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

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 ($k_{pre} < k$), 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-

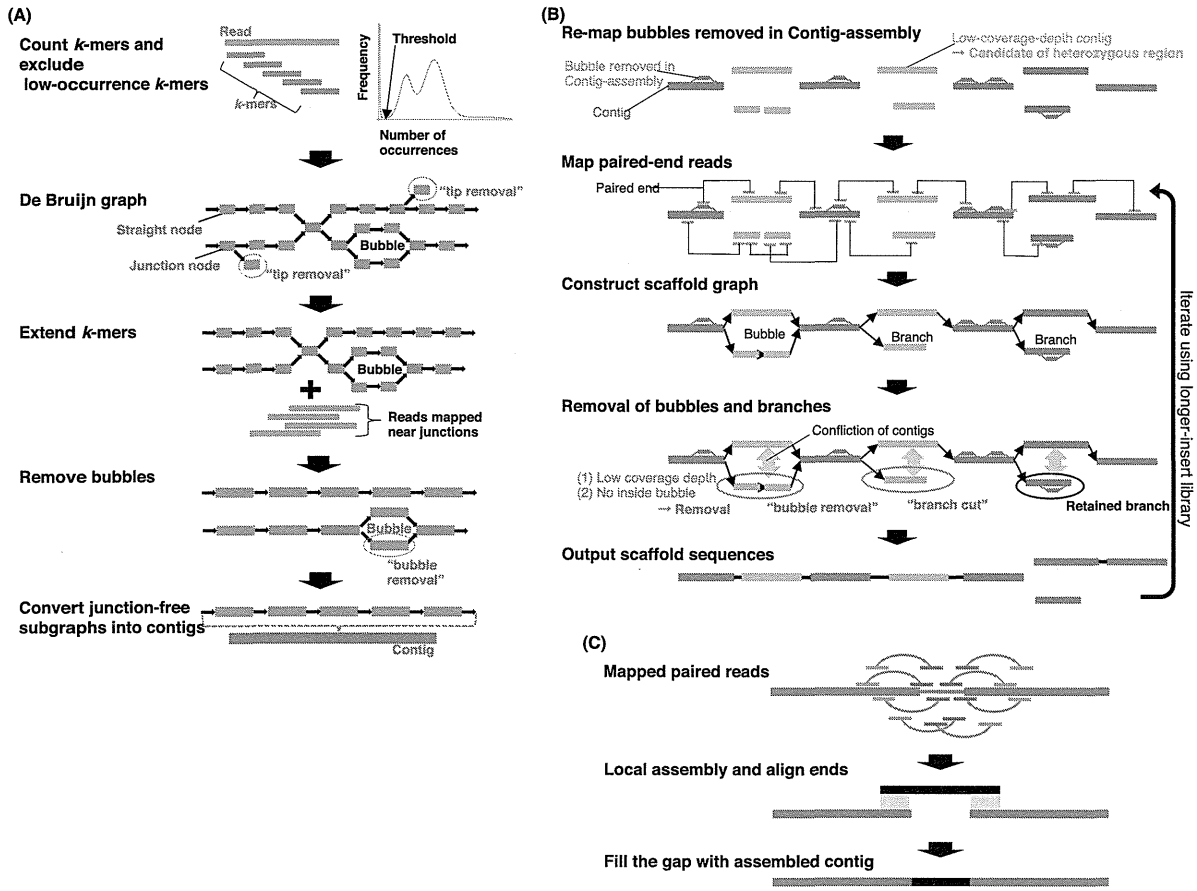

Figure 1. Schematic overview of the Platanus algorithm. (*A*) In Contig-assembly, a de Bruijn graph is constructed from the read set. Short branches caused by errors are removed by "tip removal." Short repeats are resolved by *k*-mer extension, in which previous graphs and reads are mapped to nearby *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 $2 \times c$, 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.

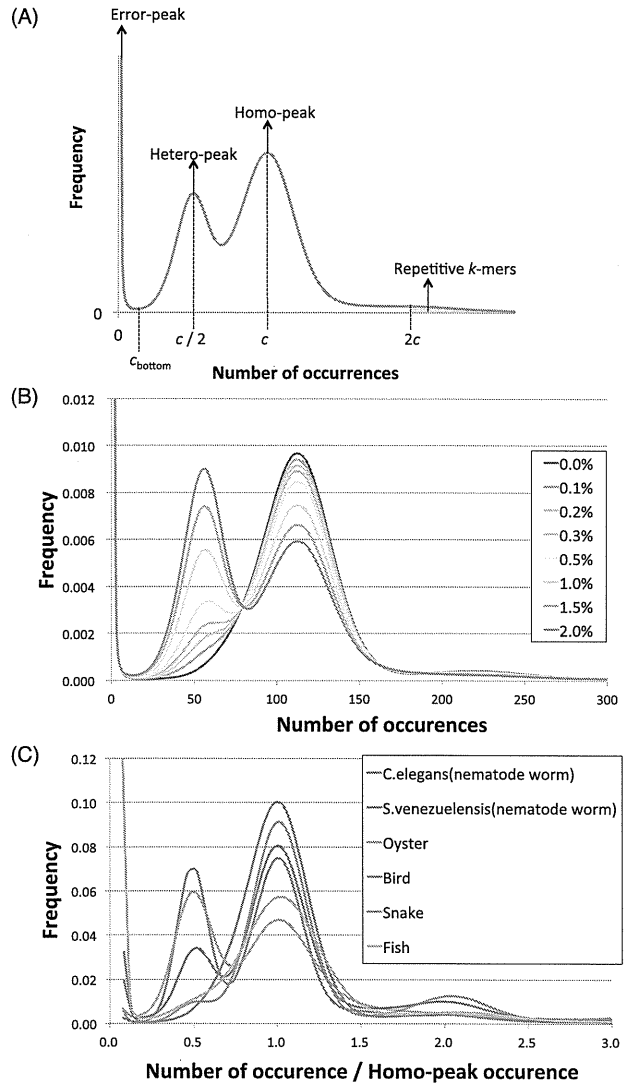

Figure 2. Distribution of the number of 17-mer occurrences. (*A*) Schematic model of the distribution of *k*-mer occurrences. This distribution is related to that shown in Table 1. (*B*) Simulated heterozygous data from *C. elegans*. (*C*) Distributions of normalized 17-mer occurrences for all species.

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.

