PyBSASeq: a novel, simple, and effective algorithm for BSA-Seq data analysis

Jianbo Zhang[1,✉] **and Dilip R. Panthee**[1,✉]

[1]Department of Horticultural Science, North Carolina State University, Mountain Horticultural Crops Research and Extension Center, 455 Research Drive, Mills River, NC 28759, USA

Bulked segregant analysis (BSA), coupled with next generation sequencing (NGS), can rapidly identify both qualitative and quantitative trait loci (QTL), and this technique is referred to as BSA-Seq here. The current methods for BSA-Seq data analysis require relatively high sequencing depth to detect major single nucleotide polymorphism (SNP)-trait associations, which leads to high sequencing cost for the organisms with a large genome. Here we developed a simple and effective algorithm for BSA-Seq data analysis and implemented it in Python, the package was named PyBSASeq. Using PyBSASeq, the likely trait-associated SNPs (ltaSNPs) were identified via Fisher's exact test, and then the ratio of the ltaSNPs and total SNPs in a chromosomal interval was used to identify the genomic regions that condition the trait of interest. The results obtained this way are similar to those generated by the current methods, but with much higher sensitivity, which makes BSA-Seq more applicable for the species with a large genome.

BSA-Seq | QTL-Seq | QTL | PyBSASeq | ltaSNP enrichment
Correspondence: ***dilip_panthee@ncsu.edu*** or ***jianbo_zhang@ncsu.edu***

Introduction

Bulked segregant analysis (BSA) has been widely utilized in the rapid identification of genetic markers associated with the trait of interest for a few decades (1, 2). The essential part of a BSA study is to construct two bulks of individuals that have contrasting phenotype (e.g. tallest plants vs. shortest plants or resistant plants vs. susceptible plants) from a segregation population. If a gene does not contribute to the trait phenotype, its alleles would be randomly segregated in both bulks; whereas if a gene is responsible for the trait phenotype, its alleles would be enriched in either bulk, e.g. one bulk has more allele *A* while the other bulk has more allele *a*. BSA was primarily used to develop genetic markers for detecting gene-trait association at its early stage (1, 2). Application of the next generation sequencing (NGS) technology to BSA eliminated the time-consuming and labor-intensive marker development and genetic mapping steps, and dramatically sped up the detection of gene-trait associations (3–18). This technique was termed either QTL-seq or BSA-Seq in different publications (5, 6, 19), we adapted the latter here because it can be applied to study both qualitative and quantitative traits.

The widely used approach in analyzing BSA-Seq data is the SNP index method (3). For each SNP, the base that is the same as in the reference genome is termed reference base (REF), and the other base is termed alternative base (ALT); the SNP index of a SNP is calculated by dividing its ALT read with the total read (locus depth, REF read + ALT read) in a bulk. The greater the Δ(SNP index) (the difference of the SNP index between bulks), the more likely the SNP contributes to the trait of interest or is linked to a gene that controls the trait. The second approach is the smoothed G-statistic method (3, 19, 20). For each SNP, a G-statistic value was calculated via G-test using the REF read and ALT read values in each bulk. The SNP with a high G-statistic value will be more likely related to the trait.

In BSA studies, bulking enriches the trait-associated alleles in either bulk, the more a gene contributes to the phenotype, the more its alleles are enriched, so are the SNPs within the gene (One bulk contains more REF read while the other bulk contains more ALT read); the SNPs flanking this gene should be enriched as well due to linkage disequilibrium, the closer the SNP to the gene, the more enrichment is achieved. Such SNPs are trait-associated SNPs. Both the SNP index method and the smoothed G-statistic method identify SNP-trait associations via quantifying the REF/ALT enrichment of a single SNP, and some of the major QTLs can be detected only with a high sequencing depth (3, 5, 20), which limits the application of BSA-Seq to the species with a large genome. Here we developed a novel, simple, and effective algorithm for analysis of the BSA-Seq data via quantifying the enrichment of trait-associated SNPs in a chromosomal interval, and a Python script, PyBSASeq, was written based on this algorithm. The sequence data of Yang et al. (3) was used to test our algorithm, and the PyBSASeq method detected more QTLs than the current methods (3, 20) even with a lower sequencing depth.

Materials and Methods

SNP calling. The raw sequences (SRR834927 and SRR834931) for BSA-Seq analysis were downloaded from NCBI using fastq-dump (https://www.ncbi.nlm.nih.gov/books/NBK158900/). Yang et al. identified six major QTLs using the sequence data and all of them are genetically verified (3). The sequences were trimmed and quality control was performed using fastp at the default setting (21). The trimmed sequences were aligned to the rice (Oryza sativa Japonica) reference genome sequence (Release 41, downloaded from https://plants.ensembl.org/Oryza_sativa/Info/Index) using BWA (22–24). SNP calling was carried out following the best practice of Genome Analysis Toolkit (GATK) (25) and GATK4 (Genome

DRAFT

Analysis Toolkit 4) tool documentation on the GATK website https://software.broadinstitute.org/gatk/documentation/tooldocs/current/. The GATK4-generated vcf file usually contains the information for two bulks; we termed them the first bulk (fb) and the second bulk (sb), respectively. The vcf file was converted to a tsv (tab separated value) file with the relevant columns (CHROM, POS, REF, ALT, fb.AD, fb.GQ, sb.AD, sb.GQ) using the GATK4 tool; Table 1 shows the first five rows of this tsv file.

SNP filtering. The GATK4-identified SNPs were filtered using the following parameters in order: 1) the SNPs with a 'NA' value in any of the above columns; 2) the SNPs with more than one ALT bases; 3) the SNPs with GQ score less than 20; 4) the unmapped SNPs or SNPs mapped to the mitochondrial or chloroplast genome.

Sliding windows. The sliding window algorithm was utilized to aid the visualization (plotting) in BSA-Seq data analysis. The window size was 2 Mb and the incremental step was 10 000 bp. Empty windows would be encountered if the amount of the SNPs is too low or the SNP distribution is severely skewed. If the total SNPs in a sliding window is zero, the value of the ltaSNP/totalSNP ratio will be replaced with the value from the previous sliding window. If the first sliding window of a chromosome is empty, the string 'empty' will be assigned to this sliding window as a placeholder that will be replaced with a non-empty value of the nearest window later.

Statistic methods.

