Resource

Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads

Rei Kajitani,[1] Kouta Toshimoto,[1,2] Hideki Noguchi,[3] Atsushi Toyoda,[3,4] Yoshitoshi Ogura,[5,6] Miki Okuno,[1] Mitsuru Yabana,[1] Masayuki Harada,[1] Eiji Nagayasu,[7] Haruhiko Maruyama,[7] Yuji Kohara,[8] Asao Fujiyama,[3,4] Tetsuya Hayashi,[5,6] and Takehiko Itoh[1]

[1]Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan; [2]AXIOHELIX Co. Ltd., Chuo-ku, Tokyo 103-0015, Japan; [3]Advanced Genomics Center, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan; [4]Center for Information Biology, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan; [5]Division of Microbial Genomics, Frontier Science Research Center, University of Miyazaki, Miyazaki 889-1692, Japan; [6]Division of Microbiology, Faculty of Medicine, University of Miyazaki, Miyazaki 889-1692, Japan; [7]Division of Parasitology, Faculty of Medicine, University of Miyazaki, Miyazaki 889-1692, Japan; [8]Genetic Strains Research Center, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

Although many de novo genome assembly projects have recently been conducted using high-throughput sequencers, assembling highly heterozygous diploid genomes is a substantial challenge due to the increased complexity of the de Bruijn graph structure predominantly used. To address the increasing demand for sequencing of nonmodel and/or wild-type samples, in most cases inbred lines or fosmid-based hierarchical sequencing methods are used to overcome such problems. However, these methods are costly and time consuming, forfeiting the advantages of massive parallel sequencing. Here, we describe a novel de novo assembler, Platanus, that can effectively manage high-throughput data from heterozygous samples. Platanus assembles DNA fragments (reads) into contigs by constructing de Bruijn graphs with automatically optimized *k*-mer sizes followed by the scaffolding of contigs based on paired-end information. The complicated graph structures that result from the heterozygosity are simplified during not only the contig assembly step but also the scaffolding step. We evaluated the assembly results on eukaryotic samples with various levels of heterozygosity. Compared with other assemblers, Platanus yields assembly results that have a larger scaffold NG50 length without any accompanying loss of accuracy in both simulated and real data. In addition, Platanus recorded the largest scaffold NG50 values for two of the three low-heterozygosity species used in the de novo assembly contest, Assemblathon 2. Platanus therefore provides a novel and efficient approach for the assembly of gigabase-sized highly heterozygous genomes and is an attractive alternative to the existing assemblers designed for genomes of lower heterozygosity.

[Supplemental material is available for this article.]

With the rapid progress in sequencing technologies, the throughput of sequencers has approached hundreds of billions of base pairs per run. Despite the drawbacks of short read lengths, a number of draft genomes have been constructed solely from these short-read data at an increasingly accelerated pace (Li et al. 2009b; Al-Dous et al. 2011; Jex et al. 2011; Kim et al. 2011; The Potato Genome Sequencing Consortium 2011; Murchison et al. 2012). The draft genome assemblies from high-throughput short reads primarily use de Bruijn-graph-based algorithms (Pevzner et al. 2001; Vinson et al. 2005; Zerbino and Birney 2008; Gnerre et al. 2011). During de novo assembly, the nodes of the de Bruijn graphs represent *k*-mers in the reads, and the edges represent (k − 1) overlaps between the *k*-mers. The graph can be simplified in a variety of ways; and as a consequence, assembled contigs or scaffolds are constructed from subgraphs lacking junctions. The most distinctive advantage of this approach is the computational efficiency that results from omitting the costly pairwise alignment steps that are required in traditional overlap-layout-consensus algorithms (Kurtz et al. 2004). The de Bruijn graph is constructed from information derived from precise *k*-mer overlaps; therefore, its calculation cost is relatively low. Although mismatches between *k*-mers caused by sequencing errors may occur, their distributions are expected to be random, such that sufficient sequence coverage would resolve the sequence error by removing the short, thin tips. Therefore, this approach is suitable for the assembly of a huge number of short reads from a massively parallel sequencer.

Despite its strong functionality, several obstacles remain in applying de Bruijn-graph-based assembly to the data from massively parallel sequencers. One of the primary difficulties to overcome is the existence of heterozygosity between diploid chromosomes (Vinson et al. 2005; Velasco et al. 2007; The Potato Genome Sequencing Consortium 2011; Star et al. 2011; Takeuchi et al. 2012; Zhang et al. 2012; Nystedt et al. 2013; You et al. 2013; Zheng et al. 2013). In cases in which a de Bruijn graph is built up from a diploid sample, different *k*-mers derived from the heterozygous

Corresponding author: takehiko@bio.titech.ac.jp

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.170720.113. Freely available online through the *Genome Research* Open Access option.

24:1384–1395 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/14; www.genome.org

regions corresponding to each homologous chromosome are created and used in the graph structures. As a result, junctions are created in the graph, which represent the borders between homozygous and heterozygous regions. This phenomenon leads to bubble structures in the graph, and most of the existing de Bruijn-graph-based assemblers attempt to simplify such structures by cutting the edge surrounding the junctions and splitting them into multiple straight graphs (Pevzner et al. 2001; Zerbino and Birney 2008; Li et al. 2010; Gnerre et al. 2011). To overcome this problem, many assemblers have developed a common solution by removing one of the similar sequences in a bubble structure with a pairwise alignment. This approach is effective for genome sequences with lower rates of nonstructural variations; however, the assembly of highly heterozygous organisms may encounter more serious problems caused by a high density of single nucleotide variants (SNVs) and structural variations (e.g., repeat sequences and coverage gaps). Algorithms to simply remove bubbles, which are used by the existing de Bruijn-graph-based assemblers, may not be sufficient to resolve these problems.

Thus, several advanced techniques have been used to sequence highly heterozygous genomes. The establishment of inbred lines is the most popular method for targeting highly heterozygous genomes, but this method is both time consuming and costly. Inconveniently, in some cases inbreeding methods can fail to eliminate high levels of heterozygosity; thus, these inbred samples can be unsuitable for use with existing whole-genome shotgun assembly methods (Zhang et al. 2012; You et al. 2013). In contrast, in the Potato Genome Project (The Potato Genome Sequencing Consortium 2011) a homozygous doubled-monoploid clone was first generated using classical tissue culture techniques and then sequenced. However, this method can also be fairly costly and is not always technically possible. Consequently, the fosmid-based hierarchical sequencing method has been increasingly used for sequencing highly heterozygous samples, such as oyster (Zhang et al. 2012), diamondback moth (You et al. 2013), and Norway spruce (Nystedt et al. 2013). Although these approaches have been successful in meeting the functional goals of each sequencing project, all are costly compared with a simple whole-genome shotgun sequencing strategy. Model organisms whose lineages have been maintained in laboratories have long been the main targets of genome sequencing. However, various wild-type organisms that may have highly heterozygous genomes are now targets; thus, a more efficient method to assemble such genomes is needed to further accelerate the genome sequencing of a wide range of organisms.

Here we describe a novel de novo sequence assembler, called Platanus, that can reconstruct genomic sequences of highly heterozygous diploids from massively parallel shotgun sequencing data. Similarly to other de Bruijn-graph-based assemblers, Platanus first constructs contigs from a de Bruijn graph and then builds up scaffolds from the contigs using paired-end or mate-pair libraries. However, various improvements (e.g., k-mer auto-extension) have been implemented to allow Platanus to efficiently handle giga-order and relatively repetitive genomes. In addition, Platanus efficiently captures heterozygous regions containing structural variations, repeats, and/or low-coverage sites; it can merge haplotypes during not only the contig assembly step but also the scaffolding step to overcome the challenge of heterozygosity. Key algorithms of Platanus and the results of the intensive evaluation of Platanas using both simulated data and real data, including those from highly heterozygous genomes and those used in the de novo assembly contest Assemblathon 2 (Bradnam et al. 2013), are described here.

Results

Algorithm overview

Platanus is divided into three subprograms—Contig-assembly (Fig. 1A), Scaffolding (Fig. 1B), and Gap-close (Fig. 1C)—similar to existing de Bruijn-graph-based assemblers (e.g., SOAPdenovo [Li et al. 2010] and Velvet [Zerbino and Birney 2008]) (see Supplemental Methods for details).

Contig-assembly

The Contig-assembly subprogram constructs de Bruijn graphs from reads, modifies the graphs, and displays the output sequences of contigs from the graph. Initially, all k_0-mers (default, $k_0 = 32$) in the reads are counted, and the de Bruijn graph is constructed from the k_0-mers. In this case, the k_0-mer and $(k_0 - 1)$ overlaps correspond to the nodes and edges, respectively. Short branches with relatively low coverage are eliminated in the so-called "tip removal" step (Supplemental Fig. 3).

To simplify the graph, Platanus increases the value of k by the step size k_{step} and iteratively reconstructs the graphs. k_{pre} is the previous k of a certain reconstruction step. When a graph of k-mer is constructed based on a graph of k_{pre}-mer ($k_{\text{pre}} < k$), k-mers (nodes) within a distance of k_{step} from the junctions are marked. Next, the k-mers are extracted from both the contigs of the k_{pre}-mer graph and reads containing marked k_{pre}-mers. In this way, repeats shorter than k can typically be resolved, and Platanus effectively excludes junctions caused by heterozygosity, short repeats, and errors. However, if k is too long, it will be difficult to ensure sufficient coverage distinguishing correct k-mers from k-mers derived from sequence errors. Using multiple k-mer sizes, Platanus uses the advantages of each k. Platanus also has unique functions to automatically determine both the maximum k-mer size and the coverage cutoff (Supplemental Figs. 4–7). This function can efficiently omit the need for manual optimization of its parameters.

Bubble removal in Contig-assembly

After the reconstruction and tip removal of the k_{max}-mer graph, the "bubbles" in the graph are removed. Bubble structures are caused by both the heterozygosity of the diploid samples and errors (Supplemental Fig. 8). A bubble is defined as a set of two straight nodes and two junction nodes at which the straight nodes are connected to the same junction in both directions. Platanus requires the following two conditions to split the straight paths surrounding a bubble structure: (1) a high identity between the two straight nodes; and (2) a low coverage depth of k-mers in the two straight nodes. The second condition is helpful to distinguish heterozygous regions from repetitive regions. The removed bubble structures are saved and utilized in the Scaffolding step. Lastly, as a result of Contig-assembly, the junction-free subgraphs constructed by these procedures correspond to the contigs.

Scaffolding

In the Scaffolding step, the orders of the contigs are determined using paired-end (mate-pair) information. Initially, Platanus maps reads to contigs based on a hash table (keys are unique k-mers on contigs; values are positions). Importantly, the bubbles removed in Contig-assembly are also considered in this step, as they are reallocated to the contigs (Supplemental Fig. 11) prior to read mapping, making it possible to detect the heterozygous contigs. The mapping method of Platanus is designed to maximize the number of accurately mapped paired-ends in highly heterozygous genomes. The mapped positions of the reads on bubbles are converted into cor-

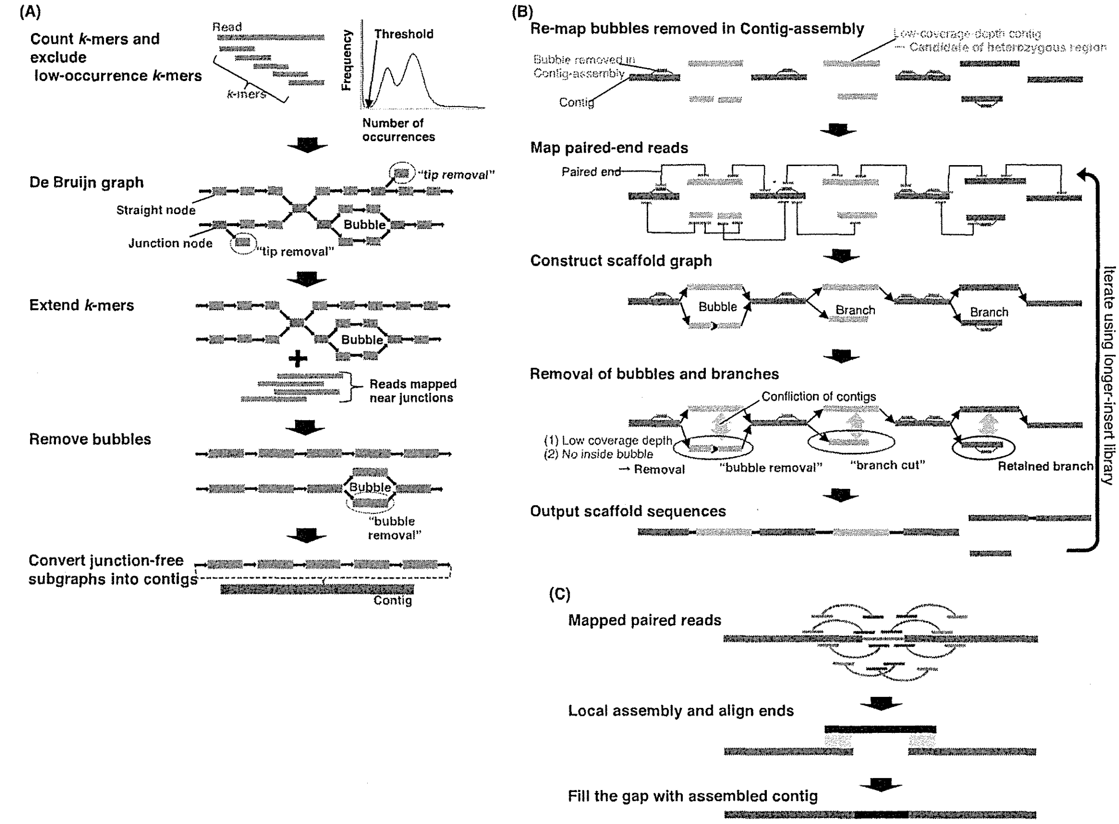


Figure 1. Schematic overview of the Platanus algorithm. (*A*) In Contig-assembly, a de Bruijn graph is constructed from the read set. Short branches caused by errors are removed by "tip removal." Short repeats are resolved by *k*-mer extension, in which previous graphs and reads are mapped to nearby *k*-mers at the junctions. Finally, bubble structures caused by heterozygosity or errors are removed. Subgraphs without any junctions represent contigs. (*B*) In Scaffolding, links between contigs are detected using paired reads. The relationship between contigs is represented by the graph. Bubbles removed in Contig-assembly are remapped on contigs and utilized for mapping of paired-end reads and detection of heterozygous contigs. Heterozygous regions are removed as bubble or branch structures on the graph by the "bubble removal" or "branch cut" step. These simplification steps are characteristic of Platanus and especially effective for assembling complex heterozygous regions. (*C*) In Gap-close, paired reads are mapped on scaffolds, and reads mapped at nearby gaps are collected for each gap. If a contig is expected to cover the gap and is constructed from collected reads, the gap is closed by the contig.

