

Supplementary Information

Sunhee Kim, Sang-Ho Chu, Yong-Jin Park, Chang-Yong Lee

1 Supplementary Figures

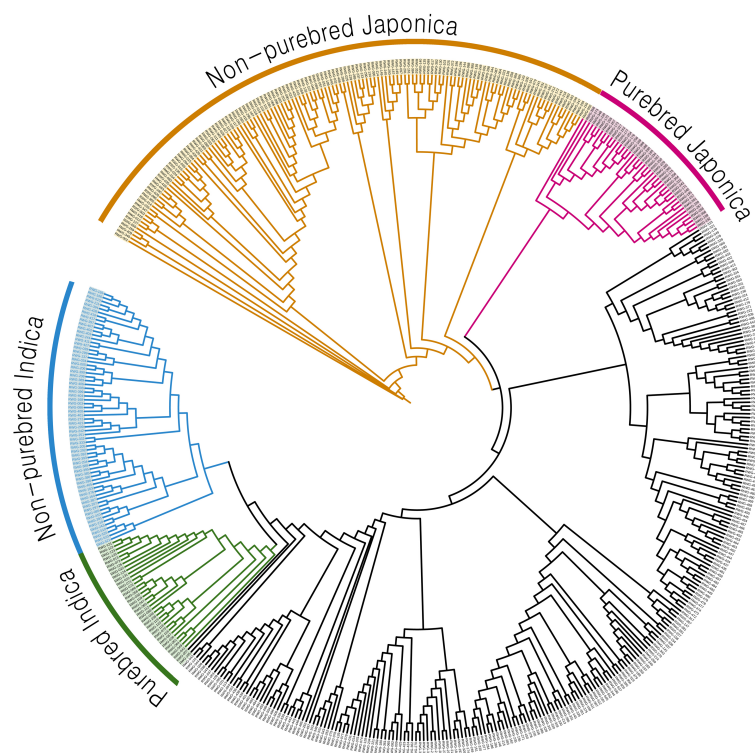


Fig. 1 The phylogenetic tree of the core accessions in KRICE was constructed using PHYLIP. The much higher resolution tree is available at https://github.com/infoLab204/rice_variety/blob/main/results/rice_circular.png.

List of typical Japonica 38 accessions					
Number	RWG_ID	Sample Name	Number	RWG_ID	Sample Name
1	RWG-152	Odae	20	RWG-256	Samcheon
2	RWG-157	Dongjin	21	RWG-257	Munjang
3	RWG-169	Unbong	22	RWG-260	Saesangju
4	RWG-170	Sinunbong	23	RWG-261	Manchu
5	RWG-171	Sambaek	24	RWG-266	Dongan
6	RWG-173	Geuru	25	RWG-267	Daesan
7	RWG-175	Sangmi	26	RWG-268	Nongho
8	RWG-176	Geumo	27	RWG-273	Hwangkeumbora
9	RWG-177	Jinbuol	28	RWG-280	Deuraechan
10	RWG-185	Juan	29	RWG-285	Seolemi
11	RWG-188	Haepyeong	30	RWG-322	Geumyoung
12	RWG-192	Ilmi	31	RWG-327	Daechan
13	RWG-195	Yeongnam	32	RWG-342	Sanduljinmi
14	RWG-199	Goami	33	RWG-345	Saegoami
15	RWG-228	Honong	34	RWG-348	Seolbaek
16	RWG-229	Joun	35	RWG-356	Ondami
17	RWG-230	Wolbaek	36	RWG-358	Unmi
18	RWG-254	Sobaeg	37	RWG-364	Joami
19	RWG-255	Sangju	38	RWG-368	Jinseolchal

