The Linkage Method: A Novel Approach for SNP Detection and Haplotype Reconstruction from a Single Diploid Individual Using Next-Generation Sequence Data

Eriko Sasaki,*,‡,[1] Ryuichi P. Sugino,[1] and Hideki Innan*,[1]

[1]Graduate University for Advanced Studies, Hayama, Kanagawa, Japan

‡Present address: Gregor Mendel Institute of Molecular Plant Biology, Dr. Bohr-Gasse 3, 1030, Vienna, Austria

***Corresponding authors:** E-mail: eriko.sasaki@gmi.oeaw.ac.at; innan_hideki@soken.ac.jp.

Associate editor: Naoko Takezaki

Abstract

When we sequence a diploid individual, the output actually comprises two genomes: one from the paternal parent and the other from the maternal parent. In this study, we introduce a novel heuristic algorithm for distinguishing single-nucleotide polymorphisms (SNPs) from the two parents and phasing them into haplotypes. The algorithm is unique because it simultaneously performs SNP calling and haplotype phasing. This approach can exploit the linkage information of nearby SNPs, which facilitates the efficient removal of haplotypes that originate from incorrectly mapped short reads. Using simulated data we demonstrated that our approach increased the accuracy of SNP calls. The haplotype reconstruction performance depended largely on the density of SNPs. Using current next-generation sequence technology with a relatively short read length, reasonable performance is expected when this approach is applied to species with an average of five heterozygous sites per 1 kb. The algorithm was implemented as the program "linkSNPs."

Key words: **SNP calling, haplotype phasing, next generation sequence, algorithm, software.**

Introduction

Next-generation sequencing (NGS) technology produces vast amounts of DNA sequences with high throughput and low costs. One of the major applications of NGS is indentifying every single difference (so-called polymorphism) between closely related genomes, such as those between close species and those between individuals within a single species. The detectable variations include single-nucleotide polymorphisms (SNPs), insertions/deletions, and copy number variations, which are essential for a wide range of research areas, including medicine, genetics, agriculture, ecology, and evolution (Wang et al. 2008; Fujimoto et al. 2010; Huang et al. 2010; Lam et al. 2010; van Bers et al. 2010). NGS-based approaches are especially powerful when a reference sequence is available so numerous short reads produced by NGS can be mapped onto the reference sequence, thereby allowing any type of variation to be identified. However, these techniques are challenging, particularly the mapping process. Various structural differences, such as indels, duplications, and inversions, between the reference genome and the sequenced genome cause the incorrect mapping of short reads, thereby resulting in incorrect calls when detecting variations such as SNPs (Li and Homer 2010).

In this study, we introduce a novel heuristic algorithm for improving the accuracy of SNP calling based on the exploitation of linkage information. When NGS is applied to a diploid individual, the output sequence actually comprises two haploid genomes. Thus, all of the SNPs are derived from either the maternal or paternal parents and they are completely linked in each lineage. This linkage information can be very helpful for identifying the true source region of a short read, especially when it contains multiple homologous regions (only one is the true origin whereas the other is a paralogous region created by duplication). We demonstrated that our algorithm delivers reasonably good SNP detection performance. More importantly, however, our algorithm produces haplotype information, which should be useful in many applied situations. We refer to this algorithm as the linkage method.

Our algorithm is different from standard SNP calling and haplotype phasing because it performs both of these processes simultaneously. By contrast, the standard procedure is conducted step-by-step. SNP calling is performed with SNP calling programs such as SAMtools (Li et al. 2009a) and GATK (McKenna et al. 2010) using short read data that has already been mapped onto a reference genome. The basic algorithm used by these programs screens the mapped data to search for sites in the mapped reads where the nucleotides differ from the reference sequence. SNP calling is conducted site-by-site (Li et al. 2009a, 2009b; McKenna et al. 2010), i.e., all sites are independently treated and information is not considered from nearby SNPs. If necessary, the called SNPs can be used to phase haplotypes.

By contrast, our algorithm performs SNP calling and haplotype phasing simultaneously. Like other algorithms, our method is based on short reads that have been mapped onto a reference genome using another program (e.g., BWA [Li and Durbin 2009]), but our algorithm does not directly move to SNP calling. First, our method observes the segregation patterns of the linked SNPs and connects them as

haplotypes. When a diploid is sequenced, there should be two major haplotypes, but there may be other haplotypes from paralogous regions, which should not be used during correct SNP calling. Our algorithm is effective for removing the haplotypes created by incorrectly mapped reads, which improves the efficiency of SNP calling, as shown below.

The concept of haplotype phasing using linkage information is not entirely new. For example, the read-back phasing algorithms of GATK (McKenna et al. 2010) perform haplotype phasing where the raw short reads are used in the process. A similar process is also employed by the haplotype improver (HI; Long et al. 2009), although those algorithms are used for phasing haplotypes in a population sample, not a single individual. Furthermore, the haplotype phasing process is independently performed after SNP calling is completed in these methods, and linkage information is not incorporated in SNP calling.

Algorithm

The algorithm is illustrated in figure 1. The algorithm is based on paired-end short reads that are mapped onto a reference genome, as shown in figure 1*A*. We assume that the data comprise a large number of short reads of α bp. The paired reads are connected by thin broken lines. First, the nucleotides are identified where at least one read has different nucleotides compared with the reference genome and these sites are referred to as variable sites. For each variable site in figure 1*A*, the reference alleles are shown in blue and the others in red or yellow (yellow indicates a second alternative allele, if necessary).

The algorithm uses a window analysis approach where a window of α bp is moved along the reference genome in a 5′ to 3′ direction. The window stops when the first site in the window is a variable site. In the example shown in figure 1*A*, the window is located where site b is the first site in the window and the analysis focuses on variable sites downstream of this site. The algorithm uses information from all reads that overlap with the window. In addition, if the window overlaps with the 5′ reads, their pairs (3′ reads) are also used. In the example shown in figure 1*B*, there are 41 reads with five variable sites within the window (sites b–f) and one (site g) in the paired reads. First, variable sites where an alternative nucleotide (shown in red) is fixed in all reads are recognized as certain differences based on the reference and excluded from the following haplotype phasing. In this example, five variable sites remain after this screening process because of one fixed difference (site d).

