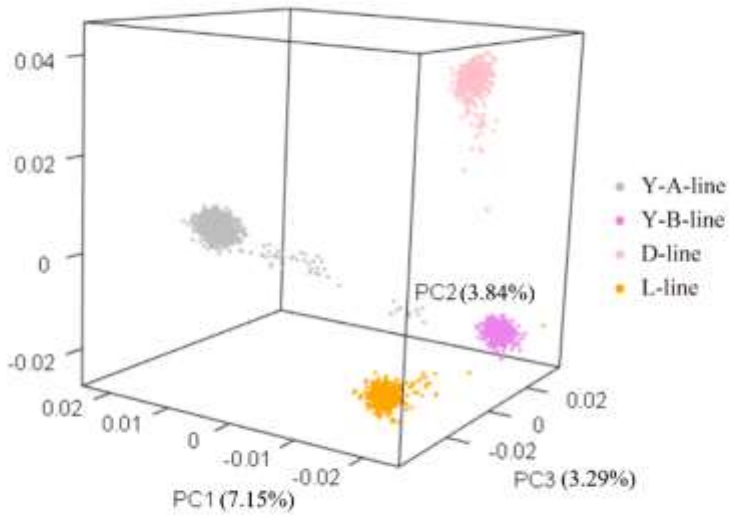
A comprehensive evaluation of factors affecting accuracy of pig genotype imputation using a single or multi-breed reference population

Kai-li ZHANG, Xia PENG, Sai-xian ZHANG, Hui-wen ZHAN, Jia-hui LU, Sheng-song XIE, Shu-hong ZHAO, Xin-yun LI*, Yun-long MA*

Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction, Ministry of Education, Key Laboratory of Swine Genetics and Breeding, Ministry of Agriculture, College of Animal Science and Technology, Huazhong Agricultural University, 430070 Wuhan, Hubei, P.R. China.

*Correspondence: Yunlong.Ma@mail.hzau.edu.cn and xyli@mail.hzau.edu.cn

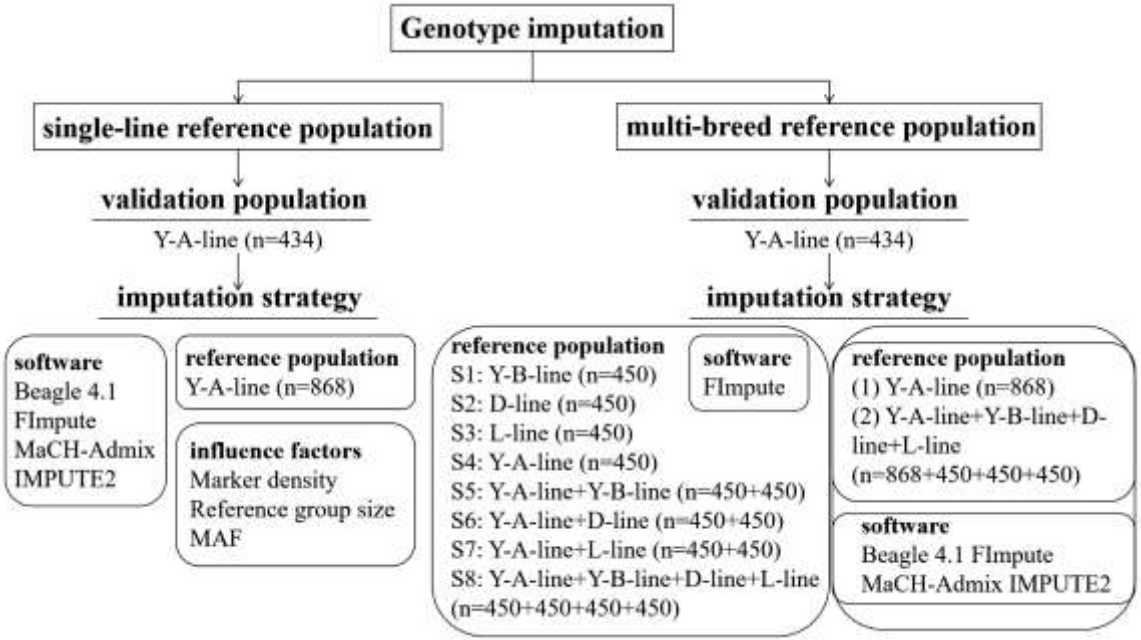


Appendix A. Principal component analysis. Yorkshire A line (Y-A-line), Yorkshire B line (Y-B-line), Duroc (D-line), Landrace (L-line).

Appendix B. The genotype concordance rate and Pearson correlation of Beagle when the *Ne* is 50, 150, 500, 1000 and 1000000 (the default value) when 95% SNPs were masked in the validation population.

				Imputation accuracy	
				Beagle 4.1	
Reference population	Validation population	Missing	*Ne*[1]	CR[2]	PC[3]
868 Y-A-line	434 Y-A-line	95%	50	0.9704	0.9636
			150	0.9711	0.9644
			500	0.9710	0.9644
			1000	0.9712	0.9645
			1000000 (default)	0.7536	0.6074

[1]*Ne*, effective population size. [2]CR, Concordance rate. [3]PC, Pearson correlation coefficient, P-value < 1e-18.



