2024 년 2 월
석사학위 논문

Re-examination of genome imputation analysis of whole genome sequencing data

조선대학교 대학원

글로벌바이오융합학과

서 윤 종

Re-examination of genome imputation analysis of whole genome sequencing data

전유전체서열분석(WGS)자료의 유전체 대치분석 필요성 재고

2024 년 2 월 23 일

조선대학교 대학원

글로벌바이오융합학과

서 윤 종

Re-examination of genome imputation analysis of whole genome sequencing data

지도교수 김 정 수

이 논문을 이학 석사학위 신청 논문으로 제출함.

2023 년 10 월

조선대학교 대학원

글로벌바이오융합학과

서 윤 종

서윤종의 석사학위논문을 인준함

위원장 김 석 준 (인)

위　원 김 규 민 (인)

위　원 김 정 수 (인)

2023 년 12 월

조선대학교 대학원

CONTENTS

LIST OF TABLES ····· **II**

LIST OF FIGURES ····· **III**

ABSTRACT ····· **IV**

I. INTRODUCTION ····· **1**

I-1. Whole genome sequence and genotype array ····· 1

I-2. Genome imputation ····· 1

I-3. Related studies ····· 2

I-4. Research purpose ····· 3

II. MATERIALS AND METHODS ····· **4**

II-1. Genotype array data ····· 4

II-2. WGS data ····· 4

II-3. Reference panels for genome imputation ····· 5

II-4. Genome imputation ····· 7

II-5. Performance measure of imputation result ····· 7

III. RESULTS ····· **9**

III-1. Study overview ····· 9

III-2. Imputation result ····· 11

III-3. Imputation performance ····· 14

III-4. Correction of genotyped SNPs ····· 19

III-5. Genome imputation with WGS ····· 21

IV. DISCUSSION ····· **24**

Ⅴ. 초 록 ····· **26**

VI. REFERENCES ····· **28**

VII. APPENDIX ····· **30**

LIST OF TABLES

Table 1. Reference panels for genome imputation ····· 6

Table 2. Classification for performance evaluation ····· 8

Table 3. Imputation result ····· 12

LIST OF FIGURES

Figure 1.Study overview ····· 10

Figure 2. Overall imputation quality ····· 13

Figure 3. Difference between WGS MAF and imputed MAF ····· 15

Figure 4. Recall by MAF ····· 16

Figure 5. Precision by MAF ····· 17

Figure 6. Concordance by MAF ····· 18

Figure 7. Genotype correction in each reference panel ····· 20

Figure 8. Non-overlapping SNPs with WGS ····· 22

Figure 9. Utilization of genome imputation in deep WGS ····· 23

ABSTRACT

Re-examination of genome imputation analysis of whole genome sequencing data

Yoonjong Seo

Advisor: Prof. Jungsoo Gim, Ph.D.

Department of Integrative Biological Sciences

Graduate School of Chosun University

Genome imputation analysis is the standard procedure in genetic analysis for exploring associations between the genome and various phenotypes. However, despite the utility and importance of genome imputation analysis, many genetically homogeneous minority populations exist in small proportions in the reference panel, and only limited performance evaluation studies have been conducted.

In this study, we analyzed how well the imputation results approximate whole-genome sequence (WGS) using Koreans as an example of a genetically homogeneous minority population, utilizing both a large dataset of 2,253 whole-genome sequencing and genotype array data for more accurate and meaningful performance assessment.

For the imputation, we selected four reference genome panels, considering the characteristics of each panel commonly used in the field: a Korean reference panel, Haplotype Reference Consortium (HRC), 1000 Genome, and Trans-Omics for Precision Medicine (TOPMed).

As expected, the results using the Korean reference panel outperformed all other reference panels in terms of all performance metrics. Particularly, it exhibited overwhelming accuracy, especially for variants with a minor allele frequency (MAF) of less than 1%, when compared to other reference panels. When using the pipeline from the Michigan Imputation Service, we

observed cases where the called genotypes were corrected based on the imputed genotypes. In these cases, the Korean reference panel showed the lowest errors in genotype correction compared to the other panels. In the genome imputation results using the Korean reference panel with the best performance, we identified variants that were not called in the WGS data. Among these, 34.7% were determined to be filtered variants that did not meet quality threshold criteria during the WGS variant calling process.

The outstanding performance of genome imputation using the Korean reference panel in the genetically homogeneous minority population of Koreans highlights the importance of developing ethnic-specific reference panels for the full utilization of genome imputation analysis. This also suggests new applications of genome imputation in Deep WGS.

I. INTRODUCTION

I-1. Whole Genome Sequence and genotype array

To generate genotype data for the analysis of the association between diseases and the genome, two methods are typically used: Whole Genome Sequencing (WGS) and Genotype array. WGS captures most of Single Nucleotide Variants (SNVs) and short INDELs, and it provides good accessibility to rare variants [1]. However, it remains expensive and places a significant burden on analysis resources [1, 2]. On the other hand, the genotype array detects only the genetic variations of interest within the entire genome but has limited access to rare variants. It is cost-effective and requires fewer analysis resources. In addition, ungenotyped variants can be inferred via genome imputation. In many large-scale studies, genotype array is still efficient. Consequently, genotype imputation has become a standard procedure for genetic analysis of the association between the genome and various phenotypes [3].

I-2. Genome imputation

In order to understand genome imputation, it is necessary to know the background on which the genotype array was created. First, when designing probes used for genotyping, not all SNPs are considered, but haplotypes are classified by linkage disequilibrium (LD) block, For each LD block, a representative SNPs called tag SNPs are selected and designed as a probe. Considering this process in reverse, if there is reference data that can estimate the LD structure, the genotype of the ungenotyped SNP can be inferred using the tag SNP information and reference data. This process is called genome imputation [3]. Many researches have been conducted on the methodology of genome imputation analysis based on various algorithms, and many high-performance tools based on the hidden-Markov algorithm that are freely available have been developed [4]. The typical construction of an efficient pipeline in current

practice involves two main steps: pre-phasing, which is the haplotype estimation process, and imputation, which is the inference of genotypes based on the determined haplotypes [5].

The performance of genome imputation is influenced by various factors, with the sample size of the reference genome panel and the ethnic similarity between the input data being particularly crucial [6]. The advancement of sequencing technologies has led to the generation of large-scale genomic data. As emphasis on genetic diversity has increased in reference panels, Large-size multi-ethnic reference panels have been developed [7]. Most of these panels are publicly available and can be easily downloaded and used for analysis, either directly or through imputation web tools.

I-3. Related studies

Sample size and ethnic similarity have been reported to play a crucial role in genome imputation, and high-resolution large-scale reference panels with ensured ethnic similarity are expected to be particularly important for the accurate inference of rare variants Although genome imputation cannot fully approximate WGS, previous study has demonstrated genome imputation performance approximating WGS with specific MAF thresholds (MAF $\geq$ 0.14% in African ancestry, MAF $\geq$ 0.11% in Hispanic/Latinx ancestry, and MAF $\geq$ 0.84% in Finnish ancestry), depending on the selection of genotyping arrays, reference genome panels, and sample ancestry [8].

A recent study showed the result of imputation of low coverage WGS with quality equivalent to high coverage sequence using large-scale reference sequence data. So, as the available sequence data are increased, the utility of genome imputation is on the rise [9].

However, most publicly available reference panels are European-centric, and performance evaluations have also been studied using European data. So, genetically homogeneous populations with less public data available such as East Asians have the

smallest proportion in the composition of large multi-ethnic reference panels. Only limited performance evaluation studies have also been conducted.

I-4. Research purpose

In most multi-ethnic reference panels, the proportion of Asians is quite low, of which very few East Asians exist. And other East Asian-specific reference panels also have very small sample sizes [10]. So, performance evaluation study in genetically homogeneous minority population has been limited. Fortunately, a publicly available Korean reference panel has recently been released. In this study, I conducted a performance evaluation analysis using the whole-genome sequencing (WGS) and genotype array data of 2,253 Korean individuals as an example of a genetically homogeneous minority population. And confirmed the benefits of high-performance genome imputation using large-size reference panels with ethnic similarity.

II. MATERIALS AND METHODS

II-1. Genotype array data

Genotype array data was called from KoreanChip [11] array platform with buccal and blood samples collected from 8K Korean subjects.

Quality control of the 8K Korean genotype array was performed using PLINK 1.9 [12]. Samples with a missing SNP rate exceeding 5% and a heterozygosity rate deviating by 3 standard deviations from the mean were excluded from the analysis. To remove batch effects, samples located outside the cluster were eliminated using both Multidimensional Scaling (MDS) and Principal Component Analysis (PCA). Additionally, duplicate or closely related samples, as well as those with gender mismatches, were removed. SNPs with low call rates and and Hardy-Weinberg Equilibrium test p-values lower than 1e-6 were excluded.

Next, I extracted 2,253 individuals for whom WGS data were available. Ultimately, samples from 2,253 people and approximately 600K SNPs were used for analysis.