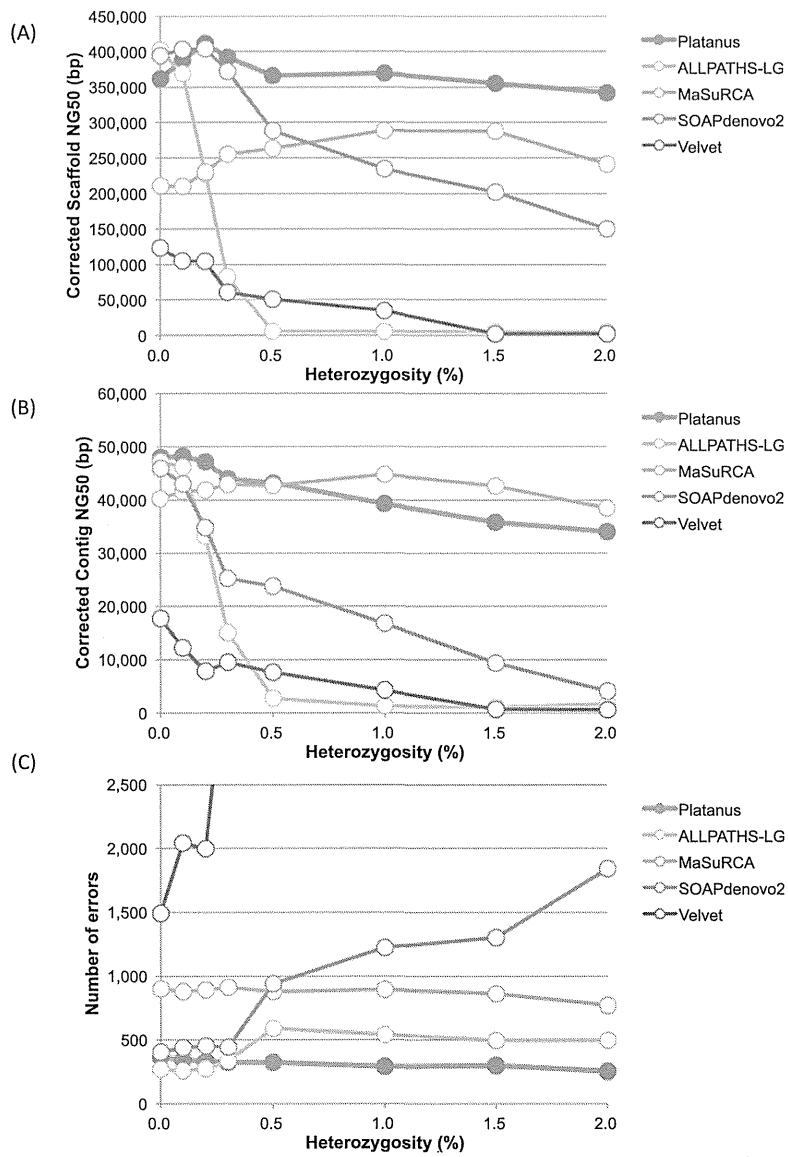

Figure 3. Results of the benchmarks of heterozygosity simulations (*C. elegans*). (*A*) Corrected scaffold-NG50 calculated by GAGE. (*B*) Corrected contig-NG50. (C) Number of errors reported by GAGE. Errors are defined as inversion, relocation, or translocation.

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

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.

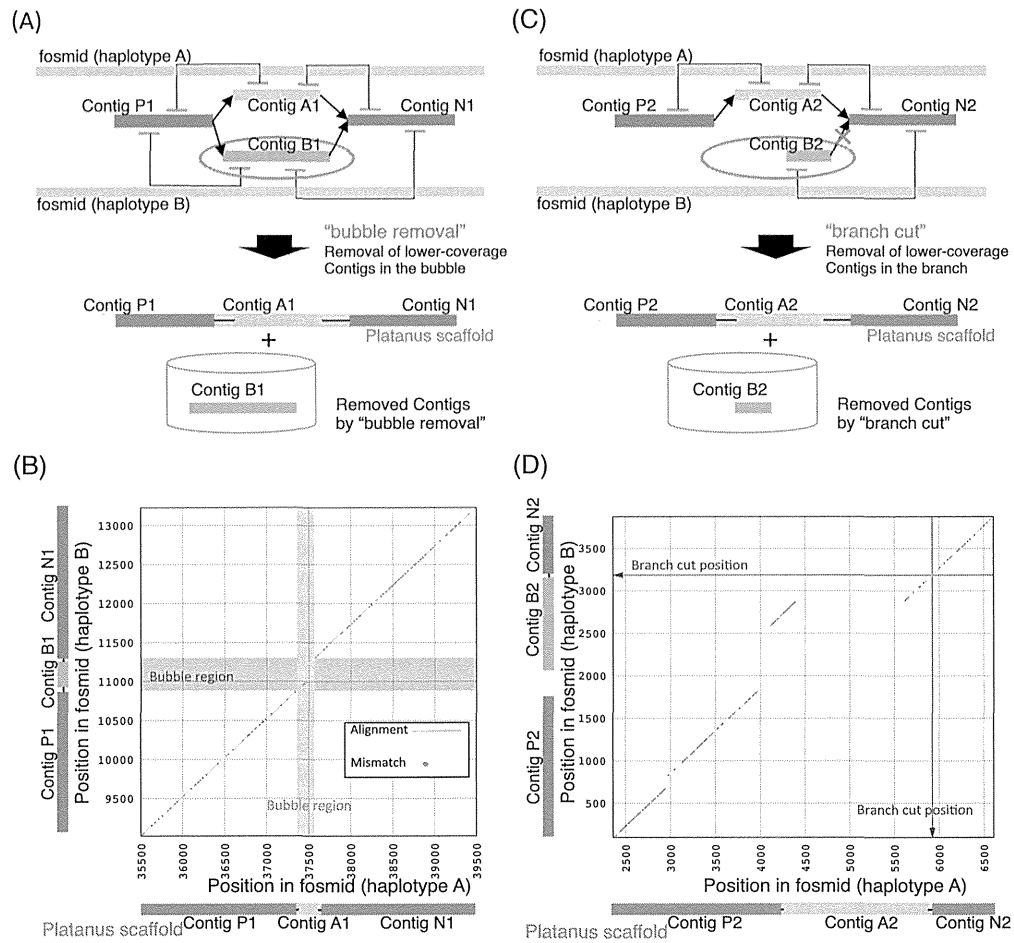

Figure 4. Example of a heterozygous region resolved by "bubble removal" and "branch cut." (*A*) Schematic model of "bubble removal" in Platanus scaffolding. (*B*) Alignment dot plot between two fosmids. Green lines and red dots indicate alignments and mismatches, respectively. Red and blue boxes indicate 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 valida-

tions 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.