Figure 1*C* shows the segregation pattern at these five sites. The 37 reads are classified into 15 distinct patterns, after excluding four reads that do not cover any variable sites (represented by empty boxed for all sites at the bottom of the leftmost panel in figure 1*C*). From the 15 distinct patterns, we can parsimoniously assemble five "major" haplotypes such that each of the 15 patterns is consistent with at least one of the five major haplotypes. Then, we compute pre-scores of these five major haplotypes, which are based on the frequencies of compatible haplotypes with some penalties due to missing data and alignment gaps. The pre-score for the i^{th} major haplotype is denoted by s_i, which is given by

$$s_i = \sum_{j=1}^{m} \frac{l - k_j}{l \cdot g_j}, \quad (1)$$

where m is the number of haplotypes that are consistent with the i^{th} major haplotype and l is the number of variable sites. For the first major haplotype in figure 1*C*, $m = 23$ and $l = 5$. k_j is the number of missing sites in l variable sites, which are represented by open boxes in figure 1*C*. g_j is the penalty for alignment gaps, i.e., the number of gaps between the read and reference genomes.

Next, the algorithm excludes any singletons from subsequent analysis because of their unreliability. The fourth major haplotype is excluded in the example shown in figure 1*C*. The sites e and f are "non-variable sites" so they are excluded. Thus, there are only four major haplotypes with three variable sites (b, c, and g). The relative frequencies of the alleles can be computed for these three sites, e.g., for the first site b, the frequency of the allele identical to the reference (blue boxes) is eight while those of the two alternative alleles (red and yellow boxes) are 10 and 2, respectively. Therefore, their relative frequencies to the most common allele (the first alternative allele shown by red boxes) are given by $r = 0.8$, $r' = 1$ and $r'' = 0.2$, respectively. Similarly, the relative frequencies of the reference and alternative alleles at site c are $r = 0.8$, $r' = 1$ and $r'' = \text{NA}$, respectively, while those for site g are $r = 1$, $r' = 0.69$ and $r'' = \text{NA}$. Using this information, we can compute a weighting factor for the i^{th} major haplotype, w_i, which is defined as the lowest value among the relative frequencies of the alleles in the l variable sites. In the example shown in figure 1*C*, the first major haplotype comprises the reference alleles at all three sites so we have $w_1 = \min[0.8,0.8,1] = 0.8$, while $w_2 = \min[0.8,0.8,0.69] = 0.69$ because it possesses the alternative allele at site g. The third haplotype has the alternative alleles at all three sites, where w_3 is given by $\min[1,1,0.69] = 0.69$. The fourth haplotype has the second alternative alleles only at site b, where w_4 is given by $\min[0.2,\text{NA},\text{NA}] = 0.2$. Using these weighting factors, we obtain the final scores for the major haplotypes as follows:

$$s_i' = w_i \times s_i. \quad (2)$$

This score is quite small if it includes very rare alleles, which are possibly due to sequencing and/or mapping errors (e.g., w_4).

Based on this score, the four major haplotypes in figure 1*C* are ranked. The top one with the highest score (8.160) is assigned as the Rank 1 haplotype, the third one with the second highest score (4.554) is assigned as the Rank 2 haplotype, while the others with the third to lowest scores (2.484, 0.080) are assigned as Rank 3. When applied to a diploid individual, the process usually produces two major haplotypes (Ranks 1 and 2) and additional major haplotypes appear occasionally, probably due to sequencing errors and/or mapping errors caused by paralogous regions.

