DNA Research, 2017, 24(4), 397–405
doi: 10.1093/dnares/dsx012
Advance Access Publication Date: 11 May 2017
Full Paper

OXFORD

Full Paper

Heap: a highly sensitive and accurate SNP detection tool for low-coverage high-throughput sequencing data

Masaaki Kobayashi[1], Hajime Ohyanagi[1,2], Hideki Takanashi[3], Satomi Asano[1], Toru Kudo[1], Hiromi Kajiya-Kanegae[3], Atsushi J. Nagano[4,5,6], Hitoshi Tainaka[3], Tsuyoshi Tokunaga[7], Takashi Sazuka[8], Hiroyoshi Iwata[3], Nobuhiro Tsutsumi[3], and Kentaro Yano[1,*]

[1]Bioinformatics Laboratory, Department of Life Sciences, School of Agriculture, Meiji University, Kanagawa 214-8571, Japan, [2]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Thuwal 23955-6900, Kingdom of Saudi Arabia, [3]Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan, [4]Faculty of Agriculture, Ryukoku University, Shiga 520-2194, Japan, [5]PRESTO, Japan Science and Technology Agency, Japan, [6]Center for Ecological Research, Kyoto University, Shiga 520-2113, Japan, [7]EARTHNOTE Co. Ltd., Nago, Okinawa 905-1152, Japan, and [8]Bioscience and Biotechnology Center, Nagoya University, Aichi 464-8601, Japan

*To whom correspondence should be addressed. Tel. +81 44 934 7046. Fax. +81 44 934 7046. Email: kyano@isc.meiji.ac.jp

Edited by Prof. Hiroyuki Toh

Received 15 June 2016; Editorial decision 15 March 2017; Accepted 20 April 2017

Abstract

Recent availability of large-scale genomic resources enables us to conduct so called genome-wide association studies (GWAS) and genomic prediction (GP) studies, particularly with next-generation sequencing (NGS) data. The effectiveness of GWAS and GP depends on not only their mathematical models, but the quality and quantity of variants employed in the analysis. In NGS single nucleotide polymorphism (SNP) calling, conventional tools ideally require more reads for higher SNP sensitivity and accuracy. In this study, we aimed to develop a tool, Heap, that enables robustly sensitive and accurate calling of SNPs, particularly with a low coverage NGS data, which must be aligned to the reference genome sequences in advance. To reduce false positive SNPs, Heap determines genotypes and calls SNPs at each site except for sites at the both ends of reads or containing a minor allele supported by only one read. Performance comparison with existing tools showed that Heap achieved the highest *F*-scores with low coverage (7X) restriction-site associated DNA sequencing reads of sorghum and rice individuals. This will facilitate cost-effective GWAS and GP studies in this NGS era. Code and documentation of Heap are freely available from https://github.com/meiji-bioinf/heap (29 March 2017, date last accessed) and our web site (http://bioinf.mind.meiji.ac.jp/lab/en/tools.html (29 March 2017, date last accessed)).

Key words: single nucleotide polymorphism (SNP), next-generation sequencing (NGS), restriction-site associated DNA sequencing (RAD-seq), genome-wide association studies (GWAS), genomic prediction (GP)

1. Introduction

DNA polymorphisms including single nucleotide polymorphisms (SNPs), insertions and deletions (INDELs), among inbred lines or collected individuals have been employed to identify loci associated with traits or to predict phenotypes by genotypes. To identify genes associated with phenotypes and implement high-throughput breeding, genome-wide association studies (GWAS) and genomic prediction (GP) studies have been performed with next-generation sequencing (NGS) data[1–4]. In GWAS and GP studies, SNPs are widely used as genetic markers. SNPs are determined by mapping NGS reads and consequent variant calls with conventional tools. Widely used software, such as Burrows-Wheeler Aligner (BWA) or Bowtie 2 performs alignments of NGS reads to reference genomes[5,6], whereas SAMtools/BCFtools or Genome Analysis Toolkit (GATK) call variants based on alignments produced during the mapping steps[7–11]. The accuracy of both mapping and variant calling should be key factors for the effectiveness of GWAS and GP.

SAMtools/BCFtools and GATK accurately call SNPs if provided with enough number of reads. When read coverage (depth) is 20X or more, SNPs have been called with sufficient sensitivity in human genome resequencing[12]. It is also reported that SNP calling becomes difficult under low read coverage (7X or lower)[12]. In large-scale GWAS or GP studies, the number of NGS reads in each individual tends to be small in order to genotype as many individuals as possible with a limited budget. In these cases, methodologies that accurately detect large number of SNPs based on a relatively low number of reads are paramount for an efficient implementation of GWAS and GP studies.

Restriction-site associated DNA sequencing, or genotype-by-sequencing (RAD-seq/GBS), is an economical strategy to identify genome wide polymorphisms[13]. While whole genome shotgun (WGS) sequencing methods provide sequencing information from the whole genome, the RAD-seq/GBS methods confine sequencing regions to the fragments neighbouring particular restriction sites. To be precise, in RAD-seq/GBS, regions neighbouring particular restriction sites are sequenced in combination with high-throughput sequencing methods and with subsequent mapping and variant calling steps on reference genome sequences to identify genomic polymorphisms. For example, RAD-seq/GBS reads sets of *Hin*dIII which covers only 0.24% of the chicken genome on average (i.e. coverage = 0.0024x) could show 3× or more depth of coverage on the RAD fragments[14]. Compared with WGS sequencing methods, the RAD-seq/GBS strategy under the same sequencing cost generally assures a relatively higher coverage at each sequencing region, even when the total number of reads is lower in each sample (e.g. inbred lines, individuals, or libraries) with the RAD-seq/GBS strategy.

Stacks is a specialized tool for identifying SNPs from RAD-seq/GBS reads[15,16]. In the Stacks pipeline, identical reads, which are called 'stack', are collected from each sample. Then, SNPs among samples are identified by comparing among stacks obtained from multiple samples. Stacks are capable of calling many SNPs compared to SAMtools/BCFtools and GATK with low coverage RAD-seq/GBS reads. However, Stacks can call false positive SNPs more frequently than other methods (as is shown in this work).