DNA Research, 2015, 22(1), 101–107
doi: 10.1093/dnares/dsu043
Advance Access Publication Date: 26 November 2014
Full Paper

Full Paper

A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster

Atsushi Iguchi[1,*], Sunao Iyoda[2], Taisei Kikuchi[3], Yoshitoshi Ogura[4,5], Keisuke Katsura[5], Makoto Ohnishi[2], Tetsuya Hayashi[4,5], and Nicholas R. Thomson[6,7]

[1]Department of Animal and Grassland Sciences, Faculty of Agriculture, University of Miyazaki, Miyazaki 889-2192, Japan, [2]Department of Bacteriology I, National Institute of Infectious Diseases, Tokyo 162-8640, Japan, [3]Division of Parasitology, Department of Infectious Diseases, Faculty of Medicine, University of Miyazaki, Miyazaki 889-1692, Japan, [4]Division of Microbiology, Department of Infectious Diseases, Faculty of Medicine, University of Miyazaki, Miyazaki 889-1692, Japan, [5]Division of Bioenvironmental Science, Frontier Science Research Center, University of Miyazaki, Miyazaki 889-1692, Japan, [6]Pathogen Genomics, The Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK, and [7]Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

*To whom correspondence should be addressed. Tel/Fax. +81 985-58-7507. E-mail: iguchi@med.miyazaki-u.ac.jp

Edited by Dr Katsumi Isono

Received 16 September 2014; Accepted 3 November 2014

Abstract

The O antigen constitutes the outermost part of the lipopolysaccharide layer in Gram-negative bacteria. The chemical composition and structure of the O antigen show high levels of variation even within a single species revealing itself as serological diversity. Here, we present a complete sequence set for the O-antigen biosynthesis gene clusters (O-AGCs) from all 184 recognized *Escherichia coli* O serogroups. By comparing these sequences, we identified 161 well-defined O-AGCs. Based on the *wzx*/*wzy* or *wzm*/*wzt* gene sequences, in addition to 145 singletons, 37 serogroups were placed into 16 groups. Furthermore, phylogenetic analysis of all the *E. coli* O-serogroup reference strains revealed that the nearly one-quarter of the 184 serogroups were found in the ST10 lineage, which may have a unique genetic background allowing a more successful exchange of O-AGCs. Our data provide a complete view of the genetic diversity of O-AGCs in *E. coli* showing a stronger association between host phylogenetic lineage and O-serogroup diversification than previously recognized. These data will be a valuable basis for developing a systematic molecular O-typing scheme that will allow traditional typing approaches to be linked to genomic exploration of *E. coli* diversity.

Key words: *E. coli*, O-antigen biosynthesis gene cluster, horizontal gene transfer, O serogroup, genomic diversity

1. Introduction

Cell-surface polysaccharides play an essential role in the ability of bacteria to survive and persist in the environment and in host organisms.[1] The O-antigen polysaccharide constitutes the outermost part of the lipopolysaccharide (LPS) present in the outer membrane of Gram-negative bacteria. The chemical composition and structure of the O-antigen exhibit high levels of variation even within a single species.[2–5] This observation is corroborated by the huge serological

variation of somatic O antigens. Currently, the O serogrouping, sometimes combined with H (flagellar) antigens and K (capsular polysaccharide) antigens, is a standard method for subtyping of *Escherichia coli* strains in taxonomical and epidemiological studies. In particular, identification of strains of the same O serogroup is a prerequisite to start any actions for outbreak investigations and surveillance.

Thus far, the World Health Organization Collaborating Centre for Reference and Research on *Escherichia* and *Klebsiella* based at the Statens Serum Institut (SSI) in Denmark (http://www.ssi.dk/English.aspx) has recognized 184 *E. coli* O serogroups. It is generally believed that the O serogrouping of *E. coli* strains provides valuable information for identifying pathogenic clonal groups, especially for public health surveillance. For example, O157 is a leading O serogroup associated with enterohemorrhagic *E. coli* (EHEC) and is a significant food-borne pathogen worldwide.[6,7] Other important EHEC O serogroups include O26, O103, and O111.[8] The Shiga toxin-producing *E. coli* O104:H4 was found responsible for a large human food-borne disease outbreak in Europe, 2011.[9] Another notable example is strains of serogroup O25; extended-spectrum beta lactamase (ESBL)-producing, multidrug-resistant *E. coli* O25:H4 has emerged worldwide to cause a wide variety of community and nosocomial infections.[10]

In *E. coli*, the genes required for O-antigen biosynthesis are clustered at a chromosomal locus flanked by the colanic acid biosynthesis gene cluster (*wca* genes) and the histidine biosynthesis (*his*) operon. Generally, the O-antigen biosynthesis genes fall into three classes: (i) the nucleotide sugar biosynthesis genes, (ii) the sugar transferase genes, and (iii) those for O-unit translocation and chain synthesis (*wzx*/*wzy* in the Wzx/Wzy-dependent pathway and *wzm*/*wzt* in the Wzm/Wzt-dependent ABC transporter pathway).[11] To date, >90 types of O-antigen biosynthesis gene cluster (O-AGC) sequences have been determined, with the majority derived from major human and animal pathogens.[12] Sequence comparisons of these O-AGCs indicate a great variety of genetic structures. Several studies have provided evidence to show that horizontal transfer and replacement of a part or all of the O-AGC have caused shifts in O serogroups.[13–15] Alternatively, point mutations in the glycosyltransferase genes in the O-AGC or acquisition of alternative O-antigen modification genes, which are located outside of the O-AGC, have also been shown to result in structural alterations of O antigen and concomitant change in the serotype of the isolate.[16,17]

Genes or DNA sequences specific for each O serogroup can be used as targets for the identification of O serogroups via molecular approaches, such as PCR-based and hybridization-based methods. Such systems have already been developed by several researchers to target specific O-antigen types.[12,18–20] In particular, molecular assays targeting major O serogroups are routinely used in EHEC surveillance for clinical or food sample screening. Considering the range of diseases caused by *E. coli* strains belonging to many different serogroups, a more comprehensive and detailed O-AGC information for the complete set of *E. coli* O serogroups is of significant clinical importance for generating a rational molecular typing scheme. This molecular typing scheme, which could be performed *in silico* directly on sequence data, also offers a mechanism with which to link the ever-expanding genomic data to our extensive epidemiological and biological knowledge of this pathogen, based on O-antigen typing. Moreover, these data will also provide a much better understanding of the complex mechanisms by which a huge diversity in O serogroups have arisen. Here, we present a complete sequence set for the O-AGCs from all 184 *E. coli* O serogroups, which include recently added serogroups (O182–O187), providing a complete picture of the O-AGC diversity in *E. coli*.

2. Materials and methods

2.1. Bacterial strains, culture condition, and DNA preparation

Reference strains of all 184 recognized *E. coli* O serogroups were obtained from SSI (see Supplementary Table S1). Cells were grown to the stationary phase at 37°C in Luria–Bertani medium. Genomic DNA was purified using the Wizard Genomic DNA purification kit (Promega) according to the manufacturer's instructions.