Simulation of the REF/ALT reads. The python module numpy.random.binomial(LD, alleleFreq) was used to calculate the simulated REF/ALT reads; LD is the locus depth of a bulk (REF read + ALT read) and alleleFreq is the frequency of the ALT base in the bulk under the null hypothesis that the SNP is not associated with the trait. The module returns the simulated ALT read, and the simulated REF read is equal to LD minus the simulated ALT read. For each SNP, we obtained a pair of REF/ALT reads for each bulk, and these two pairs of REF/ALT reads can be used for Fisher's exact test, G-test, and calculation of the Δ(SNP index).

Fisher's exact test. Python module scipy.stats.fisher_exact(table) was used for Fisher's exact test; "table" contains two pairs of the REF/ALT reads (either actual or simulated) of a SNP, one from each bulk. This module returns a pair of numbers, and the second number is the p-value of the Fisher's exact test.

G-test. Python module scipy.stats.chi2_contingency(table, correction=False, lambda_='log-likelihood') was used for G-test. This module returns four numbers, and the first one is the G-statistic value.

Calculation of the Δ(SNP index) and G-statistic thresholds. For each SNP, the simulated REF/ALT reads were obtained as described above. Δ(SNP index) and G-statistic were calculated and recorded using these two pairs of simulated REF/ALT reads (one for each bulk). This process was repeated 10 000 times, the 99% confidence interval of the 10 000 Δ(SNP index) values was used as the significant threshold for the SNP index methods, and the 99.5th percentile of the 10 000 G-statistic values was used as the significant threshold for the G-statistic method.

Python script. The Python script is available on the website: https://github.com/dblhlx/PyBSASeq, and its detailed usage can be found on the website as well.

Results

Identify SNPs likely associated with the trait of interest. In BSA-Seq studies, each bulk contains many individuals that could be either homozygous (REF or ALT) or heterozygous in any SNP, and the bulk is collectively sequenced. Hence most SNPs identified via the SNP calling pipeline contains both the reference base (REF) and the alternative base (ALT) in each bulk. If a SNP belongs to or is tightly linked to a gene that contributes to the trait phenotype, a bulk will contain more REF read of the SNP while the other will contain more ALT read of the SNP due to phenotypic selection via bulking. Thus, the ALT read proportions should be significantly different between the bulks. Fisher's exact test was performed to identify such SNPs using the REF depth and the ALT depth of each SNP in both bulks. A small p-value of the Fisher's exact test indicates that the ALT proportion difference of a SNP between bulks is more likely caused by bulking, and a SNP with its p-value less than 0.01 was considered more likely associated with the trait and was termed ltaSNP here. 239 945 ltaSNPs were identified among total 1 296 234 filtered SNPs (see materials and methods section for the filter criteria), and the SNP distribution on the chromosomes was summarized in Table 2. The chromosomes 8, 1, 2, 10, and 5 contained the most ltaSNPs and had highest ltaSNP/totalSNP ratio, correlating perfectly with the chromosomes carrying the verified QTLs (3, 20).

Genomic distribution of the ltaSNPs. The ltaSNPs should cluster around the genes controlling the trait phenotype on the chromosomes due to linkage disequilibrium. Using the sliding window technique, the number of ltaSNPs was plotted on all the chromosomes to test if this was the case. It turned out the ltaSNP plot approximately matched with the major peaks in plots produced by the SNP index method and the smoothed G-statistic method (3, 20) (Figure 1A). Then we tested if any of the peaks in Figure 1A was statistically significant. There were 46 795 219 ltaSNPs (most SNPs were counted multiple times due to the sliding window algorithm) distributed among 34 763 sliding windows. Random distribution of the ltaSNPs among the sliding windows was obtained via simulation, and then the 99.5th percentile of these 34 763 values was used as the threshold to determine if a peak in Figure 1A was statistically significant. The threshold obtained this way was 1442. All the peaks corresponding to the six major QTLs identified

DRAFT

Table 1. The first five rows of the GATK4 output file

CHROM[a]	POS[b]	REF[c]	ALT[d]	834927.AD[e]	834927.GQ[f]	834931.AD[e]	834931.GQ[f]
1	29 759	C	G	0,2	6	0,2	6
1	31 071	A	G	25,39	99	33,29	99
1	31 478	C	T	27,38	99	48,32	99
1	33 667	A	G	21,46	99	39,32	99
1	34 057	C	T	29,37	99	32,31	99

[a] The chromosome the SNP is located.
[b] The position of the SNP on the chromosome.
[c] The base sequence of the SNP that is the same as the one from the reference genome.
[d] The base sequence that is different from REF.
[e] The allele depth of the SNP in the first bulk (834927) or the second bulk (834931). This column contains two numbers, the first one is the REF read and the second is the ALT read.
[f] The genotype quality of the SNP in the first bulk (834927) or the second bulk (834931).

Table 2. SNP distribution on chromosomes

	ltaSNP	totalSNP	ltaSNP/totalSNP
Chr1	52 045	159 905	0.325
Chr2	48 860	124 284	0.393
Chr3	3483	45 401	0.077
Chr4	3735	62 335	0.060
Chr5	15 451	101 719	0.152
Chr6	7621	158 809	0.048
Chr7	12 658	127 456	0.099
Chr8	54 307	131 510	0.413
Chr9	1698	57 272	0.030
Chr10	28 642	97 764	0.293
Chr11	5212	182 193	0.029
Chr12	6233	47 586	0.131
Genome-wide	239 945	1 296 234	0.185

by Yang et al. (3) were well above this threshold, and additional statistically significant peaks on chromosomes 2, 3, 6, 7, 8, and 12 were observed as well (Figure 1A).

Genomic distribution of the ratio of ltaSNPs and total SNPs. The above approach would work perfectly if the SNPs were evenly distributed along the chromosomes. This assumption was obviously not true; as shown in Figure 1A (the blue curve), there were nearly zero SNPs in some sliding windows while there were more than 15 000 SNPs in other sliding windows. If a gene that conditions the trait is located in a region with fewer SNPs, it would be missed using the above approach. Thus, we used the ratio of ltaSNPs and total SNPs in a chromosomal region for measuring the ltaSNP enrichment; the ltaSNP/totalSNP ratios were plotted for all the chromosomes (Figure 1B), and the plot pattern matched very well with that produced by the smoothed G-statistic method (3, 20). The most obvious difference between Figure 1A and figure 1B was the first peak on chromosome 2 and the peaks on chromosomes 3, 6 and 9; these regions contained fewer SNPs, but the ltaSNPs enrichment was relatively high.