responding contig positions (Supplemental Fig. 11). The insert size of each library is estimated from pairs mapped to the same contig, and links between the contigs are detected using pairs that are situated in different contigs. Links between contigs are represented as a graph in which the contigs and links correspond to the nodes and edges, respectively. In this case, two contigs are considered to be linked if the number of read pairs bridging the contigs exceeds the threshold *n*. The contigs are finally combined into scaffolds to the extent that conflicts occur. Scaffolding then continues using each library, ranging from short- to long-insert libraries.

"Bubble removal" and "branch cut" in scaffold graph

The procedures for the removal of bubbles ("bubble removal") and short branches ("branch cut") are applied in Scaffolding (Supplemental Figs. 15, 16). Compared with other assemblers, these graph simplification steps in Scaffolding are unique to Platanus and are especially effective in assembling complex heterozygous regions. In these steps, bubbles and branches are primarily derived from highly heterozygous regions (i.e., regions with high SNV densities and/or structural variations), and Platanus constructs each haplotype as separate contigs. Platanus recognizes bubbles or branches derived from the heterozygous regions based on the following information: (1) coverage depth; (2) identity with other contigs; and (3) bubble structures constructed in Contig-assembly. The first and second conditions are similar to the conditions of bubble removal in Contig-assembly. The third condition means that Platanus assumes that the target genome is diploid and therefore does not allow for triple or higher-ordered heterozygote alleles. In the following section describing the assembly of the real data from heterozygous samples, we provide an example of a highly heterozygous region assembled by these algorithms.

Gap-close

Finally, in the Gap-close step, reads are mapped on scaffolds to collect those covering each gap. Each set of reads is assembled locally, and the resulting contigs are used to close the gaps (Supplemental Figs. 18, 19). Both the de Bruijn graphs from multiple *k*-mer sizes and the overlap-layout-consensus algorithm are used in the Gap-close step.

Benchmarks overview

A summary of the assemblies of all species targeted in this study is provided in Table 1. In all benchmarks, the contiguity of the as-

Table 1. Summary of the assemblies

Species	Genome size (Mbp)	Insert sizes of the paired end libraries (bp)	Insert sizes of the mate pair libraries (bp)	Sequence depth of paired ends (x)	Heterozygosity (%)	Peak occurrence of Homozygous 17-mer[a]	Hetero-peak-height/Homo-peak-height[a]	Repetitive 17-mer fraction[a]	Scaffold NG50 Platanus (bp)	Largest scaffold NG50 except Platanus (bp); (assembler's name)	
C. elegans (nematode worm)	100.3	230, 420	4.7k	139.6	0.00	113	0.0704	0.236	478,744	**507,513**	(SOAPdenovo2)
					0.10				490,975	**497,363**	(SOAPdenovo2)
					0.20				**535,328**	489,092	(SOAPdenovo2)
					0.30				**545,914**	460,620	(MaSuRCA)
					0.50				**497,387**	475,513	(MaSuRCA)
					1.00				**511,190**	466,806	(MaSuRCA)
					1.50				**516,958**	472,079	(MaSuRCA)
					2.00				**580,832**	351,406	(MaSuRCA)
S. venezuelensis (nematode worm)	57.7	200, 450	3.4k	133.4	0.93	111	0.955	0.289	**274,622**	176,206	(MaSuRCA)
Crassostrea gigas (oyster)	565.7	170–800	2–20k	122.5	0.92	98	1.27	0.471	**381,943**	154,144	(ALLPATHS-LG)
Melopsittacus undulates (bird)	1085.2	220–800	2–40k	107.9	0.46	91	0.424	0.313	**21,684,294**	17,716,398	(ALLPATHS-LG [ALLPATHS])
Boa constrictor constrictor (snake)	1431.5	400	2–10k	92.3	0.17	77	0.108	0.436	**17,165,953**	4,536,273	(SGA [SGA])
Maylandia zebra (fish)	915.0	180	2.5–40k	52.5	0.15	41	0.194	0.441	2,371,946	**4,850,564**	(Newbler, ALLPATHS-LG, Atlas, Phrap [BCM-HGSC])

Sequence depths are calculated for the preprocessed data, which were entered as inputs of assemblers. Heterozygosity (≥0.1%) of *Caenorhabditis elegans* was simulated in silico, whereas the other values of heterozygosity were estimated by paired-end mapping. The preprocess step includes trimming the adaptor sequences and low-quality regions. NG50 is the length for which the collection of all sequences of that length or longer contains 50% of the estimated genome size. Bold numbers indicate the largest scaffold NG50.

[a] Schematic representations of each indicator are shown in Figure 2A. Precisely duplicated repetitive 17-mer occurrences are more than double the occurrence of the homozygous peak. Let n_{all}, n_{error}, and n_{repeat} be the number of all 17-mers, 17-mers whose occurrences are less than c_{bottom}, and 17-mers whose occurrences are greater than 2 × c, respectively. c and c_{bottom} correspond to those in Figure 2A. Estimated genome size equals $(n_{all} - n_{error})/c$. Repetitive 17-mer fraction equals $n_{repeat}/(n_{all} - n_{error})$.

sembly result was measured using the NG50 value, which represents the length at which the collection of all sequences of that length or longer contains 50% of the genome size. NG50 values were calculated for both the scaffolds and contigs. According to the GAGE study (Salzberg et al. 2012), we define a gap as Ns ≥ 3 bp, and contigs are derived from splitting the scaffolds by defined gaps. For species for which reference genomes have not been sequenced, we performed assembly validation using fosmids or BACs. In this validation, we first constructed one-to-one relationships between the fosmids/BACs and the scaffolds and then summed the alignment lengths. The resulting sum is called the "top-hits-length" and is used as the validation score (see Methods for details). In addition, we counted the number of "contained" fosmids/BACs, 90% of the lengths of which were at least covered by one scaffold. The other evaluation criteria are described in each section of Results.

First, we generated simulated heterozygous data from the Illumina HiSeq 2000 sequence reads for the nematode (*Caenorhabditis elegans*) and investigated the effect of heterozygosity on the de novo genome assembly. Second, we applied the assemblers to the real-world data from a heterozygous nematode worm (*Strongyloides venezuelensis*). Third, we performed a test using the data from the oyster (*Crassostrea gigas*) genome (Zhang et al. 2012), which is heterozygous, large, and highly repetitive. Finally, we assembled the data of a bird (*Melopsittacus undulatus*), a snake (*Boa constrictor constrictor*), and a fish (*Maylandia zebra*), which were produced for Assemblathon 2.

To investigate the characteristics of each genome, we performed 17-mer frequency analysis using paired-end reads. In this analysis, the level of heterozygosity is represented by the height difference of two peaks, with left- and right-hand peaks denoting heterozygous and homozygous regions, respectively (Fig. 2A). Essentially, the greater the degree of heterozygosity, the greater the size of the left-hand peak; thus, our data demonstrate that *S. venezuelensis* and the oyster are highly heterozygous species compared with other organisms tested here (Fig. 2B,C; Table 1). In addition, the genome size of each species and proportions of precisely duplicated repetitive regions were estimated (Table 1). In short, we observed that (1) the genome sizes and repeat contents of nematode worms are low; (2) the oyster genome is the most repetitive among those investigated; and (3) the three Assemblathon 2 samples have relatively large genome sizes, ranging from 0.9 to 1.5 Gbp, and low or intermediate levels of heterozygosity.

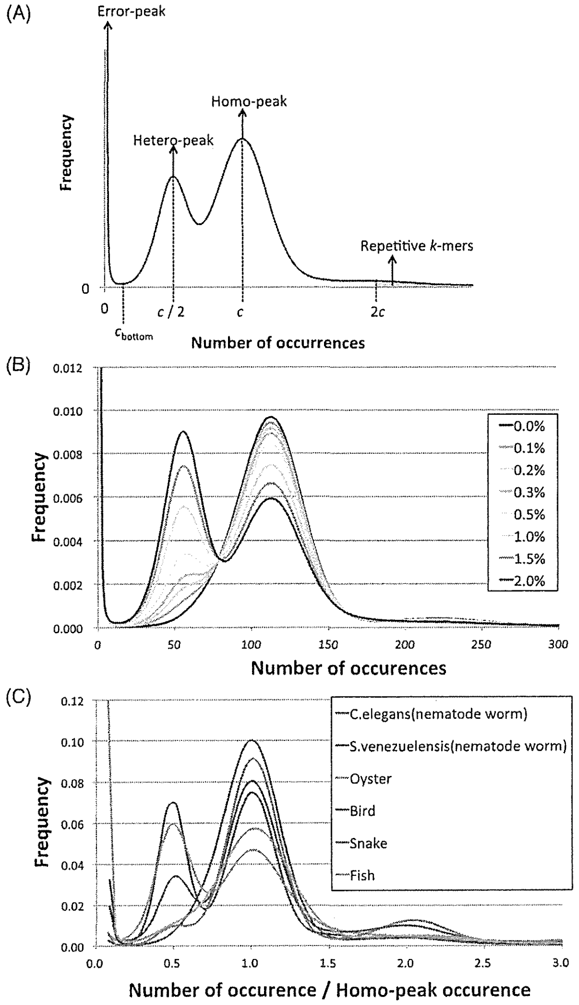


Figure 2. Distribution of the number of 17-mer occurrences. (*A*) Schematic model of the distribution of *k*-mer occurrences. This distribution is related to that shown in Table 1. (*B*) Simulated heterozygous data from *C. elegans*. (*C*) Distributions of normalized 17-mer occurrences for all species.

Assemblers for comparisons

We compared Platanus (version 1.2.1) with other major assemblers, including ALLPATHS-LG (Gnerre et al. 2011) (version 44837), MaSuRCA (Zimin et al. 2013) (version 2.0.4), Velvet (Zerbino and Birney 2008) (version 1.2.07), and SOAPdenovo2 (Luo et al. 2012) (version 2.04). When the assembly test of human chromosome 14 was performed in the GAGE study (Salzberg et al. 2012), these assemblers recorded the largest scaffold NG50 values and were ranked first through fourth, respectively.

ALLPATHS-LG, Velvet, and SOAPdenovo2 all use de Bruijn-graph-based algorithms. Velvet was first developed for the assembly of small genomes, whereas ALLPATHS-LG and SOAPdenovo2 were customized for large eukaryotic genomes. In the benchmarks, we optimized SOAPdenovo2 and Velvet for *k*-mer length, the most important parameter. ALLPATHS-LG was implemented with a default *k*-mer length of 96 in accordance with the manual instructions. We also optimized other options of these assemblers relating to the resolution of the heterozygous regions. SOAPdenovo2 possesses a parameter termed "mergeLevel" (-M) that was tested in two ways: the "-M 1" (default) and "-M 3" modes. ALLPATHS-LG was run in the diploid mode (see Supplemental Methods for details).

MaSuRCA was developed based on the Celera assembler (Myers et al. 2000) and uses an overlap-layout-consensus approach. Although this approach is time consuming, it can overcome the repeat sequences, errors, low-coverage regions, and small structural differences caused by heterozygosity. Certain improvements in MaSuRCA have been implemented to handle high-throughput data from such platforms as Illumina. MaSuRCA was run with the default settings except that the option related to memory usage was changed.

Simulations of heterozygosity using *C. elegans* data

We performed the assembly benchmark against the simulated heterozygous data. We resequenced the genomic DNA of the nematode

C. elegans (with a genome size of 100 Mbp), using Illumina HiSeq 2000. Next, the data were processed in silico, and simulated Illumina read sets were generated with various levels of heterozygosity (0.1%–2.0%) (see Methods). By mapping original paired-end reads onto the reference genome (The *C. elegans* Sequencing Consortium 1998), the raw heterozygosity of *C. elegans* was estimated to be 1.85×10^{-3}% (see Methods). Therefore, the effects of the intrinsic heterozygosity were expected to be low enough to use these simulated data sets to investigate how different levels of heterozygosity affect the assembly.

