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

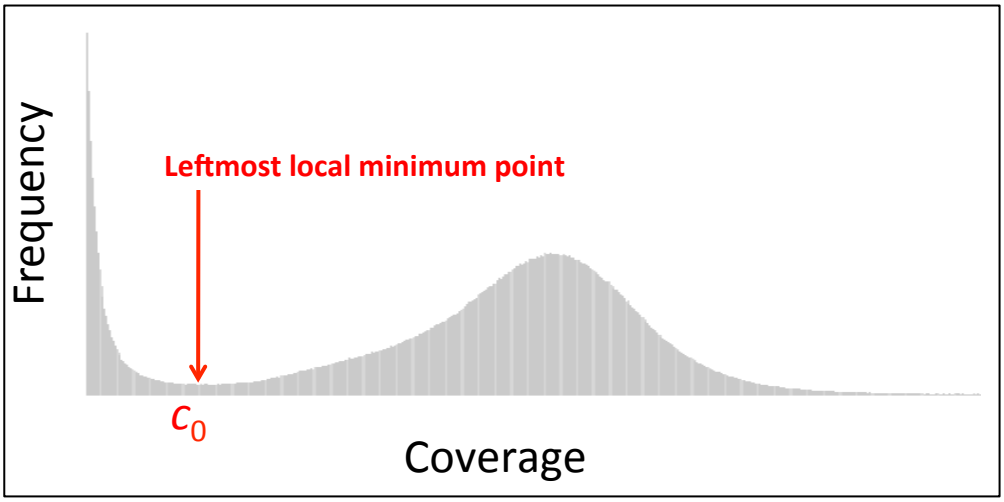
Supplemental Materials

Supplemental Methods

Algorithm Detail

k-mer coverage distribution (Contig-assembly)

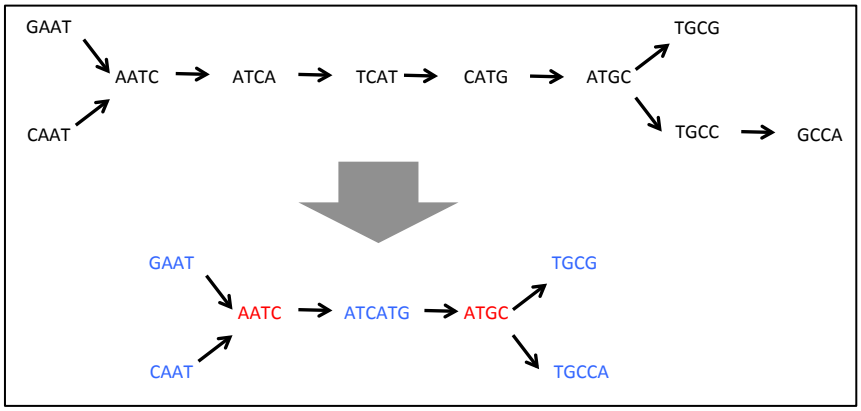
Initially, *k*-mers (default, initial $k = 32$) in reads are counted, excluding k_0-mers that contain non-ATGC characters. Counting is performed using a hash table and temporary files. Here, a k_0-mer and its reverse complement are not distinguished. The size of the hash table is fixed and limited memory space is required until the frequency distribution of *k*-mer coverage (the number of occurrences) is completed. For the *k*-mer frequency distribution, the coverage that corresponds to the leftmost local minimum frequency is detected using a size:7 smoothing window (Supplemental Figure 1). This coverage value is used as the threshold c_0.



Supplemental Figure 1. *k*-mer frequency distribution and c_0

De Bruijn graph construction (Contig-assembly)

