A comprehensive imputation-based evaluation of tag SNP selection strategies

Dat Thanh Nguyen[*]
Center for Biomedical Informatics
Vingroup Big Data Institute
Hanoi, Vietnam
n.dat@outlook.com

Hieu Quang Dinh
Massachusetts Institute of Technology
Cambridge, MA, USA
hieudinh@mit.edu

Giang Minh Vu
Center for Biomedical Informatics
Vingroup Big Data Institute
Hanoi, Vietnam
v.giangvm1@vinbigdata.org

Duong Thuy Nguyen
Center for Biomedical Informatics
Vingroup Big Data Institute
Institute of Genome Research
Vietnam Academy of Science and Technology
Hanoi, Vietnam
v.duongnt43@vinbigdata.org

Nam Sy Vo
Center for Biomedical Informatics
Vingroup Big Data Institute
College of Engineering and Computer Science
VinUniversity
Hanoi, Vietnam
v.namvs@vinbigdata.org

***Abstract*—Regardless of the rapid development of sequencing technology, single nucleotide polymorphism (SNP) array has been widely used for many large-scale genomic studies due to its cost-effectiveness. Recently, in parallel with the advancement in imputation strategies, several genotyping platforms for various species have been developed. Despite the importance of imputation accuracy in SNP array design, to the best of our knowledge, there are no systematic studies for evaluating tag SNP selection methods based on this metric. In this paper, using the leave-one-out cross-validation approach on the 1000 genome high-coverage dataset, we comprehensively evaluated four well-known tag SNP selection algorithms based on imputation accuracy. Our results showed that although all widely used methods for SNP array design can provide reasonable imputation accuracy, pairwise linkage disequilibrium based tag SNP selection algorithm achieves the best performance. Our pipelines for running evaluated algorithms and leave-one-out cross-validation are available for public use at https://github.com/datngu/TagSNP_evaluation.**

***Index Terms*—Tag SNP selection, SNP array design, genotyping imputation, linkage disequilibrium**

I. Introduction

Genomic variants can be genotyped using various technologies, including genome-wide single nucleotide polymorphism (SNP) array and whole-genome sequencing (WGS). While WGS is able to capture all genomic variants in the genome, SNP arrays, which have several advantages, such as cost-effectiveness, reliability of the technology, and light computational requirement, are being used extensively [1]. However, due to capacity limitation of the SNP arrays, array-based genomic studies often required genotyping imputation to infer missing variants. This analysis routine significantly increases the number of variants for association tests by predicting the genotypes at the SNPs that are not directly genotyped in the study sample. Performance of the imputation is dependent on three main factors including imputation algorithms [2], imputation reference panel [3], [4], and the design of the genotyping arrays [5].

Recent years have witnessed the rapid development of several genotyping platforms for various species focusing on imputation optimization [6]–[11]. This trend is also popular in human. For instance, numerous genotyping arrays were introduced recently to improve the power of genome-wide association study at the population-specific level. Indeed, population-specific genotyping arrays such as the UK Biobank Axiom Array [12], the Axiom-NL Array [13], the TWB Array [14], the Axiom China Kadoorie Biobank Array [15], the Japonica and Japonica NEO Arrays [16], [17], and the Axiom KoreanChip [18] have been successfully implemented.

Two main strategies for selecting SNPs to place on microarrays, called tag SNPs, includes optimizing genomic distance with the equidistant principle and maximizing linkage disequilibrium (LD) [6]–[11], [16]–[23]. The genomic distance-based method, which is commonly referred to as equidistance (EQ), is mainly employed in animal sciences to choose SNPs based on physical intervals (in pb) along chromosomes. This strategy can be set to select tag SNPs either uniformly or to optimize for minor allele frequency (MAF) in each genomic window [11], [23]. In contrast, the LD-based method, which is used in both human and animal sciences [11], [24], utilizes pairwise LD information with the greedy approach to maximize LD coverage [17], [22], [25]. A typical algorithm using this strategy weights each SNP candidate by the number of neighbor SNPs that pass a specific LD threshold. The SNP with the highest number of neighbors is selected as tag SNP for the current round, and all neighbor SNPs are removed from the target set. These steps are iterated until the desired number of tag SNP is reached, or LD r^2 threshold is not satisfied by the remaining SNPs [19], [25]. In addition, the LD-based approach can be extended to multi-markers, which prioritizes tag SNPs based on multi-marker LD computation [26]–[28].