List of typical Indica 33 accessions					
Number	RWG_ID	Sample Name	Number	RWG_ID	Sample Name
1	RWG-006	CT9993-5-10-1-M	18	RWG-131	Suhyeonghando
2	RWG-013	Jejubukjeju-2002-521	19	RWG-021	Saduchu
3	RWG-022	Seon	20	RWG-071	Chih-Tsao-He
4	RWG-030	Baekgokna	21	RWG-072	Hsiang-Ha-Tsan
5	RWG-047	Bai Cyue Hwa Lue	22	RWG-048	Yangmyeon
6	RWG-056	Waikyakushi	23	RWG-238	Wase Gingbouzu
7	RWG-060	Wu Cyue	24	RWG-119	San-Li-Shun
8	RWG-068	Yung Yuen Chuen Zim	25	RWG-118	Tsong-Gan-Shun
9	RWG-073	Cheongdo-Donggok-4	26	RWG-120	Yang-Sha-Tsan
10	RWG-109	Taichung-Woo-Tsan	27	RWG-067	Xi Gua Bai
11	RWG-112	Di Jiao Wu Jian	28	RWG-070	Qua 77 Wuan-Dau
12	RWG-121	Chungdo 23	29	RWG-116	Nian Chi Shi
13	RWG-122	Chungdo Hwayang 12	30	RWG-036	Chiem Chanh
14	RWG-123	Chungdo Hwayang 14	31	RWG-049	Belle Patna
15	RWG-124	Sungju 3	32	RWG-061	Binato
16	RWG-125	Jangsung 1	33	RWG-064	Hawm Supan
17	RWG-126	Soonchun 5			

List of introgressed Japonica 34 accessions					
Number	RWG_ID	Sample Name	Number	RWG_ID	Sample Name
1	RWG-001	Namseon 52	18	RWG-161	Jopyeong
2	RWG-004	Suwon 301	19	RWG-174	Inwol
3	RWG-011	Jejubukjeju-2002-340	20	RWG-180	Jeoginju
4	RWG-050	Pyeongbuk 3	21	RWG-181	Naepung
5	RWG-111	Xao Bai Mang Sue Dao	22	RWG-183	Seon
6	RWG-117	Sung Pan Tao	23	RWG-187	Sobi
7	RWG-129	Danyang 9	24	RWG-191	Seolhyangchal
8	RWG-134	Namseon 126	25	RWG-212	Goun
9	RWG-138	Jinbu	26	RWG-218	Gangbaek
10	RWG-141	Dongjinchal	27	RWG-231	Gangchan
11	RWG-142	Gopum	28	RWG-241	Milseong
12	RWG-144	Haiami	29	RWG-246	Seolag
13	RWG-147	Chilbo	30	RWG-247	Samnam
14	RWG-150	Junam	31	RWG-263	Sura
15	RWG-154	Hwaseong	32	RWG-287	Saellmi
16	RWG-158	Nagdong	33	RWG-334	Migwang
17	RWG-159	Giho	34	RWG-357	Undoo Byeo

List of introgressed Indica 34 accessions					
Number	RWG_ID	Sample Name	Number	RWG_ID	Sample Name
1	RWG-010	Jejubukjeju-2002-171	18	RWG-205	Taebaeg
2	RWG-003	Suweon 255	19	RWG-206	Gaya
3	RWG-005	Suweon 347	20	RWG-207	Baegyung
4	RWG-034	Bala	21	RWG-208	Cheongcheong
5	RWG-039	IR38	22	RWG-299	IR72
6	RWG-041	Mala	23	RWG-411	BP603
7	RWG-044	Red Rice	24	RWG-416	IR841-85-1-1-2
8	RWG-054	Hatadani	25	RWG-417	Jasmine 85
9	RWG-085	Milyang 50	26	RWG-426	BP066
10	RWG-086	Yeongpung	27	RWG-427	BP444
11	RWG-137	IR40	28	RWG-491	Tubtim Chumphae
12	RWG-163	Hanareum	29	RWG-492	Hom Mali Nin Surin
13	RWG-164	Dasan	30	RWG-493	Khao Dawk Mali 105
14	RWG-165	Milyang 23	31	RWG-494	Pathumthani 1
15	RWG-166	Samgang	32	RWG-524	Basmati
16	RWG-168	Tongil	33	RWG-525	Organic Basmati
17	RWG-204	Anda	34	RWG-526	Basmati2 (California)

Fig. 2 (Top) The list of purebred Japonica and purebred Indica samples to call the actual variants. (Bottom) The list of non-purebred Japonica and non-purebred Indica samples to call the predicted variants.