Under the null hypothesis that the SNPs were not associated with the trait, resampling was utilized to obtain the threshold to determine which peak in Figure 1B was statistically significant. 300 SNPs were randomly selected from the SNP dataset. For each SNP in this sample, its simulated REF read and ALT read were calculated as detailed in the materials and methods section, and then the simulated REF read and ALT read from both bulks were used to perform Fisher's exact test. A SNP with its p-value less than 0.10 was considered a ltaSNP (A high cut-off p-value results in a high threshold), and the ratio ltaSNP/totalSNP (total SNPs was the sample size) in the sample was recorded. This process was repeated 10 000 times, and the 99.5th percentile of these 10 000 values was used as the significant threshold for the detection of peak-trait associations. The threshold obtained this way was 0.127. In addition to the six major QTLs (two of them on chromosome 2) verified in the work of Yang et al. (3), a new peak on chromosome 2, 3, 6, or 9 was also well above the threshold (Figure 1B); among these peaks, the peak on chromosome 6 or 9 was below the threshold in Figure 1A because they were located in the region with fewer SNPs. All the SNPs on chromosome 8 were above the threshold, which was probably caused by linkage disequilibrium because the chromosome contained a major QTL nearby the end of the long arm and a minor QTL nearby the end of the short arm.

Sequencing depth affected the detection of SNP–trait association. The sequencing depth of the SNPs (locus depth, REF depth + ALT depth) is 62 folds in the first bulk and 74 folds in the second bulk in the work of Yang et al. (3), and the maximum LD is 163 in the first bulk while the maximum LD is 206 in the second bulk. It would be very costly to achieve such high sequencing depth in the species with a large genome. Thus, we wanted to know how sequencing depth affects the detection of SNP-trait associations. We further filtered the SNP dataset with different LD values to obtain SNP datasets with a lower sequencing depth, and then plotted the ltaSNP/totalSNP ratio distribution as above. LD value filtering did not change the overall SNP distribution pattern (Figure S1). The plotting patterns of ltaSNP/totalSNP were very similar at different LD levels. However, the ltaSNP/totalSNP ratio was decreased dramatically with the decreasing LD depth while the threshold did not change much. Hence, more and more minor SNP-trait associations were missed, and more background noise was observed in the ltaSNP/totalSNP plots (Figure 2). When LD $\leq$ 40 (the average LD was 30 in the first bulk and 25 in the

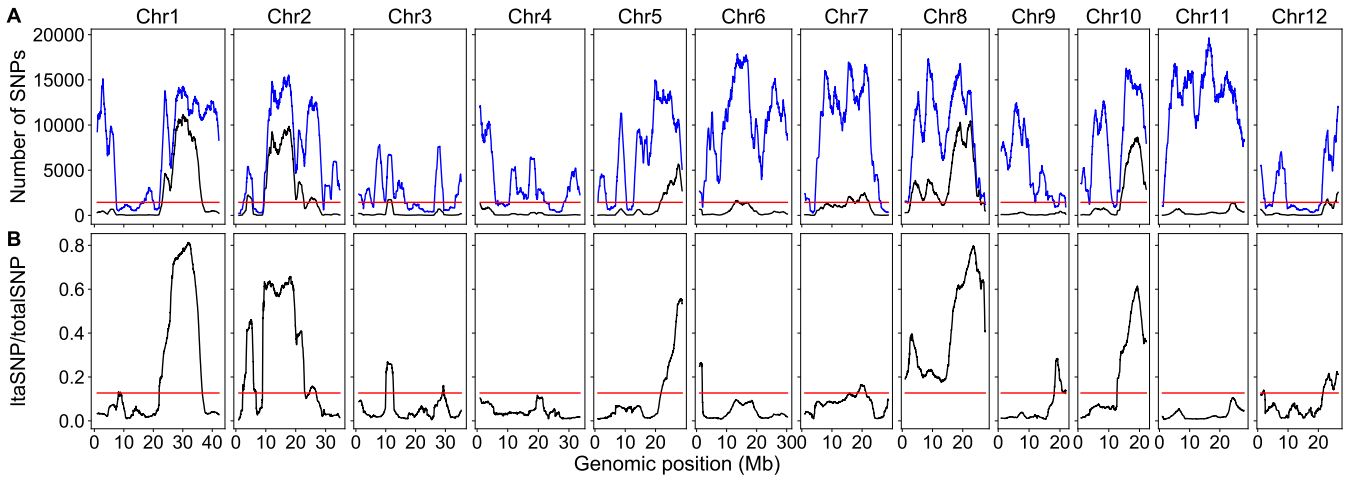

Figure 1. Genomic distributions of the SNPs and the ratios of ItaSNPs and total SNPs. The red horizontal lines are the thresholds. **(A)** The ItaSNPs (black) and total SNPs (blue). **(B)** The ratio of ItaSNPs and total SNPs.

