Imputation from low-pass whole-genome sequencing data with GLIMPSE2 versus TOPMed

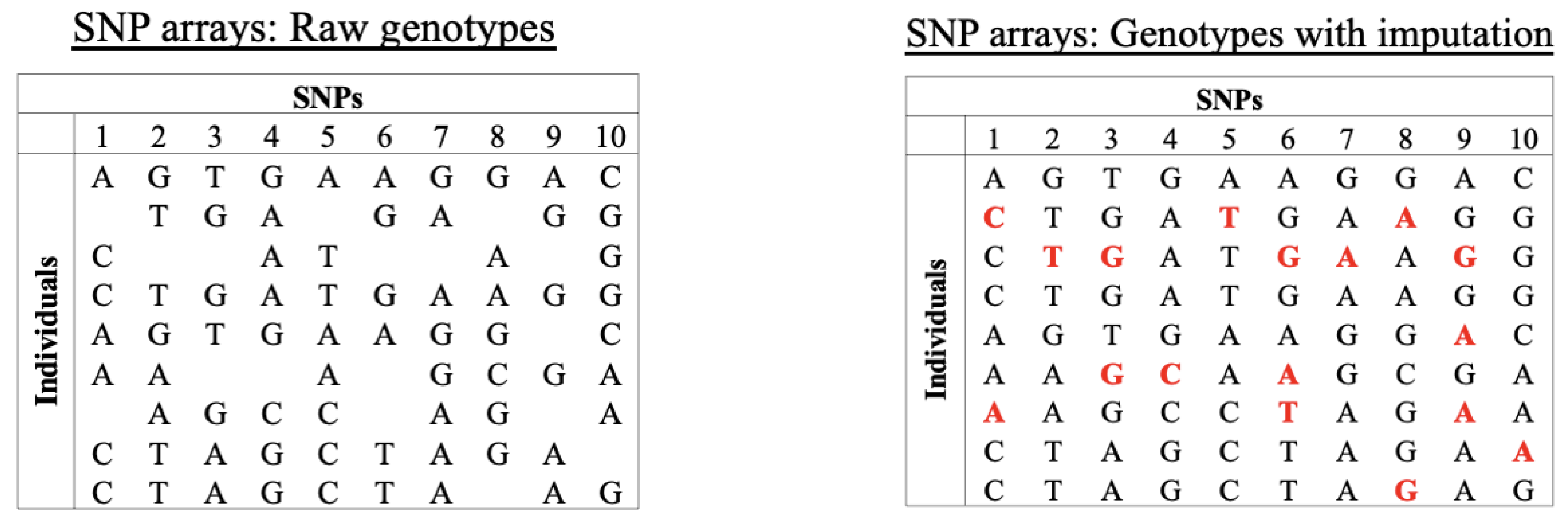
Zinhle Cindi[1], Yuki Bradford[1], Phumla Sinxadi[2,3], David W. Haas[4,5], Marylyn D. Ritchie[1,6]

[1]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Division of Clinical Pharmacology, Department of Medicine, University of Cape Town, Cape Town, South Africa
[3]SAMRC/UCT Platform for Pharmacogenomics Research and Translation, South African Medical Research Council, Cape Town, South Africa
[4]Vanderbilt University Medical Center, Nashville, TN 37232, USA
[5]Meharry Medical College, Nashville, TN 37208, USA
[6]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA

Perelman School of Medicine
University of Pennsylvania

Introduction

Genotyping arrays followed by imputation (**Figure 1**) have traditionally been the most prevalent method to assay human genetic variation. The introduction of DNA sequencing (**Figure 2**) has allowed for the detection of novel genetic variants across the entire allele frequency spectrum.