Despite the importance of tag SNP selection in SNP array design, available comparative studies of tag SNP selection strategies have been carried out on small datasets and primarily focused on evaluating pair-wide LD coverage [29], [30]. Furthermore, evaluating tag SNPs selected from either LD coverage or genomic distance remains challenging. It is unclear which approach would yield better imputation accuracy, a gold standard for SNP array assessment [5], [22]. In this work, we introduce an imputation-based evaluation for four widely-used algorithms in SNP array design - TagIt , FastTagger, EQ_uniform, and EQ_MAF. These algorithms select tag SNPs based on pairwise LD, multi-marker LD, EQ uniform, and EQ with MAF optimization, respectively [11], [23], [25], [28]. Using the leave-one-out cross-validation approach, we evaluated each algorithm's performance on genomic data of four super populations, obtained from The 1000 Genomes Project (1kGP)-high coverage dataset [20]–[22], [31].

II. Method

A. Genomic data prepossessing

Phased variant callsets in variant calling format (VCF), which contain 2504 unrelated samples originated from 1kGP phase 3 release and expanded 698 related individuals, are obtained from The International Genome Sample Resource (IGSR) data portal [31]. Only unrelated samples are included in the analysis, and they are assigned to their super population according to IGSR's annotation. This analysis only has chromosome 10 due to computational reasons, and only biallelic SNPs with minor allele frequency (MAF) $\geq 1\%$ are kept for tag SNP selection and imputation evaluation. Finally, four major populations are analyzed in our study, including East Asian (EAS), European (EUR), South Asian (SAS), and Americas (AMR) with 504, 503, 489, and 347 individuals, respectively.

B. Tag SNP selection algorithms

We compare four current widely used approaches for SNP array designing, including TagIt, FastTagger, EQ_uniform, and EQ_MAF [11], [23], [25], [28]. TagIt, widely used in human SNP array design, is a tag SNPs selection algorithm based on pairwise LD information [17], [22], [25]. From a set of targeted SNPs, TagIt weights each SNP candidate by the number of neighbor SNPs (in a specific genomic distance) that have square pairwise LD r^2 greater or equal to a specific cutoff, e.g., 0.8. The SNP with the highest score is selected, and its neighbor SNPs are removed from the targeted set. The process is iterated until reaching the desired number of tag SNP, or no more SNP satisfies the LD r^2 threshold [25]. FastTagger uses a fast implementation of the multi-marker LD approach, which reduces the number of tag SNPs selected while maintaining high genomic coverage. In brief, the multi-marker LD approach finds association rules of one and multiple SNPs, termed multi-marker r^2 statistics, and uses this information to identify tag SNPs [27], [28], [32]. The major bottleneck of this approach is the computational burden; for instance, standard algorithms using multi-marker models usually fail to run on chromosomes containing more than 100k SNPs. FastTagger, on the other hand, resolves this problem by employing several techniques to reduce running time and memory consumption and makes this approach became more scalable [28].

The EQ strategy selects tag SNP according to the equidistant principle that involves dividing chromosomes into certain intervals with equal genomic length [6]–[11], [23]. While the EQ_uniform assumes that genetic variants are uniformly distributed and selects them based on physical intervals (in pb) along chromosomes [23], the EQ_MAF adjusts for minor allele frequency (MAF) in each genomic window [11]. For each interval, the SNP with the highest MAF, or further to the left, in case of equivalent MAF, was chosen as representative of the interval [11].

C. Performance evaluations

In our study, imputation accuracy and imputation coverage are employed as evaluation metrics in replacement of LD coverage [5], [20], [22]. Imputation accuracy is determined through the leave-one-out cross-validation approach (Figure 1). More precisely, imputation is performed individually for each sample using Minimac4 with a reference panel that does not have the sample itself [2]. Selected tag SNPs are denoted as 'genotyped', and other sites are set as missing. For each SNP, squared Pearson's correlation is then calculated from imputation estimated dosages to the true genotypes in the original VCF file. Imputation coverage is calculated as the proportion of SNP with imputation r^2 that are greater or equal to a specific threshold, e.g., 0.8. Imputation accuracy is reported as overall imputation accuracy by computing mean squared Pearson's correlation of all SNPs or binned imputation accuracies of various discrete minor allele frequency bins. In addition, running time of each algorithms is also evaluated in this study.

D. Parameter setting and running algorithms

To facilitate comparison between methods, LD cutoff is set at 0.8 for TagIt and FastTagger min_r2_1; FastTagger min_r2_2, and min_r2_3 are set to 0.9, 0.95 respectively, as recommended by the authors. Since EQ_uniform, EQ_MAF, and FastTagger select more tag SNPs than TagIt does, the number of selected tag SNPs from each population for all algorithms are set to the number of SNP selected by TagIt. Because methods require different input data formats, customized scripts and data pre-processing steps are needed. As TagIt requires pre-computed LD pairwise and MAF information, we used Plink v1.9 and vcftools to obtain such data respectively [33], [34]. Pairwise LDs are computed within a maximum genomic distance of 1 megabase (MB) and minimum LD r2 cutoff of 0.8 [33] and MAFs are extracted with vcftools [34]. A customized script is then used to obtain the final Tagit's input. In the meanwhile, another customized script is used