2.2. O-AGC sequences and comparative analyses

One hundred and eight *E. coli* O-AGC sequences were determined by Sanger-based capillary sequencing and/or Illumina MiSeq sequencing from PCR products covering O-AGCs (Supplementary Table S1). The O-AGC regions of the reference strains were amplified by PCR using 10 ng of genomic DNA as template with the Tks Gflex DNA polymerase (Takara Bio Inc.) by 25 amplification cycles for 10 s at 98°C and for 16 m at 69°C, and with a combination of three forward primers (TATGCCAGCGGCACCAAACG, ATACCGGCGATGAAAGCC, and GCGGGTGGGATTAAGTCTCT) designed on the *hisFI* genes and two reverse primers (GTGATGCAGGAATCCTCTGT and CCACGCTAATTACGCCATCTT) designed on the *wcaM* genes, or strain-specific primers designed based on the draft genome sequences determined using the MiSeq system from reference strains. Identification and functional annotation of the CDSs were performed based on the results of homology searches against the public, non-redundant protein database using BLASTP. The sequences reported in this article have been deposited in the GenBank database (accession no. AB811596-AB811624, AB812020-AB812085, and AB972413-AB972425). The other 76 *E. coli* O-AGC sequences were obtained from public databases. For a list of accession numbers, see Supplementary Table S1.

2.3. Phylogenetic analysis

Multilocus sequence typing (MLST) was carried out according to the protocol described on the *E. coli* MLST website (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli), and the phylogenetic relationships of reference strains were analysed based on the concatenated sequences (3,423 bp) of seven housekeeping genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*) used for MLST. Multiple alignments of DNA and amino acid sequences were constructed by using the CLUSTAL W program.[21] Phylogenetic trees were constructed by using the neighbour-joining algorithm using the MEGA4 softrware.[22]

3. Results

3.1. Genetic structures of the O-AGCs from all *E. coli* O serogroups

Of the 184 known O serogroups, 76 complete O-AGC sequences were obtained from public databases. The sequence of the other 108 O-AGC was determined in this study from *E. coli* O-serogroup reference strains (Supplementary Table S1). Our analysis of these sequences confirmed several previously observed characteristics of O-AGCs in *E. coli* (Supplementary Fig. S1). In brief, O-AGCs are located between the *wca* and *his* operons. This region contains three housekeeping genes: *galF* (encoding UTP-glucose-1-phosphate uridylyltransferase), *gnd* (6-phosphogluconate dehydrogenase), and *ugd* (UDP-glucose 6-dehydrogenase), and most genes for O-antigen biosynthesis in each cluster are directly flanked by *galF* and *gnd*/*ugd*, while *gne* (UDP-GalNAc-4-epimerase) and *wzz* (O-antigen chain

length determination protein) located immediately outside of the region between *galF* and *gnd/ugd* (see Supplementary Fig. S1). The exceptions for this are the O-AGCs for O serogroups O14 and O57, which contain no O-antigen genes at the typical locus. However, it is known that the *E. coli* O14 reference strain Su4411-41 shows an O rough phenotype and lacks the O-AGC.[23] For O57, a further analysis is also required to investigate the presence of O-antigen structure in the LPS of the reference strain. Our data revealed that the O-AGCs located between *galF* and *gnd* ranged in size from 4.5 kbp (O155, including four genes) to 19.5 kbp (O108, including 18 genes).

3.1.1. Nucleotide sugar biosynthesis genes

Genes required for the deoxythymidine diphosphate (dTDP)-sugar biosynthesis pathway (*rmlBDAC*) to synthesize dTDP-L-rhamnose (dTDP-L-Rha), the precursor of L-Rha, were widely distributed in the O-AGCs (conserved in 56 O-serogroup O-AGCs; see Supplementary Fig. S1). The *vioAB* operon, for the biosynthesis of dTDP-*N*-acetylviosamine (dTDP-VioNAc), the precursor of VioNAc, was present in three O-serogroup O-AGCs; the *fnlABC* operon for the synthesis of uridine diphosphate (UDP)-*N*-acetyl-L-fucosamine (UDP-L-FucNAc), the precursor of L-FucNAc, was in 11 O-serogroup O-AGCs; the *fnlA-qnlBC* genes for the synthesis of UDP-*N*-acetyl-L-quinovosamine (UDP-L-QuiNAc), the precursor of L-QuiNAc, were in four O-serogroup O-AGCs; the *nnaDBCA* genes for synthesis of cytidine monophosphate (CMP)-*N*-acetylneuraminate (CMP–NeuNAc), the precursor of *N*-acetylneuraminic acid (Neu5Ac or sialic acid), were found in six O-serogroup O-AGCs (Supplementary Fig. S1). In addition, a gene set comprising seven genes putatively involved in the synthesis of di-*N*-acetyl-8-epilegionaminic acid (8eLeg5Ac7Ac) were found in three O-serogroup O-AGCs. For at least 49 O serogroups, gene sets for nucleotide sugar biosynthesis were not found in their O-AGCs (Supplementary Fig. S1), suggesting that, in these serogroups, nucleotide sugars required for O-antigen biosynthesis were synthesized by pathways encoded by the genes located outside of the O-AGCs.

3.1.2. Glycosyltransferase

Each O-AGC contained two to six genes encoding putative glycosyltransferases for synthesizing O-antigen subunits and a total of 611 glycosyltransferase genes identified in all O-AGCs. Pfam analysis revealed that at least 25 types of glycosyltransferase-related domains were found in the 611 glycosyltransferase genes (Supplementary Table S2). 'Glycosyl transferases group 1' (PF00534) and 'Glycosyl transferase family 2' (PF00535) were the most widely distributed domains, which were found in 216 and 253 genes, respectively. Except for the five genes belonging to 'Glycosyltransferase family 52' (PF07922), which were found in five of the six *nnaDBCA*-containing O-AGCs (O24, O56, O104, O131, and O171), there were no relationships between the type of glycosyltransferase-related domain and the gene set for sugar synthesis in each O-AGC.

3.1.3. O-antigen subunit translocation and chain synthesis

All O-AGCs carried either *wzx/wzy* or *wzm/wzt* gene pairs. Of the 182 O-AGCs (the above-mentioned O14 and O57 were excluded from the 184 clusters analysed in this study), 171 carried the *wzx/wzy* genes, and the other 11 carried the *wzm/wzt* genes (Supplementary Fig. S1 and Table S1). Detailed sequence comparisons of the *wzx/wzy* and *wzm/wzt* genes are described below.