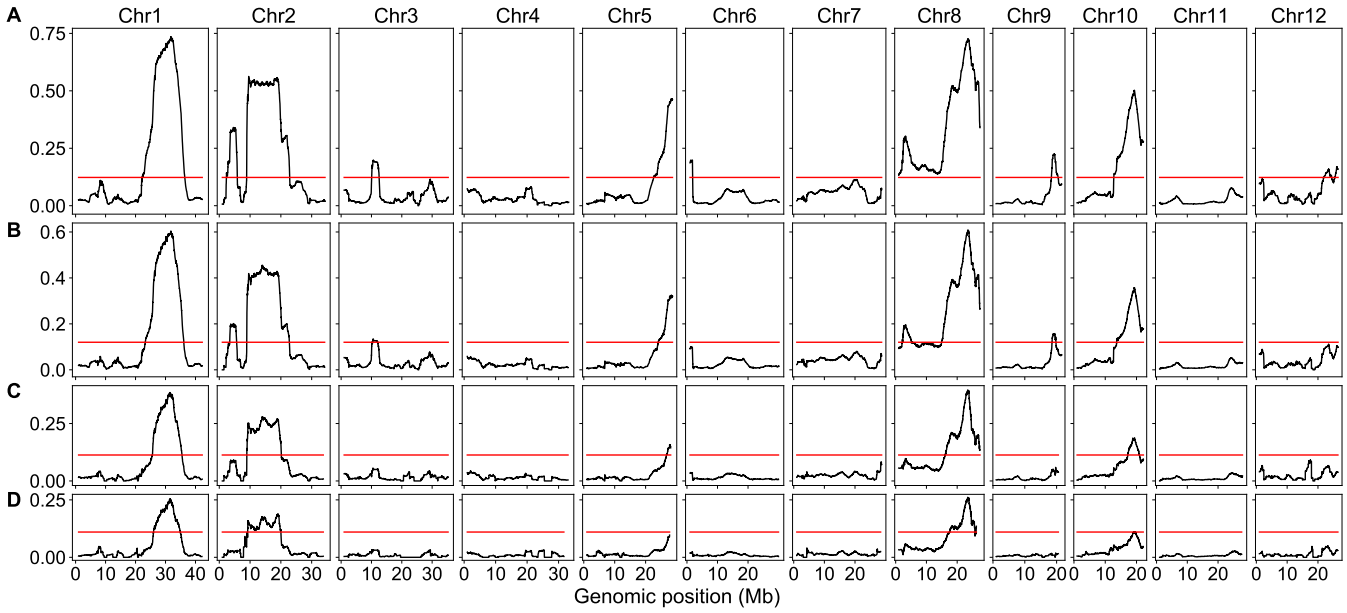

Figure 2. Genomic distribution of the ratio of ItaSNPs and totalSNPs at different sequencing depth levels. The red horizontal lines are the thresholds. **(A)** LD $\leq$ 80. **(B)** LD $\leq$ 60. **(C)** LD $\leq$ 40. **(D)** LD $\leq$ 30.

second bulk), all the major QTLs can still be detected. However, the major QTL on chromosome 5 was missed when LD $\leq$ 30 (the average LD was 23 in the first bulk and 18 in the second bulk).

Sensitivity comparison. The above data indicated that the PyBSASeq approach had higher detection power. However, different methods were used to generate the SNP datasets (3, 20), which might lead to different detection sensitivities. To rule out this possibility, we implemented the SNP index method and the G-statistic method in Python, and tested all the three methods with the same SNP dataset. First, we tested if the results of Yang et al. and Mansfeld and Grumet can be replicated using our scripts. As in the studies mentioned above, the SNP dataset was filtered with the following criteria: fb.GQ $\geq$ 99, sb.GQ $\geq$ 99, fb.LD $\geq$ 40, sb.LD $\geq$ 40, fb.LD+sb.LD $\geq$ 100, and fb.LD+sb.LD $\leq$ 400. Although the SNP datasets were generated in different ways (GATK4 vs. GATK vs. Samtools) and no smoothing was applied in our scripts, the results, including the plot patterns, the G-statistic values, and the Δ(SNP index) values and its confidence intervals, were very similar (3, 20), and the positions of the peaks/valleys matched almost perfectly between different approaches (Figure S2). A non-parametric method was used to calculate the threshold in the smoothed G-statistic method by Yang et al. and Mansfeld and Grumet, and different approaches were used to remove the G-statistic values from the QTL regions; thus the thresholds were a little different in these studies, so was the QTL detection results (3, 20). In our G-statistic script, we used simulation for threshold

calculation (see the materials and methods), and the thresholds obtained this way were consistent across all the chromosomes and was less conservative than the previously reported approaches. Using the high sequencing depth SNP subset, similar results were obtained by both the SNP index method and the G-statistic method: the six major QTLs and a minor QTL on chromosome 2 were detected (Figure S2). However, the PyBSASeq approach had the highest sensitivity using the same filtering criteria, and it can detect more minor QTLs than other methods even if the whole SNP dataset was used (Figure 1B, Figure 2, and Figure S2).

As in PyBSASeq, we also tested how the sequencing depth affected the detection of the SNP-trait associations in these two methods. Using the whole SNP dataset as in the PyBSASeq method, the SNP index method had very low detection power, the major QTL on chromosome 5 was missed and the peak (valley) representing the major QTL on chromosome 10 was barely beyond the threshold; the QTLs on chromosomes 2 and 10 were missed when LD $\leq$ 80 and all the QTL were missed when LD $\leq$ 40 (Figure 3). For the G-statistic method, with the whole set of the SNPs, all the 6 major QTLs can be detected; the QTLs on chromosomes 2, 5, and 10 were missed when LD $\leq$ 60 and all the QTLs were missed when LD $\leq$ 40 (Figure 4).

Discussion and Conclusions

Both the SNP index method and the smoothed G-statistic method identify SNP-trait association by measuring REF/ALT enrichment of a single SNP; whereas the PyBSASeq method identifies SNP-trait association by measuring ltaSNP enrichment in a chromosomal region. The average number of SNPs was 7155 in a sliding window, much higher than the average locus depth of the SNPs (62 in the first bulk and 74 in the second bulk), which could be why PyBSASeq has much higher statistical power.

Detection of the SNP-trait association via PyBSASeq is determined by the ltaSNP/totalSNP ratio and the significant threshold. Our results indicated that the ltaSNP/totalSNP ratio decreased dramatically with the decreasing sequencing depth (locus depth, LD), while the threshold seemed not vary much (Figure 2). We performed simulation to have a clearer understanding how altering sequencing depth affects the threshold (Figure S3). Here the threshold is defined as the 99.5th percentile of 10 000 simulated ltaSNP/totalSNP ratios at a certain LD level. The simulation revealed that increasing sequencing depth led to increased threshold when the sequencing depth was low (less than 20) and the thresholds largely fluctuated between 0.10-0.14 when the sequencing depth was greater than 20. This, however, does not mean that more significant SNP-trait associations can be detected at a low sequencing depth, because the ltaSNP/totalSNP ratio was low at a low sequencing depth (Figure 2). Table S1 further explains why the ltaSNP/totalSNP ratio is low when the sequencing depth is low. This table contains the Fisher's exact test p-values for different REF/ALT combinations at LD=7; only the very extreme cases (High REF proportion in one bulk and high ALT proportion in the other bulk) have their p-values less than 0.01, and thus can be considered as ltaSNPs. With increasing sequencing depth, the proportion of the ltaSNP cells becomes higher, e.g. 6 of 64 cells have their p-values less than 0.01 at ld=7, whereas 10 of 81 cells have their p-values less than 0.01 at LD=8. The maximum threshold is 0.14 in the simulation; hence, it is safe to just use 0.14 as the significant threshold to save computing time when the sequencing depth is high.