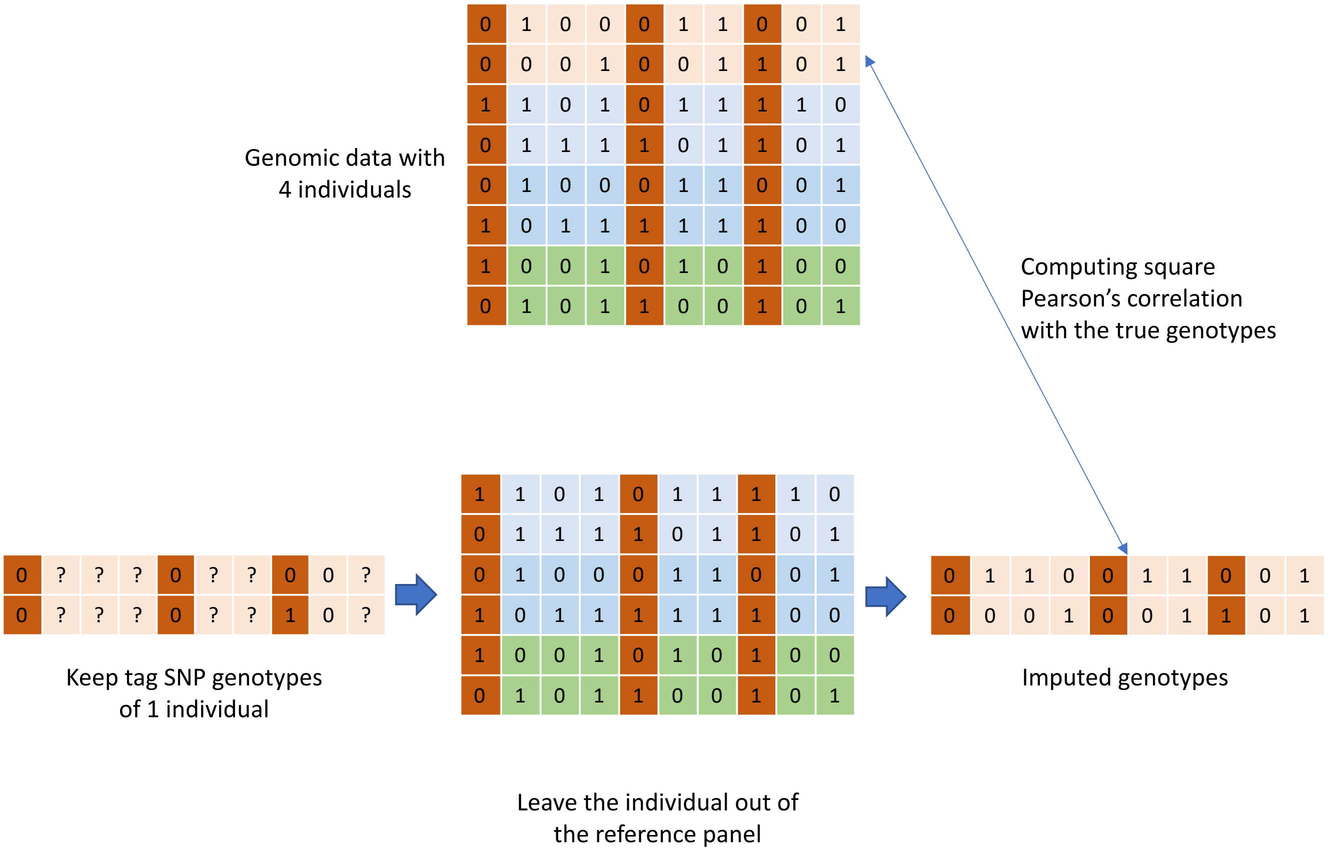

Fig. 1. Leave-one-out cross-validation strategy. Tag SNPs selected are denoted as 'genotyped', and other sites are set as missing. Imputation is performed individually for each sample with the exclusion of itself from the reference panel. Squared Pearson's correlation is then calculated from imputation estimated dosages to the true genotypes from the original VCF file for each SNP.

to obtain FastTagger required input that are matrixes of haplotypes and their corresponding MAF information. Regarding EQ_uniform, and EQ_MAF, genomic positions and MAFs are extracted with bcftools as input for tag SNP selections. All runs are implemented in a workstation using Intel Core i9-9900K CPU at 3.60GHz/core under Ubuntu 18 OS. All tools are run in single-core, each core has 16 GB memory; the reported time is in core-hours. The performance evaluation workflow is implemented in a set of scripts that is available for public use at https://github.com/datngu/TagSNP_evaluation.

