \quad (1)$$

where r is imputation r^2 (described later), l is LD r^2 between tag SNP and tagged SNP, ($l \in (0 : 1]$), m_{tag} is MAF of tag SNP, ($m_{tag} \in (0 : 0.5]$), m_{tagged} is MAF of tagged SNP

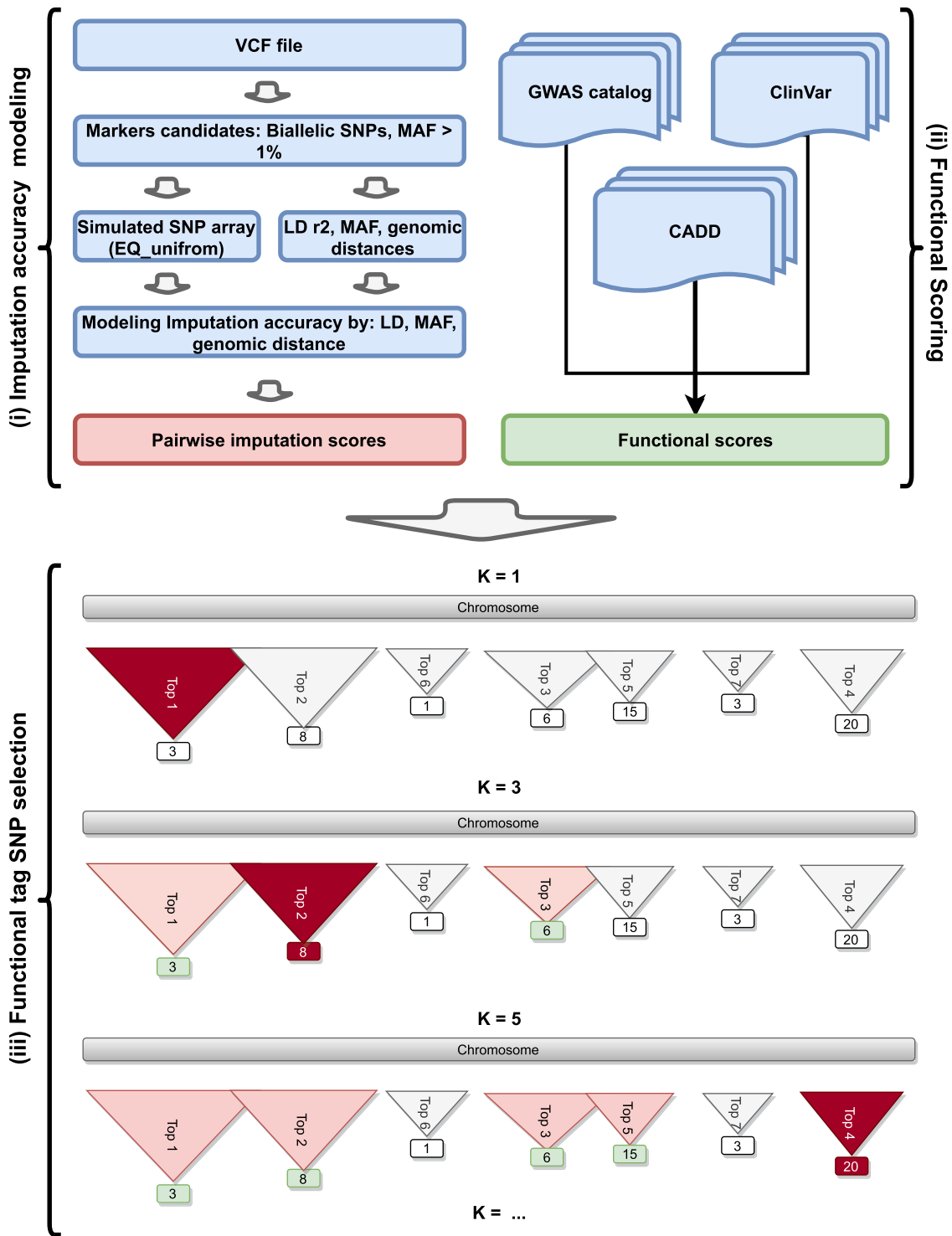


Figure 1. Overview of LmTag. (i) Imputation accuracy modeling, this includes modeling imputation accuracy metric as a function of LDs, MAFs and genomic distances. (ii) Functional scoring, this includes steps of weighting functional scores of SNPs based on public databases. (iii) Functional tag SNP selection, imputation capability of each SNP is represented as triangles while functional scores are showed in the lower rectangles. When $K = 1$, the beam search algorithm becomes the best-first search that select SNP with the highest estimated imputation performance - colored bold red triangles. When $K > 1$, the algorithm selects top K SNPs with the highest estimated imputation performances - colored light pink triangles, the functional scores in these SNPs - colored light green is weighted to find the highest functional SNPs as tag SNPs - colored bold red triangles.

($m_{tagged} \in (0 : 0.5]$), d is genomic distance between tag SNP and tagged SNP, ($d \in N$).

In this model, untyped SNPs are assumed to be tagged by the highest LD SNP in the tag SNP set. The relationships among pairwise LD r^2 , MAF and genomic

distance are established by simulation. In detail, a theoretical naive SNP array is created followed by imputation accuracy scores computation for corresponding tagged SNPs. The corresponding information including LD and genomic distance are then extracted before being

used to estimate parameters for the linear model. It is noted that more complicated models may provide better performance. However, these improvements are marginal. We opt to use the linear model due to its efficiency, simplicity and explainability, and set it as the default setting. Further details of model selection and optimization can be found in the supplementary document.

Because LmTag employs the greedy framework in tag SNP selection, the number of tag SNPs selected by LmTag is expected to be comparable with the number of tag SNP selected by standard greedy algorithms such as TagIt when they become saturated. Theoretically, the model parameters are sensitive with the size of the simulated array. Thus, we run a standard greedy tag SNP selection algorithm, TagIt (<https://github.com/statgen/TagIt>) [31], with default parameters (LD r^2 threshold is 0.8, and MAF threshold is 0.01) to estimate scaffold sizes k for each chromosome to make the simulation as realistic as possible. We denote the input containing n SNPs as $A = \{SNP_1, SNP_2, SNP_3, \dots, SNP_n\}$. We then sort them by their genomic positions and uniformly sub-sample k SNPs as a tag SNP set $T = \{SNP_1, SNP_2, \dots, SNP_k\}$. The remaining $n - k$ SNPs are labeled as a tagged SNP set $G = \{SNP_{k+1}, SNP_{k+2}, \dots, SNP_n\}$. Imputation accuracy scores for tagged SNPs $\in G$ are computed with a leave-one-out internal validation approach [6, 27]. Specifically, imputation is performed individually for each sample with the exclusion of itself from the reference panel with Minimac4 v1.0.2 [3]. Tag SNPs $\in T$ are denoted as ‘genotyped’ and the sites $\in G$ are set as missing. The imputation accuracy r_i for each tagged SNP $i \in G$ is represented by the concordance rate, e.g., squared Pearson’s correlation coefficient which we term imputation r^2 to make a distinction from LD r^2 , between imputed genotype dosages in (0-2) and masked ground truth genotypes in (0, 1, 2).