In this study, we developed a new tool, Heap, that identifies numerous SNPs with high accuracy from RAD-seq/GBS or WGS sequencing reads. Heap is capable of detecting SNPs with high sensitivity, which is the ratio of correctly identified SNPs (True positives) to all the existing SNPs (True positives + False negatives), and high positive predictive values (PPVs), which is the ratio of True positives to all the identified SNPs (True positives + False positives), from even reads with a low coverage aligned to the reference genome sequences. To confirm the performance of Heap, we identified SNPs from RAD-seq/GBS reads and calculated an accuracy index *F*-score, which is the harmonic mean between sensitivity and PPV, of SNP calling in 17 inbred sorghum lines and 4 inbred rice lines.

2. Materials and methods

2.1. Software details

2.1.1. Algorithms of Heap

We designed Heap to identify SNPs from short read sequences of diploid species. In Heap analysis, short read sequences must be aligned to reference genome sequences in advance, and information on aligned reads, which is stored in either Sequence Alignment/MAP (SAM) format files or the binary version of SAM (BAM) format files[7], must be employed (Fig. 1A).

After importing information on read alignments from a SAM or BAM file, Heap performs read filtering to obtain high quality reads (Fig. 1B). By default setting of Heap, reads with a phred scaled mapping quality score (MAPQ) below 20 are removed. Bases with a phred scaled quality score in base calling below 13 are also eliminated from the search scope of valid SNP sites. Heap also trims both ends of each read before mining SNPs, because it is empirically observed that these regions contain many base calling errors. By default setting, both ends with $n = 2$bp are not counted for SNP searches. Moreover, flanking regions within $i = 5$bp of each INDEL site are also removed in SNP calling. To determine the default setting of the n and i values, we examined the empirical distribution of *F*-scores under multiple conditions (for *F*-scores, see Performance comparison among SNP calling tools with RAD-seq reads in sorghum in results and discussions) (Supplementary Fig. S1). The highest *F*-score was obtained under around $n = 2$–3, while *F*-score reached a plateau under around $i = 4$–5. Therefore, we determined that $n = 2$ and $i = 5$ as default values.

Next, Heap determines each sample's genotype in every site that passes quality filtering (Fig. 1C). On each nucleotide site, the allele frequency is calculated from the number of nucleotide bases (A, T, C or G) aligned on it. Heap ignores any sites with three or more allele variants, since any allele possibilities above 2 on diploid organisms are likely due to sequencing errors. To determine the zygosity of each allele, a binomial test with allele frequency is performed (the null hypothesis H_0: the allele is heterozygous). When the *P* value of the binomial test is <0.05, the null hypothesis is rejected thus the zygosity on the site is determined as homozygous. Conversely, when the P value is ≥ 0.05 and the minor allele is supported by two or more NGS reads, the zygosity is determined as heterozygous. When neither is the case, the genotype is included in subsequent analyses as missing genotype (./.). This genotyping step is repeated for all samples.

Heap then performs SNP calling by comparing the genotypes between the sample (e.g. an inbred line) and the reference genome (Fig. 1D). The information on all SNPs between the sample and the reference genome is stored in a Variant Call Format (VCF) file. In a VCF file, reference allele, first alternative allele, second alternative allele, and missing is presented as '0', '1', '2', and '.' in the genotype field, respectively. Finally, to determine SNPs among all samples, the VCF files for all samples are merged in a single VCF file by BCFtools (Fig. 1E).

2.1.2. Software implementation and requirements

Heap is implemented in C++ programming languages and is able to be installed on a Linux/UNIX-like operating system including, but

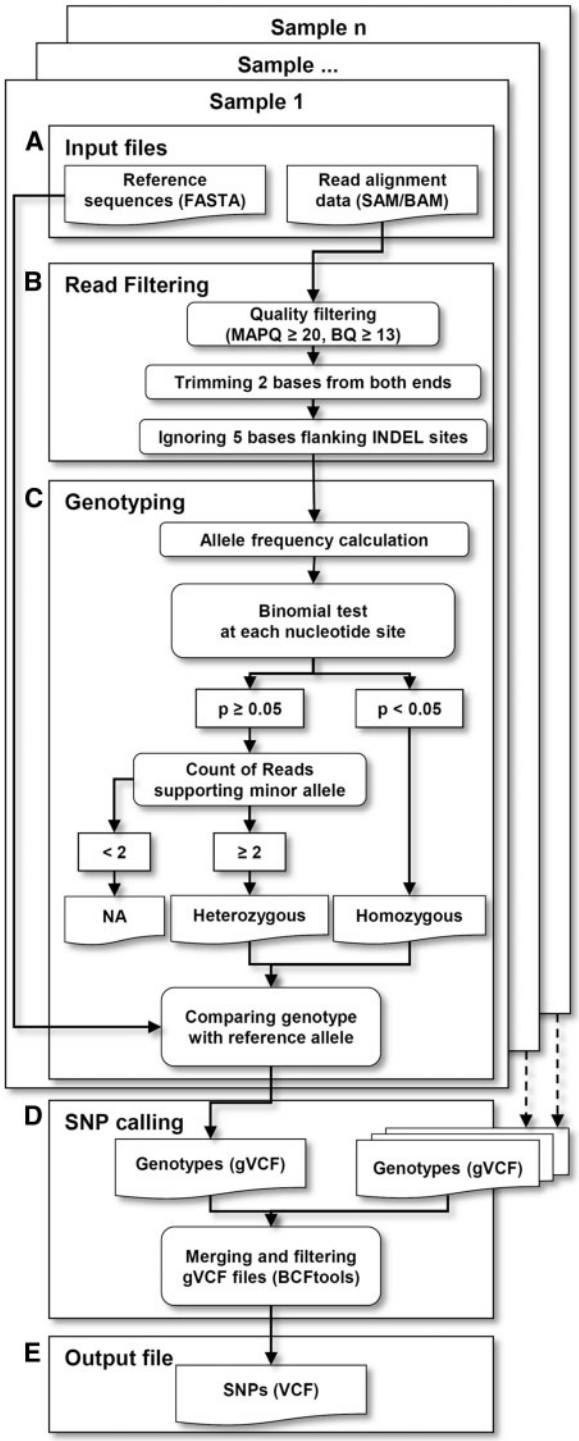

Figure 1. Workflow of Heap algorithm.

not limited to, CentOS, Ubuntu and Mac OS X. The following additional software and libraries are required: Boost C++ libraries, BCFtools (≥ ver. 1.2), HTSlib (≥ ver. 1.2), and g++ compiler (≥ ver. 4.1.2). All steps of Heap algorithm can be invoked with a single command line. For the usage and options of Heap, see the README text file. Heap is released under the GNU General Public License Version 3 (see the COPYING text file).