In Figure 3, Supplemental Figure 22, and Supplemental Table 2, the corrected scaffold NG50, the corrected contig NG50, the numbers of errors, and other statistical information of scaffolds (≥500 bp) obtained by each assembler tested are shown. The corrected scaffold NG50 was computed after breaking assembled sequences at each misassembled (structural difference) point detected by the GAGE benchmark program by comparison with the reference genome. According to these benchmarks, heterozygosity has a strong impact on both the corrected scaffold and the contig NG50 of the existing de Bruijn-graph-based assemblers (SOAPdenovo2, ALLPATHS-LG, and Velvet) (Fig. 3A,B). These values sharply decreased in the interval of 0.0%–0.5% compared to the decrease in the interval 0.5%–2.0%. We therefore hypothesize that 0.5% marks the critical point of heterozygosity that determines the seriousness of the effects on these three de Bruijn-graph-based assemblers. For SOAPdenovo2 and Velvet, the numbers of identified errors also increased relative to the level of heterozygosity (Fig. 3C; Supplemental Table 2F). In contrast, only a slight reduction in the corrected scaffold NG50 values from Platanus was observed. No significant reduction was observed in the corrected scaffold NG50 values from MaSuRCA, but the number of errors was approximately twofold greater in MaSuRCA than in Platanus for all heterozygosity levels.

When the heterozygosity values were 0.0% and 2.0%, the scaffold NG50 values of the initial Platanus contigs (the outputs of Contig-assembly step) were 12,345 bp and 3840 bp, respectively, illustrating that the Contig-assembly step of Platanus was strongly influenced by the heterozygosity. Indeed, the bubble-removal algorithms in the de Bruijn graphs have been implemented in other assemblers; thus, it would appear that Platanus does not possess an advantage in this step. However, the NG50 values of the final scaffolds of Platanus were significantly greater than those from the other assemblers (478,744 bp [heterozygosity: 0.0%] and 580,832 bp [heterozygosity: 2.0%]), indicating that Platanus was able to effectively overcome the high heterozygosity in the scaffolding step.

Next, we investigated the per-base accuracy of the scaffolds according to the numbers of mismatches (SNPs) and indels (<5 bp) reported in the GAGE evaluations of the *C. elegans* data in the absence of simulated heterozygosity (Table 2). The raw heterozygosity of the *C. elegans* genome was estimated to be 1.85×10^{-3}%, and the expected number of variants was estimated to be less than 1850. The higher-than-predicted numbers obtained are likely due to errors in the assemblies. For both the numbers of mismatches and indels, the number generated by Platanus displayed the lowest value (thereby indicating the fewest errors), from which we infer that the scaffolds had the best per-base accuracy. There may be a tradeoff between the per-base accuracy and the 'N' rate because the number of mismatches and indels is reduced when an assembler has the tendency to report less confidential regions as 'N's. The 'N' rate of Platanus was the middle value (third) among the five assemblers assessed, and Platanus did not

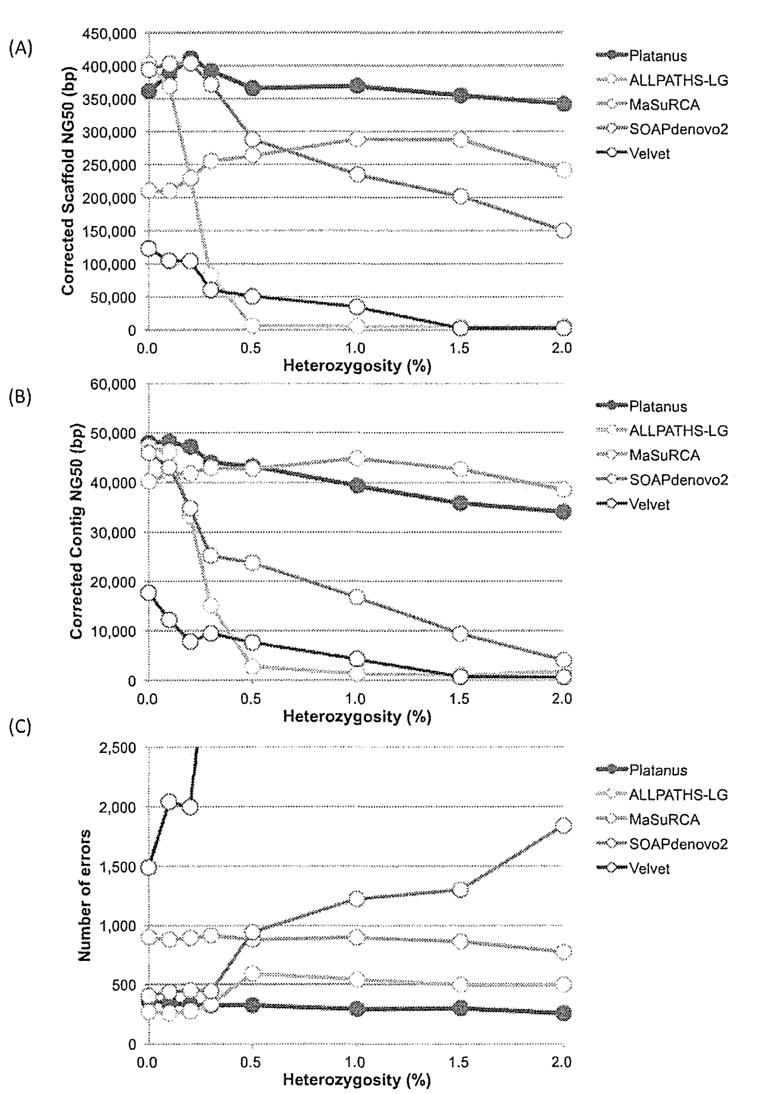


Figure 3. Results of the benchmarks of heterozygosity simulations (*C. elegans*). (*A*) Corrected scaffold-NG50 calculated by GAGE. (*B*) Corrected contig-NG50. (*C*) Number of errors reported by GAGE. Errors are defined as inversion, relocation, or translocation.

Table 2. Mismatches, small indels, and the 'N' rate in *C. elegans* (heterozygosity 0.0%) assembly

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Number of mismatches	4534	5762	15,521	16,650	16,941
Number of indels (<5 bp)	3352	5125	9142	5236	5102
Rate of 'N' (%)	1.40	2.63	0.77	0.43	3.33

Mismatches and indels correspond to SNPs and indels (<5 bp) reported by GAGE, respectively. Rates of 'N' (an ambiguous base) are measured for all scaffolds.

decrease the number of mismatches and indels at the cost of its 'N' rate. In contrast, MaSuRCA and SOAPdenovo2 recorded lower 'N' rates but considerably higher numbers of mismatches (more than three times the number reported by Platanus). In contrast to the scaffold NG50 values, the contig NG50 values of Platanus were not much greater than those of the other assemblers. However, Platanus produced the fewest mismatches and small indels, implying that it constructs highly accurate contigs using a relatively conservative approach in contig assembly.

Assembly of real data from the highly heterozygous nematode *S. venezuelensis*

The heterozygosity of *S. venezuelensis* was estimated to be 0.927% by mapping paired-end reads on fosmid sequences. According to 17-mer frequency analysis (Table 1), the number of precisely duplicated repeats in *S. venezuelensis* (0.289) is comparable to that of *C. elegans* (0.236). This similarity indicates that *S. venezuelensis* is useful for investigating the effect of real heterozygosity on de novo assemblies.

We measured scaffold NG50 values using the estimated genome size of 57.7 Mbp derived from the 17-mer analysis (Table 3). Platanus produced the largest scaffold NG50, confirming its effectiveness for real heterozygous data. Compared with the 1.0%-heterozygous *C. elegans* data (Fig. 3; Supplemental Table 2), the obtained scaffold-NG50/Platanus-scaffold-NG50 ratios were smaller for all other assemblers (Supplemental Table 5). This observation implies that true heterozygous data consist of complex variations that were not simulated in the *C. elegans* tests and that Platanus was able to successfully resolve such variants. We provide an example of complex variant resolution in the following paragraph. Next, we performed assembly validation by aligning eight fosmid sequences (a total of 272,981 bp) to the scaffolds (Table 3). Platanus displayed the largest top-hits-lengths, and all fosmids were contained within the relevant scaffolds, confirming that no large misassembly occurred at least in these fosmid regions. If there is an inaccurate sequence or a gap in regions covered by fosmids, a top-hits-lengths value may decrease because an unaligned region appears. Although fosmids covered the genome partially, this result implies that Platanus' scaffolds possess higher accuracy and/or fewer gaps compared with those produced by the other assemblers.

We further performed a fine evaluation of Platanus' scaffolds using two fosmid pairs, each representing the two haplotypes at a single locus. As noted in the section describing the algorithm overview, we anticipated that Platanus predominantly extends the assembly using a characteristic simplification of the scaffold graph. In the Platanus Scaffolding step, the bubble and branch structures from the heterozygous regions were removed by "bubble removal" and "branch cut" functions, respectively. Platanus should also execute these procedures in the two regions covered by the fosmid pairs.

First, we provide an example of "bubble removal" in Scaffolding (Fig. 4A,B) using a dot plot analysis within the nucmer alignment program. For the alignment of two fosmids covering the region where the bubble was removed (Fig. 4B), a 209-bp indel was present with 2.09% heterozygosity level. The scaffold generated by Platanus (ContigP1–ContigA1–ContigN1) was correctly aligned to one of the fosmids, corresponding to the diagonal line shown in Supplemental Figure 24A. We replaced the bubble region contig (ContigA1) in the scaffold with the removed contig sequence (ContigB1), and the resulting scaffold (ContigP1–ContigB1–ContigN1) was aligned to the fosmid of another haplotype with no gap (Supplemental Fig. 24B). These results indicate that Platanus correctly resolved the region containing a relatively large indel, many SNVs, and several small indels that existed simultaneously using the bubble-removal routine. Second, we provide an example of "branch cut" (Fig. 4C,D). As in the "bubble removal" example, we aligned the two fosmids covering the position of the branch cut (Fig. 4D). This algorithm was designed to resolve heterozygous regions in which the bubble structures do not appear in graphs due to complex variants, repeats, or low coverage depth. Three indels were apparent, with sizes of 126 bp, 715 bp, and 1206 bp and with a high heterozygosity (1.93%). The scaffold sequence (ContigP2–ContigA2–ContigN2) could be aligned to one fosmid of the pair (Supplemental Fig. 25), and the removed branch (size: 1217 bp; ContigB2) matched the other fosmid, confirming the correctness of Platanus' resolution. Platanus may derive its advantage by using

Table 3. Statistics and validations of *S. venezuelensis* assemblies

		Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Assembly statistics	Total (≥500 bp)	58,503,663	61,205,926	66,053,722	52,677,856	63,982,183
	Number of scaffolds (≥500 bp)	2560	9608	4876	3383	11,696
	Scaffold NG50 (bp)	274,622	16,765	176,206	87,219	17,006
	Contig NG50 (bp)	71,357	2008	84,739	48,010	1946
Fosmid validation	Top-hits-lengths (bp)	272,164	69,792	256,848	270,392	78,159
	Average identity (%)	99.42	99.31	99.39	98.72	99.31
	Number of contained fosmids	8	0	7	8	0

For the fosmid validation, eight fosmids (total: 272,981 bp) were aligned to the scaffolds using nucmer and delta-filter (programs in MUMmer package). One-to-one relationships between fosmids and scaffolds were constructed according to the longest alignment for each fosmid, and the sum of these alignment lengths (top-hits-length) was calculated. "Contained fosmid" refers to a fosmid that is 90% covered by a single scaffold.