Pairwise LDs are calculated using Plink v1.9 within a maximum genomic distance of 1 megabase (MB), and minimum LD r^2 cutoff of 0.2 [39]. Allele frequencies are computed and extracted with bcftools v1.10.2 (<https://github.com/samtools/bcftools>). To simplify the linear model, we assume that each tagged SNP’s genotype is inferred based on the sole tag SNP that has the highest LD r^2 . Thus, we find the best tag SNP $i \in T$ for each SNP $j \in G$ that has the most LD with the targeted tag SNP j to extract relevant information including LD pairwise l_{ij} , MAF m_i, m_j , and genomic distance d_{ij} . Together with imputation scores r_i estimated from the previous step, these data are then used to estimate parameters for the linear model (1).

SNP prioritization with high-functional scores

In general, markers are functionally ranked based on biological evidence and genome-wide predicted functional scores simultaneously. In the current implementation, SNP positions matched to the GWAS catalog and the ClinVar databases [36, 37] are functionally ranked as in the highest score category. For non-biological evidence SNPs,

we use CADD scores [38] to prioritize functional SNPs to make sure all SNPs are functionally scored. The CADD scoring system is a widely used metric that effectively prioritizes causal variants in genetic analyses, especially in highly penetrant contributors to severe Mendelian disorders. CADD integrates more than 60 genomic features based on DNA sequence, for examples gene model annotations, evolutionary constraint, epigenetic measurements and functional predictors into a single score by a machine learning model. In addition to the comprehensive use of genomic features, two other key advantages of the CADD model include the genome-wide estimation and the interpretability for each estimate. CADD scores are computed for all approximately 9 billion possible single-nucleotide variants (SNV) across the human genome. For interpretability, the scores are transformed into ‘PHRED-scaled’ to provide a relative ranking system between SNVs at genome-wide coverage. Regardless of the details of the annotation set and model parameters, CADD scores can be interpreted simply as follows: a scaled score of 10 or greater equivalent to a raw score in the top 10% of all possible reference genome SNVs, and a score of 20 or greater indicates a raw score in the top 1%, and so on [40].

Functional tag SNP selection

Similar to most LD based tag SNP selection methods [24–26, 31, 41], we employ a greedy approach for computational efficiency. However, there are two main differences in our algorithm. Firstly, we use estimated pairwise imputation r^2 scores for ranking SNP candidates instead of using pairwise LD r^2 like conventional methods. Specifically, for each pair of SNPs, imputation score r^2 for each SNP is estimated independently by using coefficients derived from the established linear model and the corresponding LD r^2 , its MAF, mate’s MAF, and genomic distance between the two SNPs. Given two SNPs, SNP_i , and SNP_j with LD r^2 (SNP_i, SNP_j) = l_{ij} , MAF $SNP_i = m_i$, MAF $SNP_j = m_j$, and genomic distance (SNP_i, SNP_j) = d_{ij} . Their estimated imputation scores \hat{r}_i , and \hat{r}_j are calculated as:

- $\hat{r}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot l_{ij} + \hat{\beta}_2 \cdot m_i + \hat{\beta}_3 \cdot m_j + \hat{\beta}_4 \cdot d_{ij}$
- $\hat{r}_j = \hat{\beta}_0 + \hat{\beta}_1 \cdot l_{ij} + \hat{\beta}_2 \cdot m_j + \hat{\beta}_3 \cdot m_i + \hat{\beta}_4 \cdot d_{ij}$

where $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, and $\hat{\beta}_4$ are estimated from the linear model (1). Secondly, LmTag employs the beam search [35] instead of the best-first search strategy like other algorithms. The main advantages of the beam search is allowing us to prioritize highly functional SNPs. In details, we introduce a tuning parameter K in the algorithm to select tag SNPs with high functional scores. LmTag algorithm starts with an empty tag SNP set T , a tagged SNP set G , and n input SNP candidates $A = \{SNP_1, SNP_2, SNP_3, \dots, SNP_n\}$. For each iteration, the algorithm includes two main steps as follows:

1. Imputation scoring.

Each $SNP_i \in A$ is scored as s_i , which is sum of estimated imputation $r^2 \hat{r}_j$ of all its neighboring $SNP_j \in A$

given that pairwise LD $r^2 l_{ij}$ is equal to or greater than a specific cut-off c :

$$s_i = \sum_{j=1}^n \hat{r}_j; \text{ If } l_{ij} \geq c, \text{ and } j \neq i. \quad (2)$$

2. Tag SNP selection with beam search.

Our approach considers the functional term of each marker in tag SNP selection by adapting the beam search algorithm [35]. In brief, beam search is a heuristic searching algorithm used to solve combinatorial optimization problems. This approach employs a truncated branch-and-bound searching procedure, where only the most promising K nodes (instead of all nodes) at each level of the search tree are evaluated and retained for further branching; K is the so-called beam width [42]. We consider top K SNPs with highest imputation scores as a candidate list of tag SNPs. Then, the search branching is extended to functional scores, i.e. the SNP with the highest functional score in this list is subsequently chosen as a tag SNP_{*t*}. This SNP is subsequently moved from the candidate set A into tag SNP set T , and SNP_{*t*}'s neighboring SNPs (satisfying pairwise LD r^2 cut-off) are moved into tagged SNP set G . Overall, both the selected tag SNP and its associated tagged SNPs are removed from the candidate set A .

These steps are iterated until either A is empty or no pair $(\text{SNP}_i, \text{SNP}_j) \in A$ satisfying the condition $l_{ij} \geq c$ could be found. Finally, the tag SNP set A , and their associated tagged SNP set G are exported.

Datasets

We evaluate the performance of LmTag in both in-house generated and public datasets, including data from the 1000 Vietnamese Genomes Project (1KVG) pilot phase and data of three super populations from the 1000 Genomes Project samples re-sequenced by New York Genome Center (1KGP-NYGC) [43]. The genomic data of the 1KVG pilot phase were obtained from 504 unrelated Vietnamese population individuals (VNP), including 208 males and 296 females. Their genomes were sequenced at coverage 30x with 150bp paired-end reads using an Illumina NovaSeq 6000 system. Variant calling was performed using the DRAGEN pipeline [44] with the GRCh38 patch release 13 reference genome [45]. Quality check and filtering were performed with bcftools v1.10.2, and phasing was performed with SHAPEIT v4.1.3 to obtain the phased genotypes in Variant Call Format (VCF) [46]. Phased genotype data in VCF format of 1KGP NYGC high coverage are obtained from The International Genome Sample Resource (IGSR) data portal. We include only unrelated samples belonging to East Asian (EAS), European (EUR), and South Asian (SAS) in the analysis. These samples are assigned to their super population according to IGSR's annotation. All genomic data are

reprocessed with bcftools v1.10.2 to keep only biallelic SNP with MAF > 1%.