Appendix C. Flow chart of the entire imputation scheme design. The S1-S8 were 8 scenarios set for the reference population when studying the influence of multi-breed reference populations on the imputation accuracy. Yorkshire A line (Y-A-line), Yorkshire B line (Y-B-line), Duroc (D-line), Landrace (L-line).

Appendix D. The genotype concordance rate and Pearson correlation coefficient of three-fold cross-validation when the relationship between imputation accuracy and the marker density of validation population was investigated.

Reference population	Validation population		Imputation accuracy							
			Beagle 4.1		FImpute		IMPUTE2		MaCH-Admix	
the first fold		Missing[1]	CR[2]	PC[3]	CR	PC	CR	PC	CR	PC
868 Y-A-line	434 Y-A-line	20%	0.9985	0.9981	0.9980	0.9976	0.9921	0.9793	0.9985	0.9982
		45%	0.9981	0.9978	0.9974	0.9969	0.9893	0.9752	0.9982	0.9978
		70%	0.9968	0.9961	0.9950	0.9939	0.9793	0.9621	0.9970	0.9963
		95%	0.9737	0.9676	0.9612	0.9512	0.8345	0.7687	0.9761	0.9707
		99%	0.8194	0.7407	0.7859	0.6934	0.7188	0.5820	0.8381	0.7817
the second fold		Missing	CR	PC	CR	PC	CR	PC	CR	PC
868 Y-A-line	434 Y-A-line	20%	0.9971	0.9963	0.9965	0.9954	0.9898	0.9738	0.9970	0.9962
		45%	0.9966	0.9957	0.9957	0.9944	0.9866	0.9685	0.9966	0.9956
		70%	0.9946	0.9931	0.9927	0.9905	0.9767	0.9582	0.9947	0.9932
		95%	0.9686	0.9607	0.9547	0.9426	0.8309	0.7640	0.9712	0.9643
		99%	0.8106	0.7300	0.7762	0.6823	0.7146	0.5808	0.8286	0.7708

the third fold		Missing	CR	PC	CR	PC	CR	PC	CR	PC
		20%	0.9990	0.9988	0.9985	0.9981	0.9923	0.9737	0.9991	0.9989
		45%	0.9988	0.9985	0.9980	0.9975	0.9902	0.9718	0.9989	0.9986
868 Y-A-line	434 Y-A-line	70%	0.9975	0.9970	0.9957	0.9947	0.9808	0.9606	0.9978	0.9972
		95%	0.9760	0.9701	0.9629	0.9525	0.8396	0.7700	0.9784	0.9733
		99%	0.8251	0.7481	0.7954	0.7049	0.7246	0.5864	0.8424	0.7848

[1]Percentage of markers masked of validation population. [2]CR, Concordance rate. [3]PC, Pearson correlation coefficient, P-value < 1e-18.

Appendix E. The specific results of genotype concordance rate (CR), Pearson correlation coefficient (PC) (P-value < 1e-18), and their standard error (SE) when the relationship between imputation accuracy and the mark density of validation population was investigated in three replicates (n=3).

			Imputation accuracy							
			Beagle 4.1		FImpute		IMPUTE2		MaCH-Admix	
Reference population	Validation population	Missing	CR±SE	PC±SE	CR±SE	PC±SE	CR±SE	PC±SE	CR±SE	PC±SE
868 Y-A-line	434 Y-A-line	20%	0.9989±	0.9987±	0.9981±	0.9977±	0.9932±	0.9806±	0.9986±	0.9983±
			0.00038	0.00049	0.00007	0.00007	0.00110	0.00119	0.00003	0.00006
868 Y-A-line	434 Y-A-line	45%	0.9981±	0.9978±	0.9974±	0.9968±	0.9892±	0.9750±	0.9982±	0.9978±
			0.00003	0.00003	0.00006	0.00007	0.00012	0.00012	0.00003	0.00003
868 Y-A-line	434 Y-A-line	70%	0.9968±	0.9961±	0.9951±	0.9940±	0.9791±	0.9620±	0.9970±	0.9963±
			0.00003	0.00003	0.00003	0.00003	0.00025	0.00070	0.00003	0.00000
868 Y-A-line	434 Y-A-line	95%	0.9727±	0.9664±	0.9586±	0.9477±	0.8337±	0.7674±	0.9754±	0.9700±
			0.00080	0.00096	0.00242	0.00329	0.00108	0.00120	0.00065	0.00078
868 Y-A-line	434 Y-A-line	99%	0.8132±	0.7325±	0.7761±	0.6796±	0.7177±	0.5765±	0.8319±	0.7726±
			0.00328	0.00501	0.00549	0.00758	0.00092	0.00327	0.00340	0.00514

Appendix F. The specific results of genotype concordance rate and Pearson correlation coefficient when studying the effect of reference population size on imputation accuracy.