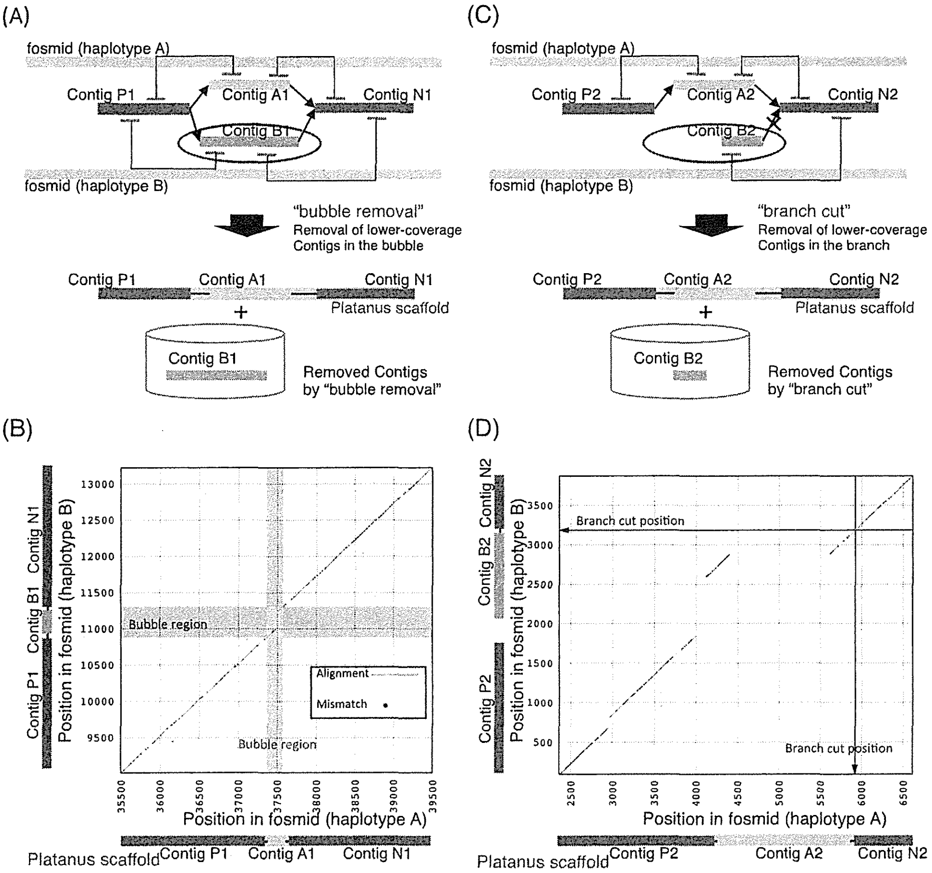

(A)
fosmid (haplotype A)
Contig P1
Contig A1
Contig N1
Contig B1
fosmid (haplotype B)
"bubble removal"
Removal of lower-coverage Contigs in the bubble
Contig P1 Contig A1 Contig N1
+
Platanus scaffold
Contig B1
Removed Contigs by "bubble removal"
(C)
fosmid (haplotype A)
Contig P2
Contig A2
Contig N2
Contig B2
fosmid (haplotype B)
"branch cut"
Removal of lower-coverage Contigs in the branch
Contig P2 Contig A2 Contig N2
+
Platanus scaffold
Contig B2
Removed Contigs by "branch cut"
(B)
Contig P1 Contig B1 Contig N1
Position in fosmid (haplotype B)
Bubble region
Alignment
Mismatch
Bubble region
Position in fosmid (haplotype A)
Platanus scaffold Contig P1 Contig A1 Contig N1
(D)
Contig P2 Contig B2 Contig N2
Position in fosmid (haplotype B)
Branch cut position
Branch cut position
Position in fosmid (haplotype A)
Platanus scaffold Contig P2 Contig A2 Contig N2

Figure 4. Example of a heterozygous region resolved by "bubble removal" and "branch cut." (*A*) Schematic model of "bubble removal" in Platanus scaffolding. (*B*) Alignment dot plot between two fosmids. Green lines and red dots indicate alignments and mismatches, respectively. Red and blue boxes indicate the regions corresponding to the bubbles. (*C*) Schematic model of "branch cut" in Platanus scaffolding. (*D*) Alignment dot plot between two fosmids. Green lines and red dots indicate alignments and mismatches, respectively. The blue arrow indicates the position corresponding to the root of the branch.

the improved scaffolding algorithms typified by the preceding examples to assemble such complex regions, resulting in higher scaffold NG50 numbers in the simulated data from *C. elegans*, in which only SNVs and small indels were simulated.

Because the fosmids only partially covered the genome, we also investigated the distribution of heterozygosity across the entire genome. As a complete or draft genome of *S. venezuelensis* has not yet been published, we used the Platanus' assembly, which demonstrated the largest scaffold NG50 in the reference sequences. SNVs and small indels on the scaffolds were detected by mapping paired-end reads (see Methods), and heterozygosity was calculated for every 1-kbp nonoverlapping window. The average heterozygosity was 0.950%, and the resulting distribution of heterozygosity is shown in Supplemental Figure 26. Compared with the 1.0%-heterozygous *C. elegans* data, the *S. venezuelensis* data had an uneven distribution of heterozygosity. This uneven distribution may be another cause of the observed different statistics between the real data and the simulated data. The fact that the proportion of low heterozygosity regions is greater in *S. venezuelensis* than in the 1.0%-heterozygous *C. elegans* might make assemblies easier, but small scaffold NG50 rates were actually produced by other assemblers. To investigate the reason for this observation, we measured the intervals of 1-kbp windows with high levels of heterozygosity (≥1.0%), and our results suggest that the average length of these intervals was not very long (1930 bp). Consequently, regions of low heterozygosity were bordered by highly heterozygous regions, creating a mosaic structure of both high and low heterozygosity. This mosaic structure may have contributed to the small scaffold NG50 produced by the other assemblers.

Real data from the highly heterozygous and repetitive oyster genome

We input whole-genome shotgun data sequenced in the Oyster Genome Project into the assemblers. The heterozygosity of the oyster genome was estimated to be 0.923% by mapping paired-ends to eight BACs a total of 1,081,613 bp in length. The 17-mer frequency analysis (Table 1) indicated that both the genome size and repeat content of the oyster genome are larger than those of the nematodes. In addition to being highly heterozygous, the oyster is also a suitable model organism for testing the scalabilities and performances of the repetitive sequences. Similar to the process for *S. venezuelensis*, the scaffold NG50 values for the oyster were measured based on the estimated genome size, and validations were performed using the eight BAC sequences (Table 4). For the scaffold NG50 and BAC validation, Platanus' scaffold NG50 and top-hits-length exceeded those of the other assemblers. Velvet and MaSuRCA crashed during the execution of the runs (RAM: 512 GB; CPU: 32). Velvet is not scalable for use with large eukaryotic genomes in the GAGE benchmark (*Bombus impatiens*). MaSuRCA ran for more than 1 mo in real time (using 32 threads) but stopped as a result of an error. Although this assembler is customized for Illumina data, this result is indicative of the time-consuming nature of the overlap-layout-consensus algorithm, which is unsuitable for organisms with a large-sized genome such as the oyster. We also compared the assembling result in this study with sequences assembled by the fosmid-based hierarchal methods produced in the Oyster Genome Project. Remarkably, the values from Platanus were comparable to these fosmid-based reference sequences.

We also investigated whether Platanus' scaffolds could substitute reference sequences during post-assembly analysis. Thus, we investigated the coverage of the transcript sequences. Reads from all the RNA-seq data in the Oyster Genome Project were assembled into contigs (RNA-contigs) using Trinity (Supplemental Methods; Grabherr et al. 2011). We then mapped RNA-contigs whose lengths exceeded 500 bp. Using BLAT (Kent 2002), "top-hits-lengths" were calculated in the same manner as in the BAC validation, and the number of mapped RNA contigs with alignments of the top hit showed ≥90% coverage and ≥90% identity (Table 4). The average identities of top-hit alignments were also calculated. The top-hits-length, mapped RNA-contig numbers, and average identities produced by Platanus were the best of the three whole-genome-based assembly results and were comparable to the results from the fosmid-based reference sequence. These findings demonstrate that Platanus' assembly results are sufficient for practical usage in gene annotation for highly heterozygous genomes. In addition, we counted the number of mapped RNA-seq contigs without any 'N'-bases in the alignment between the RNA-contigs and assembled genome sequences. The result is shown in Table 4 as the "Number of mapped RNA-contigs ('N' free alignment)." Even in this benchmark, Platanus showed results that were nearly identical to the fosmid-based results, although its contig NG50 was the smallest. This result suggests that Platanus' contigs are sufficient for gene annotation.

Assembly of the Assemblathon 2 data

Finally, we applied Platanus to larger genomes and compared its assembly with additional methods to confirm its versatility. We demonstrated the assemblies of three species (bird, snake, and fish) during Assemblathon 2. In this contest, sequence reads were opened and each team freely chose their methods, including the preprocess steps, assemblers, and machines. By mapping the reads to genomic sequences (bird and snake: fosmids; fish: Platanus' scaffolds), we estimated the heterozygosity of the bird, snake, and fish genomes to be 0.463%, 0.165%, and 0.147%, respectively. Consequently, these species are not suitable for testing the assembly of highly heterozygous (>0.5%) samples. Nevertheless, the Assemblathon 2 benchmark has several benefits. First, the assembly protocols of other teams were assumed to be highly optimized. For many teams, the participants were themselves the authors of the assembly tools, decreasing the likelihood that their optimization methods would be insufficient. Second, these three species all have relatively large genome sizes (0.9–1.4 Gbp in length), making it possible to test Platanus' capacity to assemble giga-order-size genomes.

A summary of the results for this section is provided in Table 1, and detailed results are provided in Supplemental Table 7. For the bird and snake, fosmid data (a total of 1,035,129 bp and 378,186 bp, respectively) are available, and we validated the resulting assemblies in the same manner as for the *S. venezuelensis* and oyster assemblies. Platanus recorded the highest values for both the scaffold NG50 (bird: 21,684,294 bp; snake: 17,165,953 bp) and "top-hits-length" of fosmid validation. For the snake assembly in particular, the scaffold NG50 of Platanus was unexpectedly large, more than three times the size of the second largest value. According to the 17-mer frequency analysis (Fig. 2; Table 1), the snake genome is rich in repetitive 17-mers and has sufficient coverage depth compared to that of the fish genome. In the fish assemblies, the scaffold NG50 of Platanus (2,371,946 bp) was the fifth largest of 17 entries. When limited to a single program's results, the scaffold NG50 of Platanus was second, behind that of ALLPATHS-LG. One important feature of the fish data is the low coverage depth (52.5×) of their paired-end reads, which most likely reduced Platanus' scaffold NG50 value.

Table 4. Statistics and validations of the oyster assemblies using BAC and RNA-contigs

		Platanus	ALLPATHS-LG	SOAPdenovo2	Fosmid-based reference
Assembly statistics	Total (≥500 bp)	684,614,954	655,152,639	859,413,081	557,340,816
	Number of scaffolds (≥500 bp)	36,091	18,238	67,846	6432
	Scaffold NG50 (bp)	381,943	154,144	116,321	392,835
	Contig NG50 (bp)	9011	12,025	11,719	26,430
BAC validation	Top-hits-length (bp)	864,992	752,977	851,083	750,984
	Average identity (%)	96.48	96.41	96.28	96.92
	Number of contained BACs	3	2	2	1
RNA-seq validation	Top-hits-length (bp)	42,801,107	38,060,320	40,846,500	42,241,208
	Average identity (%)	98.48	98.34	98.47	98.52
	Number of mapped RNA-contigs	30,700	28,152	30,230	30,150
	Number of mapped RNA-contigs ('N' free alignment)	28,452	25,914	27,092	28,520

For the BAC validation, eight BACs (total: 1,081,613 bp) were aligned to the scaffolds using nucmer and delta-filter (programs in MUMmer package). One-to-one relations between BACs and scaffolds were constructed according to the longest alignment for each BAC, and the sum of these alignment lengths (top-hits-length) was calculated. "Contained BAC" refers to a BAC that is 90% covered by a single scaffold. RNA-contigs (number: 40,503; total: 56,540,774 bp) were aligned to the scaffolds using BLAT. One-to-one relations between RNA-contigs and scaffolds were constructed according to the longest alignment for each RNA-contig, and the total of those alignment lengths (top-hits-length) was calculated. "Mapped RNA-contig" refers to a RNA-contig that is 90% covered by a single scaffold.

Time and peak memory usage

The execution times (real and CPU) and peak memory usages are shown in Table 5. The execution environment is conducted with 32 threads of an Intel Xeon 2.27 GHz CPU with 512 GB RAM. SOAPdenovo2 exhibited the fastest performance in real time for nematodes, whereas Platanus exhibited the fastest performance for the oyster, which has a larger genome size and a greater number of repeats. Notably, MaSuRCA, which is based on the overlap-layout-consensus algorithm, had a considerably longer run time than the de Bruijn-graph-based assemblers. Although SOAPdenovo2 and Velvet were optimized for certain parameters, their execution times did not include the iteration for optimizations and therefore consumed more time for the benchmarks.

Discussion

Although heterozygosity poses a challenge to genome assembly, its effects on genome assembly have never been systematically evaluated. To our knowledge, our simulation of heterozygosity (0.0%–2.0%) using *C. elegans* data is the first attempt to address this issue. All of the de Bruijn-graph-based assemblers tested, except for Platanus, showed dramatically reduced scaffold NG50 values when the heterozygosity was >0.5%. MaSuRCA, the overlap-layout-consensus–based assembler, did not undergo a sharp decrease in its scaffold NG50 in our simulation. However, in assembling real data from various organisms, Platanus was superior, as shown by its scaffold NG50 values that were much larger than those from MaSuRCA, possibly due to the presence of more complex variants in the actual data set. Furthermore, MaSuRCA required excessive execution time for assembly; for example, more than 1 mo in real time (using 32 threads) was required to assemble the oyster data. The oyster genome is ~0.5 Gbp, and de Bruijn-graph-based methods, such as Platanus, can efficiently handle the data from much larger genomes. ALLPATHS-LG exhibited the best performance with overlapping paired-ends (insert size: 180 bp) and a long-jump library (insert size: ~10 kbp), which is consistent with the results of the present study. ALLPATHS-LG's scaffold NG50 was relatively large in the oyster test, for which library insert sizes ranged from 180 to 20 kbp; however, its scaffold NG50 was inferior to that of Platanus. An additional advantage of Platanus is that it does not require the manual optimization of any parameters. In fact, Platanus was executed using the default parameters in all tests performed in this study. In contrast, we needed to iteratively execute SOAPdenovo2 and Velvet with various *k*-mer sizes (21–91), as both substantially depend on this parameter. For example, dependent on the *k*-mer sizes used, SOAPdenovo2's scaffold NG50 for *S. venezuelensis* varied from 4479 to 87,219 bp.