II-2. WGS data

WGS data was sequenced at 30X depth on Illumina Novaseq6000 using whole blood from 2,253 Korean subjects. The Truseq PCR-Free Prep library kit was used for the sequencing library, and VCF results were obtained using BWA-mem [13] for alignment and GATK4 [14] for variant call.

gVCF files which contain reference information were used for the comparison analysis. We combined the gVCF files present at the sample level into a single file. Then, we filtered for variants with a “PASS” status in the “FILTER” column, indicating high-quality variants and generated a single compressed gVCF file which has only genotype information for these variants using Bcftools 1.17 [15].

II-3. Reference panels for genome imputation

Four reference panels were selected among the conventionally used reference panels considering each ethnicity and genome size (Table 1). 1000 Genom [16] (East Asian) was selected as the East Asian specific reference panel, Haplotype Reference Consortium (HRC) [7] and Trams-Omics for Precision Medicine (TOPMed) R2 [17] which is a largest multi-ethnic panel were selected as the large-size multi-ethnic reference panel. As a Korean reference, I create a Korean reference panel using 3,330 Korean WGS data. After pre-phasing the VCF file, I tried various processes such as VCF modification and missing genotype processing. However, due to resource limitations such as file size and processing time, The analysis progressed at an exceedingly slow pace. Fortunately, a Korean reference panel consisting of data from approximately 4,700 Koreans was developed and released. So, Korean Imputation Service (KIS) Phase1 Panel was selected. This is a largest Korean reference with sample size 4.7K [18].

Table 1. Reference panels for imputation

Reference panel	Ancestry	Sample size	Genome size
Korean Reference Panel	Korean	4,799	38M
1000 Genome-East Asian	East Asian	525	49M
Haplotype Reference Consortium	Multi-ethnic	32,470	39M
TOPMed R2	Multi-ethnic	97,256	308M

II-4. Genome imputation

Genotype array QC was processed by Plink 1.9. QCed Plink BED format files were converted to VCF and re-aligned swapped alleles. And then, I generated compressed VCF separated by chromosome for the genome imputation using bcftools (version 1.17).

I used the genome imputation pipeline from Michigan Imputation Service [19] consisting of Eagle (version 2.4) [20] for the pre-phasing and Minimac4 [5] for the imputation. Imputation was performed using three online tools: Korean Imputation Service (KIS), Michigan Imputation Service (MIS), and TOPMed Imputation Service (TIS). After imputation, imputed variants were lifted over to hg38 build for the comparison with WGS data using Picard [21]. Only Imputed variants with R2 score of 0.8 or higher passed QC, and only bi-allelic SNPs were considered in the analysis.

II-5. Performance measure of imputation result

To measure the accuracy of imputation, whole-genome sequencing data were used as the truth data set. Imputed SNPs were categorized into five groups: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), and Misclassified (MC) (Table 2). And then, the following performance metrics were calculated by MAF.

$$\text{Recall} = TP / (TP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Concordance} = TP / (TP + FP + MC)$$

Table 2. Classification for performance evaluation

Classification	Truth Sample (WGS)	Imputed Sample
True Positive (TP)	ALT	ALT
True Negative (TN)	REF	REF
False Positive (FP)	HOM_REF	HET_REF_ALT or HOM_ALT
False Negative (FN)	ALT	REF
Misclassified (MC)	HET_REF_ALT	HOM_ALT
	HOM_ALT	HET_REF_ALT

III. RESULTS

III-1. Study overview

I imputed the genotype array data of 2,253 Koreans using four different reference panels. Then, I assessed the imputation quality and the performance against WGS genotypes as the true genotypes (Figure 1).

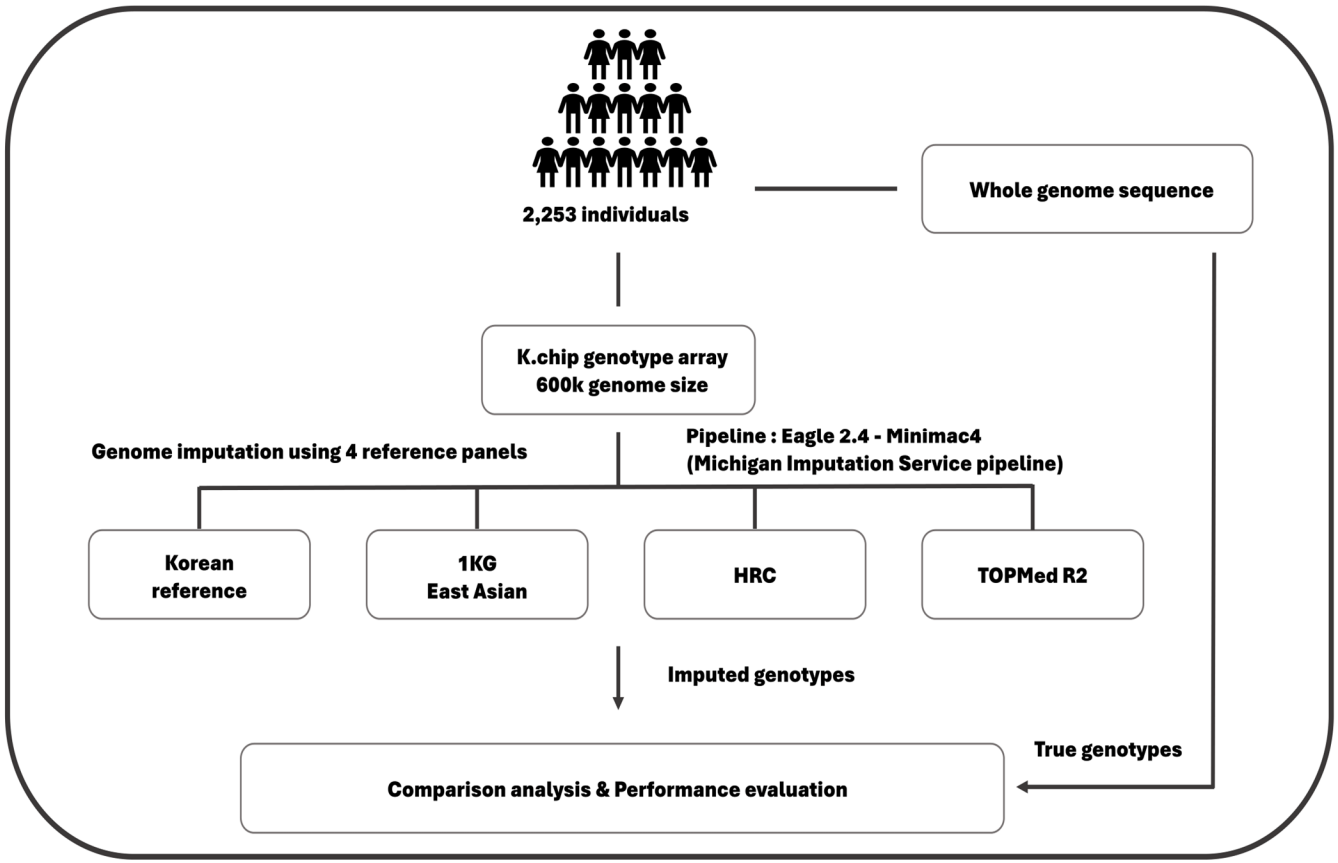

Figure 1. Study overview

III-2. Imputation result

The imputed genome size was similar to the size of each reference panel. When using the largest reference panel, TOPMed R2, there were overwhelmingly more variants present. However, after R2 score filtering, when using the Koren reference, the loss rate of SNP and INDEL variants was approximately 50%, with a majority of variants passing through the filtering. In the case of TOPMed R2, the loss rate for SNP variants was 97%, indicating that most of the variants were imputed with low quality. The results for the 1000 Genomes and HRC panels showed similar loss rates, with approximately 12.8% and 15% respectively, and the number of variants was also similar after filtering (Table 3).

Using the R2 score as a correlation metric between the reference panel and imputed variants, I analyze the median R2 score for each reference panel according to Minor Allele Frequency (MAF) bins and the proportion of well-imputed SNPs ($R2 \geq 0.8$) within each MAF bin. When using the Korean reference panel, the proportion of high-quality SNPs among those with a frequency lower than 1% was the highest at 25%, and the median R2 value was also the highest at 0.53. However, in the results from other reference panels, most of the rare variants had low quality. For SNPs with a frequency higher than 1%, the Korean reference panel also showed the best quality, while the 1000 Genomes reference panel had the lowest quality (Figure 2).