3.2. Grouping the O-AGCs by sequence

On the basis of sequences and genetic structures of the entire O-AGC regions, in addition to 145 unique O-AGCs from different *E. coli* O serogroups, the O-AGCs from 37 O serogroups could be placed into 16 groups (named Gp1–Gp16) with the members of each group having identical or very similar O-AGC genes (mostly sharing ≥95% DNA sequence identity) (Fig. 1). This included nine groups with members of different serogroups but which carried identical O-AGC gene sets (Gp1–Gp9) and one group, Gp10, where two strains (O13 and O129) of the three-member group carried an identical O-AGC gene set (sharing 98.3–99.9% DNA sequence identity) (Fig. 1). The reason(s) why they belong to different O serogroups even though they have identical O-AGCs are discussed in the Discussion section. Indels or exchange of one or more genes was also shown to explain the differences between O135 and other members of Gp10 and members Gp11–Gp16, which otherwise carried highly conserved orthologous genes (summarized in Fig. 1). Simple insertions of insertion sequence (IS) elements containing one or two transposase genes were found in three groups without any gene disruption: an IS*629* insertion in O18ab of Gp12, IS*Ec11* in O164 of Gp13, and IS*1* in O62 of Gp14. IS element-associated replacement of the right-end portion of the O-AGC had occurred in three groups, Gp14, Gp15, and Gp16, resulting in the replacement (or deletion) of glycosyltransferase gene(s). Exchange of the *wzx* gene had also occurred in Gp16. These data suggest that IS elements are important drivers for generating O-antigen biosynthesis gene replacement and therefore diversity.

3.3. Diversity and specificity of the *wzx/wzy* or *wzm/wzt* genes among the *E. coli* O-AGCs

As previously proposed,[12] most *wzx/wzy* or *wzm/wzt* orthologues showed high levels of sequence diversity and their sequences were unique to each O-AGC or O-AGC group described above (Fig. 2 and Supplementary Fig. S2). DNA sequence identities of the closest pairs were <70%, except for the O96/O170 pair, the *wzx* genes of which showed 86% DNA sequence identity. Within the 16 O-AGC groups, the orthologous *wzx/wzy* or *wzm/wzt* genes also showed high sequence conservation (≥95% DNA sequence identity, but mostly ≥97% identity), except for Gp16 that shared only the *wzy* gene (Fig. 2).

3.4. Phylogenetic relationships of *E. coli* O-serogroup reference strains

Based on the concatenated nucleotide sequences of seven housekeeping genes used for MLST, we determined the evolutionary relationships of all *E. coli* O-serogroup reference strains (Fig. 3). This analysis revealed that the members of five groups sharing the common O-AGCs (Gp8, Gp10, Gp11, Gp14, and Gp15) and two members (O17 and O77) of Gp9 were found in closely related lineages. However, the members of other groups (and three members of Gp9) were found in distinct evolutionary lineages. For example, O20 and O137, both carrying the Gp1 O-AGC, were found in two distinct lineages, each belonging to phylogroups A and E/D, respectively, and five serogroups (O17/O77, O44, O73, and O106) belonging to Gp9 were found in multiple lineages (A, E/D, and B1).

The systematic phylogenetic analysis of all *E. coli* O-serogroup reference strains further revealed that one-quarter of the reference strains (46/184) belonged to a single clonal group (≥99.9% sequence identity), which was represented by sequence type (ST) 10 and its very close relatives in phylogroup A (Fig. 3 and Supplementary Fig. S3). Additionally, three clonal groups containing five or more reference

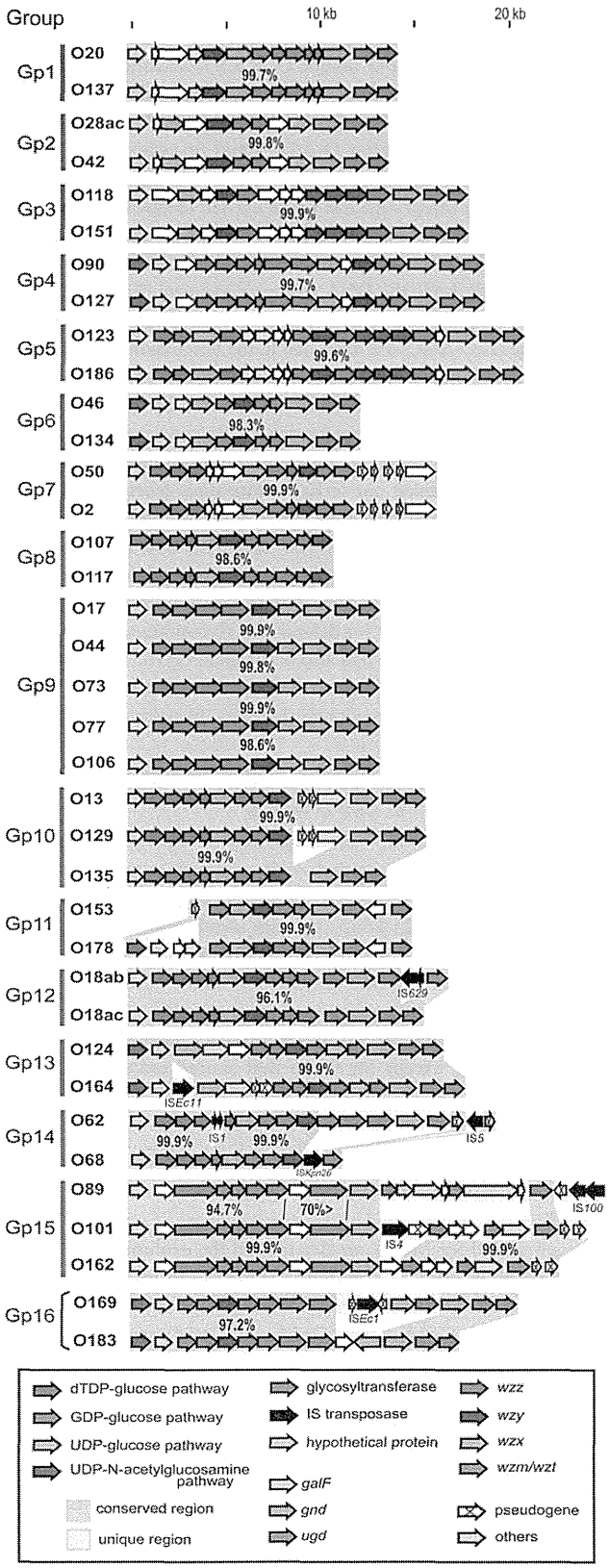

Figure 1. Sixteen *Escherichia coli* O-AGC groups identified in this study. Group members have different O serogroups in each group, but these share nearly identical or highly similar genetic organizations. Group names (Gp) are indicated at the left side. DNA sequence identities (%) between group members are indicated in each group.