2.2. Wet lab procedures

2.2.1. DNA extraction from sorghum and rice plants

DNA samples were extracted from 17 inbred lines of *Sorghum bicolor* (GULUM ABIAD, A-6129, AGIRA, LR 399, CRIOLLO CABEZA APRETADA, B2AR3043, RTx430, B3Tx2817, ZIRA-EL-SABI, ITALIAN, CAPRICORN, SOR 1, 15065, SAP-155, RCV, SC 56, and SIL-05) and 4 inbred lines of *Oryza sativa* (Omachi, Yamada Nishiki, Hitomebore, and Kameji). Methods of DNA extraction are described in Supplementary Methods.

2.2.2. RAD-seq of sorghum

We performed RAD-seq in the 17 inbred sorghum lines using an Illumina HiSeq 2000 platform with 100 cycles with single end layout. Detailed methods for DNA libraries construction and RAD-seq are described in Supplementary Methods. These RAD-seq data have been submitted to DDBJ Sequence Read Archive (DRA) (DRR045054-DRR045070).

2.2.3. WGS sequencing of sorghum

WGS sequencing was performed in the 17 inbred sorghum lines using an Illumina HiSeq 2000 platform with 100 cycles with paired end layout. Detailed methods for DNA libraries construction and WGS sequencing are described in Supplementary Methods. These WGS sequencing data have been submitted to DRA (DRR045071-DRR045087).

2.2.4. RAD-seq of rice

We performed RAD-seq in the 4 inbred rice lines using an Illumina HiSeq 2000 platform with 50 cycles with single end layout. Detailed methods for DNA libraries construction and RAD-seq are described in Supplementary Methods. These RAD-seq data have been submitted to DRA (DRR045088-DRR045091).

2.2.5. WGS sequencing data of rice

Besides the sequencing data mentioned above, WGS sequencing data of the 4 inbred rice lines (DRR000719, DRR000720, DRR003652, DRR003655, DRR004451, DRR004452, DRR004453, DRR003648, DRR003649, and DRR003658) were downloaded from DRA. Each of them corresponds to each of 4 inbred lines mentioned in RAD-seq of rice.

2.3. Benchmarking

2.3.1. Preprocessing of reads

In order to correctly map the WGS sequencing reads and the RAD-seq reads to the reference genome sequences, we performed adapter trimming and quality filtering as described previously[17]. After quality control by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (29 March 2017, date last accessed)), we trimmed adaptor sequences by cutadapt (https://cutadapt.readthedocs.io/en/stable/ (29 March 2017, date last accessed)). Low-quality reads were also filtered out by an empirically optimized custom Perl script as the following: (i) both ends of each read must have QV ≥10 (if not, the end base with QV <10 is trimmed away until QV ≥10 is exposed); (ii) each read must have average QV ≥17 (if not, the read is discarded); (iii) final length of each read must be ≥20 bp (if not, the read is discarded); (iv) each read must have low-quality positions (QV <10) no more than 10% of final length (if not, the read is discarded); and (v) each read must not contain any N bases (if not, the read is discarded). The Perl script for filtering out low-quality reads are available from our website (http://bioinf.mind.meiji.ac.jp/lab/en/download/readPreprocessingScripts.tar.gz (29 March 2017, date last accessed)).

2.3.2. Mapping of reads to the reference genome sequences

We downloaded the genome sequences of *S. bicolor* (cv. BTx623) (Sbicolor_v2.1) and *O. sativa* (Japonica group, cv. Nipponbare) (Os-Nipponbare-Reference-IRGSP-1.0) from Phytozome (http://www.phytozome.net/ (29 March 2017, date last accessed)) and the Rice Annotation Project Database (RAP-DB) (http://rapdb.dna.affrc.go.jp/ (29 March 2017, date last accessed)), respectively[18–21]. We aligned the preprocessed reads on the reference genome sequences using BWA with the default options. We then performed realignment of the reads of neighbouring INDELs using the RealignerTargetCreator and IndelRealigner commands of GATK (with default settings). Finally, we selected uniquely mapped reads with the X0:i:1 and X1:i:0 tags in the optional fields of the SAM files.

2.3.3. SNP calling with RAD-seq reads

To compare performance of Heap (ver. 0.7.8), Stacks (ver. 1.29), SAMtools/BCFtools (ver. 1.1), and GATK (ver. 3.2) under similar conditions, we called SNPs with RAD-seq reads as follows.

Analysis with Heap, Stacks and GATK, SNPs supported by 3 or more reads were used. In SAMtools/BCFtools and GATK analysis, SNPs with low genotype quality scores (GQs) (GQ <13) were excluded.

In Heap analysis, we executed 'heap' command with the parameter of a minimum read depth (-d 3). Consequently, Heap provides an SNP list, which contains SNPs between samples.

In Stacks analysis, to obtain SNPs between samples, we executed by the pipeline via the ref_map.pl. To obtain SNPs supported stacks, we executed the ref_map.pl with the parameter of a minimum stack depth (-m 3). SNPs sites were obtained by the 'populations' command of Stacks in VCF format.

In SAMtools/BCFtools analysis, we ran the 'mpileup' command of SAMtools and the 'call' command of BCFtools. To describe GQs in FORMAT fields of the VCF file for subsequent filtering, the 'call' command was performed with the format fields parameter (-f GQ).