Table 3. Imputation result

Reference panel	Imputed		R2 ≥ 0.8	
	SNP	INDEL	SNP	INDEL
Korean	31,363,307	5,615,139	15,816,953	2,793,667
1KG_EAS	43,280,933	3,233,367	5,575,408	638,433
HRC	38,568,539	None	5,995,467	None
TOPMed R2	256,013,554	19,750,125	7,916,785	588,975

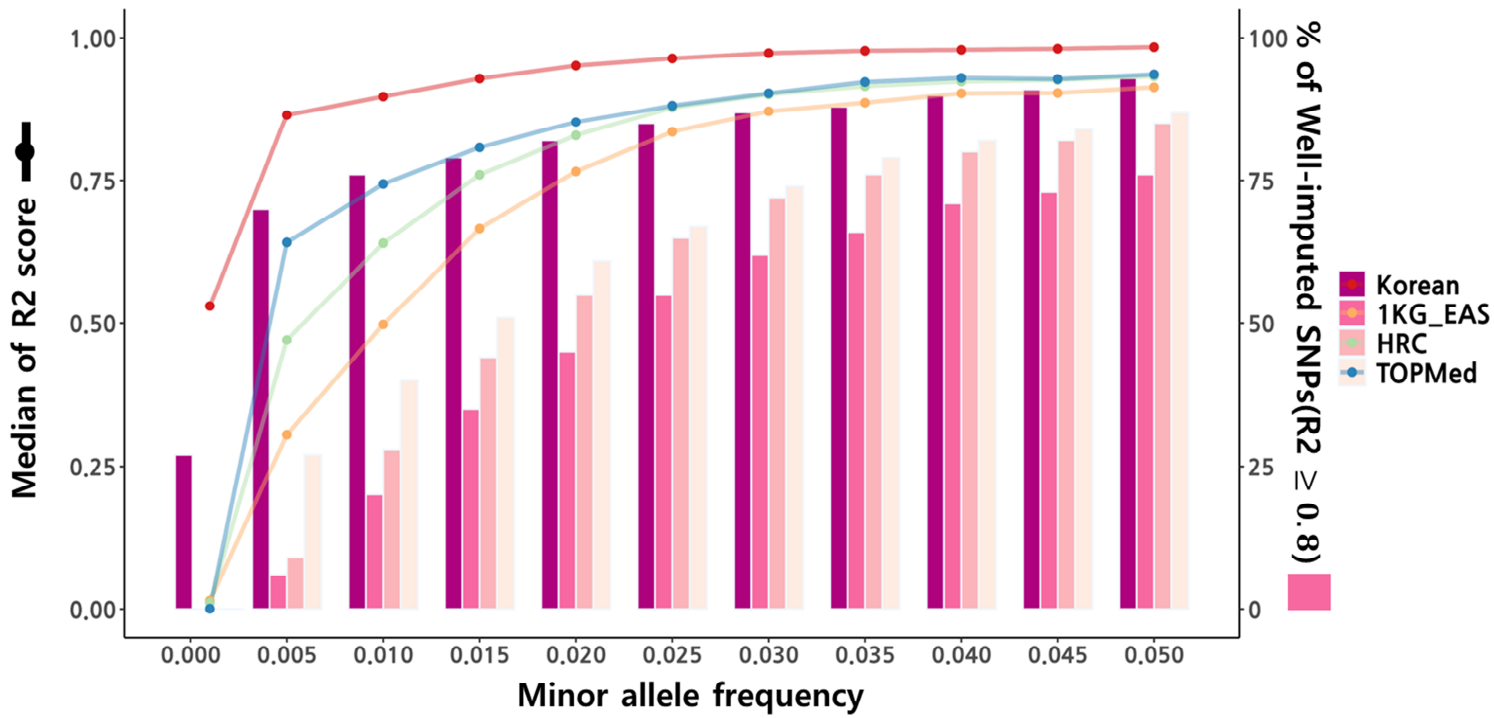

Figure 2. Overall imputation quality

III-3. Imputation performance

I analyzed how well the imputed genome approximates WGS using Minor Allele Frequency (MAF) (Figure 3) and three performance metrics: recall (Figure 4), precision (Figure 5), and concordance (Figure 6).

In the case of MAF, the study examined the differences between the imputed SNP MAF and the true WGS MAF for each SNP position. The x-axis shows the absolute difference in MAF, and the y-axis shows the number of variants. Only well-imputed SNPs with an R2 score of 0.8 or higher were used in the analysis. When there is a high degree of ethnic similarity between the reference panel and the input data, the differences in MAF are smaller. As expected, the Korean reference panel showed the smallest differences in results, while the other three panels were similar to each other but had larger differences compared to the results from the Korean reference panel.

To make these differences more evident, I specifically evaluated performance for rare variants with MAF lower than 1%. Similar to the imputation quality, when using the Korean reference panel, I observed significantly improved accessibility for rare variants. It demonstrated the least variation across all performance metrics and, on average, outperformed the results obtained using other reference panels. In the results obtained using TOPMed R2, which was the second-best in terms of quality, it was observed that the proximity to WGS was the lowest, in contrast to the imputation quality.