PyBSASeq can detect all the major QTLs when the average locus depth was 30 in the first bulk and 25 in the second bulk; whereas the other methods missed all the QTL detection at the same LD levels. Because of the high sensitivity of the PyBSASeq method, SNP-trait associations can be detected at reduced sequencing depth, leading to reduced sequencing cost. Thus, BSA-Seq can be more practically applied to the species with a larger genome.

ACKNOWLEDGEMENTS

Jianbo Zhang was supported by the grant IOS-1546625 awarded to Dilip R. Panthee by the National Science Foundation. We are grateful to Yang et al. for generating the sequencing data and making it available to the public. The manuscript was prepared using a modified version of the Henriques lab bioRxiv LaTex template.

Bibliography

1. James J. Giovannoni, Rod A. Wing, Martin W. Ganal, and Steven D. Tanksley. Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. *Nucleic Acids Research*, 19(23):6553–6568, December 1991. ISSN 0305-1048. doi: 10.1093/nar/19.23.6553.
2. R. W. Michelmore, I. Paran, and R. V. Kesseli. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences of the United States of America*, 88(21):9828–9832, November 1991. ISSN 0027-8424.
3. Zemao Yang, Daiqing Huang, Weiqi Tang, Yan Zheng, Kangjing Liang, Adrian J. Cutler, and Weiren Wu. Mapping of Quantitative Trait Loci Underlying Cold Tolerance in Rice Seedlings via High-Throughput Sequencing of Pooled Extremes. *PLOS ONE*, 8(7):e68433, July 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0068433.
4. Jian Song, Zhen Li, Zhangxiong Liu, Yong Guo, and Li-Juan Qiu. Next-Generation Sequencing from Bulked-Segregant Analysis Accelerates the Simultaneous Identification of Two Qualitative Genes in Soybean. *Frontiers in Plant Science*, 8:919, 2017. ISSN 1664-462X. doi: 10.3389/fpls.2017.00919.
5. Hiroki Takagi, Akira Abe, Kentaro Yoshida, Shunichi Kosugi, Satoshi Natsume, Chikako Mitsuoka, Aiko Uemura, Hiroe Utsushi, Muluneh Tamiru, Shohei Takuno, Hideki Innan, Liliana M. Cano, Sophien Kamoun, and Ryohei Terauchi. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *The Plant Journal: For Cell and Molecular Biology*, 74(1):174–183, April 2013. ISSN 1365-313X. doi: 10.1111/tpj.12105.
6. Fabien Duveau, Brian P. H. Metzger, Jonathan D. Gruber, Katya Mack, Natasha Sood, Tiffany E. Brooks, and Patricia J. Wittkopp. Mapping small effect mutations in Saccharomyces cerevisiae: impacts of experimental design and mutational properties. *G3 (Bethesda, Md.)*, 4(7):1205–1216, April 2014. ISSN 2160-1836. doi: 10.1534/g3.114.011783.
7. Harry Klein, Yuguo Xiao, Phillip A. Conklin, Rajanikanth Govindarajulu, Jacob A. Kelly, Michael J. Scanlon, Clinton J. Whipple, and Madelaine Bartlett. Bulked-Segregant Analysis Coupled to Whole Genome Sequencing (BSA-Seq) for Rapid Gene Cloning in Maize. *G3 (Bethesda, Md.)*, 8(11):3583–3592, 2018. ISSN 2160-1836. doi: 10.1534/g3.118.200499.
8. Anastasie Sigwalt, Claudia Caradec, Christian Brion, Jing Hou, Jacky de Montigny, Paul Jung, Gilles Fischer, Bertrand Llorente, Anne Friedrich, and Joseph Schacherer. Dissection of quantitative traits by bulk segregant mapping in a protoploid yeast species. *FEMS yeast research*, 16(5), 2016. ISSN 1567-1364. doi: 10.1093/femsyr/fow056.
9. Josh Clevenger, Ye Chu, Carolina Chavarro, Stephanie Botton, Albert Culbreath, Thomas G. Isleib, C. C. Holbrook, and Peggy Ozias-Akins. Mapping Late Leaf Spot Resistance in Peanut (Arachis hypogaea) Using QTL-seq Reveals Markers for Marker-Assisted Selection. *Frontiers in Plant Science*, 9:83, 2018. ISSN 1664-462X. doi: 10.3389/fpls.2018.00083.
10. Asuka Kodama, Ryouhei Narita, Makoto Yamaguchi, Hiroshi Hisano, Shunsuke Adachi, Hiroki Takagi, Taiichiro Ookawa, Kazuhiro Sato, and Tadashi Hirasawa. QTLs maintaining grain fertility under salt stress detected by exome QTL-seq and interval mapping in barley. *Breeding Science*, 68(5):561–570, December 2018. ISSN 1344-7610. doi: 10.1270/jsbbs.18082.
11. Huaiyong Luo, Manish K. Pandey, Aamir W. Khan, Jianbin Guo, Bei Wu, Yan Cai, Li Huang, Xiaojing Zhou, Yuning Chen, Weigang Chen, Nian Liu, Yong Lei, Boshou Liao, Rajeev K. Varshney, and Huifang Jiang. Discovery of genomic regions and candidate genes controlling shelling percentage using QTL-seq approach in cultivated peanut (Arachis hypogaea L.). *Plant Biotechnology Journal*, December 2018. ISSN 1467-7652. doi: 10.1111/pbi.13050.