CADD v1.6 database [40], release version 2021-07-08 of the GWAS catalog [36], and the ClinVar database [37] are downloaded and filtered to obtain functional scores for each population. Finally, processed genomic data of four populations and their associated functional annotations are used in our analysis, including VNP, EAS, EUR and SAS, which comprise 504, 504, 503 and 489 individuals respectively. Details of datasets can be found in Table 1. Due to limited computational resources, our analyses are performed on chromosome 10, but the results should be generalizable to all chromosomes.

Performance evaluation

We compare LmTag against commonly used methods in SNP array design including TagIt [31], FastTagger [25], EQ_uniform (uniform tag SNP selection based on genomic distances) [29] and EQ_MAF (MAF-optimized tag SNP selection based on genomic distances) [30] using various metrics including imputation accuracy and functional enrichment. We also compare imputation accuracies of tag SNPs selected by LmTag against those of tag SNP sets from various commercial genotyping arrays. By this way, we explore potential applications of LmTag in designing genotyping arrays.

In terms of current methods for genotyping array design, the first two methods optimize imputation accuracy by maximizing LD while the later methods select makers based on the equidistant principle. The distance-based methods are widely used in animal SNP array designs that involve dividing chromosomes into certain intervals with equal genomic length [29, 47, 48] and further optimized toward MAF [30, 49, 50]. For each interval, the SNP with the highest MAF is selected as representative of all SNPs in the interval [30]. TagIt is a typical greedy algorithm selecting tag SNPs based on pairwise LD information widely used in human SNP array designs [19, 27, 31]. Meanwhile, FastTagger is a fast implementation of the multi-marker LD approach, which reduces the number of tag SNPs selected while still maintaining high genomic coverage. In brief, the multi-marker LD approach methods find association rules of one SNP with multiple SNPs, termed multi-marker r^2 statistics, and use this information to find tag SNPs [25, 33, 51]. Details on comparing these methods can be found in previous reports [52].

Evaluation metrics are based on imputation accuracy and functional prioritizing. Imputation accuracy is measured as squared Pearson's correlation of imputed dosages estimated through a leave-one-out internal validation and the 'true genotypes.' In details, selected tag SNPs are denoted as 'genotyped,' and other sites are set as missing. For each SNP, squared Pearson's correlation is calculated from imputation 'estimated dosages' (0–2) to the 'true genotypes' (0,1,2) in the original VCF file [6, 26, 27]. An overall imputation value is defined as mean imputation r^2 of all markers in the population. Functional prioritizing is evaluated based

Table 1. Datasets are used in this study

Populations	Number of samples	Total markers	GWAS markers	ClinVar markers
VNP	504	382 700	5064	1590
EAS	504	405 234	5160	1617
SAS	489	486 024	5868	1876
EUR	503	456 166	6168	1814

on CADD scores and their corresponding percentiles among all SNPs, and the relative proportion of GWAS and ClinVar markers which is defined by the number of GWAS and ClinVar markers in the tag SNP sets over their corresponding number in the examined populations. These parameters are defined as follow:

$$P = (1 - 10^{-\frac{Q}{10}}) \times 100, \quad (3)$$

$$p_g = \frac{n_g}{N_g} \times 100, \quad (4)$$

and

$$p_c = \frac{n_c}{N_c} \times 100; \quad (5)$$

where P is percentile ranking of CADD score; Q is its original scores in 'PHRED-scaled'; N_g and N_c are total GWAS and ClinVar markers in each populations; n_g , n_c are number of GWAS and ClinVar markers in selected tag SNP sets; and p_g , p_c are their corresponding proportions.

For comparison between methods, the LD cutoff is set at 0.8 in LD-based methods, including LmTag, TagIt and FastTagger. FastTagger requires further LD settings for `min_r2_2`, and `min_r2_3` that are set as 0.9, and 0.95 respectively, as recommended by the authors. LmTag is further ran with several K values varying from 1 to 2000 to examine the relationship between imputation accuracy and functional SNP inclusion. Functional scores of selected tag SNPs by the other tag SNP selection methods are also computed for comparison. To enable a fair and comprehensive evaluation, tag SNPs are selected corresponding to multiple cutoffs ranging from 8000 to 32 000 in all populations.

Results

LmTag improves functional enrichment in tag SNP selection

The summary results of functional enrichment in tag SNP selection of LmTag, EQ_uniform, EQ_MAF, TagIt, FastTagger and baseline (mean functional score and proportion of biological evidenced markers in all SNPs in the population) are shown in Figure 2, 3, and Table S.2, S.3. LmTag is evaluated with various beam width parameters $K=1, 10, 20, 30, 50, 100, 200, 500, 1000, 1500$ and 2000 denoted as LmTag_K1, LmTag_K10, LmTag_K30, LmTag_K50, LmTag_K100, LmTag_K200, LmTag_K500, LmTag_K1000, LmTag_K1500 and LmTag_K2000, respectively. The results are collected from all four populations EAS, EUR, SAS and VNP under the 32 000 tag SNPs setting.

In general, LmTag shows a significant improvement in functional prioritization with almost zero imputation performance trade-off. Particularly, in comparison with the baseline and other methods, LmTag obtains significant improvements with approximately 2-folds (at $K=200$), and 3-folds (at $K = 2000$) in terms of selection GWAS and ClinVar markers; and yet increases averagely 15–17%, and 23–27% CADD score percentile ranking in term of selection population-wide variants as tag SNPs with K setting at 200, and 2000 respectively.