III. Result

A. Imputation accuracy

In general, all evaluated algorithms obtain high imputation accuracy, with most of the overall imputation accuracy are higher than 0.8. Details of the performance are reported in Table I. With overall imputation accuracies of 0.87,0.90, 0.88, and 0.88 in EAS, EUR, SAS, and AMR, TagIt is the best performer, followed by EQ_uniform. In contrast, FastTagger and EQ_MAF are consistently the worst. We further evaluate the performance of these methods by comparing imputation accuracy for various MAF bins as represented in Figure 2. The curves of TagIt, EQ_uniform, EQ_MAF, and FastTagger are consistent with their overall imputation accuracy. The curves for TagIt are consistently above the ones for other methods. It is noteworthy that FastTagger outperforms EQ_MAF at low MAF but is worse than EQ_MAF at high MAF. The Figure 2 also indicates that imputation accuracy decreases as MAF bins move to lower frequency, and this trend is consistent for all algorithms. For instance, in the population of EUR, TagIt, EQ_uniform, EQ_MAF, and FastTagger yield imputation accuracy of 0.94, 0.92, 0.92, and 0.89 in the MAF bin (0.2:0.5], respectively, while the corresponding values in the MAF bin (0.01:0.05] are 0.81, 0.76, 0.72, and 0.73.

B. Imputation coverage

Regarding imputation coverage, TagIt also achieved the highest performances followed by EQ_uniform, while the

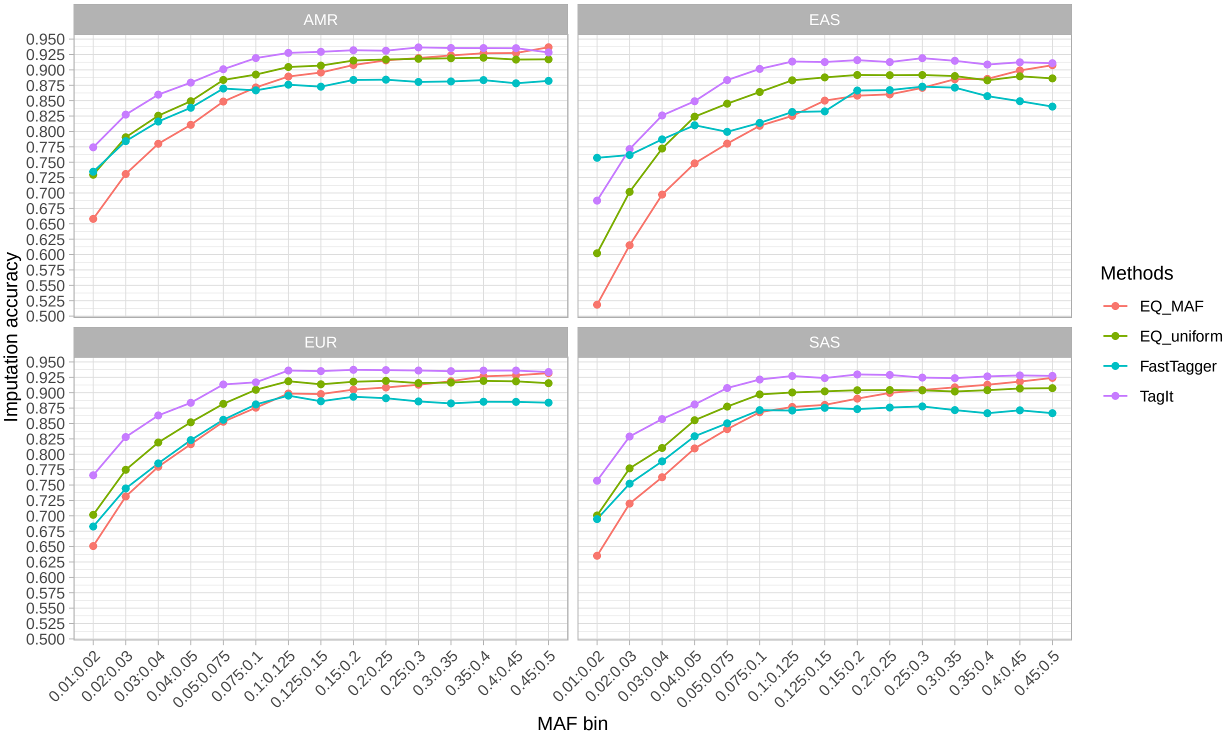

Fig. 2. Imputation accuracy curves of TagIt, FastTagger, EQ_uniform, and EQ_MAF measured by leave-one-out cross-validation. The x-axis shows various MAF bins, and the y-axis shows the mean imputation accuracy of SNPs corresponding to each MAF bin in the x-axis.