Platanus merges haplotype sequences into a single contig/scaffold, resulting in mosaic sequences of both haplotypes. By adopting this approach, Platanus can achieve remarkably longer scaffolds. An alternative strategy for addressing highly heterozygous data involves the separate construction of each haplotype (haplotype assembly method), which has been applied to *Ciona intestinalis* (Kim et al. 2007) (heterozygosity: 1.2%; scaffold N50: 37.9 kbp) and *Ciona savignyi* (Vinson et al. 2005; Small et al. 2007) (heterozygosity: 4.6%; scaffold N50: 496 kbp) (note that both projects used the Sanger sequencing method). These results suggest that longer haplotype sequences are constructed for higher variant densities. Why should heterozygosity be high for the construction of longer haplotype assemblies? The explanation is simple: To construct a haplotype assembly, the linkage information between neighboring SNVs or indels should be resolved. This linkage information requires neighboring SNVs or indels to be almost covered with one read or pair of reads by one DNA fragment. If the linkage information is broken by a long nonheterozygous region, the haplotype assembly will be disrupted at that point. As described for the assembly of the *S. venezuelensis* genome, the regions in which no sequence variation was observed within a 1-kbp window encompass 11.8% of the entire genome. This observation suggests that if haplotype assembly is adapted to *S. venezuelensis,* the results will be very poor. The *C. savignyi* haplotype assembly may represent a rather exceptional case of a successful run of a genome with extremely high heterozygosity and the use of long Sanger reads. We thus propose that the merging method is suitable for the assembly of most heterozygous samples.

Although Illumina reads are often described as "short reads," they have advantages regarding their throughput and accuracy. In the bird assembly for Assemblathon 2, Platanus' scaffold NG50 was the highest, exceeding those of other strategies that utilize other types of sequence data (Roche 454 and/or PacBio). It should be noted that the conditions are not equivalent regarding the cost and coverage depths for each data type, and thus, it cannot be conclusively stated that Illumina data are the most suitable for

Table 5. Run time and peak memory usage

	C. elegans			*S. venezuelensis*			Oyster		
	CPU time	Real time	Peak memory (GB)	CPU time	Real time	Peak memory (GB)	CPU time	Real time	Peak memory (GB)
Platanus	588,408 sec (163 h)	23,966 sec (7 h)	20.0	238,767 sec (66 h)	10,431 sec (3 h)	19.8	2,485,919 sec (691 h)	114,107 sec (32 h)	98.2
ALLPATHS-LG	648,721 sec (180 h)	62,844 sec (17 h)	129.6	424,661 sec (118 h)	26,515 sec (7 h)	73.1	3,860,440 sec (1072 h)	306,899 sec (85 h)	322.7
MaSuRCA	802,214 sec (223 h)	64,055 sec (18 h)	72.9	748,571 sec (208 h)	118,230 sec (33 h)	70.1		Crashed	
SOAPdenovo2	86,605 sec (24 h)	6873 sec (2 h)	36.1	53,453 sec (15 h)	5449 sec (2 h)	16.6	2,254,545 sec (626 h)	248,160 sec (69 h)	148.4
Velvet	23,191 sec (6 h)	4727 sec (1 h)	35.0	19,442 sec (5 h)	3639 sec (1 h)	38.2		Crashed	

Environment: Processor: Intel(R) Xeon(R) CPU X7560 2.27 GHz. Number of processors: 32. RAM: 512 GB. All programs were executed in the multithread mode using 32 threads. The run times were measured by the GNU time, and the peak memory usages were recorded every 0.1 sec using the "ps" command. SOAPdenovo2 was run with GapCloser.

de novo assembly. Therefore, whole-genome shotgun short-read (Illumina) data remain a strong candidate for the strategy of de novo assembly, particularly for the assembly of large and highly heterozygous genomes. In this study, all data except the fish have >90× sequence coverage depths of paired-ends reads (Table 1). There is the possibility that each assembler has optimal coverage depth, and we performed the benchmark test using reduced amount of sequence data for *C. elegans* (heterozygosity: 0%, 1%, 2%) (Supplemental Fig. 28; Supplemental Table 14). In summary, Platanus indicated the largest corrected scaffold NG50 for heterozygous data whose coverage depth >100× but was sensitive to the downsampling effect. This result corresponds to the small scaffold NG50 of Platanus in the test of the fish. Consequently, the optimal coverage depth for Platanus is probably >100×, which may be suitable for the increasing throughput of sequencers.

Fosmid-based assembly has recently been introduced as an effective and economic method for highly heterozygous genomes (Zhang et al. 2012). However, this method may require many more sequence reads compared to the whole-genome shotgun strategy. For instance, if the fosmid library is constructed to have a depth of 10× against the genome size and each fosmid is sequenced to a depth of 100×, the total required reads may be as much as 1000× the genome size. In the Diamondback Moth Genome Project (You et al. 2013) and Oyster Genome Project (Zhang et al. 2012), paired-end reads with a coverage depth of 2170× (total reads: 855 Gbp) and 690× (total reads: 390 Gbp) against the genome size were produced to assemble the fosmids, respectively. In addition, whole-genome shotgun reads were separately produced, and these data were also used in those projects. Therefore, if highly heterozygous genomes could be assembled from whole-genome shotgun data alone, the cost would be expected to decrease significantly. When a project targets many genomes of nonmodel and/or wild-type samples, such as the Genome 10K Project (Genome 10K Community of Scientists 2009), Platanus is especially helpful because it does not require inbreeding, which is often the bottleneck of the project.

Finally, it should be noted that even in samples with a heterozygosity of <0.5%, such as the *C. elegans* data (0.0%–0.3% heterozygosity) and the Assemblathon 2 data, Platanus produced the largest scaffold NG50 and/or the best validation results. This result indicates the great versatility of Platanus; its effectiveness is not restricted to highly heterozygous samples.

Methods

Data for benchmarks

C. elegans reference sequences: NC_001328.1, NC_003279.6, NC_003280.8, NC_003281.8, NC_003282.6, NC_003283.9, and NC_003284.7
Oyster genomic reads: SRA040229
Oyster reference sequences: AFTI01000000
Oyster BACs: GU207451.1, GU207446.1, GU207415.1, GU207462.1, GU207436.1, GU207459.1, GU207449.1, and GU207460.1
Oyster RNA-seq: GSE31012 (Gene Expression Omnibus)
Bird (Assemblathon 2) genomic reads: ERA200248, ERA201590, and ERA250291
Snake (Assemblathon 2) genomic reads: ERA198728, ERA199152, and ERA250292
Fish (Assemblathon 2) genomic reads: SRA026860
Fosmid sequences (VFR) and assembly results related to Assemblathon 2: Downloaded from the website of Assemblathon 2 (http://assemblathon.org/assemblathon2)

Construction of simulated sequencing data sets with various rates of heterozygosity

Simulated heterozygous diploid chromosome sequences were constructed from the reference genome sequences by randomly introducing substitutions and indels (with a substitution:indel ratio of 9:1). The reads from HiSeq 2000 were mapped to the reference genome of *C. elegans* using Bowtie 2 (Langmead and Salzberg 2012), and the positions of the reads were determined. Approximately 50% of the mapped reads were transformed into the sequence of the simulated heterozygous chromosome. For each simulated heterozygous site, the rate of the transformed reads followed a normal distribution. Linkages between variants were simulated because the transformations were performed as a unit of paired reads.

Variant calling and estimations of heterozygosity

We called variants using Bowtie 2 and SAMtools (Li et al. 2009a). Paired ends were mapped on the *C. elegans* reference genome using Bowtie 2. Mapping was initially performed using a single-end mode. A read was excluded if it had multiple best hits or if the edit distance of the best hit was greater than 5. The insert sizes were counted for each of the pairs whose reads were mapped on the same scaffold with a reasonable direction. Pairs whose insert sizes were within the mean (±2 × standard deviation) were used for the analysis, and the remainders were excluded. The mapping results were merged using SAMtools. In this case, PCR-duplicate reads were removed (samtools rmdup).

When the mapping results were merged, base-quality filtering was performed (minimum: 30, set in the –Q option of "samtools mpileup"). For variant calling, the minimum coverage was 20 and the maximum coverage was twice the average. Sites closer than 100 bp to either the gaps ('N') or ends were also excluded. Finally, we searched the remaining regions. The variants were counted if rates of variant reads were in the range of 0.25 to 0.75.

To ensure that this method correctly computes heterozygosity, we applied it to simulated heterozygous data (Supplemental Table 3). Because we filtered out reads with a minimum edit distance of 5, over-filtering occurred in 2% of the heterozygous data and the rates were underestimated, whereas data with heterozygosity rates ≤1.5% were successfully analyzed. Therefore, we assumed that the low heterozygosity calculated for the *C. elegans* genome was reliable. For data on *S. venezuelensis*, oyster, bird, and snake, we applied the same methods to estimate heterozygosity, mapping the reads on fosmids or BACs. For the fish, reads were mapped on the scaffolds of Platanus because neither a fosmid nor a BAC was available.

Validation of assemblies using fosmid or BAC

We used three programs (nucmer, delta-filter, and show-coords) in the MUMmer package (Kurtz et al. 2004). First, each fosmid (BAC) was aligned (queried) to scaffolds using nucmer. Second, the results from nucmer (out.delta) were filtered using delta-filter with the –g switch (one-to-one global alignment, not allowing for rearrangements). Third, the filtered results were entered as input to show-coords, and the coordinates of the resulting alignments were determined. Finally, we picked up alignments that represented the longest length (top-hit) for each fosmid (BAC) and summed those lengths. This sum was referred to as the "top-hits-length." The one-to-one relations can be used to exclude overestimations of the alignment length from the redundant scaffolds. The top-hits-length decreases when the scaffolds contain errors and gaps. Note that 'N' regions were not counted as 'hit.' Thus, this value summarizes the quality of the scaffolds. Fosmids (BACs) with top-hits-lengths of at least 0.9 times their length were defined as "contained."

Data access

The newly sequenced *C. elegans* and *S. venezuelensis* genomic reads for this study were submitted to the DDBJ Sequence Read Archive (DRA; http://trace.ddbj.nig.ac.jp/dra/index_e.html) under accession numbers DRA000967 and DRA000971, respectively. Platanus is freely available at http://platanus.bio.titech.ac.jp/. All of the benchmark data sets are available from http://platanus.bio.titech.ac.jp/platanus_benchmark.

Acknowledgments

We thank all the members of the laboratories involved in this project for their helpful discussions. This work was supported by KAKENHI (Grant-in-Aid for Scientific Research on Innovative Areas, No. 22125008), KAKENHI (Grant-in-Aid for Scientific Research [B], No. 24310142), and KAKENHI for Innovative Areas "Genome Science" (No. 221S0002) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Author contributions: Project design and coordination: T.I., F.A., Y.K., and T.H. Algorithm development: R.K., H.N., and T.I. Program development: R.K. and K.T. Benchmark: R.K., M.O., K.T., and T.I. Genome sequencing: E.N., Y.O., H.M., T.H., A.T., A.F., Y.K., M.Y., M.H., and T.I. Management and dissection: A.F., T.H., Y.K., and T.I. Imaging: R.K. and T.I. Writing: R.K., T.H., and T.I.

References

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J, et al. 2011. *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat Biotechnol* **29:** 521–527.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, et al. 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *Gigascience* **2:** 10.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282:** 2012–2018.

Genome 10K Community of Scientists. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered* **100:** 659–674.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci* **108:** 1513–1518.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29:** 644–652.

Jex AR, Liu S, Li B, Young ND, Hall RS, Li Y, Yang L, Zeng N, Xu X, Xiong Z, et al. 2011. *Ascaris suum* draft genome. *Nature* **479:** 529–533.

Kent WJ. 2002. Blat—the BLAST-like alignment tool. *Genome Res* **12:** 656–664.

Kim JH, Waterman MS, Li LM. 2007. Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Res* **17:** 1101–1110.

Kim EB, Fang X, Fushan AA, Huang Z, Lobanov AV, Han L, Marino SM, Sun X, Turanov AA, Yang P, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* **479:** 223–227.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5:** R12.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al. 2009b. The sequence and *de novo* assembly of the giant panda genome. *Nature* **463:** 311–317.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20:** 265–272.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1:** 1–18.

Murchison EP, Schulz-Trieglaff OB, Ning Z, Alexandrov LB, Bauer MJ, Fu B, Hims M, Ding Z, Ivakhno S, Stewart C, et al. 2012. Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell* **148:** 780–791.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al. 2000. A whole-genome assembly of *Drosophila*. *Science* **287:** 2196–2204.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497:** 579–584.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci* **98:** 9748–9753.

The Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* **475:** 189–195.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, et al. 2012. GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res* **22:** 557–567.

Small KS, Brudno M, Hill MM, Sidow A. 2007. Extreme genomic variation in a natural population. *Proc Natl Acad Sci* **104:** 5698–5703.

Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A, et al. 2011. The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477:** 207–210.

Takeuchi T, Kawashima T, Koyanagi R, Gyoja F, Tanaka M, Ikuta T, Shoguchi E, Fujiwara M, Shinzato C, Hisata K, et al. 2012. Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. *DNA Res* **19:** 117–130.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, et al. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2:** e1326.

Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, Anderson S, Mesirov JP, Satoh N, Satou Y, Nusbaum C, et al. 2005. Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Res* **15:** 1127–1135.

You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, et al. 2013. A heterozygous moth genome provides insights into herbivory and detoxification. *Nat Genet* **45:** 220–225.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829.

Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, Yang P, Zhang L, Wang X, Qi H, et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490:** 49–54.

Zheng W, Huang L, Huang J, Wang X, Chen X, Zhao J, Guo J, Zhuang H, Qiu C, Liu J, et al. 2013. High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nat Commun* **4:** 2678.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA assembler. *Bioinformatics* **29:** 2669–2677.

Received December 6, 2013; accepted in revised form April 21, 2014.

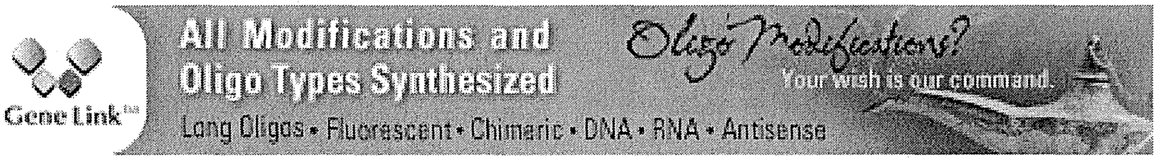
CSH PRESS

GENOME RESEARCH

Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads

Rei Kajitani, Kouta Toshimoto, Hideki Noguchi, et al.

Genome Res. 2014 24: 1384-1395 originally published online April 22, 2014
Access the most recent version at doi:10.1101/gr.170720.113



細胞進化の証人たち 細胞進化モデル生物図鑑

監修 永宗 喜三郎 国立感染症研究所 寄生動物部

第12回（最終回）

動物が動植物に入り込む！

自由生活から寄生生活への進化モデル：糞線虫

菊地泰生[1]，丸山治彦[2]

宮崎大学大学院医学獣医学総合研究科

[1] E-mail：taisei_kikuchi@med.miyazaki-u.ac.jp，[2] E-mail：hikomaru@med.miyazaki-u.ac.jp

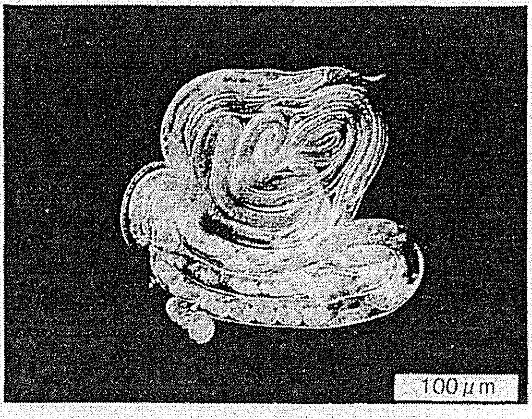
線虫は一般にはなじみがなく目立たない存在だが，動物の中では最も成功したグループの1つである．有名な線虫学者ネイサン・コブが「世界は線虫という薄い膜で覆われている」という言葉を残したくらい[1]，深海を含む海洋から陸上の土壌中に至るまで，膨大な種数の線虫が数え切れないくらいたくさん生息している[2, 3]．生活様式は当然のことながら多彩で，バクテリアや動植物の死骸，沈殿物などを餌として自由生活を営むもの，植物に寄生するもの，動物に寄生するものなど様々である．寄生という生活様式は，線虫門では少なくとも15回は獲得されたとする説もある[4]．本モデル生物図鑑シリーズでは，細菌や藻類が様々な宿主細胞に共生を始めた歴史が示されてきたが，線虫も，過去から現在に至るまで，いろいろなグループが多細胞の動植物への侵入寄生を試みている．線虫は，動物が自由生活から寄生生活へと進化した跡をたどる，格好のモデルを提供してくれるのである．

このような視点から我々が注目しているのが糞線虫類である．糞線虫類は脊椎動物の消化管に寄生する線虫で，種によってヒトやラット，ウシやヘビなど種々の宿主に寄生する[5]．じつは糞線虫には，数ある寄生虫の中でもユニークな特徴がある．それは宿主に寄生して産卵する寄生世代と，宿主外の土壌中で成熟して交尾，産卵する自由生活世代の両方があることである．寄生世代のメスが生んだ卵からは自由生活世代のオスとメスが発育し，自由生活世代のメスが生んだ卵からは寄生世代のメスが発育する〔寄生世代にオスはなく，メスは単為生殖で産卵する（図1）〕．以上のような生活史の特徴から，糞線虫類は，自由生活線虫から動物寄生性線虫への過渡期にあるグループではないかと考えられている[6]．この仮定が正しければ，糞線虫と糞線虫周辺の線虫類の詳細な比較ゲノム・比較生物学的研究によって，どのように自由生活線虫が「動物寄生能力」を獲得したのか再現できるはずである．

糞線虫類は，属全体では40〜50種あるとされるが，現在ゲノム解読が進められているのは，ラットを自然宿主とするネズミ糞線虫とベネズエラ糞線虫，ヒトに感染する糞線虫，そしてウシやヒツジなど反芻獣に寄生する乳頭糞線虫の4種である．これに近縁の線虫2種（パラストロンジロイデスとラブディトファネス）を加えて，比較ゲノム研究が進行中である．これまでに構築されたゲノムは，ポスト*C. elegans*としての利用が期待できるほどの高い品質で，高精度な比較解析を可能にしている（ウェルカムトラストサンガー研究所，ブリストル大学，宮崎大学の共同研究）．

細胞進化モデル生物 File 12

ベネズエラ糞線虫

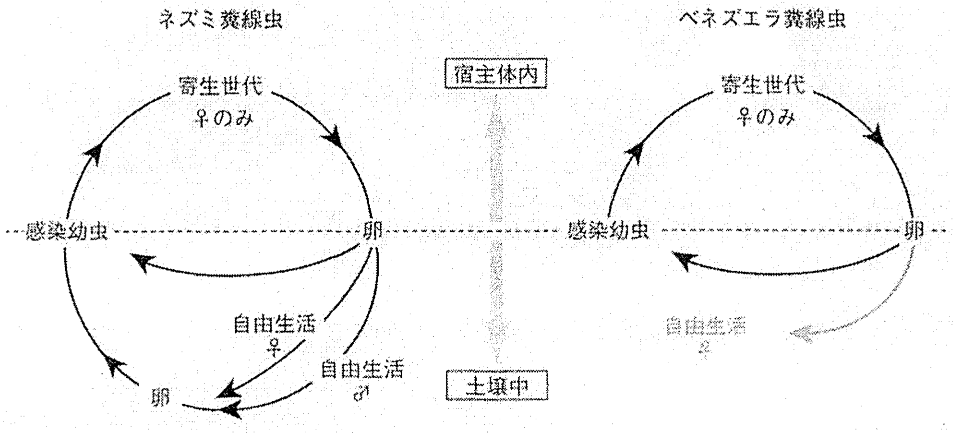


［DATA］

学名：*Strongyloides venezuelensis*

体長：約2 mm

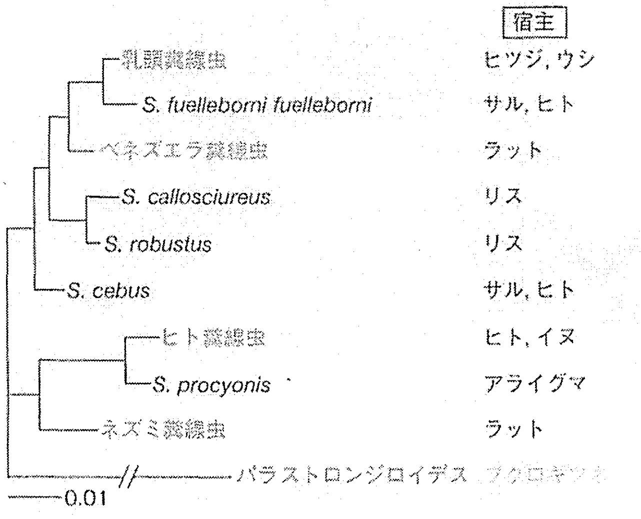
分布：世界中のドブネズミ（上部小腸に寄生）



［図1］ 糞線虫の生活史
ネズミ糞線虫とベネズエラ糞線虫の生活史．ほとんどの糞線虫はネズミ糞線虫型で，自由生活世代と寄生世代を交互に繰り返す．ベネズエラ糞線虫は事実上自由生活世代を欠いている．

我々は最近，比較ゲノムプロジェクトに含まれている4種の糞線虫と他のいくつかの糞線虫の類縁関係を，200種類以上の相同遺伝子の比較によって明らかにした．その結果，ヒト糞線虫とネズミ糞線虫が1つのグループを，乳頭糞線虫とベネズエラ糞線虫が別のグループを形成していることがわかった（図2）．染色体数も，ヒト糞線虫とネズミ糞線虫が2n＝6で，乳頭糞線虫とベネズエラ糞線虫は2n＝4と分かれており，塩基配列による系統解析を支持する[7]．ネズミ糞線虫とベネズエラ糞線虫はどちらもラットを自然宿主としているが，どうやら系統的に近いわけではなく，それぞれ独立にラットへの寄生能力を獲得したようである．

我々の研究室では，特にベネズエラ糞線虫に焦点を当てて研究に取り組んでいる．それは，ベネズエラ糞線虫は他の糞線虫と違って自由生活世代の成虫をまったくと言っていいほど出さないからである．寄生世代のメスから産出された虫卵から発育するのはほぼ全部が感染型の幼虫で，すべてが感染して寄生型になるメスである（図1）．ごく稀に自由生活世代のメスは観察できるが，オスはまったく出現しない[7]．これは，ベネズエラ糞線虫は今まさに自由生活世代を捨てつつあるのであり，純粋な寄生虫として「離陸」しようとしているのだと解釈できないだろうか．自由生活が出現しない理由をネズミ糞線虫との詳細な比較で明らかにすることを手始めに，ゲノムに記された自由生活から寄生生活への進化プロセスを解明していくことが目下の我々の目標である．



［図2］ 糞線虫と近縁種の系統関係
200個以上の遺伝子を用いて作製した精度の高い系統図．赤い字で示した種ではゲノムプロジェクトが進行中．糞線虫では，雌性生殖器の構造も，ヒト糞線虫とネズミ糞線虫は直線型，乳頭糞線虫とベネズエラ糞線虫はらせん型というように分かれている．

文献

1) Cobb NA: Nematodes and their relationships (Yearbook of United States Department of Agriculture): pp.457-490, 1914
2) De Ley P: WormBook (2006) DOI: 10.1895/wormbook.1.41.1
3) Kumar S, et al: Worm (2012) 1: 42-50
4) Blaxter M, et al: Parasitology (2014) 25: 1-14
5) Viney ME, et al: WormBook (2007) DOI: 10.1895/wormbook.1.141.1
6) Dorris M, et al: Int J Parasitol (2002) 32: 1507-1517
7) Hino A, et al: Parasitology (2014) in press

Acta Tropica 141 (2015) 178–183

Contents lists available at ScienceDirect

Acta Tropica

journal homepage: www.elsevier.com/locate/actatropica

Detection of active schistosome infection by cell-free circulating DNA of *Schistosoma japonicum* in highly endemic areas in Sorsogon Province, the Philippines☆

Naoko Kato-Hayashi[a], Lydia R. Leonardo[b], Napoleon L. Arevalo[c], Ma. Nerissa B. Tagum[d], James Apin[e], Lea M. Agsolid[f], James C. Chua[b], Elena A. Villacorte[b], Masashi Kirinoki[a], Mihoko Kikuchi[g], Hiroshi Ohmae[h], Kosuke Haruki[i], Yuichi Chigusa[a,*]

[a] *Laboratory of Tropical Medicine and Parasitology, Dokkyo Medical University, Mibu 321-0293, Tochigi, Japan*
[b] *Department of Parasitology, College of Public Health, University of the Philippines Manila, 625 Pedro Gil St., Ermita, Manila 1000, the Philippines*
[c] *Center for Health Development No. 5, Department of Health, Sorsogon City, Sorsogon, the Philippines*
[d] *Municipal Health Office, Irosin, Sorsogon, the Philippines*
[e] *Municipal Health Office, Juban, Sorsogon, the Philippines*
[f] *Provincial Health Office, Sorsogon City, Sorsogon, the Philippines*
[g] *Department of Immunogenetics, Institute of Tropical Medicine (NEKKEN), Nagasaki University, Sakamoto, Nagasaki 852-8523, Japan*
[h] *Department of Parasitology, National Institute of Infectious Diseases, Toyama 1-23-1, Shinjuku-ku, Tokyo 162-8640, Japan*
[i] *Department of Clinical Laboratory, Dokkyo Medical University Koshigaya Hospital, Koshigaya 343-8555, Saitama, Japan*

ARTICLE INFO

Article history:
Available online 15 May 2014

Keywords:
Schistosomiasis japonica
Diagnosis
Active infection
Cell-free circulating parasite DNA
Serum
Urine

ABSTRACT

The current status of schistosomiasis in highly endemic areas is difficult to determine by ovum detection because of the superficially low parasite load after mass drug administration, whereas the parasite transmission rates are still high. Cell-free parasite DNA is fragments of parasite-derived DNA existing in the host's body fluids. We conducted population-based studies to test the presence of cell-free schistosome DNA in endemic areas of Sorsogon Province, the Philippines. Schistosome DNA in the serum and urine of Kato–Katz (KK)-positive subjects was detected by PCR (100% sensitivity). Schistosome DNA was also detected from KK-negative subjects (9/22 serum and 10/41 urine samples). Schistosome DNA was found to be network echogenic pattern (NW)-positive (serum 53.3%, urine 42.9%) or NW-negative (serum 25.5%, urine 20.8%) and enzyme-linked immunosorbent assay (ELISA)-positive (serum 47.1%, urine 40%) or ELISA-negative (serum 33.3%, urine 13.3%). These results indicate that cell-free schistosome DNA is a promising diagnostic marker for active schistosome infection in the case of light infection.