In GATK analysis, for each sample, we obtained genotypes at all nucleotide sites by the 'HaplotypeCaller' command. This command provides genotypes for all sites in Genomic VCF (gVCF) format, therefore, it contains both homozygous and heterozygous SNPs in each site. We then created a VCF file containing information of all SNPs between samples, which was integrated with all gVCF files by the 'GenotypeGVCFs' command using the default parameters.

To compare the SNPs derived by the multiple tools, we filtered the SNP sites, because conditions of SNP calling are different between Stacks and the other three tools (SAMtools/BCFtools, GATK, and Heap). Stacks calls SNPs at only nucleotide sites containing more than one allele among multiple samples. On the other hand, SAMtools/BCFtools, GATK, and Heap call SNPs at nucleotide sites containing one or more alternative alleles among multiple samples. For example, in three samples, a site containing genotypes '0/1', '0/1', and './.', is reported by the four tools. On the other hand, a site containing genotypes '1/1', '1/1', and './.', is not reported by Stacks but reported by the other tools. To equalize the conditions of the SNP lists derived by the multiple tools, we selected polymorphic SNP sites which contained 2 or more genotypes with more than one allele among samples, from the VCF files by a custom AWK script (http://bioinf.mind.meiji.ac.jp/lab/en/download/awkScriptToExtractPolySNPsitesFromVcf.sh (29 March 2017, date last accessed)).

2.3.4. Mapping of the WGS reads and Genotyping of the 'definitive answer genotypes'

To determine if genotypes obtained from the low coverage RAD-seq reads are correct or not at each nucleotide site, we determined more probable genotypes as 'definitive answer genotypes' by using the high coverage WGS sequencing reads. In order to establish the definitive answer genotypes at each nucleotide site in each sorghum and rice line, we aligned WGS sequencing reads to the reference genome sequences and determined the definitive answer genotypes stringently at each nucleotide site with the aligned WGS sequencing reads (Fig. 2A).

After adapter trimming and quality filtering, we aligned the WGS sequencing reads to the reference genome sequences by using BWA. We obtained bases of the mapped WGS sequencing reads using the mpileup command of SAMtools. To ensure the high-quality calls from accurately mapped reads, we used the MAPQ 20 and the minimum base quality 20 for the mpileup command. Additionally, to obtain bases that had sufficient read coverages from the mpileup result, nucleotide sites supported with 20 or more reads were extracted by the awk command (awk -F '\t' '{if($4 >=20){print $0}}'). We counted reads for each nucleotide base (A, T, G, or C) and calculated frequencies of the nucleotide bases at each nucleotide site using custom Perl scripts. We then determined each genotype according to the following rigorous conditions: (i) when the major allele frequency is ≥ 0.95, the genotype is homozygous; (ii) when the major allele frequency is ≥ 0.5 and ≤ 0.6, the genotype is heterozygous; (iii) otherwise, the genotype is not available (NA). Finally, we filtered out the monomorphic sites, which contained a single genotype among samples. A shell script and Perl scripts for making definitive answer genotypes are available from our website (http://bioinf.mind.meiji.ac.jp/lab/en/download/benchMarkingScripts.tar.gz (29 March 2017, date last accessed)).

3. Results and discussion

To compare the sensitivities and PPVs of SNP calling for Heap, Stacks, SAMtools/BCFtools, and GATK with low read coverage, we detected SNPs using low coverage RAD-seq reads by the tools in two plant species, sorghum and rice. Sensitivities and PPVs of the SNPs were calculated by comparing the SNPs detected with RAD-seq reads, and SNPs rigorously detected by a conventional strategy with more than adequate amounts of WGS sequencing reads, i.e. definitive answer genotypes. Then, we calculated the *F*-score, which represents a total performance index for sensitivity and accuracy of SNP mining.

3.1. Genotyping with WGS sequencing reads for the 'definitive answer genotypes'

Firstly, to establish the 'definitive answer genotypes', we aligned the WGS sequencing read to the reference genome sequences in sorghum and rice. An overview of the preprocessing and the mapping for the WGS sequencing reads are presented in Tables 1 and 2. Averages of the WGS read coverages were 21.5 and 38.5 in the 17 sorghum lines and in the 4 rice lines, respectively. These read coverages (around 20–40) should be sufficient to perform definitively accurate genotyping[12]. Subsequently, we conservatively determined the definitive answer genotypes at each nucleotide site. In average, 4,045 SNP sites and 2,955,720 non-SNP sites were obtained in sorghum; 122 SNP sites and 2,779,362 non-SNP sites were obtained in rice. These results indicated that the number of non-SNP sites was very larger than the SNP sites. In this situation, specificity shows a large value, and is not informative value. In fact, all specificity values were over than 0.999 in this study (data not shown). Therefore, we adopted PPV to examine rate of precisely called SNPs in all SNPs.