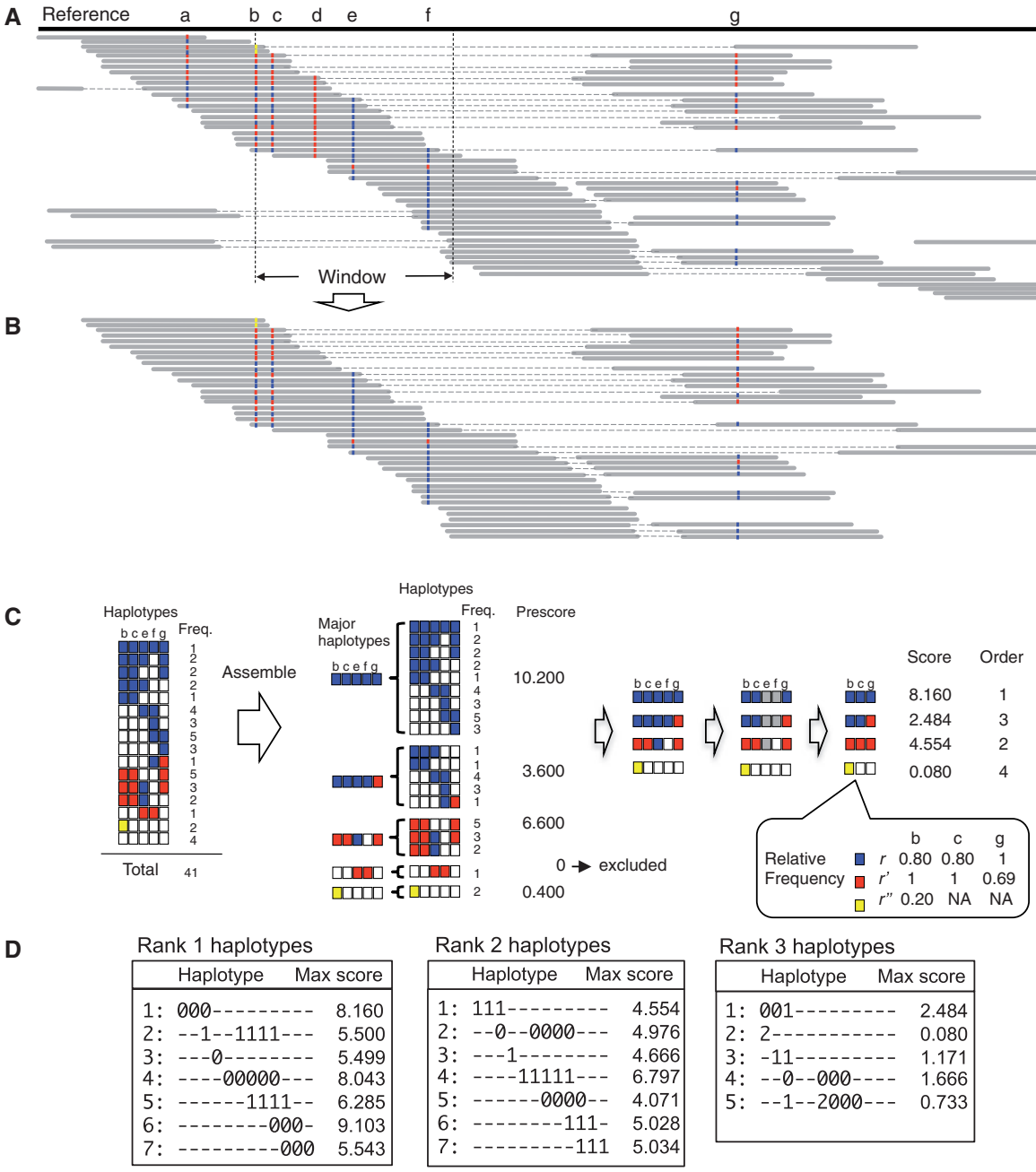

Fig. 1. Overview of the algorithm. (*A*) An example of short read data mapped onto a reference sequence. (*B*) All of the short reads used for the analysis in the example window. (*C*) Assembling the local haplotypes. (*D*) Summary of the local haplotypes. See text for details.

This process can be carried out for other windows and the results are summarized in figure 1*D*, where the haplotypes are grouped according to their ranks. The nucleotide states at the variable sites are presented by 0, 1, and 2, instead of blue, red, and yellow boxes, respectively. There are seven Rank 1 haplotypes and seven Rank 2 haplotypes. This indicates that at least seven windows are involved. It is possible that multiple windows will produce the exact same set of major haplotypes so we take the highest score (presented by Max score in fig. 1*D*). We refer to these window-based haplotypes as local haplotypes.

The haplotype phasing process is performed by concatenating these local haplotypes. We employ a simple greedy algorithm. In brief, a simple concatenation process is performed in a large number of possible ways, which takes the best result according to the concept of minimum fragment removal (MFR) (Geraci 2010). Each run of the haplotype concatenation process is conducted as follows.

i) Choose a "seed" haplotype, which should be one of the most informative of the Rank 1 haplotypes, i.e., those containing information in the largest number of variable sites. This seed haplotype is denoted by H_1 and is used to start the following concatenation process.
ii) Select another Rank 1 haplotype randomly. If this haplotype is consistent with H_1 they are concatenated and

H_1 is updated using the resulting concatenated haplotype, whereas this haplotype is discarded otherwise. This step is iterated until no more haplotypes are consistent with H_1.

iii) Continue the concatenation process using Rank 2 haplotypes. The procedure is identical to step ii and the resulting concatenated haplotype is denoted by H_1'.

iv) Pool all of the Rank 1 and 2 haplotypes that have not been used to construct H_1', which is denoted by Δ_{-1}. Select the most informative haplotype from Δ_{-1}, which is used as an alternative seed haplotype (H_2) to construct a secondary concatenated haplotype in the same way as step i.

v) Concatenate this secondary seed haplotype using other haplotypes in Δ_{-1} in a random order by following step ii, then make a concatenated haplotype H_2'. Δ_{-2} denotes the pool containing all the remaining haplotypes that were not merged with H_2'.

vi) H_1' is compared with the Rank 3 haplotypes and all the Rank 3 haplotypes that are consistent with H_1' are merged with H_1', which yields the final concatenated haplotype H_1''. In the same manner, we construct another final concatenated haplotype H_2'' using H_2' and the remaining Rank 3 haplotypes.

vii) Compute the scores for H_1'' and H_2'' as the sum of the scores of the local haplotypes that are involved (see eq. 2). The scores are denoted by S_1 and S_2, respectively.

viii) Determine the final result. If S_1 is much larger than S_2 (say, β times), then we consider that there is only one primary haplotype H_1'', which should originate from a paralogous region, and vice versa. Otherwise, H_1'' and H_2'' are considered to represent the paternal and maternal haplotypes in the heterozygous state.

This process can be repeated for a large number of replicates using all possible seed haplotypes. In this study, this process is repeated five times for each seed haplotype; hence, if there are t Rank 1 haplotypes, the total number of replications is $5 \times t$ (if $t > 30$, the top 30 informative haplotypes are used as the seed haplotypes). Next, we consider that the best result is the replicate with the lowest number of haplotypes in Δ_{-2}.

Figure 2 shows a typical output of our algorithm, i.e., a pair of relatively long haplotypes with 15 variable sites (from sites 01–15). The 16th site in some haplotypes is not connected to them successfully because the distance between the 15th and 16th sites is too long to phase. Thus, the outputs of the algorithm comprise a number of haplotype blocks. Note that phasing is feasible only when there are multiple heterozygous sites within a distance that can be covered by a single-paired read (i.e., $2\alpha + \nu$ bp, where ν is the length of the insert between the 5′ to 3′ reads), so that their coupling/decoupling information will be available.