1. Introduction

Schistosomiasis is a major parasitic disease affecting approximately 240 million people worldwide, with more than 700 million people at risk of infection (WHO Schistosomiasis, 2011). The disease results in diverse health and socioeconomic problems, ranging from atelioses to the death of the infected individuals and workforce reduction. In recent years, the disease has also become a health threat in non-endemic areas as a result of parasite importation (Clerinx and Gompel, 2011; Wichmann et al., 2013).

Most of the current diagnostic methods are concerned with detection of the presence of schistosome ova or parasite-specific antibodies by microscopy or immunological methods, respectively (Doenhoff et al., 2004). Pathological lesions are diagnosed as characteristic images by ultrasound (US) or computed tomography and confirmed by biopsy (Fig. 1). Although the current gold standard for diagnosing schistosomiasis is ovum detection, it is not applicable in the early stage of infection because of the absence of parasite ova that cause the pathological lesions. Furthermore, the symptoms and signs in the early stage of infection mimic those of various other diseases. Therefore, it is important to develop early diagnostic tools that are not dependent on the presence of ova.

Cell-free circulating nucleic acids are fragments of nucleic acids that have been liberated from cells and exist in the bloodstream,

☆ Some parts of this study were presented at the Report on the Collaborative Project, Institution of Tropical Medicine (Nekken), Nagasaki University, Japan (2010), the Proceedings of the Annual Meeting of Japanese Society of Parasitology (2011, 2013, Tokyo, Japan) and the RNAS+ Meeting (Hanoi, 2012).

* Corresponding author. Tel.: +81 282 87 2134; fax: +81 282 86 6431.
E-mail address: ychigusa@dokkyomed.ac.jp (Y. Chigusa).

http://dx.doi.org/10.1016/j.actatropica.2014.05.003
0001-706X/© 2014 Elsevier B.V. All rights reserved.

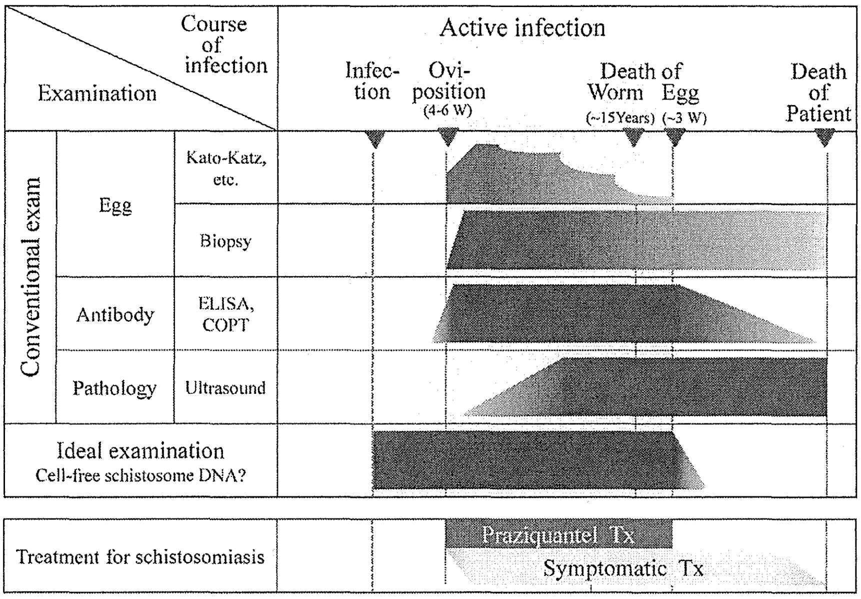


Fig. 1. Schematic diagram of diagnostic examinations and treatment of schistosomiasis. Current diagnostics for schistosomiasis are detection of parasite ova, parasite-specific antibodies and pathological lesions. All three diagnostics involve ova-related phenomena, and their results do not always reflect active schistosome infection.

urine and other body fluids. Recent studies have demonstrated that cell-free circulating nucleic acids in plasma/serum and urine are useful as molecular diagnostic tools in oncology, prenatal diagnosis, transplantation and other clinical areas (Botezatu et al., 2000; Chan et al., 2003). Furthermore, parasite-derived cell-free circulating DNA can be detected in patient plasma, urine and saliva (Pontes et al., 2002; Gal and Wainscoat, 2006; Mharakurwa et al., 2006; Nwakanma et al., 2009; Buppan et al., 2010; Parija and Khairrnar, 2007; Khairnar and Parija, 2008; Sandoval et al., 2006a; Wichmann et al., 2009, 2013; Enk et al., 2012; Lodh et al., 2013; Kato-Hayashi et al., 2013). We have demonstrated schistosome DNA in the serum and urine of infected animals 1 day after infection (Kato-Hayashi et al., 2010). These findings suggest that cell-free circulating schistosome DNA in host body fluids may indicate active schistosome infection. We conducted preliminary population-based studies to detect active schistosome infection using cell-free circulating DNA of *Schistosoma japonicum* in highly endemic areas of the Philippines.

2. Materials and methods

2.1. Study areas

The studies were conducted in villages (barangays) endemic for *Schistosoma japonicum* in Sorsogon Province, which is located on the southern Luzon Island of the Philippines. The barangays were Bacolod, Bagsangan, Bolos, Buena Vista, Bura Buran, Carriedo, Gumapia, Guruyan, and San Pedro.

2.2. Examinations

The participants were examined by abdominal US examinations, enzyme-linked immunosorbent assays (ELISAs) and Kato–Katz (KK) stool tests (Endiss et al., 2005) using single stool samples. In the first study in 2009 (Study 1), schistosome DNA detection was performed with a focus on those participants who were diagnosed with severe hepatic fibrosis. Hepatic fibrosis is characteristic of chronic schistosomiasis and is triggered by parasite ova. The hepatic lesions have a network echogenic pattern (NW) in case of infection with *S. japonicum* (US type 3) (Ohmae et al., 1992). This can indicate either a number of infections over a period of time or even long-standing infection. Twenty-three participants (21 males and two females, 16–63 years of age) were enrolled (23 serum and six saliva samples). In the second study in 2012 (Study 2), schistosome DNA detection was performed with a focus on those who lived in the Bagsangan barangay because, according to previous surveys, this barangay was highly endemic for schistosomiasis and because we expected more patients here to test positive for schistosome ova. Forty-five participants (29 males and 16 females, 6–74 years of age) were enrolled (45 urine samples). The participants were enrolled in only a single study and not both.

In parallel with the examinations, each participant was given a questionnaire regarding the symptoms of their illness that inferred schistosome infection, e.g., bloody stool, abdominal pain, haematemesis, dizziness, headache, convulsion, paralysis and speech disturbance and the experience of praziquantel (PZQ) treatment. Furthermore, they were asked about the presence of snail colonies in their daily living area and their experiences with farm work.

Before these examinations, informed consent was obtained from all of the participants. This study was approved by the Bioethics Committee of Dokkyo Medical University (approval No. 1969).

2.3. ELISA

For convenience of transportation and storage, blood samples for the ELISA test were obtained on a piece of filter paper (Advantec® Blood Sampling Paper Type I, Toyo Roshi Kaisha, Ltd., Japan) and allowed to dry at room temperature (RT). A parasite-specific antibody in the serum was tested by standard ELISA (Matsuda et al., 1984). To prepare the samples for the assay, a 3.0-mm-diameter disc of blood-soaked filter paper (approx. 4 μl of whole blood, i.e., approx. 2 μl of serum) was punched out and extracted in 600 μl of extraction buffer (1% BSA, 0.05% Tween 20, 0.5% skim milk in PBS) overnight at 4 °C. Polystyrene 96-well ELISA plates (Greiner Bio-One, Co., Ltd., Germany) were coated with 100 μl/well of *S. japonicum* soluble egg antigens (SEA, 10 μg/ml) in 0.05 M carbonate bicarbonate buffer (pH 9.6) and incubated at 37 °C and then left overnight at 4 °C. The plates were then washed

Table 1
Comparative evaluation of PCR and conventional examinations.

Study 1

			KK									PCR								
	Positive rate		+	–	Total	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	PLR (95% CI)	NLR (95% CI)	+	–	Total	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	PLR (95% CI)	NLR (95% CI)
PCR (serum)	43.5% (n = 23)	+	1	9	10	100%	59.1%	10.0%	100%	2.4	0									
		–	0	13	13	(1.3–100)	(35.4–79.3)	(0.3–44.5)	(66.1–100)	(1.4–4.0)										
NW	65.2% (n = 23)	+	1	14	15	100%	36.4%	6.7%	100%	1.6	0	8	7	15	80.0%	46.2%	53.3%	75.0%	1.5	0.4
		–	0	8	8	(1.3–100)	(17.2–59.3)	(0.2–31.9)	(51.8–100)	(1.1–2.2)		2	6	8	(44.4–97.5)	(19.2–74.9)	(26.6–78.7)	(34.9–96.8)	(0.8–2.7)	(0.1–1.7)
ELISA	73.9% (n = 23)	+	1	16	17	100%	27.3%	5.9%	100%	1.3	0	8	9	17	80.0%	30.8%	47.1%	66.7%	1.2	0.7
		–	0	6	6	(1.3–100)	(10.7–50.2)	(0.1–28.7)	(42.1–100)	(1.1–1.8)		2	4	6	(44.4–97.5)	(9.1–61.4)	(23–72.2)	(22.3–95.7)	(0.7–1.9)	(0.1–2.9)
KK	4.3% (n = 23)	+										1	0	1	10.0%	100%	100%	59.1%	0	0.9
		–										9	13	22	(0.3–44.5)	(66.1–100)	(1.3–100)	(36.4–79.3)		(0.7–1.1)

Study 2

			KK									PCR								
	Positive rate		+	–	Total	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	PLR (95% CI)	NLR (95% CI)	+	–	Total	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	PLR (95% CI)	NLR (95% CI)
PCR (urine)	31.1% (n = 45)	+	4	10	14	100%	75.6%	28.6%	100%	4.1	0									
		–	0	31	31	(28.4–100)	(59.7–87.6)	(8.4–58.1)	(83.8–100)	(2.4–7.0)										
NW	46.7% (n = 45)	+	2	19	21	50.0%	53.7%	9.5%	91.7%	1.1	0.9	9	12	21	64.3%	61.3%	42.9%	79.2%	1.7	0.6
		–	2	22	24	(6.8–93.2)	(37.4–69.3)	(1.2–30.4)	(73–99)	(0.4–3.0)	(0.3–2.6)	5	19	24	(35.1–87.2)	(42.2–78.2)	(21.8–66.0)	(57.8–92.9)	(0.9–3.0)	(0.3–1.2)
ELISA	66.7% (n = 45)	+	3	27	30	75.0%	34.1%	10.0%	93.3%	1.1	0.7	12	18	30	85.7%	41.9%	40.0%	86.7%	1.5	0.3
		–	1	14	15	(19.4–99.4)	(20.1–50.6)	(2.1–26.5)	(68.1–99.8)	(0.6–2.1)	(0.1–4.2)	2	13	15	(57.2–98.2)	(24.5–60.9)	(22.7–59.4)	(59.5–98.3)	(1.0–2.1)	(0.1–1.3)
KK	8.9% (n = 45)	+										4	0	4	28.6%	100%	100%	75.6%	0	0.7
		–										10	31	41	(8.4–58.1)	(83.8–100)	(28.4–100)	(59.7–87.6)		(0.5–1.0)

KK: schistosome ova detected by Kato–Katz stool test, PCR: cell-free circulating schistosome DNA detected by polymerase chain reaction, NW: ova-induced pathology detected by ultrasonography as network echogenic pattern, ELISA: schistosome-specific antibody detected by enzyme-linked immunosorbent assay PPV: positive predictive value, NPV: negative predictive value, PLR: positive likelihood ratio, NLR: negative likelihood ratio, CI: confidence interval.

three times with PBS–T (0.05% Tween 20 in PBS), blocked with 120 μl/well of 1% BSA in PBS–T and incubated for 15 min at RT. After being washed three times, 100 μl/well of samples was added to the plates, which were then incubated for 45 min at 37 °C. After being washed three times, horseradish peroxidase (HRP)-conjugated goat anti-human IgG (Cappel, USA) was diluted 1:2000 with 1% BSA in PBS–T, and 100 μl was added to each well, and the plates were incubated for 1 h at 37 °C. Then, 200 μl/well of substrate solution containing 0.03% of 2,2′-azino-bis(3-ethylbenzthiazoline-6-sulfonic acid) (ABTS; Sigma-Aldrich Co., USA) and 0.003% H_2O_2 in 0.1 M citrate phosphate buffer (pH 5.0) was added and the plates incubated for 1 h at RT. Optical density (OD) was measured at 415 nm. The mean OD value of the positive controls in each plate was adjusted to 1.064, and concomitantly, the rest of the original data were adjusted. OD values of $\geq$0.200 were considered positive.