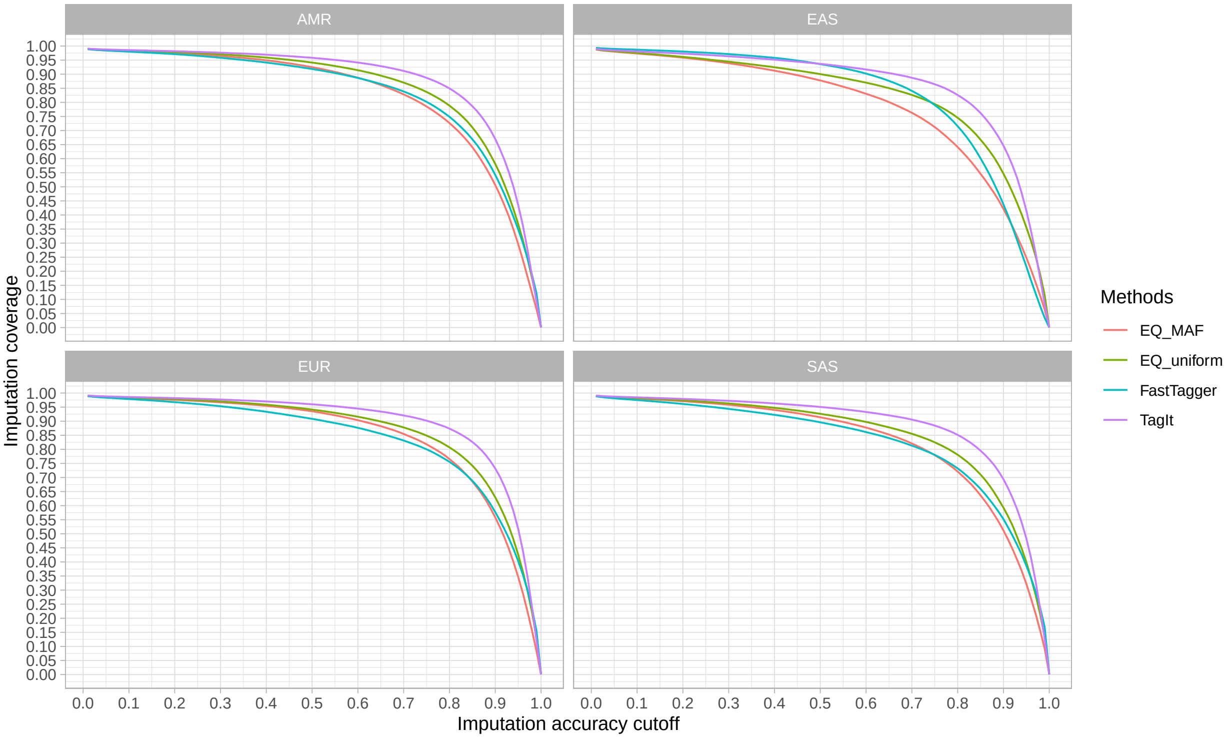

Fig. 3. Imputation coverage curves of TagIt, FastTagger, EQ_uniform, and EQ_MAF measured by leave-one-out cross-validation. The x-axis shows imputation accuracy cutoffs, and the y-axis shows the proportion of SNP with equal or higher imputation accuracy corresponding cutoff in the x-axis.

TABLE I

IMPUTATION ACCURACIES OF TAGIT, FASTTAGGER, EQ_UNIFORM, AND EQ_MAF IN VARIOUS MAF BINS, MEASURED BY LEAVE-ONE-OUT CROSS-VALIDATION.

Population	Method	Minor allele frequency bin				
		(0.01: 0.05]	(0.05: 0.1]	(0.1: 0.2]	(0.2: 0.5]	All
EAS	TagIt	0.76	0.89	0.91	0.91	0.87
EAS	EQ_uniform	0.69	0.85	0.89	0.89	0.83
EAS	EQ_MAF	0.61	0.79	0.85	0.88	0.79
EAS	FastTagger	0.77	0.81	0.85	0.86	0.83
EUR	TagIt	0.81	0.91	0.94	0.94	0.90
EUR	EQ_uniform	0.76	0.89	0.92	0.92	0.87
EUR	EQ_MAF	0.72	0.86	0.90	0.92	0.85
EUR	FastTagger	0.73	0.87	0.89	0.89	0.84
SAS	TagIt	0.80	0.91	0.93	0.93	0.88
SAS	EQ_uniform	0.75	0.89	0.90	0.90	0.85
SAS	EQ_MAF	0.69	0.85	0.88	0.91	0.83
SAS	FastTagger	0.74	0.86	0.87	0.87	0.83
AMR	TagIt	0.81	0.91	0.93	0.93	0.88
AMR	EQ_uniform	0.77	0.89	0.91	0.92	0.86
AMR	EQ_MAF	0.71	0.86	0.90	0.92	0.83
AMR	FastTagger	0.77	0.87	0.88	0.88	0.84

TABLE II

IMPUTATION COVERAGES OF TAGIT, FASTTAGGER, EQ_UNIFORM, AND EQ_MAF BY VARIOUS IMPUTATION ACCURACY THRESHOLDS IN PERCENTAGE, MEASURED BY LEAVE-ONE-OUT CROSS-VALIDATION.