Table 1. Summary of WGS sequencing reads and mapping of the reads in Sorghum

Sample	Raw reads		Preprocessed reads		Mapped reads		Uniquely mapped reads					
	Count ($\times 10^6$)	Total length (Gb)	Count ($\times 10^6$)	Total length (Gb)	Count ($\times 10^6$)	Rate (%)[a]	Count ($\times 10^6$)	Rate (%)[a]	Coverage			
									Mean	1st quartile	Median	3rd quartile
GULUM ABIAD	219.2	22.1	200.5	19.9	189.0	94.3	100.3	50.0	21.2	11	20	27
A-6129	205.1	20.7	191.9	19.1	181.1	94.4	97.0	50.5	20.2	12	19	26
AGIRA	214.0	21.6	195.4	19.4	185.5	94.9	104.5	53.5	18.7	10	18	24
LR 399	232.7	23.5	213.7	21.2	202.6	94.8	113.4	53.1	20.8	11	20	27
CRIOLLO CABEZA APRETADA	260.9	26.4	239.2	23.8	230.0	96.2	131.1	54.8	24.8	14	24	32
B2AR3043	206.7	20.9	181.4	18.4	171.9	94.8	98.5	54.3	19.1	12	18	24
RTx430	215.2	21.7	196.4	19.5	188.4	95.9	105.8	53.8	20.4	12	19	25
B3Tx2817	226.5	22.9	209.8	20.9	203.3	96.9	118.7	56.6	21.7	12	21	28
ZIRA-EL-SABI	235.7	23.8	218.1	21.7	210.3	96.4	118.0	54.1	22.9	14	22	29
ITALIAN	192.2	19.4	177.0	17.6	170.2	96.1	97.5	55.1	19.2	12	18	24
CAPRICORN	250.0	25.3	233.1	23.2	225.2	96.6	129.1	55.4	24.6	16	24	31
SOR 1	270.0	27.3	237.4	23.4	225.8	95.1	120.1	50.6	23.8	14	23	30
15065	232.4	23.5	207.2	20.5	196.7	94.9	105.9	51.1	20.8	12	20	26
SAP-155	217.6	22.0	200.4	19.8	191.7	95.6	107.9	53.9	20.9	13	20	26
RCV	216.8	21.9	197.0	19.4	188.7	95.8	108.9	55.3	20.8	12	20	26
SC 56	203.8	20.6	177.8	17.4	170.3	95.8	96.9	54.5	19.3	12	18	24
SIL-05	272.4	27.5	239.2	23.5	229.6	96.0	129.9	54.3	25.7	15	25	33
Total	3871.3	391.0	3515.5	348.6	3360.2	–	1883.3	–	–	–	–	–
Average	227.7	23.0	206.8	20.5	197.7	95.6	110.8	53.6	21.5	12.6	20.5	27.2

All read counts and lengths are shown in millions and billions, respectively.

[a]Mapping rates are calculated as the ratio of the number of the mapped reads against the number of the preprocessed reads.

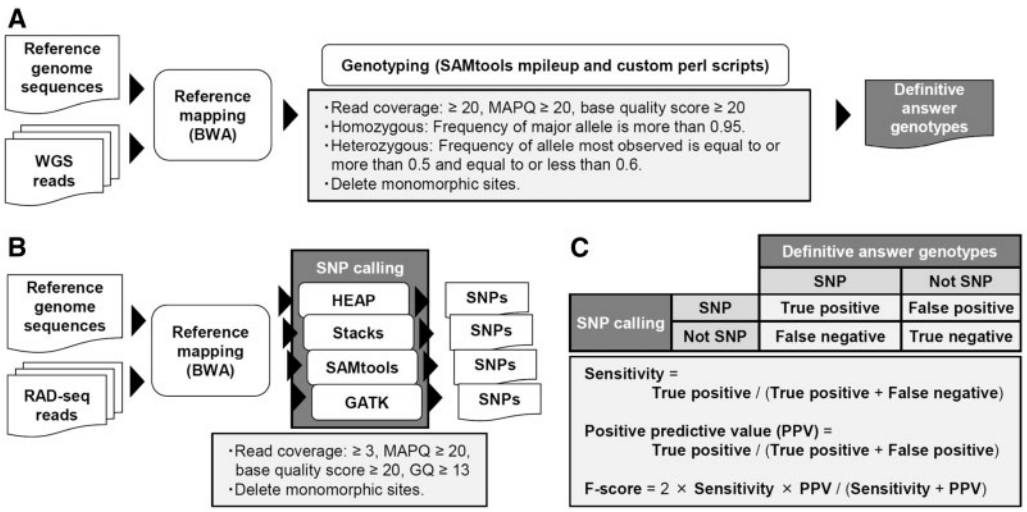

Figure 2. Schematic representation of performance comparison of SNP calling by Heap, Stacks, SAMtools/BCFtools, and GATK. (A) Schematic view of the calculation of definitive answer genotypes with WGS sequencing reads. MAPQ are measures of mapping quality. (B) Schematic view of SNP calling from RAD-seq reads employing Heap, Stacks, SAMtools, and GATK. GQ indicates genotype quality. (C) Definitions of sensitivity, positive predictive value (PPV), and *F*-score for SNP calling.

3.2. Performance comparison among SNP calling tools with RAD-seq reads in sorghum

We compared performances of Heap, Stacks, SAMtools/BCFtools, and GATK with RAD-seq reads in sorghum. At first, in 17 sorghum lines, we conducted RAD-seq and mapped the RAD-seq reads on the reference genome sequences (Table 3). The average RAD-seq read coverage in RAD-regions was around 7.4, which is low for conventional SNP calling using SAMtools/BCFtools or GATK[12]. We detected 47,901, 53,410, 30,316 and 26,728 SNP sites among the 17 inbred sorghum lines using Heap, Stacks, SAMtools/BCFtools, and GATK with the RAD-seq reads (Fig. 2B), respectively. We then calculated sensitivities and PPVs of the SNPs detected from RAD-seq

Table 2. Summary of WGS sequencing reads and mapping of the reads in rice

Sample	Raw reads		Preprocessed reads		Mapped reads		Uniquely mapped reads					
	Count ($\times 10^6$)	Total length (Gb)	Count ($\times 10^6$)	Total length (Gb)	Count ($\times 10^6$)	Rate (%)[a]	Count ($\times 10^6$)	Rate (%)[a]	Coverage			
									Mean	1st quartile	Median	3rd quartile
Omachi	297.9	22.3	260.7	18.3	257.0	98.6	185.4	71.1	41.6	27	42	56
Yamada Nishiki	218.3	19.3	184.7	15.8	181.3	98.2	140.6	76.1	37.4	30	40	47
Hitomebore	221.9	16.6	202.2	14.9	198.0	97.9	144.3	71.4	34.2	18	35	48
Kameji	233.8	20.7	219.4	18.7	214.0	97.6	154.3	70.4	40.5	27	44	54
Total	971.9	79.1	866.8	67.7	850.2	–	624.6	–	–	–	–	–
Average	243.0	19.8	216.7	16.9	212.6	98.1	156.1	72.2	38.5	25.5	40.3	51.3

All read counts and lengths are shown in millions and billions, respectively.
[a]Mapping rates are calculated as the ratio of the number of the mapped reads against the number of the preprocessed reads.