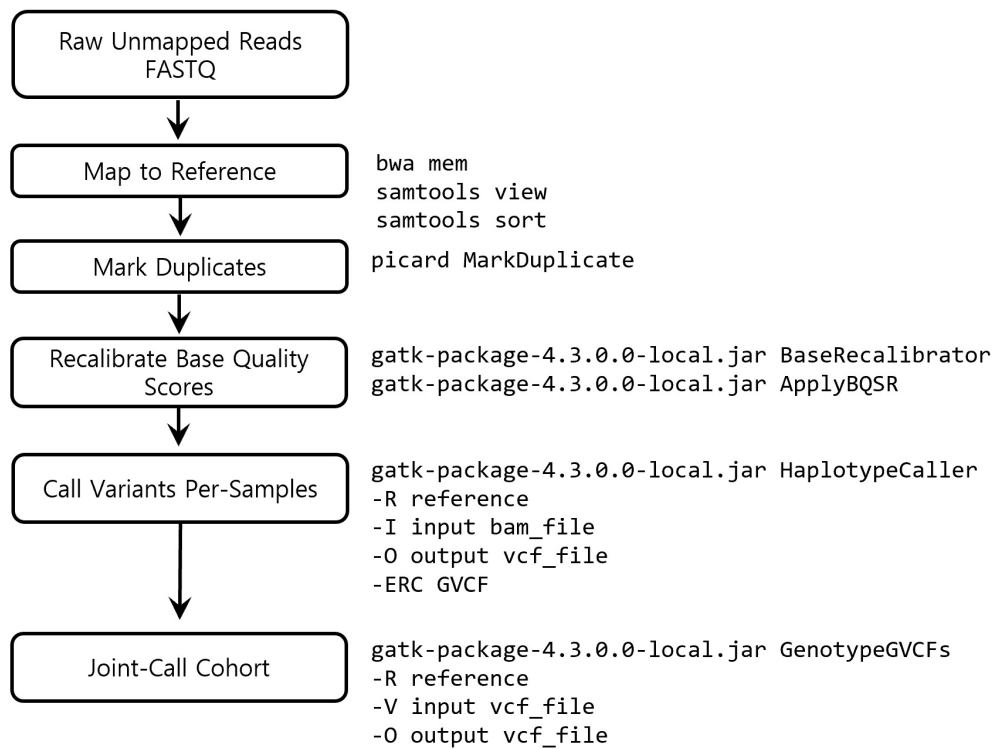


Fig. 3 A schematic of the variant calling pipeline used in this study.

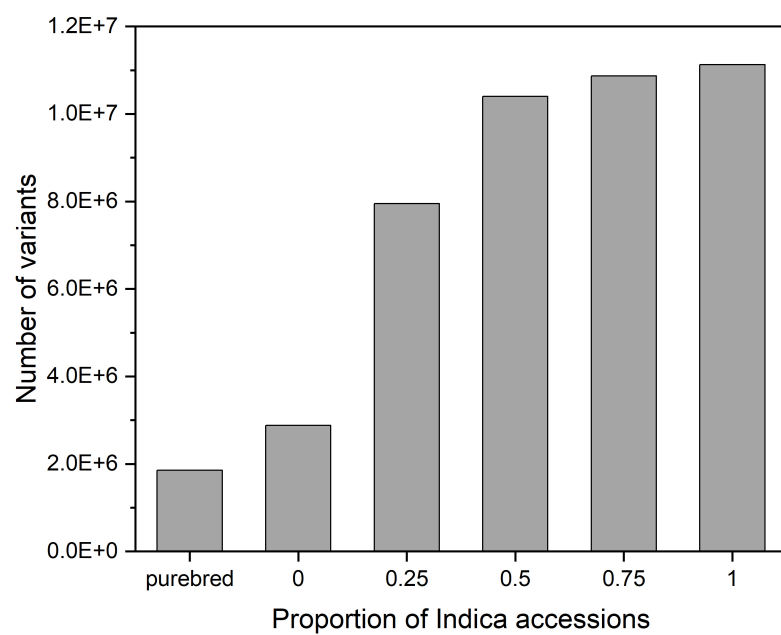


Fig. 4 The number of variants called from different models using the Japonica reference sequence in the non-coding region.