In addition to the two major haplotypes, a third haplotype is proposed. This haplotype is shorter than the two major haplotypes and is supported mainly by the Rank 2 and 3 haplotypes, while the two major haplotypes are supported by the Rank 1 and 2 haplotypes. It is suggested that this third haplotype originates from a paralogous region because it is short and hardly connected with the majority of the reads. In our algorithm, these haplotypes are excluded from SNP calling. For example, sites 09, 10, and 11 are variable sites but they are not called heterozygous SNPs because they are not variable between the two major haplotypes.

Thus, the phasing process is performed in each haplotype block, so that the total run time would be the sum of the run time of all haplotype blocks. The time complexity of the phasing process in each haplotype block can be roughly described as $O\{n \times m^2\}$, where n is the number of heterozygous sites in the block and m is the read depth. The phasing process for each block can be independently performed in a parallel way after the genome is divided into blocks.

Performance of the Algorithm

The overall performance of our method was examined using simulated data. To produce the simulated data, we assumed that a diploid genome of 10 Mb was sequenced. It was also assumed that a reliable reference sequence (haploid) was available for the same species. To generate these sequences, we used a coalescent theory-based simulator, "ms" (Hudson 2002), which generated SNP patterns using the population-scaled mutation rate (θ) and recombination rate (ρ) in the simplest demographic setting (i.e., a constant population size). The sample size was set to three, one of which was randomly assigned to the reference genome and we used the other two to make the diploid individual to be sequenced. According to the coalescent theory, the nucleotide divergence between the three sequences was expected to be identical to θ, so that $\theta = 0.01$ indicates that there was an average of 10 SNPs in a 1 kb region between any pair of the three genomes. We assumed $\theta = \rho = 0.01$ in this study, unless indicated otherwise. To incorporate the effect of duplications that were probably caused by incorrect read mapping, random regions in the sequenced genome were duplicated so the entire genome size became $\omega = 30\%$ larger than the original size (this parameter was also varied later). We selected random templates of duplicated regions (allowing overlaps), assuming that the length of each region followed a uniform distribution between 1 and 10,000 bp. These duplicated regions were inserted into random locations in the genome. During this process, divergence between the original and duplicated regions was introduced by adding random mutations in the duplicated region. The divergence level was determined as a random variable between 0% and 50%.

Next, we produced paired-end short read data from the diploid individual. Five different read lengths were used ($\alpha = 50$, 75, 100, 150 and 200 bp). The distance between paired reads was assumed to follow a normal distribution with the mean = 125 bp and standard deviation = 50 bp. Sequence errors were added to each read with a probability of 0.001 per site. The base quality was defined uniformly as 32 (phred scale: this value predicts an error rate of ~0.001 to ensure consistency with the sequence error rate), regardless of the sequence errors. These simulated short reads were mapped onto the reference genome using the BWA program (Li and Durbin 2009), which was used to apply our algorithm.