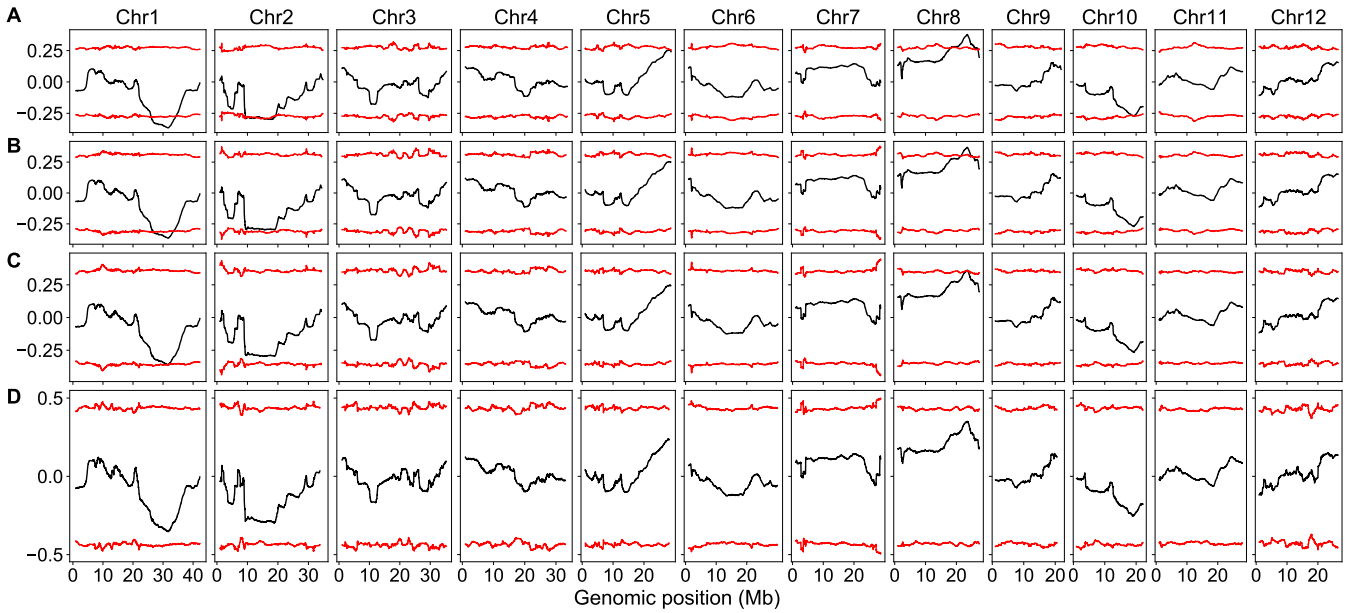

Figure 3. Chromosomal distribution of Δ(SNP index). The red curves are the confidence intervals. **(A)** All the SNPs were included. **(B)** LD ≤ 80. **(C)** LD ≤ 60. **(D)** LD ≤ 40.

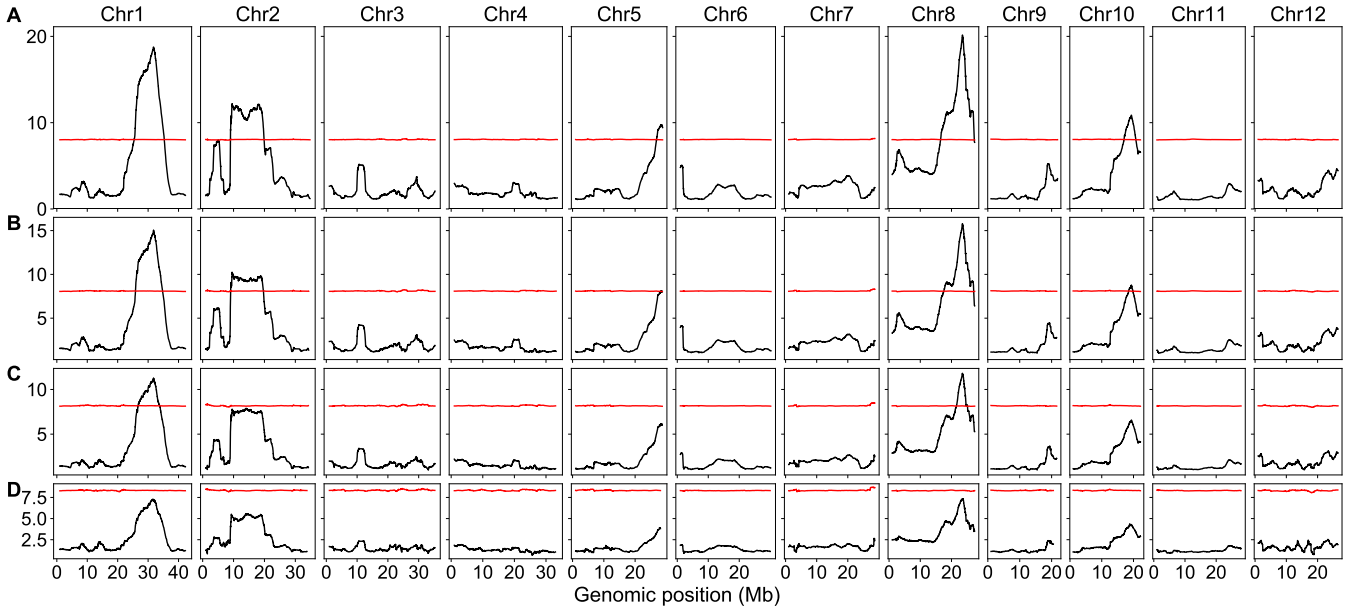

Figure 4. Chromosomal distribution of G-statistic. The red curves are the thresholds. **(A)** All the SNPs were included. **(B)** LD ≤ 80. **(C)** LD ≤ 60. **(D)** LD ≤ 40.