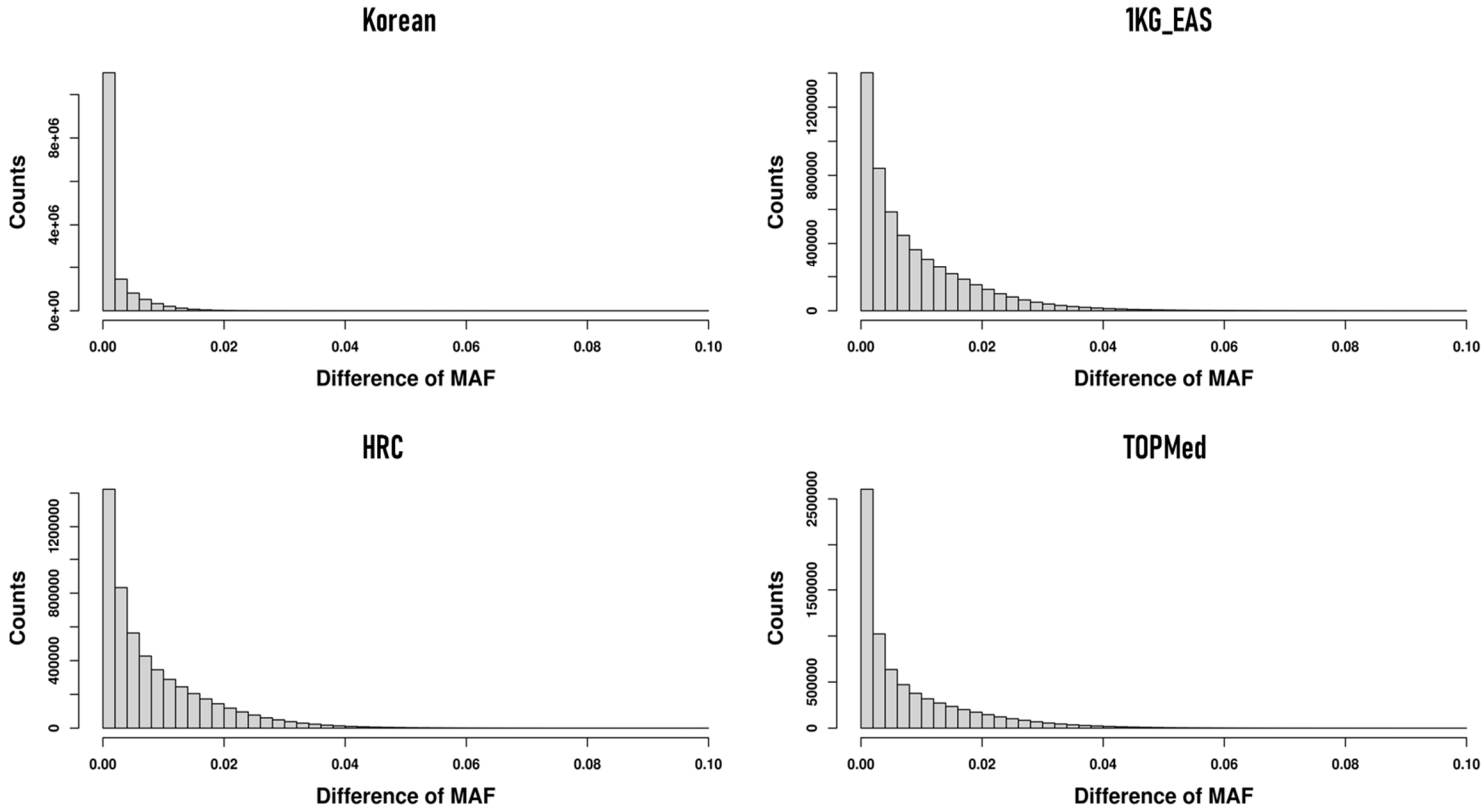

Figure 3. Difference between WGS MAF and imputed MAF

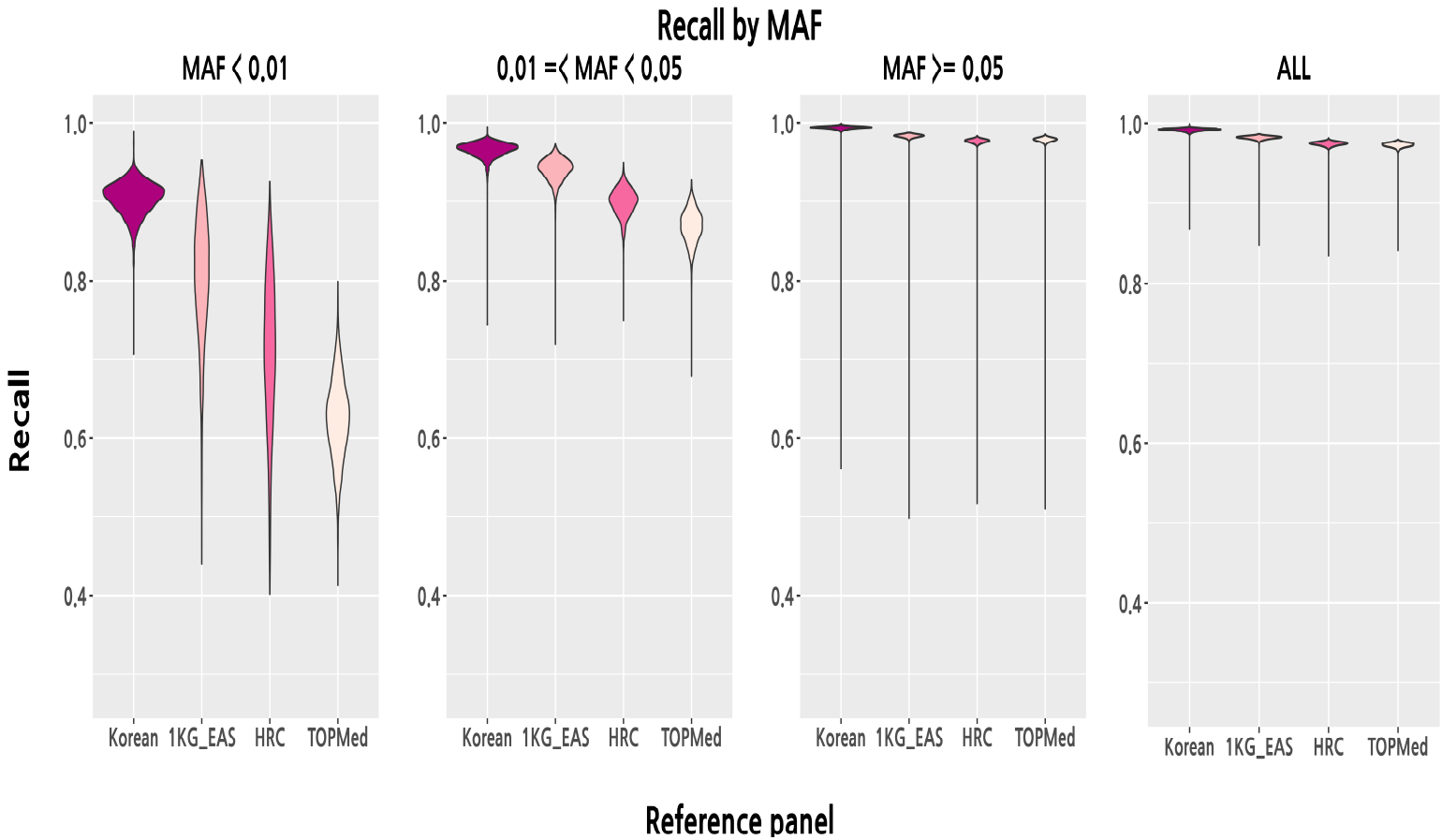

Figure 4. Recall by MAF

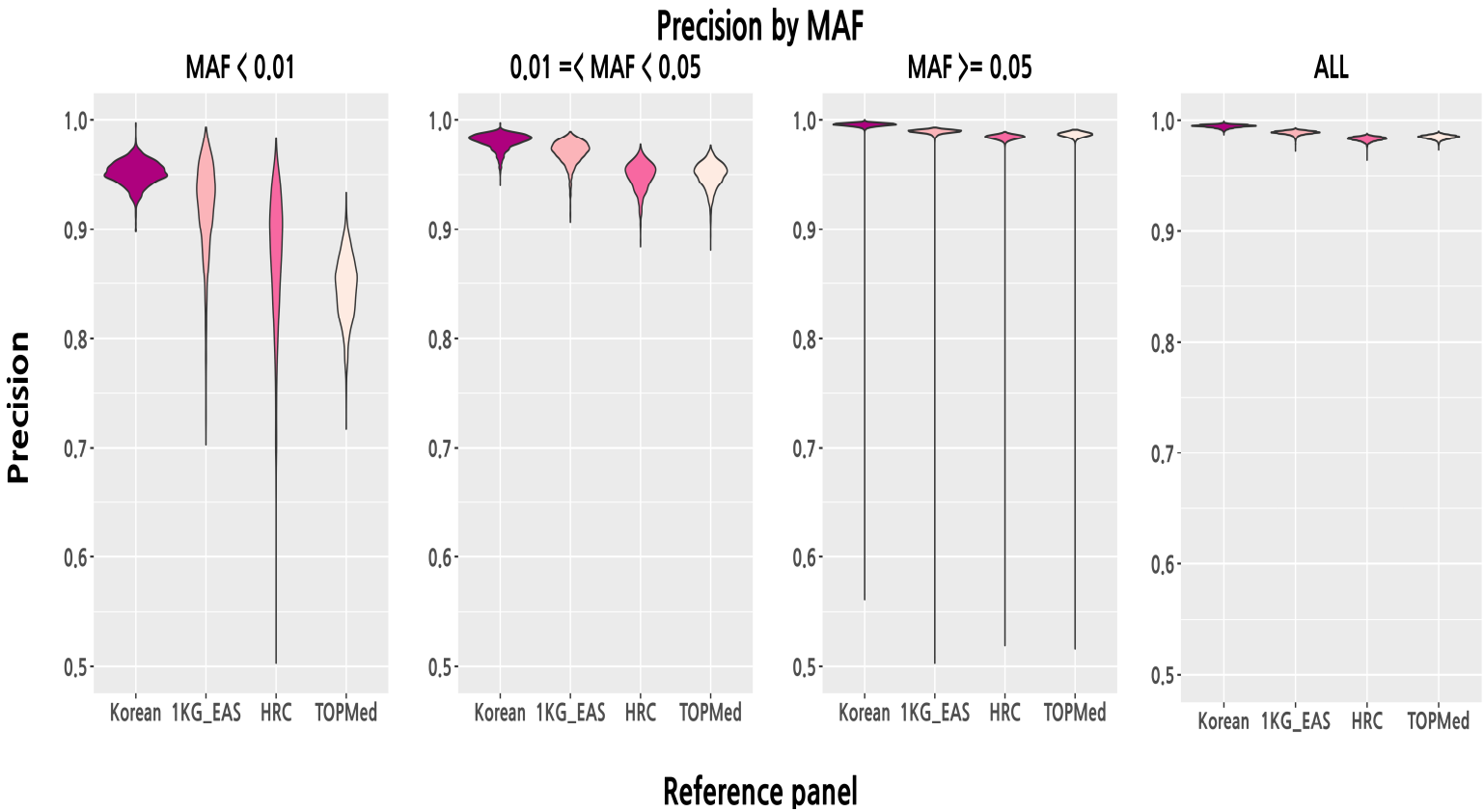

Figure 5. Precision by MAF

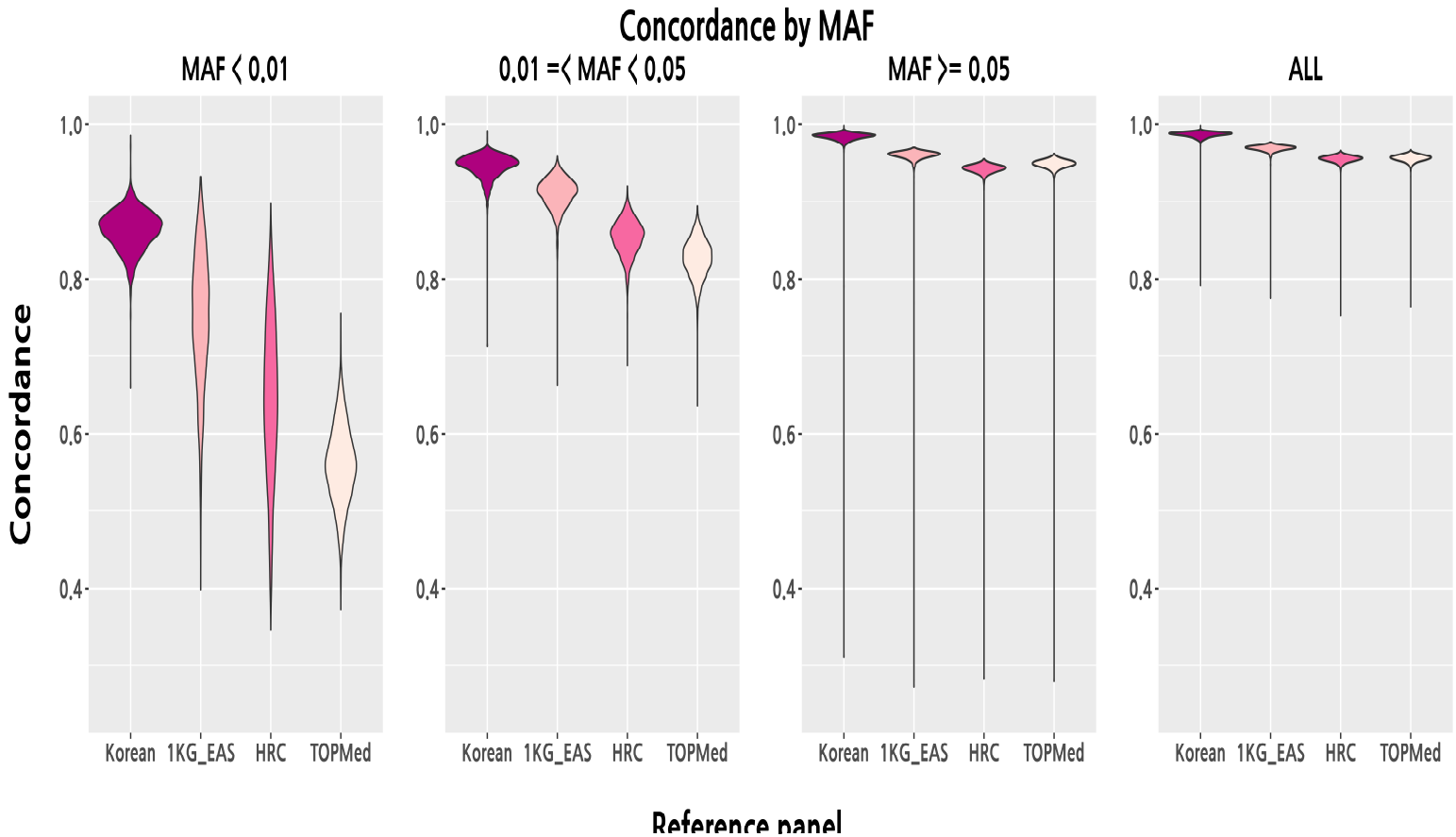

Figure 6. Concordance by MAF

III-4. Correction of genotyped SNPs

When using the Michigan Imputation Service pipeline, genotype imputation occurs not only for ungenotyped SNPs but also genotyped SNPs. During the imputation process, genotyped SNPs are corrected to appropriate genotypes by reverse estimation based on nearby imputed SNPs. This correction may involve changing correct genotypes into incorrect genotypes, which is considered as an error. I assessed and counted the genotype correction errors in the results from each panel, and found that the error count was lowest in the Korean reference panel results and highest in the HRC reference panel results (Figure 7).

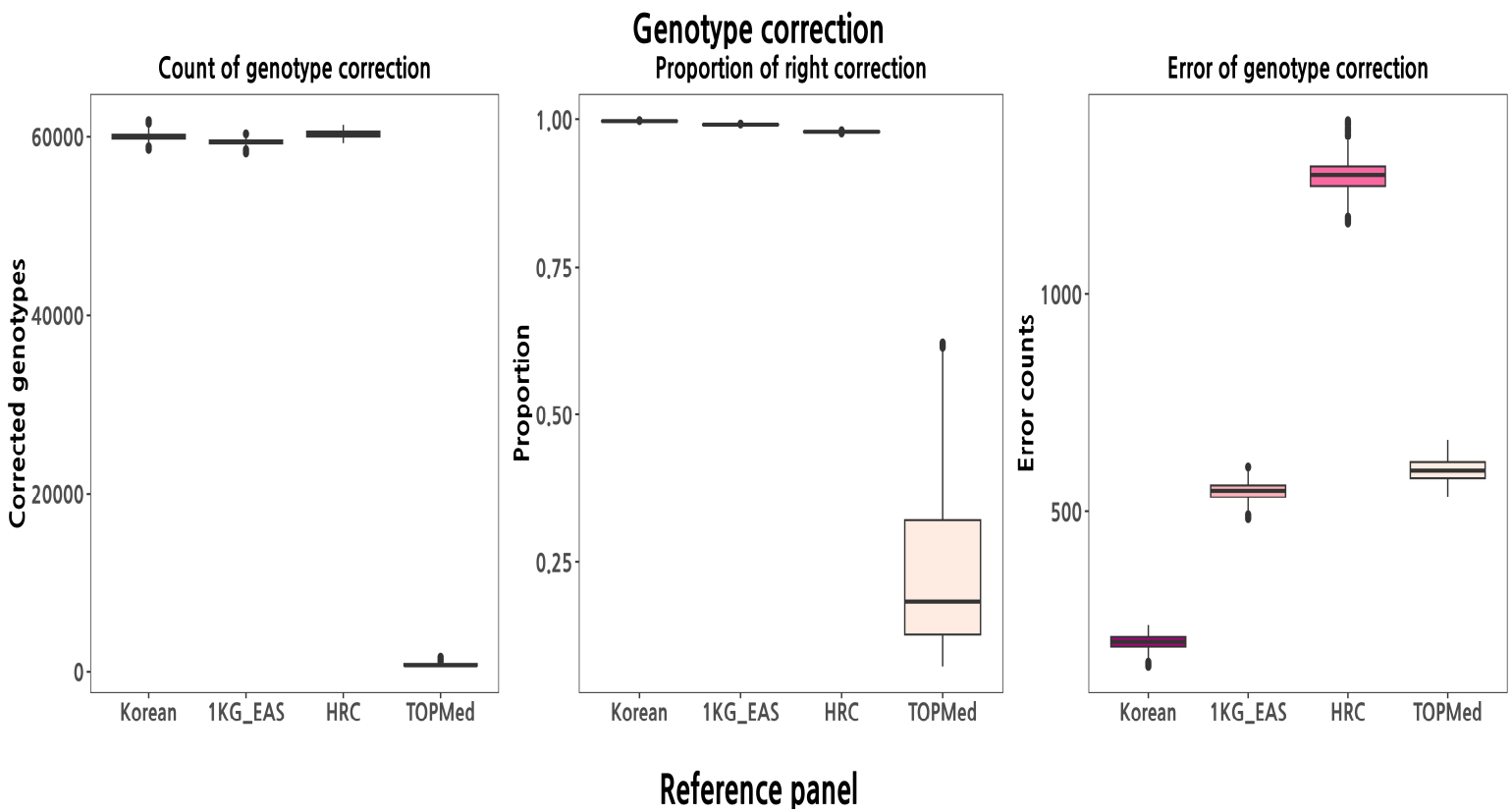

Figure 7. Genotype correction in each reference panel

III-5. Genome imputation with WGS

To determine the unique variants among the imputed variants, I overlapped SNP data from five different reference panels. Results using the Korean reference panel revealed the highest number of unique Imputed SNPs, approximately 770,000 SNPs, followed by the TOPMed R2 results with 500,000 SNPs, as the second-highest. Most of the SNPs from the Korean reference panel, which showed the closest performance to WGS, overlapped with WGS data. However, around 1.1 million SNPs were exclusively present in the Imputed SNP dataset. I analyzed unfiltered VCF data from the WGS dataset and identified that some of the SNPs unique to Korean reference panel were present in the unfiltered WGS data. Among these, 33.2% failed to pass the threshold during the GATK4 variant calling's Variant Quality Score Recalibration (VQSR) process, while 1.5% were identified as SNPs with excessive heterozygosity rates (Figure 8). This demonstrates that genome imputation can be a method to recover lost information from deep whole-genome sequencing when using an appropriate reference panel, ensuring performance (Figure 9).