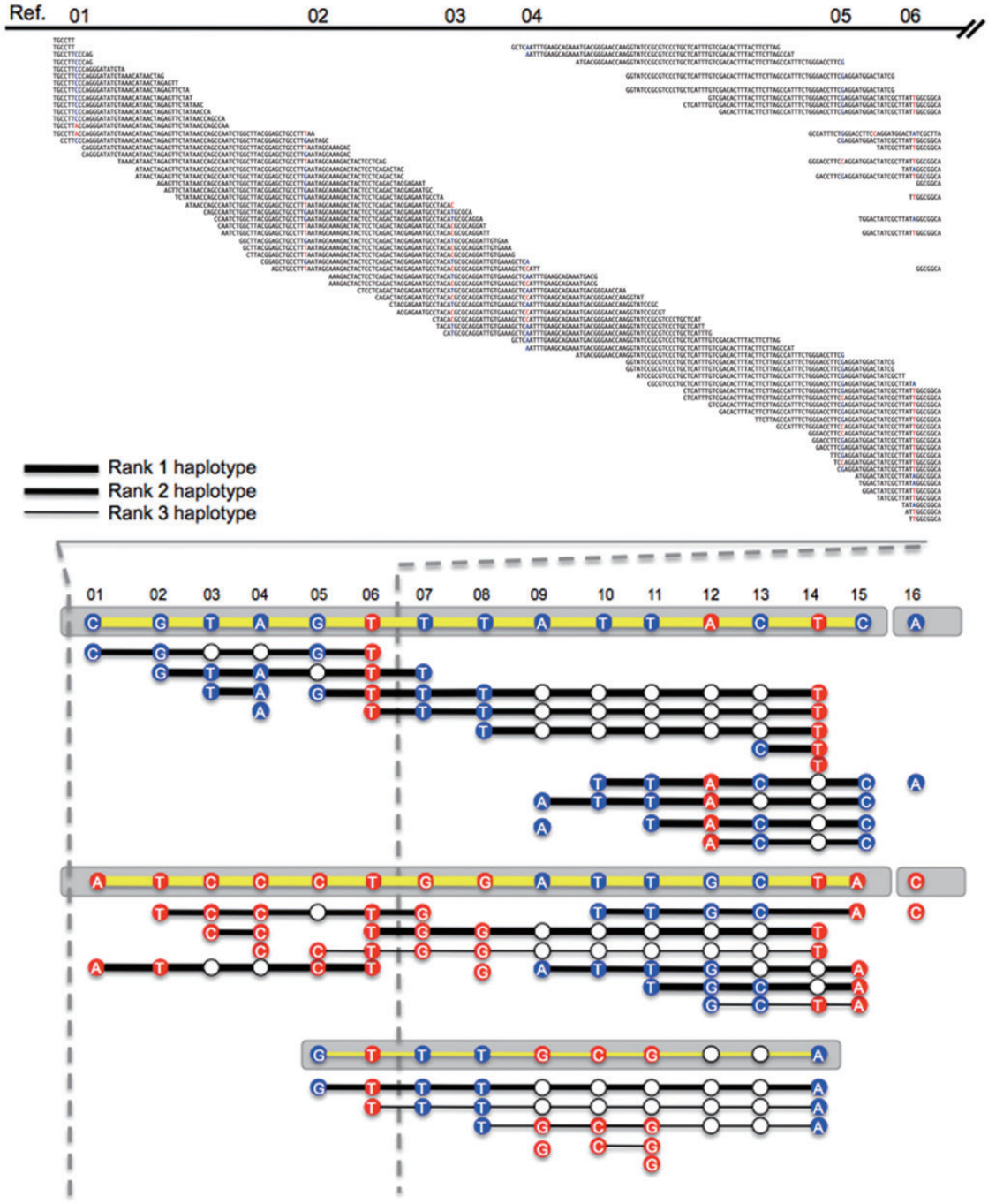

FIG. 2. An example of the phasing step of the algorithm. The two haplotypes connected with thick yellow lines are heterozygous for the region encompassing sites 01–15. In addition, another haplotype is called for the region covering sites 05–14, which originates from an external source. Local haplotypes that support each of the three haplotypes are also shown. The open circles represent sites with no information.

We assumed $\beta = 10$ in step viii. Using a fixed threshold value is a fairly standard procedure when determining homozygous and heterozygous sites, and this method works reasonably well with reasonable read depths, according to a recent review by Nielsen et al. (2011). They also suggested $\beta = 5$ because the frequency of the non-reference allele was distribute between 20% and 80% at a heterozygous site (Nielsen et al. 2011). In our algorithm, as a number of very minor haplotypes are already excluded, the distribution of the frequency of the non-reference allele is wider, i.e., approximately between 10% and 90% (data not shown). Therefore, to be consistent with this wide distribution, we assumed that $\beta = 10$, which is recommended as the default value for our algorithm.

First, we evaluated the haplotype reconstruction and SNP calling performance using the data with $\alpha = 75$. The sequence depth was assumed to be $\times 100$. We found that the entire genome was divided into 6,891 haplotype blocks (shown in italic in table 1). The average length was 801.8 bp (excluding blocks containing only one or two heterozygous sites) and the

Table 1. Effect of Read Length (α) on the Performance of the Linkage Method.

	Haplotype Blocks				Haplotype Accuracy			SNP Calling		
Read Length (α)	Number of Haplotype Blocks	Average Length	Maximum Length	Coverage	Accuracy I	Accuracy II	Total Accuracy	True Positive (Homo)	True Positive (Hetero)	False Positive
50	8711	505.7	4521	44.1	91.21	77.22	77.20	98.82	85.83	8.24
75	*6891*	*801.8*	*11215*	*55.3*	*76.49*	*76.69*	*72.20*	*99.07*	*87.57*	*4.73*
100	5405	1063.2	9598	57.5	89.34	81.74	81.74	99.29	88.66	3.68
150	3783	1728.5	15558	65.4	86.97	79.14	79.14	98.86	87.05	2.15
200	2660	2695.7	24578	71.7	81.95	72.29	72.26	99.30	86.97	1.84

maximum length was over 10 kb. These haplotype blocks covered approximately 60% of the entire genome.

The accuracy of haplotype phasing was evaluated using two metrics: haplotype accuracy I and II. Haplotype accuracy I was the proportion of haplotype blocks where the paternal and maternal haplotypes were called correctly as heterozygote. A typical error was that one was called a homozygote whereas the other was called as a haplotype from an external source (i.e., a paralogous region). Haplotype accuracy II was the proportion of haplotype blocks where all of the detected heterozygous SNPs were accurately assigned to the paternal and maternal haplotypes. The total accuracy was the proportion of haplotype blocks where these two criteria were satisfied, i.e., the proportion of haplotype blocks that did not contain any errors. We found that haplotype accuracy I and II were 76.49% and 76.69%, respectively, while the total accuracy was 72.2% (shown in italics in table 1).

SNP calling accuracy was evaluated using three metrics, the true-positive rates for the homozygous sites and heterozygous sites, and the false-positive rate of all other errors. The true-positive rates for homozygous sites and heterozygous sites were approximately 99% and 87%, respectively, while the false-positive rate was $<5\%$.

The performance depends on many parameters, including the read length (α) and depth. The effects of the read length are summarized in table 1. As α increased, the number of haplotype blocks decreased whereas the lengths of the haplotype blocks increased. This is because the genome is probably divided into small blocks if the read length is short so the linkage information is poor. Overall, the coverage increased with α. The read length did not appear to have strong effects on the haplotype accuracy. The effect on the two true-positive rates was not high during SNP calling but the false-positive rate decreased as α increased.

Table 2 summarizes the effect of the read depth when the read length was fixed at $\alpha = 75$ bp. The read depth was varied from $\times 20$ to 120. The effect on the overall performance appeared to be small, except that there may have been a weak positive correlation between the read depth and the haplotype length at a low depth.

The effects of the genomic background are explored in tables 3 and 4. Table 3 summarizes the effect of duplication, where the average length of the haplotype blocks and the haplotype coverage increased as the proportion of duplicated regions increased. This was because incorrectly mapped reads produced a number of false heterozygous sites, which help to concatenate haplotypes. As a consequence, SNP calling performance also declined with decreasing the haplotype accuracy.

The effect of θ on haplotype phasing was very high (table 4) because it determined the density of SNPs directly. Our algorithm required at least two SNPs in a paired read for haplotype phasing. Therefore, given $\alpha = 75$, the performance of haplotype phasing was not good when θ was very small. It appeared that $\theta = 0.01$ was required to produce a coverage $>50\%$. When $\theta = 0.001$ and 0.002, the coverage was $<20\%$ and the SNP calling false-positive rate was very high. When $\theta = 0.02$ and 0.05, the haplotype phasing performance was good, while the two true-positive SNP calling rates were significantly lower than that when $\theta = 0.01$. This was simply because the number of mapped reads was dramatically reduced with these high levels of divergence from the reference genome. Haplotype accuracies I and II were low when $\theta = 0.02$ and 0.05 because of the way they were defined, i.e., the haplotype accuracy was defined as the proportion of haplotypes where all heterozygous SNPs were perfectly assigned so it is well known that the haplotype accuracy declines as the average haplotype length increases (reviewed by Browning and Browning 2011). Although this metrics is used widely, it might be better to develop new ones that are robust to the haplotype length.