When K is set as 1, LmTag becomes a standard greedy algorithm with the 'best-first' search approach, i.e. no optimization is applied for selecting functional variants. In this setting, mean CADD scores, mean CADD percentiles, proportions of GWAS and ClinVar markers selected by LmTag are comparable with the baseline and other methods, as expected. The mean CADD scores of tag SNPs selected by LmTag_K1 vary from 2.92 to 2.96 across examined populations, and are in the same range with the baseline, which varies from 2.91 to 2.96. Other methods also yield comparable performances with LmTag_K1, ranging from 2.89 to 3.04. Conversion from 'PHRED-scaled' score into percentile scale shows mean CADD score percentile of LmTag_K1 and the others are equivalent with the rank from 37.97 to 39.34. In other words, under the setting of no optimization for functional SNPs, CADD scores / percentiles of tag SNP distribute equivalently regardless of the method of choice. Similarly, when considering prioritization of markers using biological evidence databases, the proportions of GWAS and ClinVar marker selected by LmTag_K1, and other methods are mostly comparable to the baseline except for GWAS marker proportions of EQ_MAF. Under the baseline scenario, the expected proportions of GWAS and ClinVar in 32 000 tag SNPs are 7.90%, 7.01%, 6.58% and 8.36% in EAS, EUR, SAS and VNP, respectively. The corresponding ranges for LmTag_K1, EQ_uniform, TagIt and FastTagger are 7.68–9.34%, 6.58–8.36%, 9.34–10.50%, respectively. Notably, the EQ_MAF method selects slightly higher proportions of ClinVar markers, from 8.74 to 11.45%, and significantly more GWAS markers ranging from 15.26 to 16.17% that are possibly explained by the detection power bias towards high-frequency variants in both clinical and association studies.

When the value of K increases, as expected, a clear improvement of functional enrichment is shown as detailed in shown in Figure 2, 3, and Table S.2, S.3. Consistently, CADD scores and proportions of GWAS and ClinVar show a strong positive correlation with the increase of K , while the overall imputation accuracies

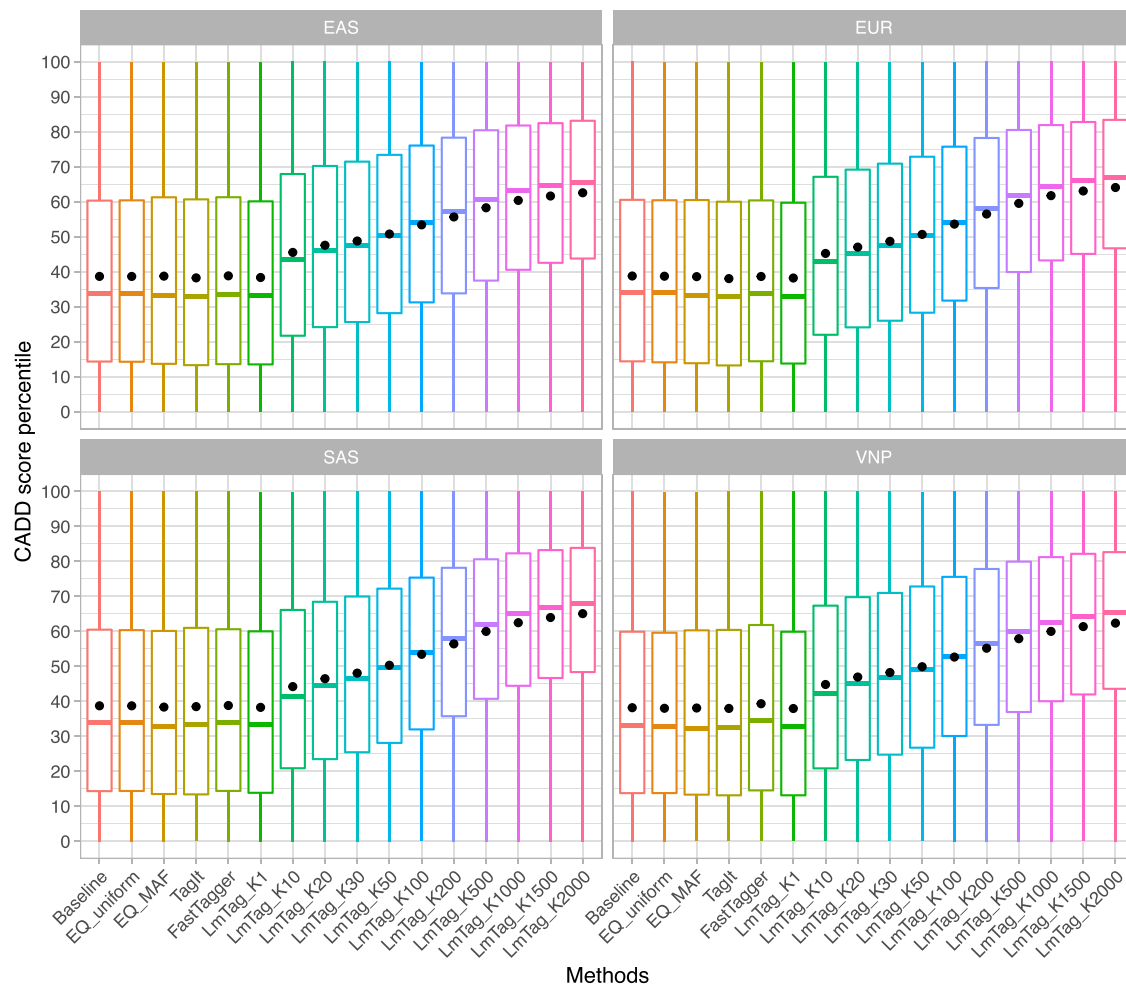


Figure 2. Mean percentile of CADD scores of tag SNP selected by LmTag (with $K=1, 10, 20, 30, 50, 100, 200, 500, 1000, 1500$ and 2000), EQ_uniform, EQ_MAF, TagIt and FastTagger. Baseline shows mean percentile of CADD scores of all input markers (32 000 SNPs) in each population.

experience very small changes as shown in Table S.4. For example, when K setting changes from 1 to 200, the overall imputation accuracy of the VNP reduces by only 0.03% (from 89.83% to 89.80%) while the functional scores of tag SNPs gain significant improvements. Mean CADD score percentile increases by 18% (55.20% at $K=200$ versus 37.20% at $K=1$). In the meantime, the GWAS and ClinVar proportions covered by 32 000 tag SNP increase more than 2-folds, both from 8.36% to 18.13%, and 18.05%, respectively. It is noted that, mean CADD score percentile values are computed by taking the average percentile ranks of all selected tag SNPs and not by directly converting from the mean of CADD ‘PHRED-scaled’ scores. When the K value is set extremely high at 2000, the functional score improvements are continued with lower rates to 62.37, 25.55 and 19.78% for mean CADD score percentile, GWAS and ClinVar proportions respectively. However, there are trade-offs in terms of imputation accuracy and computational time. In comparison to $K=200$, the mean imputation accuracies reduce 0.05–0.09% and computational cost increase linearly by 10 times in responses to the search spaces (0.6–1.17 hours when $K=200$, and 5.92–10.51 hours

when $K=2000$) in the four examined populations. Details of computation time and overall imputation accuracy changes of all populations are shown in Table S.4, S.9 and visualized in Figure S.3. Taken all together, we recommend the compromised setting of K in the range of 200–500 to obtain optimization for both imputation accuracy, functional scores and computational time.