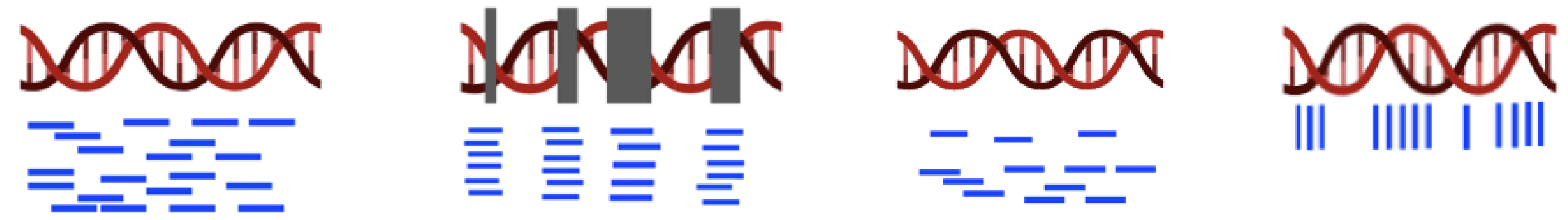
SNP arrays: Raw genotypes

Individuals	1	2	3	4	5	6	7	8	9	10
	A	G	T	G	A	A	G	G	A	C
		T	G	A		G	A		G	G
	C			A	T			A		G
	C	T	G	A	T	G	A	A	G	G
	A	G	T	G	A	A	G	G		C
	A	A			A		G	C	G	A
		A	G	C	C		A	G		A
	C	T	A	G	C	T	A	G	A	
	C	T	A	G	C	T	A		A	G

SNP arrays: Genotypes with imputation

Individuals	1	2	3	4	5	6	7	8	9	10
	A	G	T	G	A	A	G	G	A	C
	C	T	G	A	**T**	G	A	**A**	G	G
	C	**T**	**G**	A	T	**G**	**A**	A	**G**	G
	C	T	G	A	T	G	A	A	G	G
	A	G	T	G	A	A	G	G	**A**	C
	A	A	**G**	**C**	A	**A**	G	C	G	A
	A	A	G	C	C	**T**	A	G	**A**	A
	C	T	A	G	C	T	A	G	A	**A**
	C	T	A	G	C	T	A	**G**	A	G

Figure 1: Left, raw genotypes of individuals with missing data (empty cells). Right, imputed data set (red imputed genotype values).



	Whole-genome Sequencing (WGS)	Whole-exome sequencing	Low-pass WGS	Arrays
Coverage	High	High	Low	N/A
Relative cost	High	Medium	Low	Low
Novel SNP detection	✓	Limited	✓	X
Structural variant detection	✓	Limited	X	X
Copy number variant detection	✓	Limited	✓	Limited

Figure 2: Representation of sequenced regions in a genome using different approaches.

Methods

Using a test set of low-pass WGS data generated from five DNA samples, we performed imputation using both GLIMPSE2 and TOPMed panels. Unfiltered data were submitted to imputation pipelines. Concordance between output variants was assessed. We also checked for concordance between pre-imputation variant calls with GLIMPSE2 versus TOPMed. **Table 1** shows features of GLIMPSE2 and TOPMed imputation panels.

Table 1: GLIMPSE2 and TOPMed features

	GLIMPSE2	TOPMed
Reference panel	1KG Phase4	TOPMed-r3
Algorithm/Software	Gibbs sampler	Minimac4
Rare variants imputation	MAF < 0.1%	MAF < 0.1%

Results and Conclusions

There was 96.0% concordance among variants shared by datasets imputed by both GLIMPSE2 and TOPMed. **Table 2** shows imputation outcomes. With TOPMed, samples were uploaded individually onto server. This may be time consuming when imputing large datasets.