			Imputation accuracy							
			Beagle 4.1		FImpute		IMPUTE2		MaCH-Admix	
Reference population	Validation population	Missing	CR[1]	PC[2]	CR	PC	CR	PC	CR	PC
8 Y-A-line	434 Y-A-line	95%	0.7316	0.5539	0.7746	0.6412	0.6003	0.2811	0.6725	0.4537
86 Y-A-line	434 Y-A-line	95%	0.9078	0.8745	0.9070	0.8750	0.6906	0.5094	0.9104	0.8839
173 Y-A-line	434 Y-A-line	95%	0.9426	0.9265	0.9234	0.8992	0.7352	0.5891	0.9463	0.9325
434 Y-A-line	434 Y-A-line	95%	0.9624	0.9534	0.9410	0.9238	0.8170	0.7341	0.9634	0.9550
868 Y-A-line	434 Y-A-line	95%	0.9711	0.9645	0.9538	0.9411	0.8316	0.7650	0.9741	0.9684

[1]CR, Concordance rate. [2]PC, Pearson correlation coefficient, P-value < 1e-18.

Appendix G. The specific results of genotype concordance rate and correlation between imputed and observed genotypes when studying the relationship between imputation accuracy and MAF of variant being imputed.

	Imputation accuracy							
	Beagle 4.1		FImpute		IMPUTE2		MaCH-Admix	
MAF	CR[1]	PC[2]	CR	PC	CR	PC	CR	PC
[0.01, 0.03)	0.9942	0.9303	0.9922	0.9055	0.9693	0.5446	0.9948	0.9380
[0.03, 0.05)	0.9916	0.9462	0.9876	0.9197	0.9491	0.6165	0.9918	0.9474
[0.05, 0.1)	0.9869	0.9520	0.9788	0.9224	0.9147	0.6408	0.9885	0.9575
[0.1, 0.2)	0.9769	0.9532	0.9620	0.9217	0.8535	0.6638	0.9807	0.9612
[0.2, 0.3)	0.9692	0.9575	0.9502	0.9290	0.8162	0.7129	0.9737	0.9641
[0.3, 0.4)	0.9617	0.9559	0.9389	0.9267	0.7888	0.7303	0.9680	0.9634
[0.4, 0.5)	0.9616	0.9597	0.9393	0.9334	0.7819	0.7162	0.9673	0.9662

[1]CR, Concordance rate. [2]PC, Pearson correlation coefficient, P-value < 1e-18.

Appendix H. The impact of other breeds (lines) on the imputation accuracy except the main effect line. This section set three sample size gradients for other breeds or lines in the reference population, which are 50, 250 and 450, respectively. The validation population was 434 Y-A-line with 95% missing of SNPs.

		Imputation accuracy	
Reference population	Validation population	CR[1]	PC[2]
868 Y-A-line + 50 Y-B-line		0.9567	0.9409
868 Y-A-line + 250 Y-B-line	434 Y-A-line	0.9574	0.9416
868 Y-A-line + 450 Y-B-line		0.9572	0.9415
868 Y-A-line + 50 D-line		0.9570	0.9413
868 Y-A-line + 250 D-line	434 Y-A-line	0.9582	0.9429
868 Y-A-line + 450 D-line		0.9587	0.9436
868 Y-A-line + 50 L-line		0.9569	0.9414
868 Y-A-line + 250 L-line	434 Y-A-line	0.9582	0.9428
868 Y-A-line + 450 L-line		0.9582	0.9432

[1]CR, Concordance rate. [2]PC, Pearson correlation coefficient, P-value < 1e-18.

Appendix I. The specific results of imputation accuracy results of four software when imputation from multipe breeds.

Reference population	Validation population	Missing	Imputation accuracy							
			Beagle 4.1		FImpute		IMPUTE2		MaCH-Admix	
			CR[1]	PC[2]	CR	PC	CR	PC	CR	PC
868 Y-A-line	434 Y-A-line	95%	0.9708	0.9610	0.9574	0.9419	0.8488	0.7709	0.9755	0.9680
868 Y-A-line + 450 Y-B-line + 450 D-line + 450 L-line	434 Y-A-line	95%	0.9684	0.9574	0.9564	0.9400	0.7832	0.6479	0.9714	0.9624

[1]CR, Concordance rate. [2]PC, Pearson correlation coefficient, P-value < 1e-18.

Appendix J. The number of markers of reference population, the number of SNPs obtained after imputation with IMPUTE2, and missing markers compared with the reference population when studying the impact of validation population mark density.

Missing	SNPs of reference population	SNPs of imputed	SNPs of missing
20%	37933	37933	0
45%	37933	37933	0
70%	37933	37933	0
95%	37933	37239	694
99%	37933	24015	1918

Appendix K. The number of markers of reference population, the number of SNPs obtained after imputation with IMPUTE2, and missing markers compared with the reference population when studying the relationship between imputation accuracy and reference population size.

Reference population size	SNPs of reference population	SNPs of imputed	SNPs of missing
8	37933	37227	706
86	37933	37239	694
173	37933	37239	694
434	37933	37239	694
868	37933	37239	694