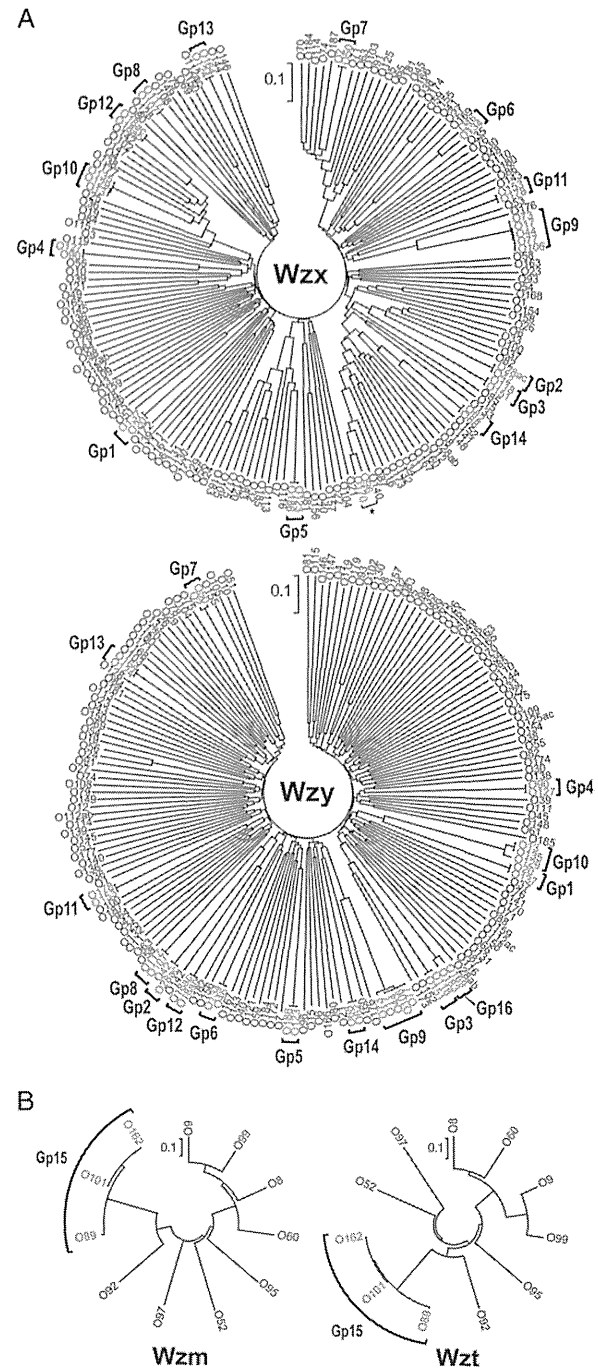

Figure 2. Phylogenetic analysis of homologues of (A) Wzx and Wzy and (B) Wzm and Wzt from *Escherichia coli* O-serogroup reference strains based on the amino acid sequences. The group names are indicated outside of trees. The pair or groups of homologues with high DNA sequence identity (≥95%, mostly ≥97%) are indicated in red. The Wzx homologues of O96 and O170, which are indicated in blue and by an asterisk, showed 86% DNA sequence identity, but in all other proteins showed low-sequence homologies to each other (<70% identity). Note that while the DNA sequence identity between the *wzx*_O46 and *wzx*_O134 in Gp6 is 99.7%, the *wzx*_O46 has a 2-bp deletion at the 3′-region, causing a frame shift.

strains were also identified in phylogroups A (ST34 and ST57) and B1 (ST300) (Fig. 3). The phylogenetic analysis also showed that the types of sugar synthesis gene sets and processing gene sets (*wzx*/*wzy* and *wzm*/*wzt*) were not limited to a specific lineage (Fig. 3).

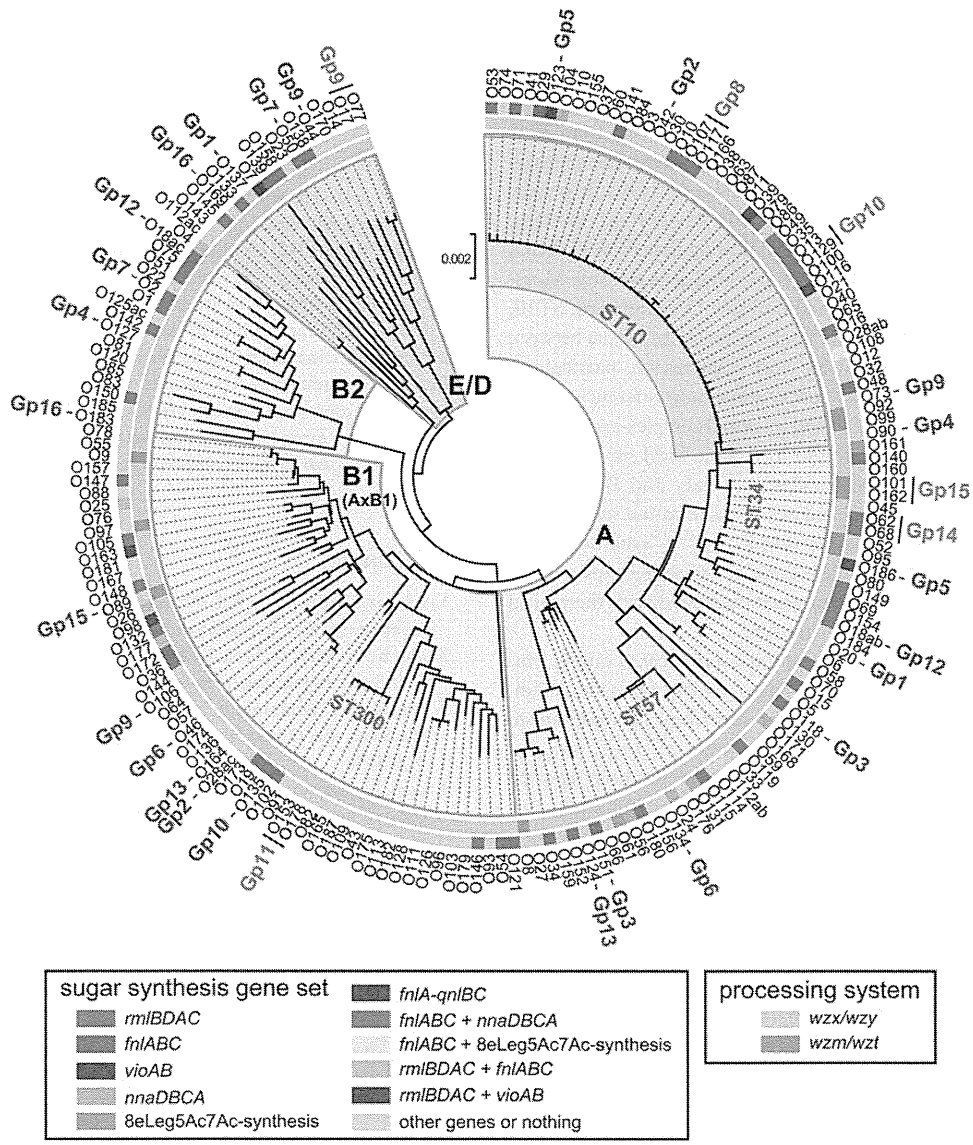

Figure 3. Correlation between the *Escherichia coli* evolutionary lineages and the distribution of O-AGCs. The phylogenetic tree was constructed based on the concatenated sequences of seven housekeeping genes from all 184 *E. coli* O-serogroup reference strains. The group names of O-AGCs (Gp1–Gp16) are indicated in the outermost region. Members in groups indicated in green were found to belong to the same or very closely related lineage, whereas members of the groups indicated in blue were found in distinct lineages. The outer circle next to the O serogroup names indicates the distribution of sugar synthesis gene sets identified in each O-AGC. The inner circle indicates the type of O-antigen processing system (*wzx/wzy* or *wzm/wzt*). Phylogenetic groups (A, B1, B2, D, and E) were determined by comparing the sequences of the strains tested with the known sequences from the ECOR collection (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli).