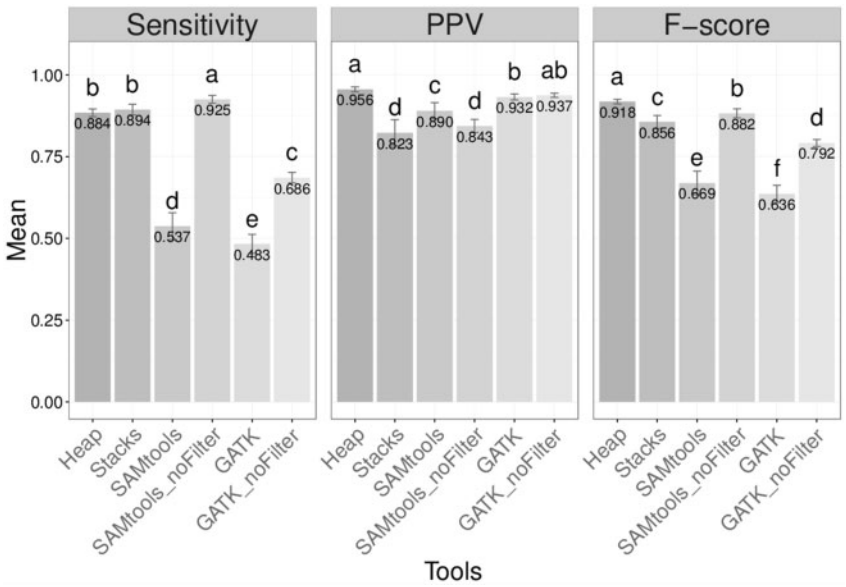

Figure 3. Performance comparison among SNP calling tools with RAD-seq reads in 17 inbred sorghum lines. Mean values of sensitivities (left), positive predictive values (PPVs) (center), and *F*-scores (right) of SNP calling by Heap, Stacks, SAMtools, and GATK from RAD-seq reads in 17 inbred sorghum lines are shown. Statistical analysis was performed using the Tukey-Kramer HSD test. Letters above the bars indicate groups that are significantly different ($P < 0.05$).

reads by referring to the definitive answer genotypes, as shown in Fig. 2C. The results indicated that Heap demonstrated a fairly high sensitivity (0.884) and the highest PPV (0.956) (Fig. 3) among the 4 tools. Additionally, we calculated F-score (*F*), which is the harmonic mean between sensitivity (*S*) and PPV (*P*) (Fig. 2C).

$$F = \frac{2SP}{S+P}$$

The *F*-score is commonly used in the field of statistical classification and employed as a measure of a test's accuracy. We found that Heap exhibited the significantly highest *F*-score (0.918) compared to the other tools according to the Tukey-Kramer honestly significant difference (HSD) test (Fig. 3). In SAMtools/BCFtools and GATK analysis, *F*-scores were low due to the low sensitivities of these tools. Here it should be noted that, in SAMtools/BCFtools and GATK analyses, not negligible amounts of SNPs with lower GQ (<13) had been eliminated by the filtering on VCF files in advance. This could lead to an underestimation of sensitivities for the 2 tools.

With the aim of conducting a fair benchmarking for each tool, we also conducted GQ filtering-free analyses of SAMtools/BCFtools and GATK. The results showed that their sensitivities increased (SAMtools/BCFtools: 0.925, GATK: 0.686), but PPVs decreased (SAMtools/BCFtools: 0.843) or did not change considerably (GATK: 0.937) in the GQ filtering-free condition (Fig. 3). Importantly, Heap still showed the highest *F*-score (0.918) in this benchmarking, while the *F*-scores of GQ filtering-free SAMtools/BCFtools and GATK were 0.882 and 0.792, respectively. These results demonstrated that Heap is the best performance tool for SNP calling from RAD-seq reads, showing sufficient sensitivity and accuracy among the 17 inbred sorghum lines.

3.3. Performance comparison among SNP calling tools with RAD-seq reads in rice

To confirm the advantages of Heap with multiple datasets, we also compared performances of the 4 tools in 4 inbred lines of rice, Omachi, Yamada Nishiki, Hitomebore, and Kameji. Prior to the

Table 3. Summary of RAD-seq reads and mapping of the reads in Sorghum

Sample	Raw reads		Preprocessed reads		Mapped reads		Uniquely mapped reads					
	Count ($\times 10^6$)	Total length (Gb)	Count ($\times 10^6$)	Total length (Gb)	Count ($\times 10^6$)	Rate (%)[a]	Count ($\times 10^6$)	Rate (%)[a]	Coverage in RAD-region			
									Mean	1st quartile	Median	3rd quartile
GULUM ABIAD	1.6	0.2	1.5	0.1	1.4	91.4	0.6	41.8	5.8	1	2	4
A-6129	1.9	0.2	1.9	0.2	1.8	92.7	0.9	45.0	7.8	1	2	7
AGIRA	2.2	0.2	2.1	0.2	2.0	94.3	1.0	47.8	8.1	1	2	7
LR 399	1.4	0.1	1.4	0.1	1.3	94.0	0.6	42.4	5.5	1	2	5
CRIOLLO CABEZA APRETADA	1.5	0.2	1.4	0.1	1.3	93.8	0.6	45.2	6.2	1	2	6
B2AR3043	2.7	0.3	2.6	0.3	2.5	93.6	1.3	48.1	9.2	1	2	7
RTx430	2.3	0.2	2.2	0.2	2.1	93.3	1.1	47.7	8.5	1	2	7
B3Tx2817	1.7	0.2	1.7	0.2	1.6	94.0	0.8	48.3	6.8	1	2	7
ZIRA-EL-SABI	2.4	0.2	2.3	0.2	2.2	94.2	1.1	48.6	7.9	1	2	7
ITALIAN	2.3	0.2	2.2	0.2	2.1	91.9	1.1	48.6	9.3	1	2	7
CAPRICORN	1.4	0.2	1.4	0.1	1.3	93.1	0.7	45.8	6.4	1	3	6
SOR 1	2.9	0.3	2.9	0.3	2.7	92.2	1.4	47.8	9.8	1	2	6
15065	1.8	0.2	1.8	0.2	1.6	92.3	0.8	45.2	6.4	1	2	5
SAP-155	2.2	0.2	2.1	0.2	2.0	93.7	1.0	48.0	7.9	1	2	6
RCV	2.0	0.2	2.0	0.2	1.9	93.9	1.0	48.8	7.4	1	2	6
SC 56	1.1	0.1	1.1	0.1	1.0	93.3	0.5	45.7	5.4	1	2	6
SIL-05	4.1	0.4	4.0	0.4	3.5	86.4	1.9	47.5	6.9	1	1	3
Total	35.3	3.6	34.8	3.3	32.1	–	16.3	–	–	–	–	–
Average	2.1	0.2	2.1	0.2	1.9	92.8	1.0	46.6	7.4	1.0	2.0	6.0

All read counts and lengths are shown in millions and billions, respectively.