LmTag demonstrates superior tagging efficiency

Regarding imputation performance, LmTag outperforms other methods in both imputation accuracy and imputation coverage. The K parameter used in this comparison is 200, while the number of tag SNPs is set at various cutoffs. Regarding imputation accuracy, LmTag is the top performer, followed by TagIt, and EQ_uniform while the worst performers are interchanged between EQ_MAF and FastTagger depending on population as reported in Table S.5, and shown in Figure 4. At the cutoff of 32 000, performance differences are substantially large between LmTag against EQ_uniform, EQ_MAF, and FastTagger but smaller against TagIt. For example, in the EAS population, LmTag obtains 87.19% overall imputation accuracy compared with 86.29%, 82.51%, 82.33%, and, 78.10% achieved

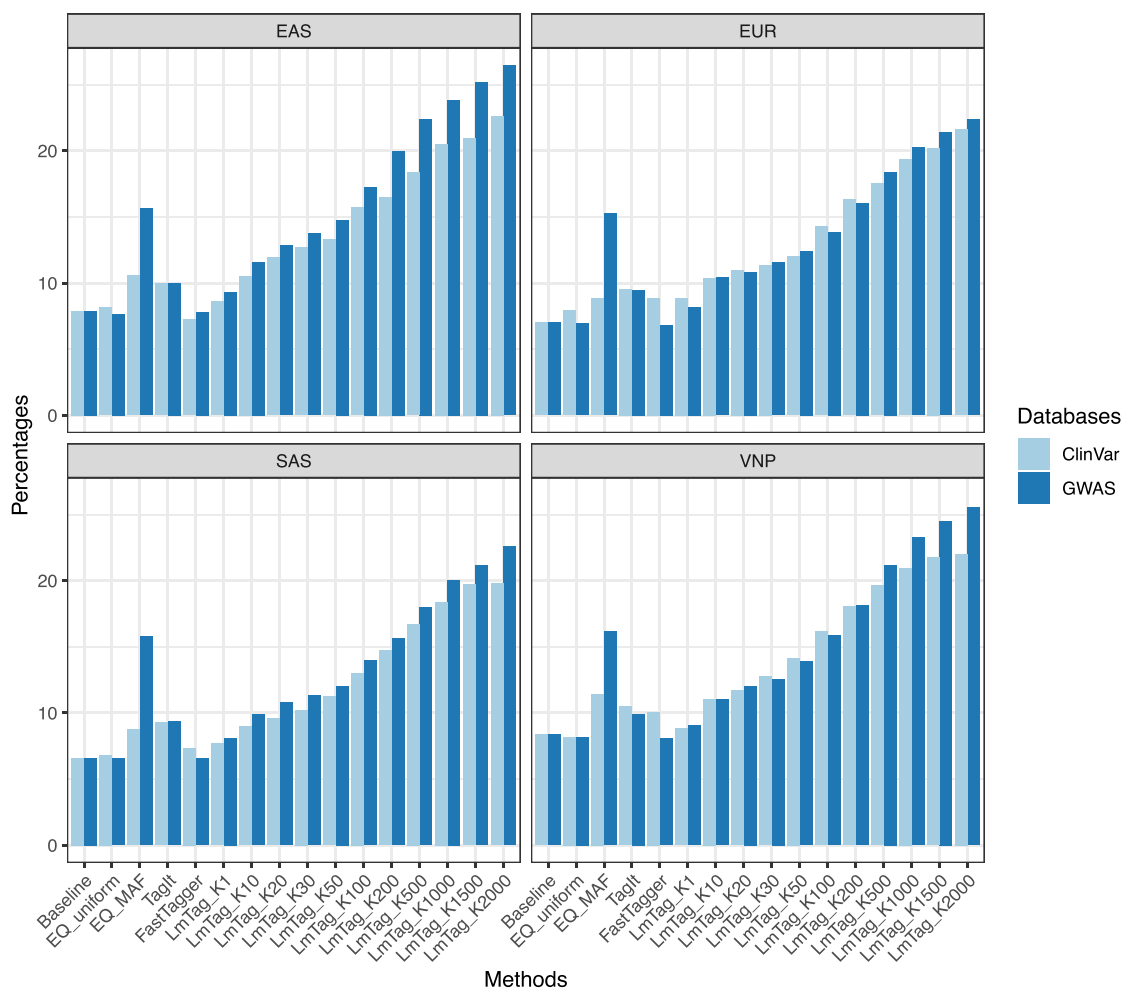


Figure 3. Percentages of GWAS and ClinVar makers covered by 32 000 tag SNPs selected by LmTag (with K=1, 10, 20, 30, 50, 100, 200, 500, 1000, 1500 and 2000), EQ_uniform, EQ_MAF, TagIt and FastTagger over the total number of GWAS and ClinVar makers in each population. Baseline shows percentages of GWAS and ClinVar markers covered over the total number of GWAS and ClinVar makers in each population corresponding 32000 tag SNP scaffold.

by TagIt, EQ_uniform, FastTagger and EQ_MAF, respectively. The same trend is also observed in EUR, SAS and VNP with 88.50%, 86.50% and 89.80% imputation accuracies achieved by LmTag_200. In terms of imputation coverage, LmTag also produces the highest performance as showed in Figure S.1. Taking imputation r^2 threshold of 80% as an example, LmTag yields the imputation coverage of 83.65%, 85.25%, 81.66% and 87.81% in EAS, EUR, SAS and VNP while the second-ranked performer obtains 82.11%, 84.08%, 80.13% and 87.04% respectively.

To examine potential effects of the number of selected tag SNPs on imputation accuracy and imputation coverage, we further evaluate overall imputation accuracy across different scaffolds by selecting top-ranked SNPs from each population with various cutoffs: 32 000, 28 000, 24 000, 20 000, 16 000, 12 000 and 8000. Details of overall imputation accuracies are reported in Table S.5. We observe that the imputation accuracy and imputation coverage increase in response to the increased number of tag SNPs selected. However, the relationship is not linear as shown in Figure 4 and Figure S.1. Nevertheless, LmTag consistently outperforms other methods across all settings. In general, the increasing rates of imputation

accuracy and imputation coverage are lower when the numbers of tag SNP is high. In other words, when the scaffolds of the SNP array contain a large enough number of SNPs, adding more tag SNPs do not significantly improve imputation accuracy and imputation coverage compared to those with small scaffolds. For example, adding 4000 tag SNPs at 12 000 tag SNPs scaffold yield approximately 8% improvement in imputation accuracy compared to the scaffolds of 8000 SNPs regardless of the method of choice. Meanwhile, increasing 4000 SNPs to the scaffold of 28 000 results in less than 2% improvement in imputation accuracy. Interestingly, we observe that imputation coverages of all methods dramatically change in response to number of tag SNPs. For example, LmTagK_200 obtains more than 80% coverage with imputation cutoff at 80% at 32 000 tag SNP. The coverage reduces significantly to 50–60% when number of tag SNPs is 8000, and even lower for EQ_MAF to 18–25%.