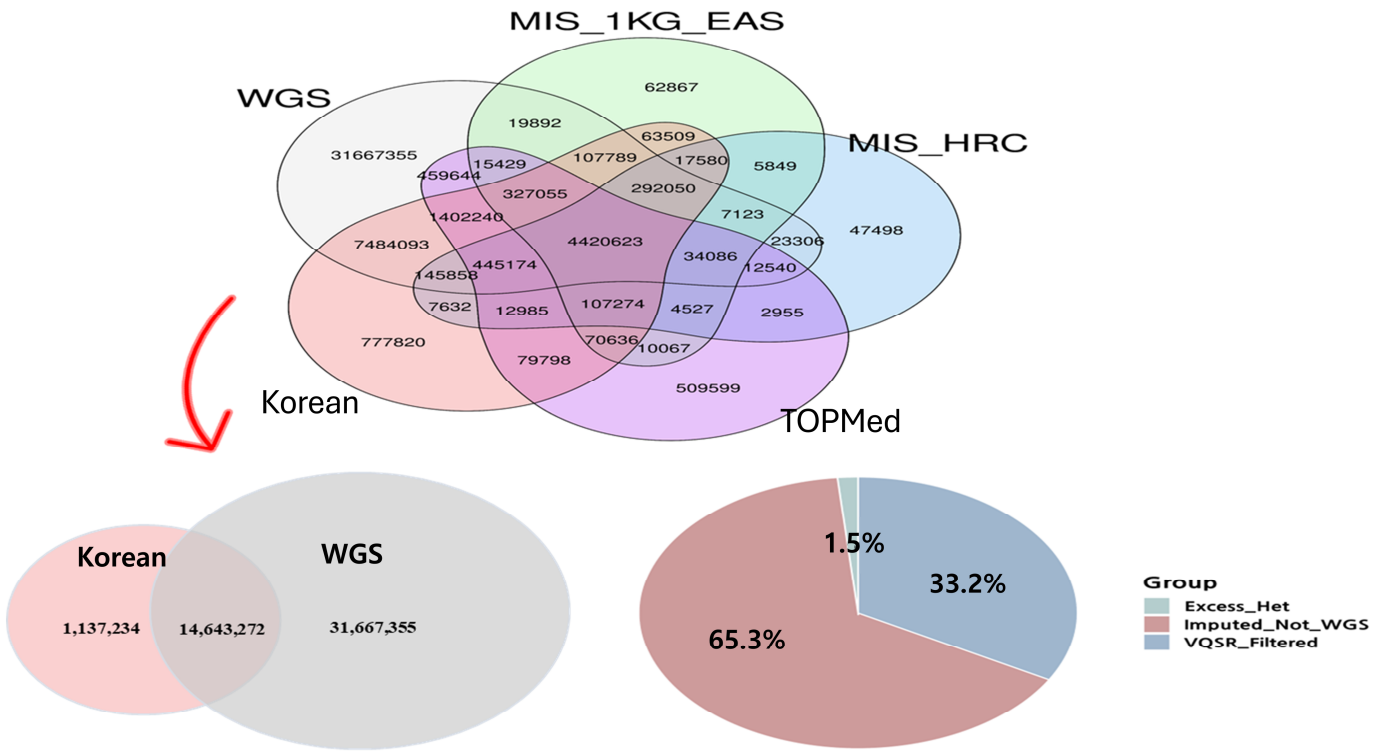

Figure 8. Non-overlapping SNPs with WGS

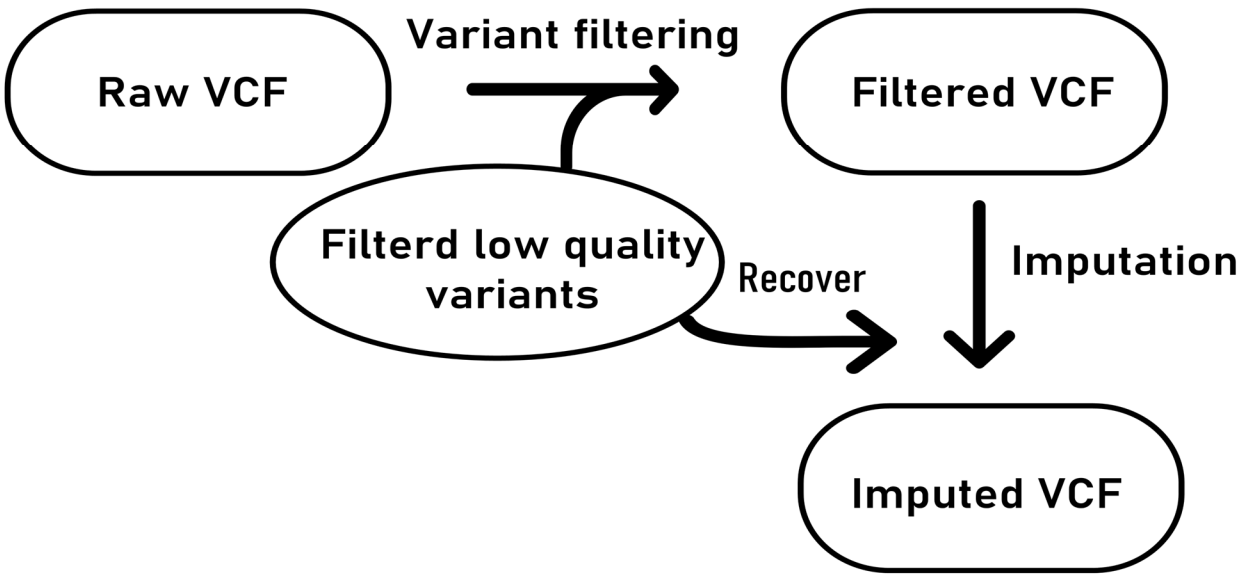

Figure 9. Utilization of genome imputation in deep WGS

IV. DISCUSSION

With the increase of large-scale sequencing data, the size of reference panels has also increased. Consequently, the performance of genome imputation has improved, and various applications are being explored. However, in populations with genetic homogeneity, such as Koreans, who are almost absent in reference panels of other ethnicities, it is expected that imputation performance may be lower compared to Europeans. This study use a large-scale dataset of Korean genome data to evaluate imputation performance and demonstrate the utility of high-performance genomic imputation.

In the imputation results, the Korean reference panel, which satisfies both the sample size of the reference panel and the overlap between the input genotype array and the reference panel, has the highest number of high-quality imputed variants, especially rare variants that are rarely found in other reference panels. there was. The results using 1000 Genome have the second highest ethnic similarity, but are considered to be of slightly lower quality than the results of the HRC reference panel due to the small sample size. The TOPMed reference panel has a sample size three times larger than the HRC panel, and because it includes South Asians, it appears that it was possible to impute more common variants than the HRC panel.

Comparison with whole-genome sequencing revealed significantly different results for genome imputation with the Korean dataset, emphasizing high racial similarity. Particularly, results using the TOPMed reference panel contradicted imputation quality assessed by R2 score. Although the correlation with the reference panel was better for the Korean panel than for the other two panels, it showed the lowest performance when compared to actual WGS data.

In the correction of genotyped SNP errors, the Korean reference panel had the least errors, similar to the 1000 Genomes and TOPMed panels, while the HRC panel had the most errors. This discrepancy is interpreted as a result of the impact of sample size since the probe design process of genotyping arrays considers variants with certain frequencies. However, beyond error count, the Korean panel demonstrated a higher rate of accurately correcting genotypes, offering a means to stabilize genotype calling error.

After the advent of WGS, the performance study of genome imputation became possible to conduct a clearer analysis by comparing WGS genotypes with the correct answer rather than evaluating performance through array masking. Previous large-scale studies have shown that genome imputation using race-specific array platforms and reference panels improved imputation quality for rare variants, showing that high-quality genomic imputation can partially replace WGS. We confirmed the possibility that high-performance genome imputation can be used in deep WGS in addition to low-depth WGS, which has low coverage. Using the Korean reference panel, we confirmed that some of the variants imputed with high accuracy were variants that did not exceed the quality threshold in WGS variant calling process. This suggests that genome imputation can be used as a method to increase WGS coverage.

So, the overwhelming performance of genome imputation analysis using the Korean reference panel in Koreans, a genetically homogeneous minority population, suggests the importance of developing a ethnic-specific reference panel for full utilization of genome imputation analysis.

V. 초 록

전유전체서열분석(WGS) 자료의 유전체 대치분석 필요성 재고

서 윤 종

지도교수 : 김 정 수

글로벌바이오융합학과

조선대학교 대학원

유전체 대치분석은 유전체와 여러가지 표현형 사이의 연관성 분석을 위한 유전형 분석에 가장 중요한 표준 절차이다. 그러나 유전체 대치분석의 활용과 중요성에 비해 유전적으로 동질한 많은 소수 인종은 참조유전체 패널의 구성부터 대치분석의 성능 평가까지 매우 제한된 연구만이 수행된 상황이다. 이 연구에서 우리는 보다 정확하고 의미 있는 성능 평가를 위해 유전적으로 동질한 소수인종의 예시인 한국인의 2,253 명 대규모 WGS 와 유전자형 어레이를 이용하여 대치의 결과가 WGS 에 얼마나 근사하는지 분석했다. 대치에 사용한 참조유전체 패널은 주로 사용되는 가용한 패널들 중 각각의 특성을 고려해 4 가지 참조 유전체 패널 한국인 참조 패널, Haplotype Reference Consortium (HRC), 1000 Genome, Trans-Omics for Precision Medicine (TOPMed)을 선정해 분석에 사용했다. 예상했듯이 모든 성능 지표 (Recall, Precision, Concordance)에서 한국인 참조 패널을 사용한 결과가 가장 성능이 좋았다. 특히, MAF 가 1% 미만인 변이에서 다른

참조패널보다 압도적으로 정확도가 높았다. Michigan Imputation Service 의 파이프라인을 이용했을 때, 대치된 유전자형 정보를 통해 이미 결정된 유전자형의 교정이 일어나는 경우가 있는데, 이러한 경우에서의 WGS 와의 유전자형 불일치 개수를 비교한 결과 한국인 참조패널에서 유전자형 교정의 오류가 가장 적었다. 성능이 가장 좋은 한국인 참조패널을 사용한 유전체 대치 결과에서 WGS 에서 호출되지 않은 변이가 존재했고, 그 중 34.7%가 WGS 변이 호출 과정에서 품질 임계 값을 만족하지 못해 필터링 된 변이임을 확인했다. 유전적으로 동질한 소수인종인 한국인에서 한국인 참조유전체 패널을 사용한 유전체 대치분석의 압도적인 성능은 유전체 대치분석의 온전한 활용을 위한 인종 특이적인 참조유전체 패널 개발의 중요성을 시사하고 있다.