12. Xiaoli Zhang, Wubin Wang, Na Guo, Youyi Zhang, Yuanpeng Bu, Jinming Zhao, and Han Xing. Combining QTL-seq and linkage mapping to fine map a wild soybean allele characteristic of greater plant height. *BMC genomics*, 19(1):226, 2018. ISSN 1471-2164. doi: 10.1186/s12864-018-4582-4.
13. Chao Zhong, Suli Sun, Yinping Li, Canxing Duan, and Zhendong Zhu. Next-generation sequencing to identify candidate genes and develop diagnostic markers for a novel Phytophthora resistance gene, RpsHC18, in soybean. *Theoretical and applied genetics*, 131 (3):525–538, March 2018. ISSN 1432-2242. doi: 10.1007/s00122-017-3016-z.
14. Kaijing Zhang, Xing Wang, Wenwei Zhu, Xiaodong Qin, Jian Xu, Chunyan Cheng, Qunfeng Lou, Ji Li, and Jinfeng Chen. Complete resistance to powdery mildew and partial resistance to downy mildew in a Cucumis hystrix introgression line of cucumber were controlled by a co-localized locus. *Theoretical and applied genetics*, 131(10):2229–2243, October 2018. ISSN 1432-2242. doi: 10.1007/s00122-018-3150-2.
15. Hongfeng Lu, Tao Lin, Joël Klein, Shenhao Wang, Jianjian Qi, Qian Zhou, Jinjing Sun, Zhonghua Zhang, Yiqun Weng, and Sanwen Huang. QTL-seq identifies an early flowering QTL located near Flowering Locus T in cucumber. *Theoretical and applied genetics*, 127 (7):1491–1499, July 2014. ISSN 1432-2242. doi: 10.1007/s00122-014-2313-z.
16. Muluneh Tamiru, Satoshi Natsume, Hiroki Takagi, Benjamen White, Hiroki Yaegashi, Motoki Shimizu, Kentaro Yoshida, Aiko Uemura, Kaori Oikawa, Akira Abe, Naoya Urasaki, Hideo Matsumura, Pachakkil Babil, Shinsuke Yamanaka, Ryo Matsumoto, Satoru Muranaka, Gezahegn Girma, Antonio Lopez-Montes, Melaku Gedil, Ranjana Bhattacharjee, Michael Abberton, P. Lava Kumar, Ismail Rabbi, Mai Tsujimura, Toru Terachi, Wilfried Haerty, Manuel Corpas, Sophien Kamoun, Günter Kahl, Hiroko Takagi, Robert Asiedu, and Ryohei Terauchi. Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. *BMC biology*, 15(1):86, 2017. ISSN 1741-7007. doi: 10.1186/s12915-017-0419-x.
17. Fei Shen, Zhenyu Huang, Baoguo Zhang, Yi Wang, Xi Zhang, Ting Wu, Xuefeng Xu, Xinzhong Zhang, and Zhenhai Hai. Mapping Gene Markers for Apple Fruit Ring Rot Disease Resistance Using a Multi-omics Approach. *G3 (Bethesda, Md.)*, March 2019. ISSN 2160-1836. doi: 10.1534/g3.119.400167.
18. Junling Dou, Xuqiang Lu, Aslam Ali, Shengjie Zhao, Lei Zhang, Nan He, and Wenge Liu. Genetic mapping reveals a marker for yellow skin in watermelon (Citrullus lanatus L.). *PloS One*, 13(9):e0200617, 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0200617.
19. Paul M. Magwene, John H. Willis, and John K. Kelly. The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing. *PLOS Computational Biology*, 7(11):e1002255, November 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002255.

20. Ben N. Mansfeld and Rebecca Grumet. QTLseqr: An R Package for Bulk Segregant Analysis with Next-Generation Sequencing. *The Plant Genome*, 11(2), 2018. ISSN 1940-3372. doi: 10.3835/plantgenome2018.01.0006.
21. Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, September 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty560.
22. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, July 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp324.
23. Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–595, March 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp698.
24. Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, https://arxiv.org/abs/1303.3997v2, March 2013.
25. Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, September 2010. ISSN 1549-5469. doi: 10.1101/gr.107524.110.

Word Counts

Statistics on word count.

```
File: BSASeq_Final.tex
Encoding: utf8
Sum count: 5127
Words in text: 4887
Words in headers: 84
Words outside text (captions, etc.): 134
Number of headers: 22
Number of floats/tables/figures: 6
Number of math inlines: 22
Number of math displayed: 0
Subcounts:
  text+headers+captions (#headers/#floats/#inlines/#displayed)
  24+11+0 (1/0/0/0) _top_
  156+1+0 (1/0/0/0) Abstract
  13+0+0 (0/0/0/0) main
  525+1+0 (1/0/0/0) Section: Introduction
  0+3+0 (1/0/0/0) Section: Materials and Methods
  172+2+9 (1/1/0/0) Subsection: SNP calling
  55+2+0 (1/0/0/0) Subsection: SNP filtering
  113+2+0 (1/0/0/0) Subsection: Sliding windows
  267+20+0 (5/0/0/0) Subsection: Statistic methods
  24+2+0 (1/0/0/0) Subsection: Python script
  0+1+0 (1/0/0/0) Section: Results
  233+9+4 (1/1/0/0) Subsection: Identify SNPs likely associated with the trait of interest
  187+5+36 (1/1/0/0) Subsection: Genomic distribution of the ltaSNPs
  416+10+33 (1/1/4/0) Subsection: Genomic distribution of the ratio of ltaSNPs and total SNPs
  233+8+0 (1/0/2/0) Subsection: Sequencing depth affected the detection of SNP-trait associati
  451+2+52 (1/2/16/0) Subsection: Sensitivity comparison
  487+3+0 (1/0/0/0) Section: Discussion and Conclusions
  1531+2+0 (2/0/0/0) Section: Bibliography
```

DRAFT

Supplementary Figures

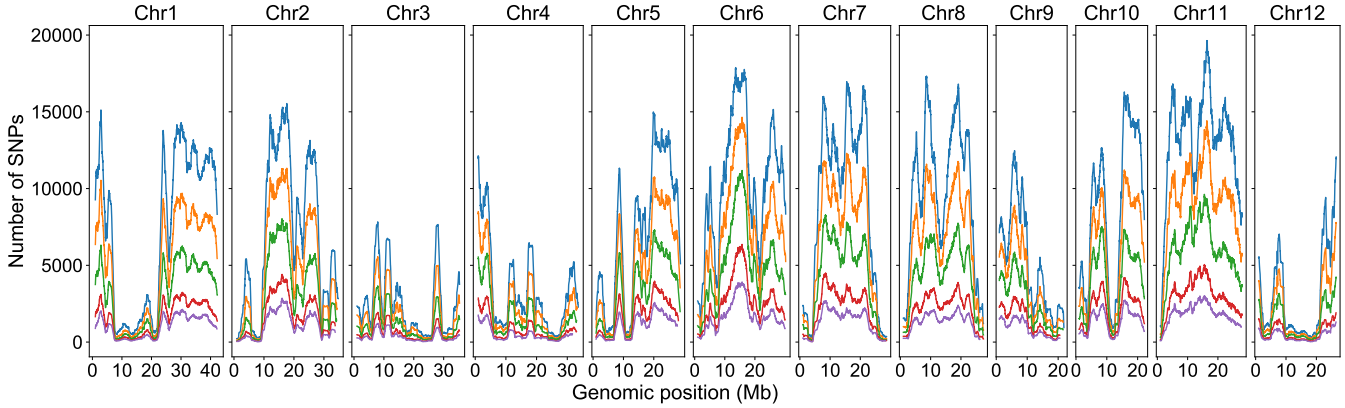

Figure S1. SNP distribution at different LD levels. Blue: All SNPs were included; orange: LD $\leq$ 80; green: LD $\leq$ 60; red: LD $\leq$ 40; purple: LD $\leq$ 30.