LmTag helps improve current genotyping arrays

To further explore potential applications of LmTag in designing genotyping arrays. We also compare imputation performances of tag SNPs selected by LmTag

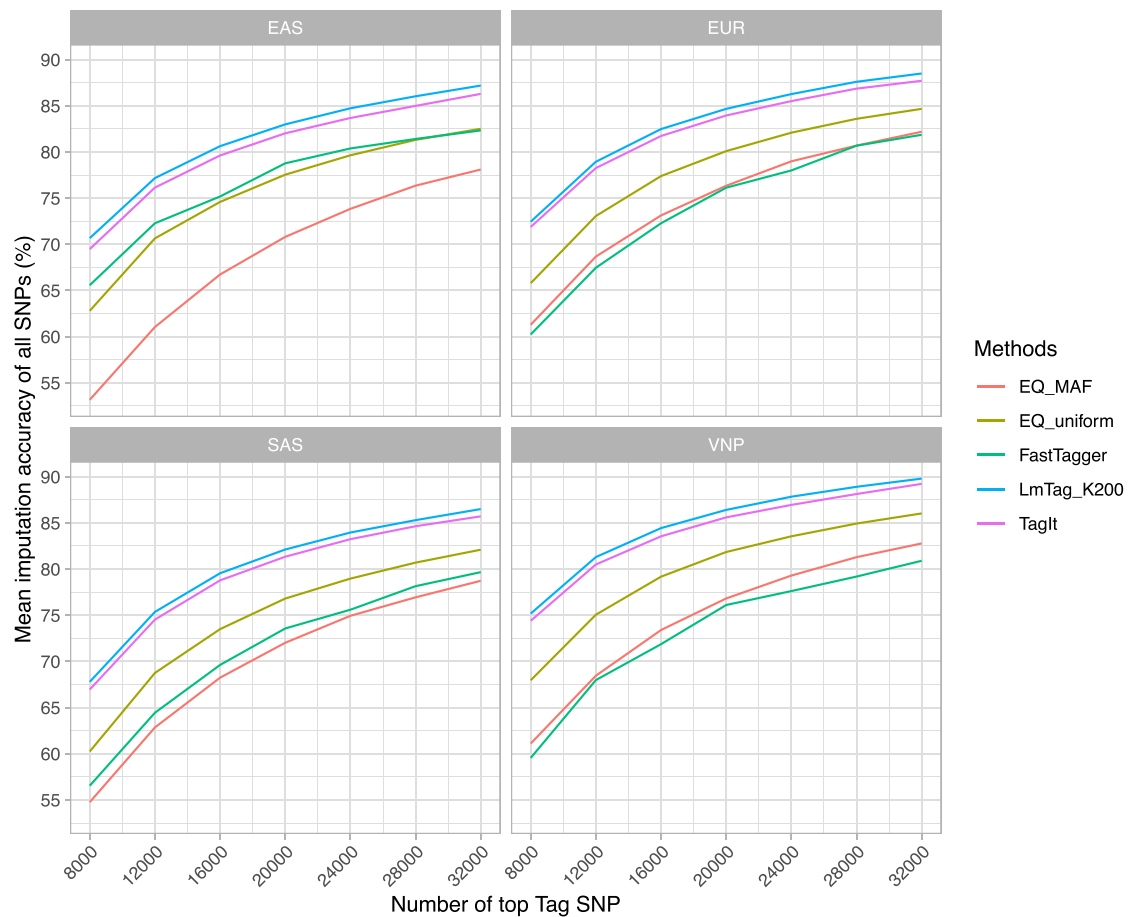


Figure 4. Overall imputation accuracies (mean imputation r^2 of all markers) for each population corresponding to multiple cutoffs ranging from 8000 to 32 000 tag SNPs selected by LmTag (with $K = 200$), EQ_uniform, EQ_MAF, TagIt and FastTagger.

(28 000, and 32 000 tag SNPs scaffolds, with $K=200$) against tag SNP sets from various genotyping arrays with sizes ranging from 30 710 to 49 191 tag SNPs in all populations. In this setting, fewer SNPs are used for LmTag compared to other arrays, as shown in Table S.1. The compared arrays include widely used arrays including Genome-Wide Human SNP Array 6.0, Axiom Genome-Wide ASI, Axiom Genome-Wide EUR, Infinium Global Screening Array v3.0; and recently developed arrays such as Axiom Precision Medicine Diversity Array, Axiom Precision Medicine Research Array; and also customized-population-specific arrays including Axiom UK Biobank Array, Axiom Japonica Array NEO. Manifests of arrays are downloaded from respective manufacturers' websites. Details of tested arrays and their corresponding number of tag SNP in chromosome 10 are reported in Table S.1. Tag SNPs in chromosome 10 are then extracted and harmonized to the UCSC hg38 reference genome coordinate with CrossMap v0.2.6 if lifted over is required to obtain final tag SNP sets [53]. Imputation performances are estimated through leave-one-out cross-validation as described previously.

The comparison yields results as shown in Table S.6, and Figure S.2. In general, LmTag's tag SNP sets outperform all compared array tag SNP sets. At 32 000 tag SNP scaffold, LmTag achieves 87.19%, 88.50%, 86.50% and

89.80% overall imputation accuracies in EUR, EAS, SAS and VNP, respectively, while the corresponding performances at 28 000 tag SNPs scaffold are 86.03%, 87.60%, 85.30% and 88.91%. We also observe that population-specific optimization and size of the tag SNP sets in the arrays are two main factors affecting imputation performances. For instance, the recently developed Axiom Japonica Array NEO [19] and the Axiom UK Biobank Array [14] are the best performers in the EAS and EUR populations with 84.70%, and 87.24% overall imputation accuracies, respectively. Besides, small size global optimization arrays such as the Infinium Global Screening Array v3.0 (30,710 tag SNPs in chromosome 10) shows the poorest performances across populations with 78.35%, 83.15%, 77.77%, and 82.81% overall imputation accuracies in EUR, EAS, SAS, and VNP, respectively. On the other hand, the Genome-Wide Human SNP Array 6.0 (49 191 tag SNPs in chromosome 10) obtains much higher performances of 81.40%, 84.64%, 82.40% and 85.69% for the same populations, respectively.