In initial de Bruijn graph construction, *k*-mers whose coverage value is c_0 or more correspond to nodes. Two *k*-mers that have (k – 1) overlap are connected by an edge. Note that each node represents one of the strands of the *k*-mer, and edges have directionality. The nodes whose outdegrees or indegrees are greater than 1 are marked as "junction nodes," while the remaining nodes are marked as "straight nodes." In order to compress the graph, adjacent straight nodes are combined (Supplemental Figure 2) and straight nodes possess multiple *k*-mers as strings. In fact, construction and compression of the graph are performed simultaneously. In this case, nodes hold values of length and coverage. Lengths of nodes are the length of strings composed of the contained *k*-mers (i.e., number of *k*-mers + *k* - 1). Coverage of nodes denotes an average coverage of the *k*-mers included. In the following section, we express sets of straight nodes, junction nodes, and edges as *S*, *J,* and *E*, respectively. For a node v, in(v), out(v), cov(v), and $|v|$ represent the indegree, outdegree, coverage, and length of v, respectively. After the initial de Bruijn graph is constructed, average coverage is calculated from nodes (*k*-mers) that are not in a bubble. A bubble is defined as a set of two straight nodes and two junction nodes, for which straight nodes are connected to the same junction in both directions. Consequently, average coverage corresponds to the *k*-mer coverage of homozygous regions. This value is used in subsequent steps.



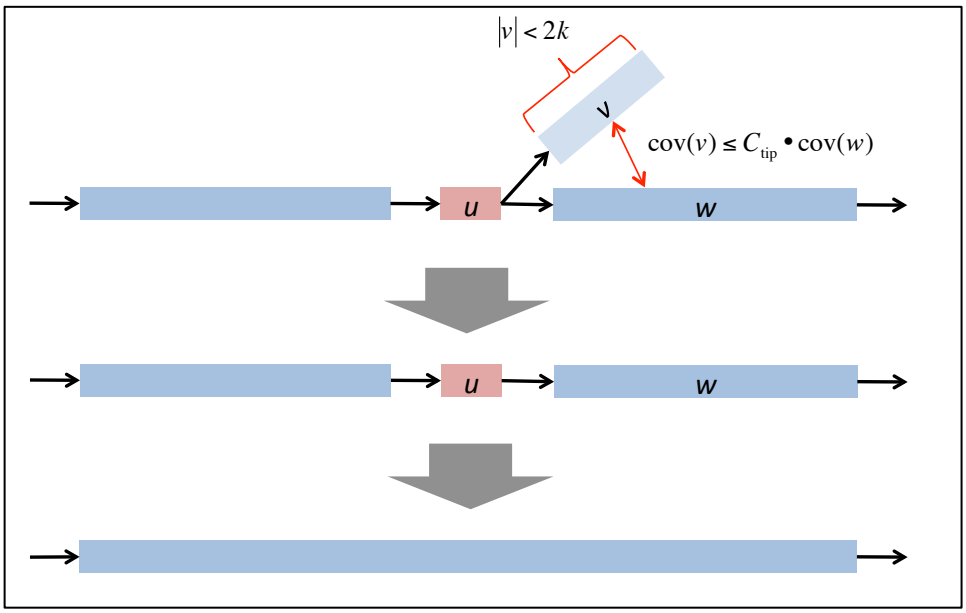
Supplemental Figure 2. Example of de Bruijn graph (*k* = 4)

Strings and arrows represent nodes and directed edges, respectively. In the upper graph, *k*-mers correspond to the node on a one-to-one basis. In the lower graph, red and blue strings represent the junction and straight nodes, respectively. In this case, straight nodes are compressed. Lengths of nodes are defined as the lengths of the corresponding strings.

Tip removal in de Bruijn graph (Contig-assembly)

To remove erroneous nodes, Platanus deletes dead-end nodes that are short and show relatively low coverage, referred to as "tips." The variables v and u represent straight and junction nodes, respectively, with v connected to u. W is the set of nodes that are connected to u in the same directions as v. C_{tip} is a constant value ranging from 0 to 1 (default: $C_{\text{tip}} = 0.5$). v is deleted if the following three conditions are satisfied (Supplemental Figure 3):

$$|v| < 2k$$
$$out(v) + in(v) = 1$$
$$cov(v) \leq C_{\text{tip}} \cdot \max_{w \in W} \text{cov}(w)$$



Supplemental Figure 3. Tip removal in contig assembly

Reconstruction of de Bruijn graph (Contig-assembly)