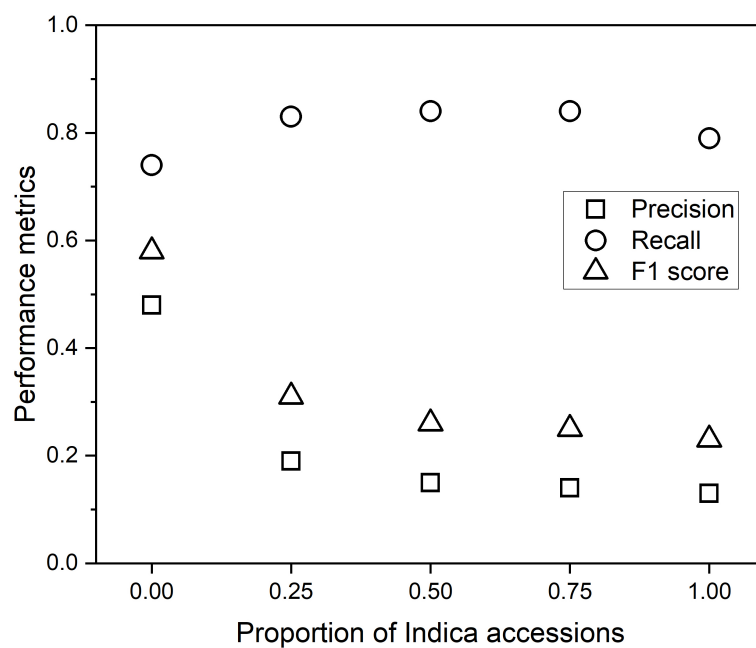


Fig. 5 A plot of the performance metrics of precision, recall, and F1 score for different models in the non-coding region using the Japonica reference. Note that the horizontal axis represents the proportion of non-Japonica (i.e., Indica) accessions in the samples.

2 Supplementary Tables

Model	0%	25%	50%	75%	100%
Number of potential false positives	1,503,557	6,410,527	8,832,738	9,304,551	9,665,015
Precision	0.48	0.19	0.15	0.14	0.13

Table 1 The number of false positives and precision from the models of different Indica-mixed proportions in the non-coding region.

(a) Japonica reference (non-coding region) Japonica-only samples		Actual variants	
		Yes	No
Predicted variants	Yes	1,373,341	1,503,557
	No	483,443	334,291,651
Metrics	P=0.48	R=0.74	F1=0.58

(b) Japonica reference (non-coding region) 25% Indica-mixed		Actual variants	
		Yes	No
Predicted variants	Yes	1,538,597	6,410,527
	No	318,187	329,384,681
Metrics	P=0.19	R=0.83	F1=0.31

Table 2 Results of the confusion matrices and their metrics constructed from variant calling in the non-coding region using the Japonica reference from (a) the Japonica-only model and (b) the 25% Indica-mixed model.

(a) Japonica reference (non-coding, before dbFP)		Actual variants	
50% Indica-mixed		Yes	No
Predicted	Yes	1,567,959	8,832,738
variants	No	288,825	326,962,470
Metrics	P=0.15	R=0.84	F1=0.26

(b) Japonica reference (non-coding, after dbFP)		Actual variants	
50% Indica-mixed		Yes	No
Predicted	Yes	1,567,959	3,588,991
variants	No	288,825	332,206,217
Metrics	P=0.30	R=0.84	F1=0.45

Table 3 Results of the confusion matrices and their metrics constructed from variant calling in the non-coding region using the Japonica reference from (a) the half-mixed sample from both varieties and (b) the same as (a) except that the variants in the dbFP are excluded.

3 Supplementary Data File

The analysis results as a file of the variant call using the Indica reference sequence can be found at

https://github.com/infoLab204/rice_variety/blob/main/results/Indica_results.xlsx.