[a]Mapping rates are calculated as the ratio of the number of the mapped reads against the number of the preprocessed reads.

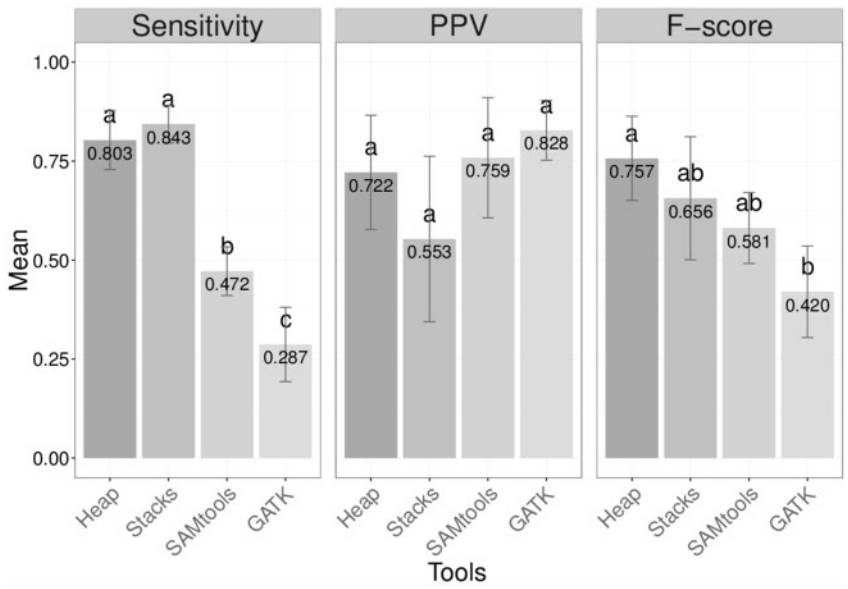

Figure 4. Performance comparison among SNP calling tools with RAD-seq reads in 4 inbred rice lines. Mean values of sensitivities (left), PPVs (center), and *F*-scores (right) of SNP calling by Heap, Stacks, SAMtools, and GATK from RAD-seq reads in 4 inbred rice lines are shown. Letters above the bars indicate groups that are significantly different ($P < 0.05$), according to the Tukey-Kramer HSD test.

SNP calling, we conducted RAD-seq and reference mapping of the RAD-seq reads (Table 4). The average RAD-seq read coverage in RAD-regions was low (9.7) for conventional SNP calling. We detected 842, 1,021, 437 and 242 SNP sites among the 4 rice lines using Heap, Stacks, SAMtools/BCFtools and GATK with the RAD-seq reads, respectively. Again, we calculated sensitivities, PPVs, and *F*-scores as mentioned above. Heap showed a high sensitivity (0.803) (Fig. 4) with a sufficiently high PPV (0.722). Here we would mention that although Stacks showed the highest sensitivity (0.843), the sensitivities of Heap and Stacks were not significantly

Table 4. Summary of RAD-seq reads and mapping of the reads in rice

Sample	Raw reads		Preprocessed reads		Mapped reads		Uniquely mapped reads					
	Count ($\times 10^6$)	Total length (Gb)	Count ($\times 10^6$)	Total length (Gb)	Count ($\times 10^6$)	Rate (%)[a]	Count ($\times 10^6$)	Rate (%)[a]	Coverage in RAD-region (average)			
									Mean	1st quartile	Median	3rd quartile
Omachi	3.6	0.2	3.6	0.2	3.5	98.4	2.4	67.3	11.2	1	2	4
Yamada Nishiki	4.2	0.2	4.2	0.2	4.1	98.1	2.8	66.9	11.7	1	1	4
Hitomebore	0.7	0.0	0.7	0.0	0.7	98.7	0.4	62.0	6.5	1	2	4
Kameji	1.9	0.1	1.9	0.1	1.8	98.4	1.3	67.2	9.2	1	2	4
Total	10.3	0.5	10.3	0.5	10.1	–	6.9	–	–	–	–	–
Average	2.6	0.1	2.6	0.1	2.5	98.4	1.7	65.9	9.7	1.0	1.8	4.0

All read counts and lengths are shown in millions and billions, respectively.
[a]Mapping rates are calculated as the ratio of the number of the mapped reads against the number of the preprocessed reads.

different, according to the Tukey-Kramer HSD test (Fig. 4). Additionally, the PPV of SNPs detected by Heap (0.722) was higher than that by Stacks (0.553). Importantly, the *F*-score of Heap (0.757) was the highest among the tools examined. These results indicate that Heap calls SNPs with a high sensitivity and high PPV with a reasonable balance between them in rice samples, too.

3.4. Performance comparison among SNP calling tools with high read coverages in sorghum

To assess the performance of Heap even with high read coverage, we identified SNPs with WGS sequencing reads among the 17 inbred sorghum lines by Heap, SAMtools/BCFtools, and GATK, and benchmarked them in sensitivities, PPVs, and *F*-scores. We did not adopt Stacks in this test, because Stacks is incompatible with WGS sequencing reads. As a result, we detected 6,153,145, 5,160,730 and 5,587,400 SNP sites using Heap, SAMtools/BCFtools and GATK, respectively. The results showed that Heap had a high *F*-score of 0.9949, which was not significantly different from the highest *F*-score of SAMtools/BCFtools (0.9952) (Supplementary Fig. S2). Compared to the *F*-scores under low read coverage, the *F*-scores of SNPs detected by SAMtools/BCFtools and GATK turned out to be considerably high (Fig. 3 and Supplementary Fig. S2). Also in Heap's case, the more read coverage available, the higher the *F*-score achieved. This result reconfirms the importance of read coverage for accurate SNP calling. However, surprisingly, the *F*-scores of SNPs detected by Heap with the high read coverage were not substantially different from that of low read coverage (Fig. 3 and Supplementary Fig. S2). These results demonstrate that Heap is applicable not only to SNP calling with high read coverage but also to that with low read coverage, with fairly high performance.