2.4. DNA extraction and PCR amplification

Because it is difficult to detect schistosome DNA from a small amount of blood on a piece of filter paper, blood was collected using a syringe, and the serum was separated by centrifugation. Urine (3.5 ml) was concentrated to 140 μl using an Amicon® Ultra-15 Centrifugal Filter Device, 100K (Merck Millipore Ltd., Ireland). The total DNA from serum (140 μl) and urine was extracted by using a QIAamp® Viral RNA Mini Kit (QIAGEN Sciences, Maryland, USA). Approximately 2 ml of the saliva sample was collected using an Oragene® DNA kit (DNA Genotek Inc., Canada) and stored at RT. DNA was extracted according to the manufacturer's instructions.

The primer pair CF (5′-GATCGTAAATTTGGA/TACTGC-3′) and CR (5′-CCAACCATAAACATATGATG-3′) was designed to detect part of the schistosome mitochondrial cytochrome c oxidase subunit 1 (CO1) gene, which is common in at least four human schistosomes (253 bp: *Schistosoma mansoni*; 254 bp: *Schistosoma haematobium*, *Schistosoma japonicum*, and *Schistosoma mekongi*) (Kato-Hayashi et al., 2010). PCR was performed in a final volume of 20 μl containing 2 μl of 10× PCR buffer, 2.5 mM $MgCl_2$, 0.2 mM aliquots of each dNTP, 0.5 U of Platinum® *Taq* DNA polymerase (Invitrogen™, USA), 0.5-μM aliquots of each primer and 2 μl of template DNA. The reactions were performed initially at 94 °C for 2 min, followed by 50 cycles of 94 °C for 30 s, 58 °C for 30 s, 72 °C for 60 s and a further 72 °C for 7 min. The PCR products were identified by electrophoresis on 2% agarose in TAE gels with 0.3 μg/ml ethidium bromide and then visualised under UV light.

2.5. Statistical analyses

All statistical analyses were performed with EZR (Saitama Medical Center, Jichi Medical University; http://www.jichi.ac.jp/saitamaHP.files/statmedEN.html; Kanda, 2013), which is a modified version of R commander (The R Foundation for Statistical Computing) and designed for implementing statistical functions frequently used in biostatistics.

3. Results

Study 1 focused on subjects who had severe hepatic fibrosis (NW-positive rate: 65.2%, 15/23). Only one participant (4.3%) tested positive for schistosome ova by the single KK test, whereas 17 participants (73.9%) tested positive by ELISA. Schistosome DNA was detected in serum samples of 10 of the 23 participants (43.5%) (Table 1). Furthermore, schistosome DNA was detected in 2 of the 6 saliva samples. In Study 2, the positive rate of single KK, NW, ELISA and PCR was 8.9% (4/45), 46.7% (21/45), 66.7% (30/45) and 31.1% (14/45), respectively (Table 1). Table 2 shows a comparison between schistosome DNA detection by the above methods. Schistosome DNA in serum and urine of KK-positive participants was detected by PCR (100% sensitivity); schistosome DNA was also detected in KK-negative participants (9 of 22 serum samples and 10 of 41 urine samples). Schistosome DNA was found to be NW-positive (serum 53.3%, urine 42.9%), NW-negative (serum 25.5%, urine 20.8%), ELISA-positive (serum 47.1%, urine 40%) and ELISA-negative (serum 33.3%, urine 13.3%). Table 3 shows the number of participants who had predictable symptoms of schistosome infection and experiences of previous PZQ treatment. The majority of participants worked on farms (95.7% and 86.7% in Study 1 and Study 2, respectively) and reported encountering snail colonies in their daily living areas (73.9% and 91.1% in Study 1 and Study 2, respectively).

Table 2
Comparison of diagnostic tests of *Schistosoma japonicum* infection.

Study 1					Study 2				
KK	NW	ELISA	PCR (serum)		KK	NW	ELISA	PCR (urine)	
			+	–				+	–
+	+	+	1	0	+	+	+	1	0
+	+	–	0	0	+	+	–	1	0
+	–	+	0	0	+	–	+	2	0
+	–	–	0	0	+	–	–	0	0
–	+	+	6	7	–	+	+	7	10
–	+	–	1	0	–	+	–	0	2
–	–	+	1	2	–	–	+	2	8
–	–	–	1	4	–	–	–	1	11
Total			10	13	Total			14	31

KK: schistosome ova detected by Kato–Katz stool test, NW: ova-induced pathology detected by ultrasonography as network echogenic pattern, ELISA: schistosome-specific antibody detected by enzyme-linked immunosorbent assay, PCR: cell-free circulating schistosome DNA detected by polymerase chain reaction.

4. Discussion

Currently, ova detection methods, such as the KK test, are still the gold standard for detection of active schistosome infection in conventional diagnostics. In the schistosomiasis-endemic areas of Southeast Asian countries, including the Philippines, mass drug administration (MDA) with PZQ is the main component of the control program (Tallo et al., 2008; WHO Schistosomiasis, 2011). More than 80% of participants in the present study had previously undergone PZQ treatment (Table 3). As confirmed in previous studies and our present study, stool examinations, especially for a single KK test, may not detect light infections in patients who have undergone MDA (Lin et al., 2008; Sinuon et al., 2010; Han et al., 2012) and in those with chronic infections. Overall, ova detection is not useful in the early stages of infection (prepatent period). Although ELISA and US examination are useful methods for schistosomiasis detection, their results do not always reflect active infection (Fig. 1). Advanced liver fibrosis detected by US is irreversible after PZQ treatment, and it takes several years for the positive ELISA results to turn negative after the treatment. Patients with chronic infections, especially those in high transmission areas, may become re-infected after treatment. In these patients, the new active infection and efficacy of PZQ cannot be evaluated by US and ELISA.

Cell-free schistosome DNA originates from worm and ova metabolites, excrement and/or cell debris. Its presence has been reported in patient serum and urine (Pontes et al., 2002; Sandoval et al., 2006a; Wichmann et al., 2009; Lodh et al., 2013; Kato-Hayashi et al., 2013). The results from studies of experimental animals have indicated that cell-free schistosome DNA is detectable in prepatent periods (Sandoval et al., 2006b; Suzuki et al., 2006; Xia et al., 2009; Kato-Hayashi et al., 2010). Accordingly, we propose that cell-free schistosome DNA is a feasible marker for detection of active schistosome infection independent of parasite ova. Although the concentration of schistosome DNA fragments in host body fluids

Table 3
Summary of questionnaire responses of participants regarding their experience of symptoms common in schistosome infection.

		Present symptoms and signs (%)								Experience of PZQ treatment (%)	Snail colony in daily living area (%)	Experience of farm work (%)
		Bloody stool	Abdominal pain	Haematemesis	Dizziness	Headache	Convulsion	Paralysis	Speech disturbance			
Study 1		0 (0)	2(8.7)	0 (0)	4(17.4)	6(26.1)	8* (34.8)	2* (8.7)	1* (4.3)	22 (95.7)	17 (73.9)	22 (95.7)
	KK+	0	0	0	0	0	0	0	0			
	PCR+	0	1	0	2	3	4	2	1			
Study 2		2(4.4)	6(13.3)	2(4.4)	15(33.3)	14(31.1)	2(4.4)	2(4.4)	1(2.2)	36 (80.0)	41 (91.1)	39 (86.7)
	KK+	1	0	0	1	1	0	0	0			
	PCR+	1	2	0	3	4	1	1	1			

KK+: number of the participants with clinical findings plus schistosome ova by Kato–Katz stool test; PCR+: number of the participants with clinical findings plus schistosome DNA by polymerase chain reaction.

* Number of participants with previous and/or present clinical findings.

may depend on parasite load, it is likely to be in minute amounts, as demonstrated by Wichmann et al. (2009), with stable results from relatively large amounts of patient serum (20 ml). Thus, the concentration and purification of sample DNA is required, except for cases of high parasite load, immediately after infection or with anthelmintic administration. In the present study, we detected schistosome DNA from relatively small amounts of samples (serum, 140 μl; saliva, approx. 250 μl; urine, 3.5 ml).

Detection of parasite-derived DNA in urine and saliva was recently reported in several studies (Mharakurwa et al., 2006; Nwakanma et al., 2009; Buppan et al., 2010; Parija and Khairrnar, 2007; Khairnar and Parija, 2008; Sandoval et al., 2006a; Lodh et al., 2013; Kato-Hayashi et al., 2013) indicating the feasibility of applying DNA detection methods to schistosomiasis. This method has the additional benefit of non-invasive collection of urine and saliva samples, resulting in the increase in cooperation of patients, and neither special equipment nor special training are required for sample collection. Moreover, using urine and saliva samples over blood reduces the biohazard risk to examiners.

Field applications of molecular-based detection of schistosomiasis mainly involve the detection of parasite ova-derived DNA from stool samples (Pontes et al., 2003; Gomes et al., 2009; Fung et al., 2012; Carneiro et al., 2013). The most frequent use of cell-free schistosome DNA is for detection of imported schistosomiasis in non-endemic countries (Sandoval et al., 2006a; Wichmann et al., 2009, 2013; Kato-Hayashi et al., 2013), and to date, very few field studies have employed this method (Lodh et al., 2013). Our study aimed at detecting active infection by targeting cell-free schistosome DNA in body fluids (serum, urine and saliva) in *S. japonicum*-endemic areas. Our results indicate a higher detection rate of active infections compared with the KK test (43.5% vs. 4.3% in Study 1; 31.1% vs. 8.9% in Study 2) (Table 1). The sensitivity of PCR was 100% for both Study 1 and 2, and the specificity was 59.1% in Study 1 and 75.6% in Study 2, based on a single KK test (Table 1). The relatively low specificities can be attributed to the verification bias and low sensitivity of the KK test. Schistosome DNA was detected from KK-negative subjects (nine of 22 serum and 10 of 41 urine samples) (Table 2). Some participants who reported neurological symptoms, including convulsion, paralysis and speech disturbance had negative results for the KK test; however, cell-free schistosome DNA was detected in their serum/urine (Table 3). Härter et al. (2014) demonstrated diagnosis of neuroschistosomiasis by schistosome DNA in the cerebrospinal fluid and serum. Cell-free schistosome DNA is not only sensitive but may also helpful in differential diagnosis, especially for cerebral schistosomiasis. It is also available for therapy evaluation after PZQ treatment (Kato-Hayashi et al., 2013).

In schistosomiasis-endemic areas, annual MDA is insufficient for disease elimination as the inhabitants are constantly at risk of infection or re-infection. Thus, a comprehensive approach including accurate diagnostics, control of snail and animal reservoir hosts, improvements in sanitation, and public health education is required. Cell-free schistosome DNA is a promising diagnostic marker for indication of active schistosome infection even in cases in which the parasite ova are difficult to detect because of PZQ intervention such as in MDA situations. The use of cell-free schistosome DNA as a diagnostic marker is necessary to obtain better information regarding the status of schistosome infections in endemic areas. Further investigation of its application in endemic areas is required.

Acknowledgements

This work was supported by the Collaborative Project, the Institute of Tropical Medicine (Nekken), Nagasaki University (22-ippan-25); Health Labour Sciences Research Grant, the Ministry of Health, Labour and Welfare, Japan (H20-Shinkosaiko-Ippan-016); and the Japan Society for the Promotion of Science (JSPS) KAKENHI (No. 24590510).

We would like to thank the staff of the Provincial Health Office, Sorsogon and the Municipal Health Office, Irosin and Juban, Sorsogon, the Philippines for their great help. We also thank the residents of the schistosomiasis-endemic barangays in Sorsogon Province, the Philippines. We also thank Dr. Ken-Ichi Manaka of the Research Support Center, Dokkyo Medical University for technical support and our laboratory coworkers, Dr. Satoru Kawai, Ms. Mayumi Ohshita, and Ms. Mayu Tanaka, at the Laboratory of Tropical Medicine and Parasitology, Dokkyo Medical University for their assistance.

References

Botezatu, I., Serdyuk, O., Potapova, G., Shelepov, V., Alechina, R., Molyaka, Y., Anan'ev, V., Bazin, I., Garin, A., Narimanov, M., Knysh, V., Melkonyan, H., Umansky, S., Lichtenstein, A., 2000. Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism. Clin. Chem. 46, 1078–1084.

Buppan, P., Putaporntip, C., Pattanawong, U., Seethamchai, S., 2010. Comparative detection of *Plasmodium vivax* and *Plasmodium falciparum* DNA in saliva and urine samples from symptomatic malaria patients in a low endemic area. Malaria J. 9, 72.

Carneiro, T.R., Peralta, R.H.S., Cristhiany, M., Pinheiro, C., de Oliveira, S.M., Peralta, J.M., Bezerra, F.S.M., 2013. A conventional polymerase chain reaction-based method for the diagnosis of human schistosomiasis in stool samples from individuals in a low-endemicity area. Mem. Inst. Oswaldo Cruz (Rio de Janeiro) 108, 1037–1044.

Chan, A.K.C., Chiu, R.W.K., Lo, Y.M.D., 2003. Cell-free nucleic acids in plasma, serum and urine: a new tool in molecular diagnosis. Ann. Clin. Biochem. 40, 122–130.

Clerinx, J., Gompel, A.V., 2011. Schistosomiasis in travellers and migrants. Travel Med. Infect. Dis. 9, 6–24.

Doenhoff, M.J., Chiodini, P.L., Hamilton, J.V., 2004. Specific and sensitive diagnosis of schistosome infection: can it be done with antibodies? Trends Parasitol. 20, 35–39.