VI. REFERENCES

1. Sazonovs, A. and J. Barrett, *Rare-variant studies to complement genome-wide association studies.* Annual review of genomics and human genetics, 2018. **19**: p. 97-112.
2. Tam, V., et al., *Benefits and limitations of genome-wide association studies.* Nature Reviews Genetics, 2019. **20**(8): p. 467-484.
3. Li, Y., et al., *Genotype imputation.* Annual review of genomics and human genetics, 2009. **10**: p. 387-406.
4. Nothnagel, M., et al., *A comprehensive evaluation of SNP genotype imputation.* Human genetics, 2009. **125**: p. 163-171.
5. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* Nature genetics, 2012. **44**(8): p. 955-959.
6. Mitt, M., et al., *Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel.* European Journal of Human Genetics, 2017. **25**(7): p. 869-876.
7. *A reference panel of 64,976 haplotypes for genotype imputation.* Nature genetics, 2016. **48**(10): p. 1279-1283.
8. Hanks, S.C., et al., *Extent to which array genotyping and imputation with large reference panels approximate deep whole-genome sequencing.* The American Journal of Human Genetics, 2022. **109**(9): p. 1653-1666.
9. Rubinacci, S., et al., *Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. bioRxiv 2022.* Google Scholar.
10. Choi, J., et al., *A whole-genome reference panel of 14,393 individuals for East Asian populations accelerates discovery of rare functional variants.* Science Advances, 2023. **9**(32): p. eadg6319.
11. Moon, S., et al., *The Korea Biobank array: design and identification of coding variants associated with blood Biochemical traits. Sci Rep 2019; 9: 1382*. PUBMED.

12. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* The American journal of human genetics, 2007. **81**(3): p. 559-575.
13. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows–Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-595.
14. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome research, 2010. **20**(9): p. 1297-1303.
15. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* Gigascience, 2021. **10**(2): p. giab008.
16. Via, M., C. Gignoux, and E.G. Burchard, *The 1000 Genomes Project: new opportunities for research and social challenges.* Genome medicine, 2010. **2**(1): p. 1-3.
17. D.Taliun., et al. *Sequencing of 53,831 diverse genome from the NHLBL TOPMed Program.* Nature. 2021, 590(7845), p.290-299
18. Hwang, M.Y., et al., *Analyzing the Korean reference genome with meta-imputation increased the imputation accuracy and spectrum of rare variants in the Korean population.* Frontiers in Genetics, 2022. **13**: p. 1008646.
19. Das, S., et al., *Next-generation genotype imputation service and methods.* Nature genetics, 2016. **48**(10): p. 1284-1287.
20. Loh, P.-R., et al., *Reference-based phasing using the Haplotype Reference Consortium panel.* Nature genetics, 2016. **48**(11): p. 1443-1448.
21. *"Picard Toolkit." 2019. Broad Institute, GitHub Repository. https://broadinstitute.github.io/picard/; Broad Institute*.

VII. APPENDIX

<Genotype array QC protocol>

R version = 4.1.1

- # require packages: library(foreach) library(doParallel)
- # IBD cutoff: pi_hat 0.2

- ## argument info
- # bfile: genome file
- # subjects_remove_list: error sample FID IID file
- # outname: output genome file prefix
- # dupliceted_ID_list: duplicated ObjectID's FID IID file, If you have duplicate subjects, you should remove them all.

- install.packages("foreach")
- install.packages("doParallel")
- library(foreach)
- library(doParallel)

- genomeQC_protocol <- function(bfile,subjects_remove_list, outname, dupliceted_ID_list=NULL){

- if(length(dupliceted_ID_list)==0){
- ## 1st QC
- # remove error sample

```
- system(paste("plink --bfile ",bfile," --remove ",subjects_remove_list," --allow-no-sex --make-just-fam --out QC0",sep=""))

- # remove low quality subject
- print("remove low quality subject")
- system(paste("plink --bfile ",bfile," --keep QC0.fam --missing --allow-no-sex --out Total_CR",sep=""))
- system(paste("plink --bfile ",bfile," --keep QC0.fam --het --allow-no-sex --out Total_HET",sep=""))

- callData <- read.table("Total_CR.imiss",header=T)
- hetData <- read.table("Total_HET.het",header=T)

- hetData <- cbind(hetData,het=((hetData[,5]-hetData[,3])/hetData[,5])*100)
- data <- merge(callData,hetData,by="FID")
- colnames(data)[2] <- "IID"

- png("low_quality_sample_before_plot.png")
- plot(data$het, data$F_MISS,xlab="Heterogyzosity",ylab="Proportion of missing SNPs")
- abline(v=c(mean(data$het)-3*sd(data$het),mean(data$het)+3*sd(data$het)),col=2,lty=2)
- abline(h=0.05)
- dev.off()

- remove <- which(data$F_MISS>0.05)
- remove <- c(remove,which(data$het>mean(data$het)+3*sd(data$het) | data$het<mean(data$het)-3*sd(data$het)))
- remove <- unique(remove)
```

- data2 <- data[-remove,]

- png("low_quality_sample_after_plot.png")
- plot(data2$het, data2$F_MISS,xlab="Heterogyzosity",ylab="Proportion of missing SNPs")
- dev.off()

- write.table(data[remove,c(1:2)],"low_quality_sampleID.txt",quote=F,row=F,col=F)

- system(paste("plink --bfile ",bfile," --keep QC0.fam --remove low_quality_sampleID.txt --allow-no-sex --make-just-fam --out QC1",sep=""))

- # SNP_pruning
- print("mds & pca analysis")
- system(paste("plink --bfile ",bfile," --keep QC1.fam --maf 0.1 --geno 0.01 --hwe 0.001 --indep-pairwise 50 5 0.2 --allow-no-sex --out snp_prune",sep=""))

- # calculate IBD
- system(paste("plink --bfile ",bfile," --keep QC1.fam --extract snp_prune.prune.in --genome --allow-no-sex --out prune_IBD",sep=""))