3.5. Relationships of the *E. coli* and *Shigella* O-AGCs

Shigella and *E. coli* belong to the same species complex[24] and many *Shigella* O antigens are known to be serologically and genetically identical or very similar to some *E. coli* O antigens, as summarized by Liu *et al.*[25] In addition to the 21 previously shown relationships, we found two additional O-AGC groups shared by *E. coli* and *Shigella*; O38 and *Shigella dysenteria* type 8 (SD8), and O169/O183 and *Shigella boydii* type 6/10 (SB6/SB10) (Supplementary Fig. S4). The O183-AGC was highly similar to the *S. boydii* types 10 cluster (sharing 98.2% DNA sequence identity). In our previous study,[26] we provisionally named a novel O serogroup for a group of Shiga toxin-producing *E. coli* strains as OSB10, which cross-reacted with *S. boydii* type 10. Sequence comparisons in this study revealed that OSB10 is not only serologically but also genetically identical to the new serogroup O183 of Gp16.

4. Discussion

Much of what we know about *E. coli* is defined at some level by O serogroups. To link genomic information to the wealth of data held in public databases, in our collective knowledge, outbreak, and disease reports and elsewhere, we endeavoured to determine whether molecular O-serogroup identification, targeting O-serogroup-specific genes (or unique sequences), was a valuable method to capture this information and maintain this important link. Not only do we show evidence supporting the effectiveness of molecular O-typing, but also we open

up the possibility of generating a molecular O-typing scheme and relate O serogroups to the underlying phylogeny of this bacterium.

By determining and comparing the sequences of O-AGCs from all known *E. coli* O serogroups, we newly defined the sequence and gene content of 145 unique O-AGCs and showed that O-AGCs from 37 O serogroups could be placed into 16 groups based on members in each group sharing nearly identical or highly similar O-AGCs. It is clear from these data that many of the grouped O-AGCs (Gp1-16) were found in distinct phylogenetic lineages indicating that these O-AGCs have been spread across this species by horizontal gene transfer. Moreover, several lineages that contained multiple O serogroups, ST10, ST34, ST57, and ST300, show that frequent exchange occurs between and within lineages. ST10 and its close relatives are particularly interesting as one-quarter of *E. coli* O-serogroup reference strains fell within this clonal group. ST10 and its clonal complex are clinically very important being recently found to include ESBL-producing *E. coli* from human and animals in Spain,[25] Italy and Denmark,[26] China,[27] and the Netherlands,[28] and in various intra-intestinal pathotypes of *E. coli*, such as enteroaggregative *E. coli*,[27,28] enterotoxigenic *E. coli*,[29,30] and EHEC.[31,32] In most cases, the O serogroups of these ST10 or ST10-related strains are unusual compared with the typical O serogroups that represent that pathotype.

Acquisition of O-antigen modification genes located on the genomes of serotype-converting bacteriophages or plasmids is also an important strategy for diversifying O-antigen structures. This mechanism has been well investigated in *Shigella flexneri*.[33,34] In *E. coli*, the O-serogroup conversion by a prophage-like element has been reported for O17 and O44,[17] which belong to Gp9 defined in this study. Another possible mechanism to generate the variation of O antigens is the mutations in the genes of the O-AGC as observed for O107 and O117,[16] which belong to Gp8. In this case, point mutations in a glycosyltransferase gene are responsible for the alteration of O-antigen structure (and thus that of O serogroup).[16] Five O-AGC groups including Gp2, Gp5, Gp7, Gp12, and Gp13 also contained differences in the amino acid sequence of their glycosyltransferases. O serogroup differences in these groups may be generated by the point mutations in glycosyltransferase genes. On the other hand, all glycosyltransferase genes in Gp1, Gp3, Gp4, Gp6, and Gp11; four strains from Gp9 (O17, O44, O73, and O77) and two from Gp10 (O13 and O129) showed 100% amino acid sequence identity. These results suggest that the serological differences between the members of these seven groups have been generated by acquisition of modification genes outside of the O-AGC as shown for O17 and O44 of Gp9.[17]

We believe that the remarkable sequence diversity observed in the *wzx/wzy* and *wzm/wzt* O-AGC genes of all known *E. coli* O serogroups appears to be sufficiently discriminative from one another to make identification of each of the known O serogroups possible. Therefore, our sequence data will serve as a valuable resource for the development of rationally designed molecular methods for O-typing as well as for detecting novel O serogroups.

In conclusion, our study provides a complete sequence set of O-AGCs of all known *E. coli* O serogroups and thus offers a full view on the genetic diversity of O-AGCs of this bacterium. In addition, the results presented suggest that horizontal gene transfer has been involved in the O serogroup diversification in *E. coli* more frequently and in a more biased or lineage-dependent fashion than previously thought.

Acknowledgements

We thank A. Akiyoshi, Y. Kato, and A. Yoshida for technical assistance.

Supplementary data

Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by Health Labor Sciences Research Grants from the Ministry of Health, Labor, and Welfare, Japan to A.I. (H25-Syokuhin-Wakate-018) and M.O. (H24-Shinkou-Ippan-012); Adaptable and Seamless Technology Transfer Program through Target-driven R&D (AS242Z00217P) from Japan Science and Technology Agency to A.I.; and a Scientific Research Grant on Priority Areas from the University of Miyazaki and the Program to Disseminate Tenure Tracking System from the Japanese Ministry of Education, Culture, Sports, Science, and Technology to A.I. (http://www.miyazaki-u.ac.jp/ir/english/index.html). This work was also supported by Wellcome Trust grant (098051). Funding to pay the Open Access publication charges for this article was provided by the University of Miyazaki, Japan.