Overall, LmTag can offer higher performance genotyping arrays with less number of tag SNPs compared to existing arrays. The imputation improvements vary from 9% compared to the Infinium Global Screening Array v3.0 in the EAS population to 1.5% compared to Axiom UK Biobank Array in the EUR population. Notably, for the

VNP population, LmTag's tag SNP sets specific for VNP appears to improve the imputation coverage the most compared to all other arrays.

Discussion and conclusions

Early genome-wide SNP arrays were usually designed by selecting tag SNPs from reference panels of predominantly European population [54]. As a result, these arrays often produce poorer performance in non-European populations [54, 55]. Using customized, small-size SNP arrays at population-specific levels has recently emerged as an extremely cost-effective genotyping solution for under-represented populations [1]. For small arrays, imputation capability is essential to increase the genotyping coverage across the genome to capture as many DNA variants as possible. In addition to imputation performance, researchers also focus on the functional aspect of tag SNPs that are used in SNP arrays, which can help with fine mapping and increase the chance to detect true causal variants associated with traits. A recent comparative study of genotyping SNP arrays [56] discussed the importance of selecting markers based on biological-evidence and CADD functional scores [40]. In this study, we introduce a novel method, LmTag, that is optimized for both imputation and inclusion of functional variants. We compare the performance of LmTag to current widely used methods including EQ_uniform, EQ_MAF, TagIt, and FastTagger; and tag SNP sets from various SNP arrays. These methods and array designs are evaluated across four different populations. The results show that LmTag not only achieves higher imputation performance than other approaches but also significantly enriches the tag SNP set with functional variants. Furthermore, results from our comparative analysis against existing SNP arrays suggest that LmTag has a high potential for designing new genotyping arrays, especially for under-represented populations.

The improvement of tagging efficiency is mainly contributed by the LmTag statistical model. Instead of utilizing solely pairwise LD information as in conventional methods such as TagIt, LmTag assesses the relationship between imputation accuracy, mirror allele frequency, pairwise LD and genomic distance, and then uses this relationship to compute imputation scores for ranking SNP candidates in tagging procedure. The model explains from 26.31% up to 44.14% imputation accuracy, depending on the genetic structure of populations. In all cases, the significant association of the model parameters with imputation accuracies is found, although the effect sizes vary across populations as shown in Table S.7. While pairwise LD, MAF of both tag SNPs and tagged SNPs positively correlate with imputation accuracy, genomic distance showed the reverse trend.

Another advantage of LmTag is the implementation of beam search that considers a secondary factor in tag SNP selection that is the functional aspect of variants by including both ClinVar, GWAS catalog, and

CADD databases simultaneously. Besides genome-wide imputation capability, the inclusion of likely functional variants can enhance the value of genotyping SNP arrays by producing key information on potential causal SNPs underlying phenotypes. For example, the UK Biobank Axiom Array [14], Japonica NEO Arrays [19] and the Axiom KoreanChip [20] applied various selection criteria to include likely functional markers in their array designs. However, these functional SNPs were selected independently from tag SNP selection procedure, i.e. no prioritization of tag SNPs regarding their biological functions was implemented. We introduce here an approach of searching for tag SNPs that are also highly functional. We believe that our proposed method will facilitate the next generation genotyping arrays that have high imputation performance as well as high biological functional potential that would facilitate post GWAS analysis such as statistical fine-mapping [34] and the elucidation of biological mechanisms underlying the relationship between genotypes and phenotypes. Notably, in this study, we demonstrate how LmTag works in human datasets and CADD scores are used as a metric to approximate functional terms. Still, in practice, users could apply the method in other species with any criteria as long as they can provide a ranking scale for each SNP. For example, in other non-model species where calling confidence of the markers is a crucial factor, the method can be adapted for marker quality scores instead of functional scores, as long as a ranking system is provided.

Key Points

- Customizing genotyping array design is emerging as a solution for under-represented populations in developing countries or non-model species.
- Imputation performance, and recently, functional content of the genotyping array are key design principles for most current platforms.
- Current tag SNP selection methods focus only on imputation accuracy and coverage, but not the functional content of the array.
- We introduce LmTag, a novel tag SNP selection method that improves both imputation performance and functional content of the array designs.
- Such arrays are particularly suitable for under-represented populations in developing countries or non-model species where sequencing investment is limited.

Availability of data and materials

The 1KGP-NYGC datasets are freely available at IGSB data portal (<https://www.internationalgenome.org>). The 1KVG pilot phase datasets are available under agreement at MASH data portal (<https://genome.vinbigdata.org/>). LmTag is available for research only purpose at: <https://github.com/datngu/LmTag>. Data and source codes to

generate figures of this study are available at: https://github.com/datngu/LmTag_data_analysis.

Authors' contributions

DTN: conceptualized and implemented the algorithm, designed experiments, analyzed and interpreted results, and drafted the manuscript. QHN contributed to the functional SNP concept development. QHN and NSV: contributed to the discussion and manuscript revision. NTD and NSV: coordinated the project and supervised the study. All authors read and approved the final manuscript.

Acknowledgments

We especially thank Nguyen T. Nguyen for his kindly help in downloading the 1KGP-NYGC datasets; Tran T.H. Tran, Mai H. Tran and the 1000 Vietnamese Genomes Project team for their support in working with the VNP dataset; Khai Q. Tran for his useful suggestions in the C++ implementation; and Nghia T. Vu for his insightful comments during the early version of this manuscript. We also thank our colleagues in Center for Biomedical Informatics, Vingroup Big Data Institute, Vietnam and Institute for Molecular Bioscience, University of Queensland, Australia for their insightful discussions during the study.

Funding

This work is funded by Vingroup Big Data Institute internal funding, and partly supported by the Vingroup Innovation Foundation (grant VINIF.DA.2020.02).