- # MDS & PCA analysis
- library(foreach)
- library(doParallel)
- core <- 2
- cl = makeCluster(core)
- registerDoParallel(cl)

```
- foreach(i=1:2) %dopar%{
- if(i==1){
- system(paste("plink --bfile ",bfile," --keep QC1.fam --extract snp_prune.prune.in --read-genome prune_IBD.genome --mds-plot 4 --cluster --allow-no-sex --out 1st_mds",sep=""))
- }else{system(paste("plink --bfile ",bfile," --keep QC1.fam --extract snp_prune.prune.in --pca --allow-no-sex --out 1st_pca",sep=""))
- }
- }

- system(paste("plink --bfile ",bfile," --keep QC1.fam --extract snp_prune.prune.in --read-genome prune_IBD.genome --mds-plot 4 --cluster --allow-no-sex --out 1st_mds",sep=""))
- system(paste("plink --bfile ",bfile," --keep QC1.fam --extract snp_prune.prune.in --pca --allow-no-sex --out 1st_pca",sep=""))

- # mds analysis
- mds <- read.table("1st_mds.mds",header=T)
- png("1st_mds_C1_C2_before_plot.png")
- plot(mds$C1,mds$C2,xlab="C1",ylab="C2")
- dev.off()

- print("Please decide the C1,C2 threshold then close the plot.")
- system("eog 1st_mds_C1_C2_before_plot.png")
- answer=""
- while (answer!="no" & answer!="n"){
- C1 <- as.numeric(unlist(strsplit(readline("C1 threshold (write minimum & maximum value of range ex;-0.5 0.05) : ")," ")))
- C2 <- as.numeric(unlist(strsplit(readline("C2 threshold (write minimum & maximum value of range ex;-0.015 -0.1) : ")," ")))
```

```
-	png("test.png")
-	plot(mds$C1,mds$C2,xlab="C1",ylab="C2")
-	abline(v=C1,col='red',lty=2)
-	abline(h=C2,col='red',lty=2)
-	dev.off()
-	system("eog test.png")
-	answer=readline("Change threshold? please answer yes(y) or no(n) : ")
-	}
-	system("mv test.png 1st_mds_C1_C2_before_plot.png")
-	ix.C1 <- which(mds$C1<= min(C1)|mds$C1>= max(C1))
-	ix.C2 <- which(mds$C2<= min(C2)|mds$C2>= max(C2))
-	ix.C <- unique(c(ix.C1,ix.C2))

-	if(length(ix.C)!=0){
-	png("1st_mds_C1_C2_after_plot.png")
-	plot(mds$C1[-ix.C],mds$C2[-ix.C],xlab="C1",ylab="C2")
-	dev.off()
-	}

-	remove_mds <- mds[ix.C,1:2]
-	write.table(remove_mds,"mds_correction.txt",row.names=F,quote=F)

-	# pca analysis
-	pca <- read.table("1st_pca.eigenvec")
-	png("1st_pca_pc1_pc2_before_plot.png")
-	plot(pca[,3],pca[,4],xlab="PC1",ylab="PC2")
-	dev.off()
```

```
- print("Please decide the pc1,pc2 threshold then close the plot.")
- system("eog 1st_pca_pc1_pc2_before_plot.png")
- answer=""
- while (answer!="no" & answer!="n"){
- pc1 <- as.numeric(unlist(strsplit(readline("pc1 threshold (write to minimum & maximum value of the range ex;-0.5 0.05) : ")," ")))
- pc2 <- as.numeric(unlist(strsplit(readline("pc2 threshold (write to minimum & maximum value of the range ex;-0.015 -0.1) : ")," ")))
- png("test.png")
- plot(pca[,3],pca[,4],xlab="PC1",ylab="PC2")
- abline(v=pc1,col='red',lty=2)
- abline(h=pc2,col='red',lty=2)
- dev.off()
- system("eog test.png")
- answer=readline("Change threshold? please answer yes(y) or no(n) : ")
- }
- system("mv test.png 1st_pca_pc1_pc2_before_plot.png")
- ix.pc1 <- which(pca[,3]> max(pc1) | pca[,3]< min(pc1))
- ix.pc2 <- which(pca[,4]> max(pc2) | pca[,4]< min(pc2))
- ix.pc <- unique(c(ix.pc1,ix.pc2))

- if(length(ix.pc)!=0){
- png("1st_pca_pc1_pc2_after_plot.png")
- plot(pca[-ix.pc,3],pca[-ix.pc,4],xlab="PC1",ylab="PC2")
- dev.off()
- }

- remove_pca <- pca[ix.pc,1:2]
- colnames(remove_pca) <- c("FID","IID")
```

- write.table(remove_pca,"pca_correction.txt",row.names=F,quote=F)
- write.table(unique(rbind(remove_mds,remove_pca)),"mds_pca_correction.txt",row.names=F,quote=F)

- system(paste("plink --bfile ",bfile," --keep QC1.fam --remove mds_pca_correction.txt --allow-no-sex --make-just-fam --out QC2",sep=""))

- # remove related sample
- print("remove related sample")
- system("awk '{if($10>0.2) print$1,$2}' prune_IBD.genome | uniq > relatedness_correction.txt")
- system(paste("plink --bfile ",bfile," --keep QC2.fam --remove relatedness_correction.txt --allow-no-sex --make-just-fam --out QC3",sep=""))

- # SEX check
- print("SEX check")
- system(paste("plink --bfile ",bfile," --keep QC3.fam --check-sex --allow-no-sex --out Check_sex",sep=""))
- system("grep PROBLEM Check_sex.sexcheck | awk '{if($4!=0) print$0}' > sex_problem.txt")
- system(paste("plink --bfile ",bfile," --keep QC3.fam --remove sex_problem.txt --allow-no-sex --make-just-fam --out QC4",sep=""))

- ## SNP QC
- # low call rate SNP
- system(paste("plink --bfile ",bfile," --keep QC4.fam --geno 0.05 --allow-no-sex --write-snplist --out snp_QC1",sep=""))

```
- # HWE
- system(paste("plink --bfile ",bfile," --keep QC4.fam --extract snp_QC1.snplist --hwe 1e-6 --allow-no-sex --write-snplist --out snp_QC2",sep=""))

- # make QC result bed file
- system(paste("plink --bfile ",bfile," --keep QC4.fam --extract snp_QC2.snplist --allow-no-sex --make-bed --out ",outname,sep=""))

- } else if(length(dupliceted_ID_list)==1){
- # remove duplicated object ID
- system(paste("plink --bfile ",bfile," --keep QC4.fam --remove ",dupliceted_ID_list," --allow-no-sex --make-just-fam --out QC5",sep=""))

- ## SNP QC
- # low call rate SNP
- system(paste("plink --bfile ",bfile," --keep QC5.fam --geno 0.05 --allow-no-sex --write-snplist --out snp_QC1",sep=""))

- # HWE
- system(paste("plink --bfile ",bfile," --keep QC5.fam --extract snp_QC1.snplist --hwe 1e-6 --allow-no-sex --write-snplist --out snp_QC2",sep=""))

- # make QC result bed file
- system(paste("plink --bfile ",bfile," --keep QC5.fam --extract snp_QC2.snplist --allow-no-sex --make-bed --out ",outname,sep=""))

- }
- }
```

<Genotype matching between imputed SNPs and WGS SNPs>

R version = 4.1.1

- args <- commandArgs(trailingOnly = TRUE)
- IMP = args[1]
- WGS = args[2]
- GARDWGSN = args[3]
- TEMP = args[4]
- res_DIR = args[5]
- library(data.table)
- library(parallel)
- library(tictoc)

- tic("Calculating_Concordance")
- print("Calculate_Start.....")

- setwd(paste(TEMP))

- ID <- read.table("/lustre/external/YJ/Imputation/IMP2
  /WGS_CHIP_2253.txt",sep = "₩t")

- chip <- as.character(unlist(read.table("/lustre/external/YJ
  /Imputation/IMP2/Imputed_sample_list.txt",sep = "₩t")))

- wgs <- as.character(unlist(read.table("/lustre/external
  /YJ/Imputation/Sequence_based_array/BU_VCF/WGS_samples.txt",sep =
  "₩t")))

- system(paste0("cat ",IMP," | awk '{print $1",'"₩t"',"$2",'"₩t"',"$",

```
    grep(ID$V2[which(ID$V1 == GARDWGSN)],
    chip)+2,"}' > ./IMP_",GARDWGSN))

- system(paste0("cat ",WGS," | awk '{print $1",'"₩t"',"$",
    grep(GARDWGSN,wgs)+1,"}' > ./WGS_",GARDWGSN))

- IMP_GT <- fread(paste0("./IMP_",GARDWGSN),sep="₩t",header = F)

- WGS_GT <- fread(paste0("./WGS_",GARDWGSN),sep="₩t",header = F)

- WGS_GT$V2 <- gsub("/","|",WGS_GT$V2)

- classification <- merge(IMP_GT,WGS_GT,by="V1",all.x=T)

- names(classification) <- c("POS","MAF","IMPUTED","WGS")

- classification$res <- rep(NA,nrow(classification))

- tmp_clss <- paste(classification$IMPUTED,classification$WGS,sep = ":")

- GT_classification <- strsplit(tmp_clss,":")

- numCores <- parallel::detectCores()-157

- myCluster <- parallel::makeCluster(numCores)

- res <- unlist(parallel::parLapply(cl = myCluster,
    X=GT_classification,function(x){
```

```
-     if(x[2] == "NA" | x[2] == "."){
-     print("0")
-     }else if(x[2] != "NA" | x[2] != "."){
-     if(x[1] == x[2]){

-     ###TP###
-     if(x[2] %in% c("1|0","0|1")){
-     print("HET_TP")
-     }else if(x[2] == "1|1"){
-     print("HOM_TP")

-     ####TN####
-     }else if(x[2] == "0|0"){
-     print("HOM_TN")
-     }

-     ####FP####
-     }else if(x[2] ==  "0|0" & x[1] != x[2]){
-     if(x[1] %in% c("0|1","1|0")){
-     print("HET_FP")
-     }else if(x[1] == "1|1"){
-     print("HOM_FP")

-     }
-     ###FN###
-     }else if(x[1] == "0|0" & x[1] != x[2]){
-     if(x[2] == "1|1"){
-     print("HOM_FN")
-     }else if(x[2] %in% c("0|1","1|0")){
```

```
-     print("HET_FN")
-     }
-     }else if(x[1] == "1|1" & x[2] %in% c("1|0","0|1")){
-     print("MC")
-     }else if(x[2] == "1|1" & x[1] %in% c("1|0","0|1")){
-     print("MC")
-     }else if(x[2] == "1|0" & x[1] == "0|1"){
-     print("HET_TP")
-     }else if(x[2] == "0|1" & x[1] == "1|0"){
-     print("HET_TP")
-     }}}))

-     classification$res <- res
-     toc()
-     print("Finished....")
-     system(paste0("rm ","./IMP_",GARDWGSN," ./WGS_",GARDWGSN))
```