References

1. Bazaka, K., Crawford, R.J., Nazarenko, E.L. and Ivanova, E.P. 2011, Bacterial extracellular polysaccharides, *Adv. Exp. Med. Biol.*, 715, 213–26.
2. Liu, B., Knirel, Y.A., Feng, L., et al. 2013, Structural diversity in *Salmonella* O antigens and its genetic basis, *FEMS Microbiol. Rev.*, 38, 56–89.
3. Stenutz, R., Weintraub, A. and Widmalm, G. 2006, The structures of *Escherichia coli* O-polysaccharide antigens, *FEMS Microbiol. Rev.*, 30, 382–403.
4. Lam, J.S., Taylor, V.L., Islam, S.T., Hao, Y. and Kocincova, D. 2011, Genetic and functional diversity of *Pseudomonas aeruginosa* lipopolysaccharide, *Front Microbiol.*, 2, 118.
5. Penner, J.L. and Aspinall, G.O. 1997, Diversity of lipopolysaccharide structures in *Campylobacter jejuni*, *J. Infect. Dis.*, 176 (Suppl. 2), S135–138.
6. Armstrong, G.L., Hollingsworth, J. and Morris, J.G. Jr. 1996, Emerging foodborne pathogens: *Escherichia coli* O157:H7 as a model of entry of a new pathogen into the food supply of the developed world, *Epidemiol Rev.*, 18, 29–51.
7. Tarr, P.I., Gordon, C.A. and Chandler, W.L. 2005, Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome, *Lancet*, 365, 1073–86.
8. Johnson, K.E., Thorpe, C.M. and Sears, C.L. 2006, The emerging clinical importance of non-O157 Shiga toxin-producing *Escherichia coli*, *Clin. Infect. Dis.*, 43, 1587–95.
9. Buchholz, U., Bernard, H., Werber, D., et al. 2011, German outbreak of *Escherichia coli* O104:H4 associated with sprouts, *N. Engl. J. Med.*, 365, 1763–70.
10. Peirano, G. and Pitout, J.D. 2010, Molecular epidemiology of *Escherichia coli* producing CTX-M beta-lactamases: the worldwide emergence of clone ST131 O25:H4, *Int. J. Antimicrob. Agents*, 35, 316–21.
11. Samuel, G. and Reeves, P. 2003, Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly, *Carbohydr. Res.*, 338, 2503–19.
12. DebRoy, C., Roberts, E. and Fratamico, P.M. 2011, Detection of O antigens in *Escherichia coli*, *Anim. Health Res. Rev.*, 12, 169–85.
13. Leopold, S.R., Magrini, V., Holt, N.J., et al. 2009, A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis, *Proc. Natl Acad. Sci. USA*, 106, 8713–8.
14. Iguchi, A., Shirai, H., Seto, K., et al. 2011, Wide distribution of O157-antigen biosynthesis gene clusters in *Escherichia coli*, *PLoS ONE*, 6, e23250.
15. Iguchi, A., Iyoda, S. and Ohnishi, M. 2012, Molecular characterization reveals three distinct clonal groups among clinical Shiga toxin-producing *Escherichia coli* strains of serogroup O103, *J. Clin. Microbiol.*, 50, 2894–900.
16. Wang, Q., Perepelov, A.V., Wen, L., et al. 2012, Identification of the two glycosyltransferase genes responsible for the difference between *Escherichia coli* O107 and O117 O-antigens, *Glycobiology*, 22, 281–7.

17. Wang, W., Perepelov, A.V., Feng, L., et al. 2007, A group of *Escherichia coli* and *Salmonella* enterica O antigens sharing a common backbone structure, *Microbiology*, **153**, 2159–67.
18. Lacher, D.W., Gangiredla, J., Jackson, S.A., Elkins, C.A. and Feng, P.C. 2014, Novel microarray design for molecular serotyping of Shiga toxin-producing *Escherichia coli* isolated from fresh produce, *Appl. Environ. Microbiol.*, 80, 4677–82.
19. Tzschoppe, M., Martin, A. and Beutin, L. 2012, A rapid procedure for the detection and isolation of enterohaemorrhagic *Escherichia coli* (EHEC) serogroup O26, O103, O111, O118, O121, O145 and O157 strains and the aggregative EHEC O104:H4 strain from ready-to-eat vegetables, *Int. J. Food Microbiol.*, **152**, 19–30.
20. Wang, Q., Ruan, X., Wei, D., et al. 2010, Development of a serogroup-specific multiplex PCR assay to detect a set of *Escherichia coli* serogroups based on the identification of their O-antigen gene clusters, *Mol. Cell Probes*, **24**, 286–90.
21. Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, **22**, 4673–80.
22. Tamura, K., Dudley, J., Nei, M. and Kumar, S. 2007, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0, *Mol. Biol. Evol.*, **24**, 1596–9.
23. Jensen, S.O. and Reeves, P.R. 2004, Deletion of the *Escherichia coli* O14:K7 O antigen gene cluster, *Can. J. Microbiol.*, **50**, 299–302.
24. Pupo, G.M., Lan, R. and Reeves, P.R. 2000, Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics, *Proc. Natl Acad. Sci. USA*, **97**, 10567–72.
25. Liu, B., Knirel, Y.A., Feng, L., et al. 2008, Structure and genetics of *Shigella* O antigens, *FEMS Microbiol. Rev.*, **32**, 627–53.
26. Iguchi, A., Iyoda, S., Seto, K. and Ohnishi, M. 2011, Emergence of a novel Shiga toxin-producing *Escherichia coli* O serogroup cross-reacting with *Shigella boydii* type 10, *J. Clin. Microbiol.*, **49**, 3678–80.
27. Olesen, B., Scheutz, F., Andersen, R.L., et al. 2012, Enteroaggregative *Escherichia coli* O78:H10, the cause of an outbreak of urinary tract infection, *J. Clin. Microbiol.*, **50**, 3703–11.
28. Okeke, I.N., Wallace-Gadsden, F., Simons, H.R., et al. 2010, Multi-locus sequence typing of enteroaggregative *Escherichia coli* isolates from Nigerian children uncovers multiple lineages, *PLoS ONE*, **5**, e14093.
29. Turner, S.M., Chaudhuri, R.R., Jiang, Z.D., et al. 2006, Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages, *J. Clin. Microbiol.*, **44**, 4528–36.
30. Nada, R.A., Shaheen, H.I., Khalil, S.B., et al. 2011, Discovery and phylogenetic analysis of novel members of class b enterotoxigenic *Escherichia coli* adhesive fimbriae, *J. Clin. Microbiol.*, **49**, 1403–10.
31. Monaghan, A.M., Byrne, B., McDowell, D., Carroll, A.M., McNamara, E. B. and Bolton, D.J. 2012, Characterization of farm, food, and clinical Shiga toxin-producing *Escherichia coli* (STEC) O113, *Foodborne Pathog. Dis.*, **9**, 1088–96.
32. Hauser, E., Mellmann, A., Semmler, T., et al. 2013, Phylogenetic and molecular analysis of food-borne shiga toxin-producing *Escherichia coli*, *Appl. Environ. Microbiol.*, **79**, 2731–40.
33. Allison, G.E. and Verma, N.K. 2000, Serotype-converting bacteriophages and O-antigen modification in *Shigella flexneri*, *Trends Microbiol.*, 8, 17–23.
34. Sun, Q., Knirel, Y.A., Lan, R., et al. 2012, A novel plasmid-encoded serotype conversion mechanism through addition of phosphoethanolamine to the O-antigen of *Shigella flexneri*, *PLoS ONE*, **7**, e46095.