References

- Tam V, Patel N, Turcotte M, et al. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2019;**20**(8):467–84.
- Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet* 2017;**101**(1):5–22.
- Das S, Forer L, Schön herr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016;**48**(10):1284–7.
- Huang J, Howie B, McCarthy S, et al. Improved imputation of low-frequency and rare variants using the uk10k haplotype reference panel. *Nat Commun* 2015;**6**(1):1–9.
- McCarthy S, Das S, Kretschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;**48**(10):1279–83.
- Nelson SC, Doheny KF, Pugh EW, et al. Imputation-based genomic coverage assessments of current human genotyping arrays. *G3: Genes, Genomes, Genetics* 2013;**3**(10):1795–807.
- Martin AR, Kanai M, Kamatani Y, et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 2019;**51**(4):584–91.
- ROSEANN E Peterson, KAROLINE Kuchenbaecker, RAYMOND K Walters, CHIA-YEN Chen, ALICE B Popejoy, SATHISH Periyasamy, MAX Lam, CONRAD Iyegbe, RONA J Strawbridge, LESLIE Brick, et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell*, **179**(3):589–603, 2019.
- Nelson MR, Wegmann D, Ehm MG, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 2012;**337**(6090):100–4.
- Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015;**526**(7571):68–74.
- Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 2019;**570**(7762):514–8.
- Duncan L, Shen H, Gelaye B, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun* 2019;**10**(1):1–9.
- Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med* 2020;**12**:1–11.
- Bycroft C, Freeman C, Petkova D, et al. The UK biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**(7726):203–9.
- Ehli EA, Abdellaoui A, Fedko IO, et al. A method to customize population-specific arrays for genome-wide association testing. *Eur J Hum Genet* 2017;**25**(2):267–70.
- Chen C-H, Yang J-H, Chiang CWK, et al. Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan biobank project. *Hum Mol Genet* 2016;**25**(24):5321–31.
- Dai J, Lv J, Zhu M, et al. Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in chinese populations. *Lancet Respir Med* 2019;**7**(10):881–91.
- Kawai Y, Mimori T, Kojima K, et al. Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals. *J Hum Genet* 2015;**60**(10):581–7.
- Sakurai-Yageta M, Kumada K, Gocho C, et al. Japonica array neo with increased genome-wide coverage and abundant disease risk SNPs. *J Biochem* 2021;**170**(3):399–410.
- Moon S, Kim YJ, Han S, et al. The Korea biobank array: design and identification of coding variants associated with blood biochemical traits. *Sci Rep* 2019;**9**(1):1–11.
- Johnson GCL, Esposito L, Barratt BJ, et al. Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001;**29**(2):233–7.
- Patil N, Bermo AJ, Hinds DA, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001;**294**(5547):1719–23.
- Sebastiani P, Lazarus R, Weiss ST, et al. Minimal haplotype tagging. *Proc Natl Acad Sci* 2003;**100**(17):9900–5.
- Carlson CS, Eberle MA, Rieder MJ, et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004;**74**(1):106–20.
- Liu G, Wang Y, Wong L. Fasttagger: an efficient algorithm for genome-wide tag snp selection using multi-marker linkage disequilibrium. *BMC Bioinformatics* 2010;**11**(1):1–12.
- Hoffmann TJ, Zhan Y, Kvale MN, et al. Design and coverage of high throughput genotyping arrays optimized for individuals of east asian, african american, and Latino race/ethnicity using imputation and a novel hybrid snp selection algorithm. *Genomics* 2011;**98**(6):422–30.
- Wojcik GL, Fuchsberger C, Taliun D, et al. Imputation-aware tag SNP selection to improve power for large-scale, multi-ethnic

- association studies. *G3: Genes, Genomes, Genetics* 2018;**8**(10):3255–67.
28. The International HapMap Consortium. The International HapMap Project. *Nature* 2003;**426**:789–96.
 29. Shashkova TI, Martynova EU, Ayupova AF, et al. Development of a low-density panel for genomic selection of pigs in Russia. *Transl Anim Sci* 2020;**4**(1):264–74.
 30. Herry F, Hérault F, Druet DP, et al. Design of low density SNP chips for genotype imputation in layer chicken. *BMC Genet* 2018;**19**(1):1–14.
 31. Weale ME, Depondt C, Macdonald SJ, et al. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene *scn1a*: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet* 2003;**73**(3):551–65.
 32. Wang W-B, Jiang T. A new model of multi-marker correlation for genome-wide tag SNP selection. In: *Genome Informatics 2008: Genome Informatics Series*, Vol. **21**. World Scientific, 2008, 27–41.
 33. Hao K. Genome-wide selection of tag snps using multiple-marker correlation. *Bioinformatics* 2007;**23**(23):3178–84.
 34. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* 2018;**19**(8):491–504.
 35. Lowerre BT. The Harpy speech recognition system. PhD thesis, 1976.
 36. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res* 2017;**45**(D1):D896–901.
 37. Landrum MJ, Lee JM, Benson M, et al. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;**46**(D1):D1062–7.
 38. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**(3):310–5.
 39. Chang CC, Chow CC, Tellier LCAM, et al. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**(1):s13742–015.
 40. Rentzsch P, Witten D, Cooper GM, et al. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;**47**(D1):D886–94.
 41. Hoffmann TJ, Kvale MN, Hesselson SE, et al. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* 2011;**98**(2):79–89.
 42. Valente JMS, Alves RAFS. Filtered and recovering beam search algorithms for the early/tardy scheduling problem with no idle time. *Comput Indus Eng* 2005;**48**(2):363–75.
 43. Byrska-Bishop M, Evani US, Zhao X, et al. High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios. bioRxiv. 2021.
 44. Miller NA, Farrow EG, Gibson M, et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Med* 2015;**7**(1):1–16.
 45. Van der Auwera GA, Carneiro MO, Hartl C, et al. (eds). From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;**43**(1):11.
 46. Delaneau O, Zagury J-F, Robinson MR, et al. Accurate, scalable and integrative haplotype estimation. *Nat Commun* 2019;**10**(1):1–10.
 47. Hayes BJ, Bowman PJ, Daetwyler HD, et al. Accuracy of genotype imputation in sheep breeds. *Anim Genet* 2012;**43**(1):72–80.
 48. Joshi R, Árnýasi M, Lien S, et al. Development and validation of 58k snp-array and high-density linkage map in Nile tilapia (*O. niloticus*). *Front Genet* 2018;**9**:472.
 49. Romain Dassonneville S, Fritz VD, Boichard D. Imputation performances of 3 low-density marker panels in beef and dairy cattle. *J Dairy Sci* 2012;**95**(7):4136–40.
 50. Qiao X, Rui S, Wang Y, et al. Genome-wide target enrichment-aided chip design: a 66 k SNP chip for cashmere goat. *Sci Rep* 2017;**7**(1):1–13.
 51. Hao K, Di X, Cawley S. Ldcompare: rapid computation of single- and multiple-marker r^2 and genetic coverage. *Bioinformatics* 2007;**23**(2):252–4.
 52. Nguyen DT, Dinh HQ, Giang Minh V, et al. (eds). A comprehensive imputation-based evaluation of tag SNP selection strategies. In: *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2021, 1–6.
 53. Zhao H, Sun Z, Wang J, et al. Crossmap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 2014;**30**(7):1006–7.
 54. Rosenberg NA, Huang L, Jewett EM, et al. Genome-wide association studies in diverse populations. *Nat Rev Genet* 2010;**11**(5):356–66.
 55. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science* 2008;**322**(5903):881–8.
 56. Verlouw JAM, Clemens E, de Vries JH, et al. A comparison of genotyping arrays. *Eur J Hum Genet* 2021;**29**:1611–1624.