Although the performance was affected by many factors, the most important appeared to be the density of heterozygous SNPs in the length covered by a typical paired-end read (i.e., $2\alpha + \nu$). In the ideal situation of "perfect mapping" where all of the genomic regions are accurately mapped with a sufficient number of reads, the density of heterozygous SNPs should depend on θ and ρ. θ simply increases the density of SNPs, while ρ affects the spatial distribution of SNPs so the distribution is more uniform with a larger ρ, and a small ρ leads to high variation in the local density according to the basic coalescent theory (Hudson 1983). Therefore, we have a rough idea of how long SNPs are phased into haplotyes, which depends on θ and ρ and the degree to which the assumptions of "perfect mapping" are violated. This is illustrated in figure 3, where we assume that $\theta = \rho = 0.01$ and $\omega = 0$. In the ideal condition, the "maximum" performance is expected; that is, haplotype phasing should be perfectly performed in a block within which the distances between all adjacent heterozygous SNPs are smaller than $2\alpha + \nu$. In figure 3, this situation is

Table 2. Effect of Read Depth on the Performance of the Linkage Method.

	Haplotype Blocks				Haplotype Accuracy			SNP-calling		
Read Depth	Number of Haplotype Blocks	Average Length	Maximum Length	Coverage	Accuracy I	Accuracy II	Total Accuracy	True Positive (homo)	True Positive (hetero)	False Positive
20	9300	513.5	9186	47.8	81.83	77.18	74.22	98.51	83.17	5.11
40	7130	766.2	11241	54.6	76.72	76	72.33	98.92	86.83	5.01
60	6984	788.7	11241	55.1	76.42	76.73	72.42	99.04	87.12	4.90
80	6881	803.4	11211	55.3	76.18	77.21	72.55	99.07	87.38	5.00
100	*6891*	*801.8*	*11215*	55.3	*76.49*	*76.69*	*72.2*	*99.07*	*87.67*	*4.97*
120	6891	802.9	13143	55.3	76.64	77.65	72.83	99.08	87.36	4.83

Table 3. Effect of the Proportion of Duplicated Regions (ω) on the Performance of the Linkage Method.

	Haplotype Blocks				Haplotype Accuracy			SNP Calling		
Proportion of Dup. Regions (ω)	Number of Haplotype Blocks	Average Length	Maximum Length	Coverage	Accuracy I	Accuracy II	Total Accuracy	True Positive (homo)	True Positive (hetero)	False Positive
0%	7698	602.3	6600	46.3	90.53	83.06	83.05	99.19	90.66	0.00
10%	6804	777.3	11053	52.9	82.64	79.94	78.06	99.10	90.30	1.78
30%	*6891*	*801.8*	*11215*	*55.3*	*76.49*	*76.69*	*72.2*	*99.07*	*87.57*	*4.73*
50%	6869	807.7	12416	55.5	75.67	76.42	71.66	99.05	86.75	5.46

Table 4. Effect of the Population-Scaled Mutation Rate (θ) on the Performance of the Linkage Method.

	Haplotype Blocks				Haplotype Accuracy			SNP Calling		
Mutation Rate (θ)	Number of Haplotype Blocks	Average Length	Maximum Length	Coverage	Accuracy I	Accuracy II	Total Accuracy	True Positive (homo)	True Positive (hetero)	False Positive
0.001	2208	429.6	10529	9.5	75.27	90.49	72.51	99.88	95.53	38.81
0.002	4223	428.6	9821	18.1	80.72	87.52	76.94	99.85	93.55	25.73
0.005	6274	554.5	11041	34.8	81.48	82.31	76.63	99.85	90.83	8.46
0.01	*6891*	*801.8*	*11215*	*55.3*	*76.49*	*76.69*	*72.2*	*99.07*	*87.67*	*4.97*
0.02	5055	1337.3	27182	67.6	69.48	69.24	64.83	87.94	68.92	2.48
0.05	6019	962.1	21108	57.9	76.91	78.29	73.67	53.38	31.36	1.94

shown in red. With the "maximum" performance, the average length of haplotype blocks increases dramatically with increasing paired-end read length (Fig. 3*A*), and the density distribution is shown as a function of the number of heterozygous SNPs in a block in figure 3*B*. A low read depth should be one of the most significant factor to reduce the performance. We demonstrated this by using the simulation results presented in tables 1 and 2. As expected, the average length of haplotype blocks decreased with decreasing the read depth (Fig. 3*A*), and the distribution of haplotype lengths was more skewed toward short ones when the read depth is lower (Fig. 3*B*).

Comparison of SNP Calling Performance Using SAMtools and GATK

The advantage of our algorithm is that it produces haplotype information from a single individual. As described above, the process is essential for removing incorrectly mapped reads. This should improve the quality of SNP calls, although it might cause loss of power in SNP detection. To examine these possibilities, we compared the SNP calling performance using two widely used programs: SAMtools (Li et al. 2009a) and GATK (McKenna et al. 2010). We compared SAMtools and GATK using the same simulated dataset (shown in tables 1–4). For SAMtools (version 0.1.12a), the "samtools pileup" command was used for SNP calling with BAQ option (Li 2011). For GATK (version 2.0-35), UnifiedGenotyper was used for SNP calling with the basic default settings, except the heterozygosity configuration was set according to the value of θ used in the simulation. The threshold for SNP calls was assumed to be SNP Quality >20 with both programs.