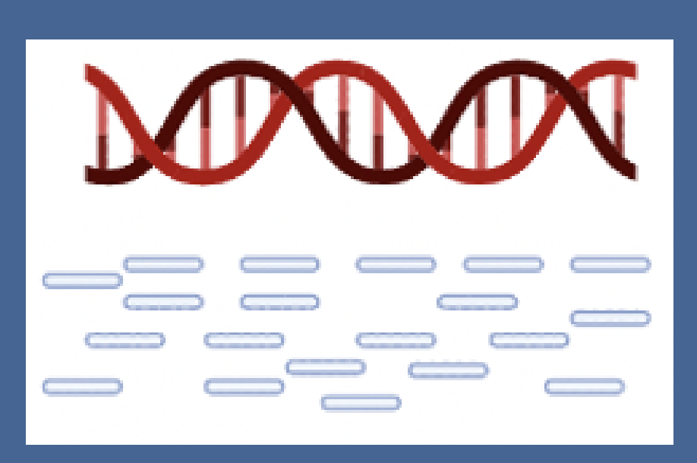
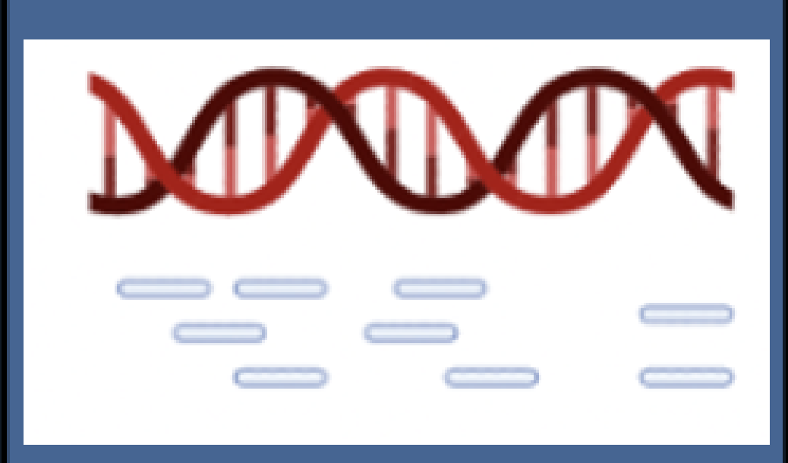
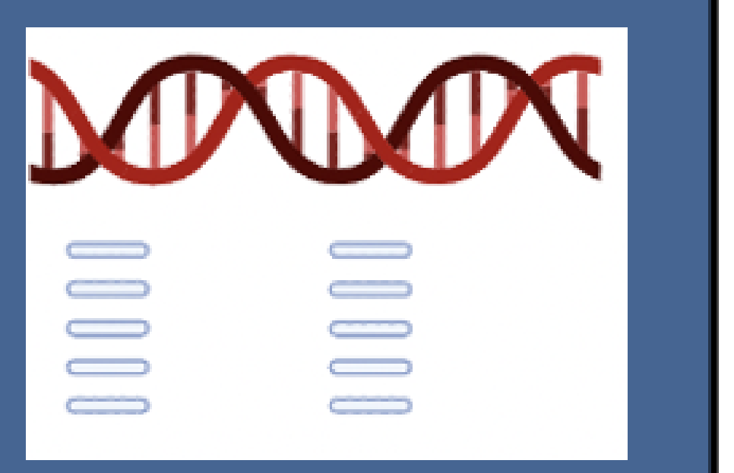
Table 2: GLIMPSE2 and TOPMed outcomes

	GLIPMSE2	TOPMed
Imputed variants	63,824,182	8,439,092
Non-missing	11,950,240	6,379,129
Concordance	88.3%	85.1%

Imputation of lp-WGS data with GLIMPSE2 yielded 7.6-times more variants than with TOPMed, indicating that TOPMed is not suitable for lp-WGS data. There was good agreement between variants generated by the two methods.

Future directions

We are interested in understanding the quality of lp-WGS with imputation in comparison to either genotype array with imputation or high depth WGS. To do this, we will evaluate the quality, coverage and concordance using 25 samples of African and 25 samples of European ancestry from the Penn Medicine Biobank (PMBB). We believe that these outcomes (**Table 3**) will provide insights in selecting the appropriate method to assay human genetic variation.

	WGS	Lp-WGS	Arrays
Sequenced regions in a genome			
Imputation method	TOPMed	GLIMPSE2	TOPMed
Reference panel	1GK Phase 4	1GK Phase 4	1GK Phase 4
Sample size	AFR = 25; EUR = 25	AFR = 25; EUR = 25	AFR = 25; EUR = 25

Acknowledgements: The authors are grateful for the support from AI077505 and the study participants.