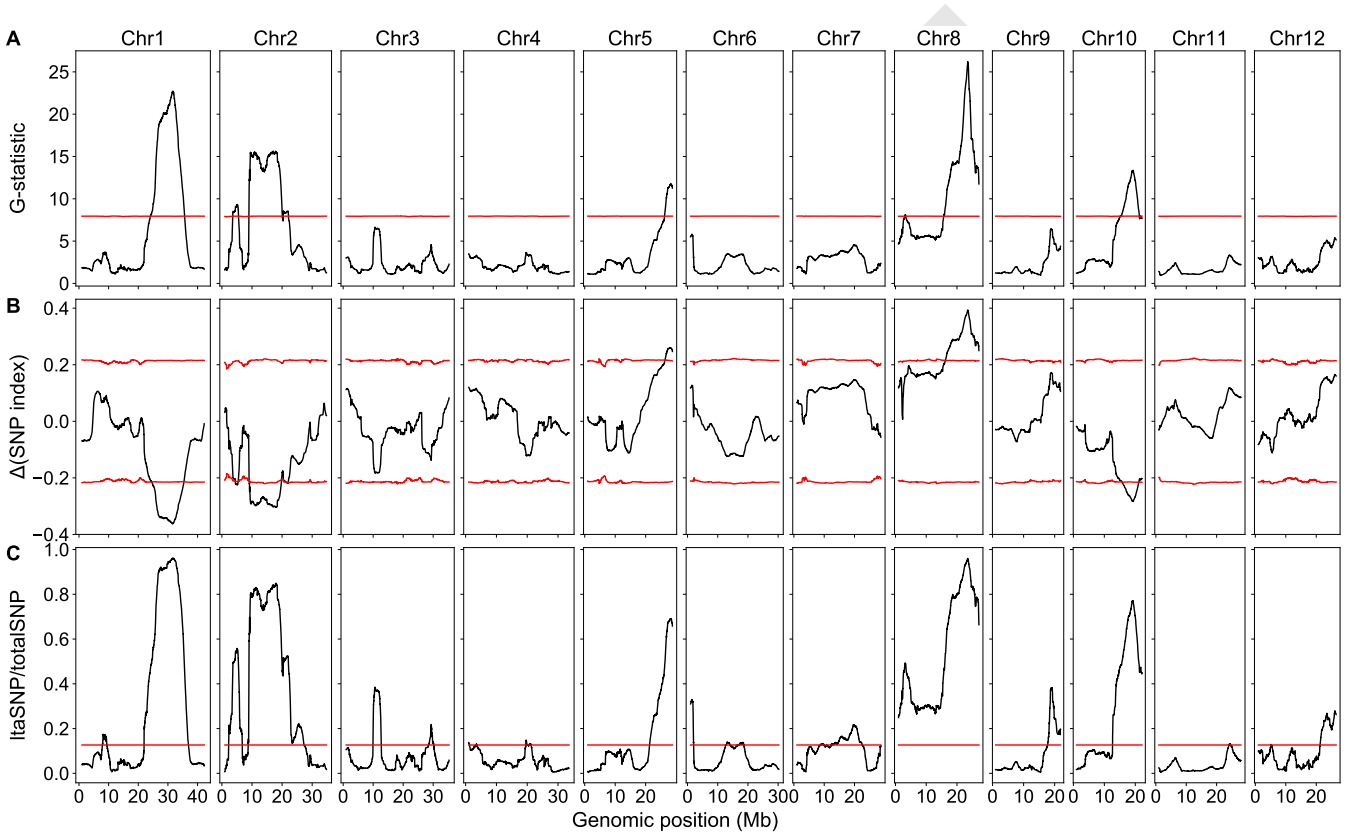

Figure S2. Replication of the SNP index method and the G-statistic method in Python. The red curves or lines indicate the thresholds. **(A)** the G-statistic method. **(B)** the SNP index method. **(C)** the PyBSASeq method included here for comparison.

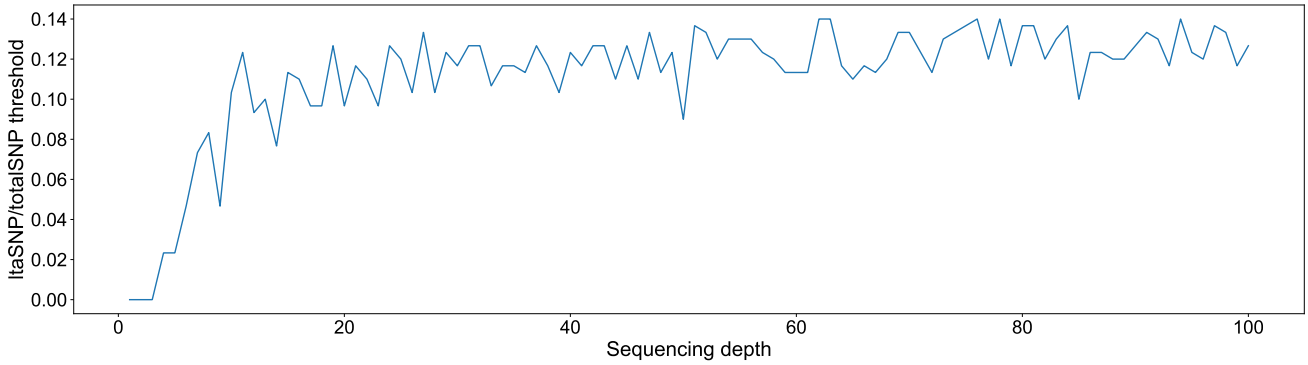

Figure S3. Simulated ItaSNP/totalSNP thresholds at different sequencing depths. Up to 25% variation was randomly introduced to the sequencing depth to mimic the real-world situation, e.g. the locus depth in either bulk could be in the range of 6–10 when the sequencing depth was 8.

Supplementary Tables

Table S1. The p-value table of the Fisher's exact test at LD=7

fb.AD \ sb.AD	0,7	1,6	2,5	3,4	4,3	5,2	6,1	7,0
0,7	1	1	0.462	0.192	0.070	0.021	0.005	0.001
1,6	1	1	1	0.559	0.266	0.103	0.029	0.005
2,5	0.462	1	1	1	0.592	0.286	0.103	0.021
3,4	0.192	0.559	1	1	1	0.592	0.266	0.070
4,3	0.070	0.266	0.592	1	1	1	0.559	0.192
5,2	0.021	0.103	0.286	0.592	1	1	1	0.462
6,1	0.005	0.029	0.103	0.266	0.559	1	1	1
7,0	0.001	0.005	0.021	0.070	0.192	0.462	1	1

DRAFT