Figure 4 summarizes the true- and false-positive rates for homozygous sites. The true positive rate with our linkage method was higher than those with SAMtools and GATK using all parameter sets, although GATK performed almost as well as the linkage method. GATK had the highest false-positive rate and the linkage method performed almost as well as SAMtools. Thus, the linkage method gave reasonably good performance when detecting homozygous SNPs

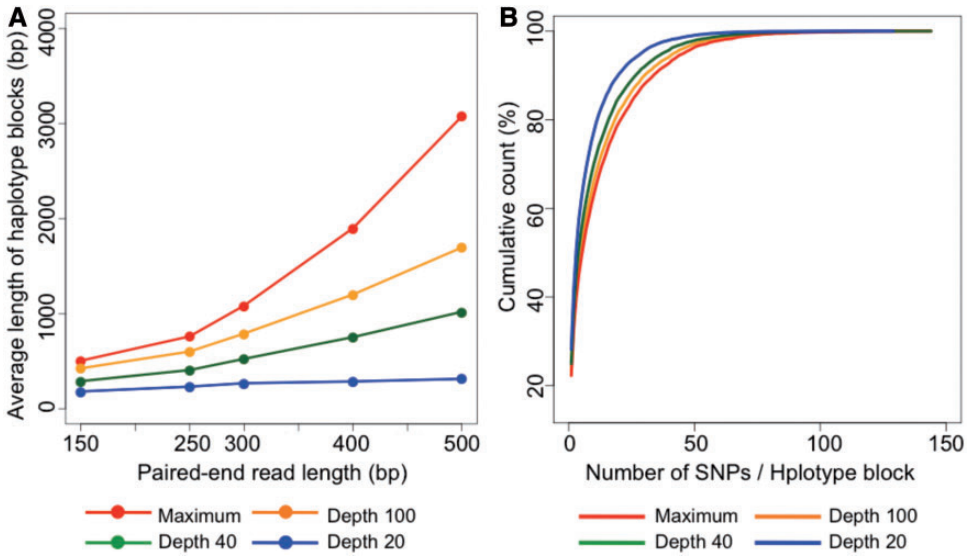

FIG. 3. Performance of the linkage method during haplotype phasing. (*A*) The average length of haplotype blocks plotted against paired-end read length, $2\alpha + \nu$. (*B*) The distribution of the number of heterozygous SNPs in phased haplotype blocks.

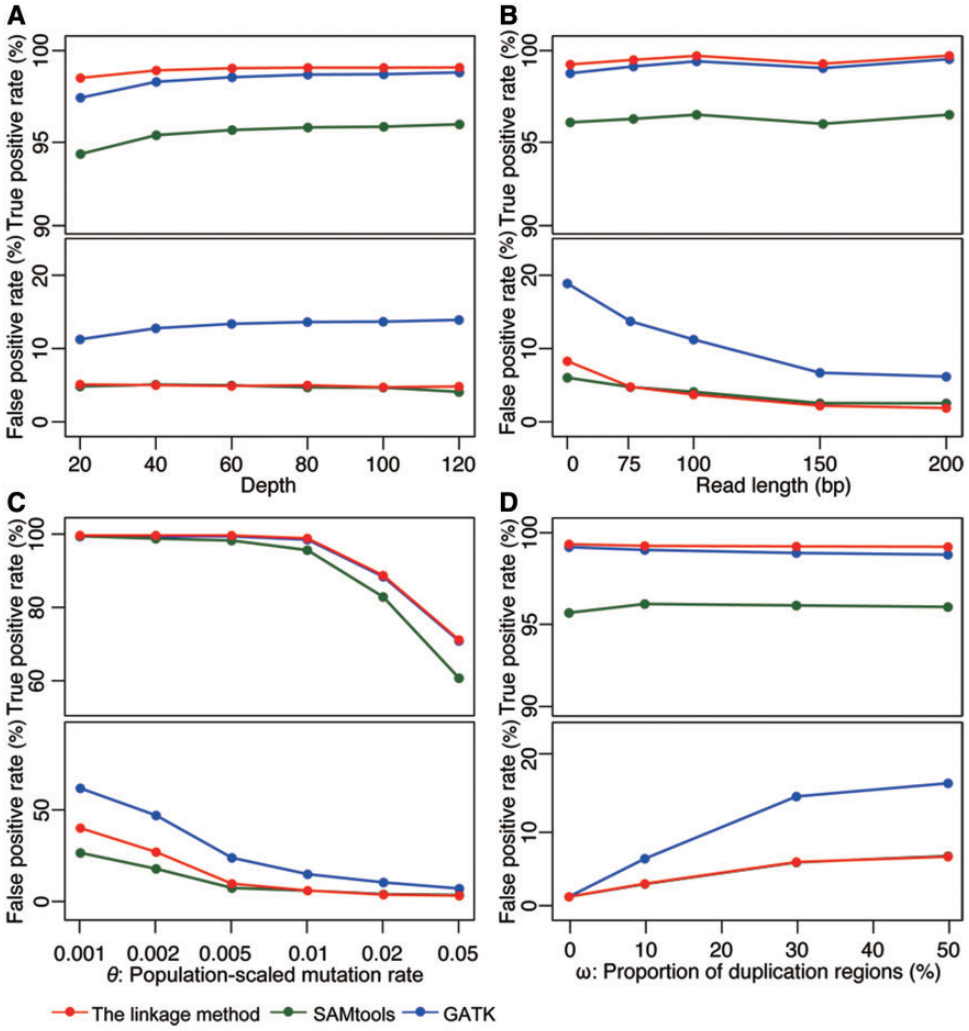

FIG. 4. SNP calling performance using the linkage method, SAMtools, and GATK. The present study investigated the effects of depth (*A*), read length (*B*), population-scaled mutation rate (*C*), and the proportion of duplicated regions (*D*).

because it had a higher true-positive rate and a lower false-positive rate. It was not easy to compare the three methods for heterozygous sites, although the linkage method and SAMtools delivered similar overall performance. GATK detected far more heterozygous SNPs than the other two but the false-positive rate was also very high (data not shown). Thus, there appeared to be a trade-off between the true- and false-positive rates so it was not easy to determine the method that performed the best.