Population	Method	Imputation accuracy threshold				
		0.9	0.85	0.8	0.75	0.7
EAS	TagIt	64.61	76.2	82.62	86.43	88.73
EAS	EQ_uniform	54.63	66.73	74.52	79.38	82.63
EAS	EQ_MAF	42.13	54.6	64.11	71.2	76.3
EAS	FastTagger	43.72	60.08	71.51	79.02	84.07
EUR	TagIt	73.19	82.61	87.35	90.18	92.02
EUR	EQ_uniform	62.95	74.18	80.68	84.88	87.78
EUR	EQ_MAF	55.82	68.58	76.54	81.78	85.45
EUR	FastTagger	57.71	68.83	75.56	80.01	83.13
SAS	TagIt	69.36	79.51	85.12	88.43	90.57
SAS	EQ_uniform	59.27	71.03	78.07	82.49	85.52
SAS	EQ_MAF	51.17	63.63	72.1	77.92	82.09
SAS	FastTagger	55.21	65.95	73.22	78.02	81.3
AMR	TagIt	66.98	78.62	84.94	88.69	91.15
AMR	EQ_uniform	58.13	71.1	78.84	83.69	87.02
AMR	EQ_MAF	50.62	64.07	72.66	78.52	82.81
AMR	FastTagger	54.23	66.89	74.87	80.18	83.85

TABLE III

RUNNING TIME OF TAGIT, FASTTAGGER, EQ_UNIFORM, AND EQ_MAF TO PERFORM TAG SNP SELECTION IN CHROMOSOME 10 POPULATION EAS, EUR, SAS, AND AMR; MEASURED BY CORE HOURS.

Method	EAS	EUR	SAS	AMR
TagIt	0.15	0.2	0.2	0.25
EQ_uniform	0	0	0	0
EQ_MAF	0.05	0.05	0.05	0.1
FastTagger	29	41	54	115

worst performers are interchanged between EQ_MAF and FastTagger, depending on population and imputation thresholds. Details of imputation coverage is showed in Table II and Figure 3. Taken the imputation accuracy threshold at 0.8 as an example, TagIt yields imputation coverage of 82.62%, 87.35%, 85.12%, and 84.94% in EAS, EUR, SAS, AMR populations while the second-ranked tool, EQ_uniform, gives imputation coverage of 74.52%, 80.68%, 78.08%, and 78.84% respectively. Using the same accuracy threshold, EQ_MAF has the poorest performance with coverage of 64.11%, 76.54%, 72.10%, and 72.66% coverage in four populations.

C. Running time

We measure the total CPU time of all methods for individual datasets, which are reported in Table III. The time is calculated from the time the methods start running until results are produced. It covers all the steps including data processing and tag SNP selection. As expected, FastTagger is the most time-consuming method. Its running time is sustainability higher than others that takes 29, 41, 54, and 115 hours to finish tag SNP selection in EAS, EUR, SAS, and AMR respectively. In contrast, EQ_uniform, EQ_MAF, and TagIt provide results in a time of no more than a half-hour in all populations.

IV. DISCUSSION AND CONCLUSION

In this study, performance of tag SNP selection methods are evaluated based on genome-wide imputation accuracy as measured by mean imputed r^2 at untyped sites rather than pairwise LD. Imputation accuracy assessment using leave-one-out cross-validation provides a real-world estimation of genomic coverage, the golden standard assessment of SNP array nowadays [5], [20]–[22]. We examined the performance of two main strategies in SNP array design, LD-based tag SNP selection and genomic-distance-based tag SNP selection, by looking at four typical algorithms - TagIt, FastTagger, EQ_uniform, and EQ_MAF [11], [23], [25], [28]. We provided a comprehensive evaluation based on the 1kGP high coverage dataset [31]. Our results indicated that all assessed algorithms are reasonable to use for imputation SNP array design, they yielded high imputation accuracies and achieved high imputation genomic coverages. It is also noticeable that LD pairwise tag SNP selection strategy (TagIt) outperformed the others. This strategy provided the best tag SNP selection performance in terms of imputation accuracy and imputation genomic coverage in all examined datasets. In addition, over-optimization approaches for MAF (EQ_MAF) and LD (FastTagger) have tended to provide poorer performance than the naive approaches that uses pairwise LD and uniform assumption. Finally, we provided a set of scripts for running the leave-one-out cross-validation for tag SNP selection algorithms to facilitate the design and evaluation of next-generation SNP array platforms.