3.5. Scope of Heap

In this study, Heap exhibited the highest *F*-score in SNP calling with low read coverage when compared with the conventional tools. Heap will contribute to reducing costs of work requiring the identification of many SNP markers among multiple samples, such as with GWAS and GP studies. Heap also demonstrated sufficiently high performance compared to the other tools in situations with high read coverage.

3.6. Conclusion

In GWAS and GP studies, a large number of SNP markers are required to detect associations between SNP markers and phenotypes. On the other hand, false positive SNPs would disturb precise association or prediction. In this study, we have developed a new tool Heap. In the low read coverage condition, we demonstrated Heap's advantages in sensitivity and PPV by calculating and benchmarking *F*-scores of SNPs as long as the sorghum and the rice datasets were used. Therefore, Heap would offer fairly reliable SNPs with special reference to GWAS and GP studies. As genomic information becomes available in many species[22–25], their SNP information serves as a useful platform on comparative functional analyses. In the future, we will be maintaining and updating the function for SNP mining in Heap.

Availability of data and program

The datasets supporting the conclusions of this article are available in the DDBJ DRA repository (http://trace.ddbj.nig.ac.jp/dra/index_e.html (29 March 2017, date last accessed)), under accession numbers DRR045054-DRR045091. The source code of Heap is freely available from the git repository (https://github.com/meiji-bioinf/heap (29 March 2017, date last accessed)) and our web site (http://bioinf.mind.meiji.ac.jp/lab/en/tools.html (29 March 2017, date last accessed)).

Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR12B5, Japan. Computations were partially performed on the NIG (National Institute of Genetics) SuperComputer Facilities hosted at NIG/ROIS (Research Organization of Information and Systems).

Conflicts of Interest

None declared.

Supplementary data

Supplementary material is available at *DNARES* online.

Funding

This work was partially supported by CREST, JST to NT, the Japan Society for the Promotion of Science (JSPS) KAENHI [Grants-in-Aid for Scientific Research on Innovative Areas (No. 26113716)], MEXT-Supported Program for the Strategic

Research Foundation at Private Universities (2014–2018), and Research Funding for Computational Software Supporting Program from Meiji University to KY.

References

1. Crossa, J., Beyene, Y., Kassa, S., et al. 2013, Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3 (Bethesda)*, **3**, 1903–26.
2. Morris, G. P., Ramu, P., Deshpande, S. P., et al. 2013, Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci.*, **110**, 453–8.
3. Jarquín, D., Kocak, K., Posadas, L., et al. 2014, Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics*, **15**, 740.
4. Sonah, H., O'Donoughue, L., Cober, E., Rajcan, I. and Belzile, F. 2015, Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnol. J.*, **13**, 211–21.
5. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–60.
6. Langmead, B. and Salzberg, S. L. 2012, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–9.
7. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–79.
8. Li, H. 2011, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–93.
9. McKenna, A., Hanna, M., Banks, E., et al. 2010, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
10. DePristo, M.A., Banks, E., Poplin, R., et al. 2011, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–8.
11. Van der Auwera, G. A., Carneiro, M. O., Hartl, C., et al. 2013, *From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 11.10.1-11.10.33.
12. Cheng, A. Y., Teo, Y.-Y., and Ong, R. T.-H. 2014, Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics*, **30**, 1707–13.
13. Baird, N. A., Etter, P. D., Atwood, T. S., et al. 2008, Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS One*, **3**, e3376.
14. Zhai, Z., Zhao, W., He, C., et al. 2015, SNP discovery and genotyping using restriction-site-associated DNA sequencing in chickens. *Anim. Genet.*, **46**, 216–9.
15. Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J. H., and De Koning, D.-J. 2011, Stacks: building and genotyping loci *de novo* from short-read sequences. *G3 Genes Genomes Genetics*, **1**, 171–82.
16. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. 2013, Stacks: an analysis tool set for population genomics. *Mol. Ecol.*, **22**, 3124–40.
17. Ohyanagi, H., Takano, T., Terashima, S., et al. 2015, Plant omics data center: an integrated web repository for interspecies gene expression networks with NLP-based curation. *Plant Cell Physiol.*, **56**, e9.
18. Goodstein, D. M., Shu, S., Howson, R., et al. 2012, Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–86.
19. Kawahara, Y., de la Bastide, M., Hamilton, J. P., et al. 2013, Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 4.
20. Sakai, H., Lee, S. S., Tanaka, T., et al. 2013, Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.*, **54**, e6.
21. Paterson, A. H., Bowers, J. E., Bruggmann, R., et al. 2009, The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–6.
22. Natsume, S., Takagi, H., Shiraishi, A., et al. 2015, The draft genome of Hop (*Humulus lupulus*), an essence for brewing. *Plant Cell Physiol.*, **56**, 428–41.
23. Sakai, H., Naito, K., Takahashi, Y., et al. 2016, The *Vigna* Genome Server, '*Vig*GS': A genomic knowledge base of the genus *Vigna* based on high-quality, annotated genome sequence of the Azuki Bean, *Vigna angularis* (Willd.) Ohwi & Ohashi. *Plant Cell Physiol.*, **57**, e2.
24. Krishnakumar, V., Kim, M., Rosen, B. D., et al. 2015, MTGD: The *Medicago truncatula* Genome Database. *Plant Cell Physiol.*, **56**, e1.
25. Ohyanagi, H., Ebata, T., Huang, X., et al. 2016, OryzaGenome: genome diversity database of wild *Oryza* species. *Plant Cell Physiol.*, **57**, e1.