It is known that the performance of UnifiedGenotyper needs to be improved (which will also improve the performance of GATK) when short read data are realigned (Nielsen et al. 2011) using RealignerTargetCreator and Indel-Realigner in the GATK package. To ensure a fair comparison, we repeated all of the above analyses using realignment according to the GATK guidelines (McKenna et al. 2010) and obtained almost identical results (data not shown).

We compared the run time of the linkage method with SAMtools and GATK for all data set. The relative run time of the linkage method to SAMtools and GATK was roughly 20–30 times for all conditions. Considering that the run time of SNP calling tools which do not call haplotypes is much shorter than phasing software (Nielsen et al. 2011), the run time of the linkage method should be reasonable.

Application to *Drosophila*

The algorithm was applied to *Drosophila melanogaster* data. We selected two random African strains (GU6 and RG3) from the Drosophila Population Genomics Project (www.dpgp.org/dpgp2/DPGP2.html, last accessed November 1, 2012). Illumina sequencing short reads (75 bp paired-end) from their haploid embryos were downloaded from www.ncbi.nlm.nih.gov/sra (last accessed November 1, 2012; SRR189120 and SRR189387, respectively). We pooled these data in equal amounts so the coverage was approximately $\times 50$. This artificially synthesized heterozygote dataset was mapped to a reference genome (release 5) using BWA (Li and Durbin 2009) and the linkage method was applied to the left arm of chromosome 2 (23,011,544 bp). We found that the haplotype phasing coverage was approximately 46% of this region, with an average length of haplotype blocks = 403.3 bp. The coverage was comparable with that expected from the simulated data with a depth of 40–60 in table 2. The numbers of homozygous and heterozygous SNP calls were 153,330 and 121,969, respectively. The expected number of homozygous SNPs would have been approximately 100,000–200,000 if we assume $\theta = 0.01$–0.02 for African populations (e.g., Andolfatto and Przeworski 2001; Glinka et al. 2003), which indicated that approximately three-fourths of the homozygous SNPs were detected and this was consistent with the simulation results in table 2.

Discussion

When we sequence a diploid individual, the output is essentially two genomes: one from the paternal parent and the other from the maternal parent. In this study, we developed a unique algorithm for distinguishing SNPs from two parents and phasing them into haplotypes. In our algorithm, the linkage method, we use linkage information from nearby SNPs, which facilitates the efficient removal of haplotypes originating from incorrectly mapped short reads, probably from paralogous regions. This process also helps to improve the quality of SNP calls, as demonstrated in figure 3. Haplotype phasing relies largely on multiple SNPs located in the same-paired reads. Therefore, the success greatly depends on the density of SNPs. Using current NGS technology with a relatively short read length, reasonable performance is expected if this method is applied to a species with a nucleotide diversity (θ) > 0.005 (i.e., an average of five heterozygous sites per 1 kb). When θ was high, the performance was slightly reduced because too many reads were lost during mapping, although future improvements of the mapping algorithm will improve this situation. For species with a reasonably high level of polymorphism, our algorithm provides useful information on haplotypes, which helps to identify the origin of each SNP in the genealogy. Another interesting application would be to allotetraploid species where it may be possible to reconstruct the genomes of the two parental species.

The algorithm was implemented as the program "linkSNPs," which was written in a shell script and Perl for Mac OSX and Linux platforms. The memory requirement is at least 2 GB. The program inputs are mapped short read data in the SAMformat (Li et al. 2008), which is the major output format used by read mapping software such as BWA (Li and Durbin 2009). Therefore, our algorithm can be integrated with the standard procedures used from mapping to SNP calling. The program is available from www.sendou.soken.ac.jp/esb/innan/InnanLab/ (last accessed June 1, 2013).

References

Andolfatto P, Przeworski M. 2001. Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. *Genetics* 158:657–665.

Browning SR, Browning BL. 2011. Haplotype phasing: existing methods and new developments. *Nat Rev Genet.* 12:703–714.

Fujimoto A, Nakagawa H, Hosono N, et al. (13 co-authors). 2010. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nat Genet.* 42:931–936.

Geraci F. 2010. A comparison of several algorithms for the single individual SNP haplotyping reconstruction problem. *Bioinformatics* 26: 2217–2225.

Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics* 165: 1269–1278.

Huang X, Wei X, Sang T, et al. (29 co-authors). 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet.* 42:961–967.

Hudson R. 1983. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol.* 23:183–201.

Hudson RR. 2002. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

Lam HM, Xu X, Liu X, et al. (16 co-authors). 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet.* 42:1053–1059.

Li H. 2011. Improving snp discovery by base alignment quality. *Bioinformatics* 27:1157–1158.

Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009a. The sequence alignment/map format and samtools. *Bioinformatics* 25:2078–2079.

Li H, Homer N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform.* 11:473–483.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18: 1851–1858.

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K. 2009b. SNP detection for massively parallel whole-genome resequencing. *Genome Res.* 19: 1124–1132.

Long Q, MacArthur D, Ning Z, Tyler-Smith C. 2009. HI: haplotype improver using paired-end short reads. *Bioinformatics* 25:2436–2437.

McKenna A, Hanna M, Banks E, et al. (11 co-authors). 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297–1303.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.* 12: 443–451.

van Bers NE, van Oers K, Kerstens HH, Dibbits BW, Crooijmans RP, Visser ME, Groenen MA. 2010. Genome-wide SNP detection in the great tit Parus major using high throughput sequencing. *Mol Ecol.* 19(Suppl 1):89–99.

Wang J, Wang W, Li R, et al. (70 co-authors). 2008. The diploid genome sequence of an Asian individual. *Nature* 456:60–65.