ACKNOWLEDGMENT

We would like to express our deepest gratitude to our colleagues at Center for Biomedical Informatics, Vingroup Big Data Institute for their very helpful suggestions and comments during our work.

REFERENCES

[1] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre, "Benefits and limitations of genome-wide association studies," *Nature Reviews Genetics*, vol. 20, no. 8, pp. 467–484, 2019.

[2] S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue *et al.*, "Next-generation genotype imputation service and methods," *Nature genetics*, vol. 48, no. 10, pp. 1284–1287, 2016.

[3] J. Huang, B. Howie, S. McCarthy, Y. Memari, K. Walter, J. L. Min, P. Danecek, G. Malerba, E. Trabetti, H.-F. Zheng *et al.*, "Improved imputation of low-frequency and rare variants using the uk10k haplotype reference panel," *Nature communications*, vol. 6, no. 1, pp. 1–9, 2015.

[4] S. McCarthy, S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood, A. Teumer, H. M. Kang, C. Fuchsberger, P. Danecek, K. Sharp *et al.*, "A reference panel of 64,976 haplotypes for genotype imputation," *Nature genetics*, vol. 48, no. 10, pp. 1279–1283, 2016.

[5] S. C. Nelson, K. F. Doheny, E. W. Pugh, J. M. Romm, H. Ling, C. A. Laurie, S. R. Browning, B. S. Weir, and C. C. Laurie, "Imputation-based genomic coverage assessments of current human genotyping arrays," *G3: Genes, Genomes, Genetics*, vol. 3, no. 10, pp. 1795–1807, 2013.

[6] R. Dassonneville, S. Fritz, V. Ducrocq, and D. Boichard, "Imputation performances of 3 low-density marker panels in beef and dairy cattle," *Journal of dairy science*, vol. 95, no. 7, pp. 4136–4140, 2012.

[7] C. Hozé, M.-N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, and P. Croiseau, "High-density marker imputation accuracy in sixteen french cattle breeds," *Genetics Selection Evolution*, vol. 45, no. 1, pp. 1–11, 2013.

[8] B. Hayes, P. Bowman, H. Daetwyler, J. Kijas, and J. Van der Werf, "Accuracy of genotype imputation in sheep breeds," *Animal genetics*, vol. 43, no. 1, pp. 72–80, 2012.

[9] A. Bouquet, K. Fève, J. Riquet, and C. Larzul, "Précision de l'imputation de génotypages haute densité à partir de puces basse densité pour des individus de race pure et croisés piétrain," *Journées Rec Porcine*, vol. 47, p. 1, 2015.

[10] X. Qiao, R. Su, Y. Wang, R. Wang, T. Yang, X. Li, W. Chen, S. He, Y. Jiang, Q. Xu *et al.*, "Genome-wide target enrichment-aided chip design: a 66 k snp chip for cashmere goat," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[11] F. Herry, F. Hérault, D. P. Druet, A. Varenne, T. Burlot, P. Le Roy, and S. Allais, "Design of low density snp chips for genotype imputation in layer chicken," *BMC genetics*, vol. 19, no. 1, pp. 1–14, 2018.

[12] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell *et al.*, "The uk biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

[13] E. A. Ehli, A. Abdellaoui, I. O. Fedko, C. Grieser, S. Nohzadeh-Malakshah, G. Willemsen, E. J. de Geus, D. I. Boomsma, G. E. Davies, and J. J. Hottenga, "A method to customize population-specific arrays for genome-wide association testing," *European Journal of Human Genetics*, vol. 25, no. 2, pp. 267–270, 2017.

[14] C.-H. Chen, J.-H. Yang, C. W. Chiang, C.-N. Hsiung, P.-E. Wu, L.-C. Chang, H.-W. Chu, J. Chang, I.-W. Song, S.-L. Yang *et al.*, "Population structure of han chinese in the modern taiwanese population based on 10,000 participants in the taiwan biobank project," *Human molecular genetics*, vol. 25, no. 24, pp. 5321–5331, 2016.

[15] J. Dai, J. Lv, M. Zhu, Y. Wang, N. Qin, H. Ma, Y.-Q. He, R. Zhang, W. Tan, J. Fan *et al.*, "Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in chinese populations," *The Lancet Respiratory Medicine*, vol. 7, no. 10, pp. 881–891, 2019.

[16] Y. Kawai, T. Mimori, K. Kojima, N. Nariai, I. Danjoh, R. Saito, J. Yasuda, M. Yamamoto, and M. Nagasaki, "Japonica array: improved genotype imputation by designing a population-specific snp array with 1070 japanese individuals," *Journal of human genetics*, vol. 60, no. 10, pp. 581–587, 2015.

[17] M. Sakurai-Yageta, K. Kumada, C. Gocho, S. Makino, A. Uruno, S. Tadaka, I. N. Motoike, M. Kimura, S. Ito, A. Otsuki *et al.*, "Japonica array neo with increased genome-wide coverage and abundant disease risk snps," *bioRxiv*, 2020.

[18] S. Moon, Y. J. Kim, S. Han, M. Y. Hwang, D. M. Shin, M. Y. Park, Y. Lu, K. Yoon, H.-M. Jang, Y. K. Kim *et al.*, "The korea biobank array: design and identification of coding variants associated with blood biochemical traits," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[19] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium," *The American Journal of Human Genetics*, vol. 74, no. 1, pp. 106–120, 2004.

[20] T. J. Hoffmann, Y. Zhan, M. N. Kvale, S. E. Hesselson, J. Gollub, C. Iribarren, Y. Lu, G. Mei, M. M. Purdy, C. Quesenberry *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*, vol. 98, no. 6, pp. 422–430, 2011.

[21] T. J. Hoffmann, M. N. Kvale, S. E. Hesselson, Y. Zhan, C. Aquino, Y. Cao, S. Cawley, E. Chung, S. Connell, J. Eshragh *et al.*, "Next generation genome-wide association tool: design and coverage of a high-throughput european-optimized snp array," *Genomics*, vol. 98, no. 2, pp. 79–89, 2011.

[22] G. L. Wojcik, C. Fuchsberger, D. Taliun, R. Welch, A. R. Martin, S. Shringarpure, C. S. Carlson, G. Abecasis, H. M. Kang, M. Boehnke *et al.*, "Imputation-aware tag snp selection to improve power for large-scale, multi-ethnic association studies," *G3: Genes, Genomes, Genetics*, vol. 8, no. 10, pp. 3255–3267, 2018.

[23] T. I. Shashkova, E. U. Martynova, A. F. Ayupova, A. A. Shumskiy, P. A. Ogurtsova, O. V. Kostyunina, P. E. Khaitovich, P. V. Mazin, and N. A. Zinovieva, "Development of a low-density panel for genomic selection of pigs in russia," *Translational animal science*, vol. 4, no. 1, pp. 264–274, 2020.

[24] R. J. Schaefer, M. Schubert, E. Bailey, D. L. Bannasch, E. Barrey, G. K. Bar-Gal, G. Brem, S. A. Brooks, O. Distl, R. Fries *et al.*, "Developing a 670k genotyping array to tag˜ 2m snps across 24 horse breeds," *BMC genomics*, vol. 18, no. 1, pp. 1–18, 2017.

[25] M. E. Weale, C. Depondt, S. J. Macdonald, A. Smith, P. San Lai, S. D. Shorvon, N. W. Wood, and D. B. Goldstein, "Selection and evaluation of tagging snps in the neuronal-sodium-channel gene scn1a: implications for linkage-disequilibrium gene mapping," *The American Journal of Human Genetics*, vol. 73, no. 3, pp. 551–565, 2003.

[26] W.-B. Wang and T. Jiang, "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, pp. 27–41.

[27] K. Hao, "Genome-wide selection of tag snps using multiple-marker correlation," *Bioinformatics*, vol. 23, no. 23, pp. 3178–3184, 2007.

[28] G. Liu, Y. Wang, and L. Wong, "Fasttagger: an efficient algorithm for genome-wide tag snp selection using multi-marker linkage disequilibrium," *BMC bioinformatics*, vol. 11, no. 1, pp. 1–12, 2010.

[29] K. M. Burkett, M. Ghadessi, B. McNeney, J. Graham, and D. Daley, "A comparison of five methods for selecting tagging single-nucleotide polymorphisms," in *BMC genetics*, vol. 6, no. 1. BioMed Central, 2005, pp. 1–5.

[30] K. Ding and I. J. Kullo, "Methods for the selection of tagging snps: a comparison of tagging efficiency and performance," *European Journal of Human Genetics*, vol. 15, no. 2, pp. 228–236, 2007.

[31] M. Byrska-Bishop, U. S. Evani, X. Zhao, A. O. Basile, H. J. Abel, A. A. Regier, A. Corvelo, W. E. Clarke, R. Musunuri, K. Nagulapalli *et al.*, "High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios," *bioRxiv*, 2021.

[32] K. Hao, X. Di, and S. Cawley, "Ldcompare: rapid computation of single- and multiple-marker r 2 and genetic coverage," *Bioinformatics*, vol. 23, no. 2, pp. 252–254, 2007.

[33] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation plink: rising to the challenge of larger and richer datasets," *Gigascience*, vol. 4, no. 1, pp. s13 742–015, 2015.

[34] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry *et al.*, "The variant call format and vcftools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.