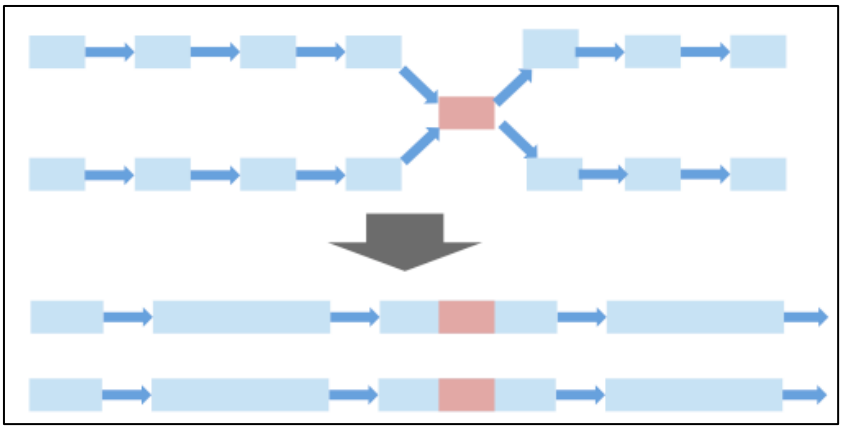
To resolve repeats that are shorter than reads, Platanus increases k and reconstructs the de Bruijn graph (Supplemental Figure 4). In de Bruijn graph-based assemblies, if k is large, repeats shorter than k can generally be resolved, but low-coverage sequences cannot be captured, a rationale to utilize the advantages of each k. k_0 and k_{step}, the step size of k, are constant values (default, k_0 = 32 and k_{step} = 10). Maximum k, k_{max}, is calculated according to the average k_0-mer coverage value and read length (described below). k_{pre} is previous k in certain reconstruction step. Roughly speaking, a k-mer graph consists of straight nodes of k_{pre}-mer graph and reads that are mapped near the junction nodes of k_{pre}-mer graph. The detailed procedure to reconstruct the k-mer graph from k_{pre}-mer graph consists of the following steps:

(1) The strings of straight nodes in k_{pre}-mer graph are extended in both directions (Supplemental Figure 5). The extensions are continued until paths are forked or extra lengths are reached ($k - k_{\mathrm{pre}}$).
(2) k-mers in extended strings are extracted. Coverage of k-mers from a straight node v is weighed as below:

$$\mathrm{cov}(v) \cdot \frac{r - k + 1}{r - k_{\mathrm{pre}} + 1}$$

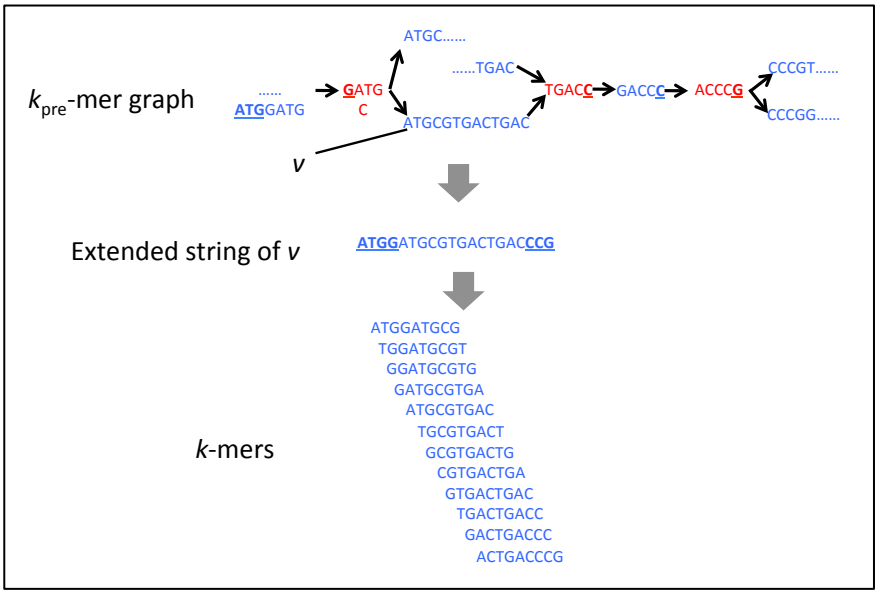
where r denotes average read length.
(3) For the k_{pre}-mer graph, k_{pre}-mers within distance of ($k - k_{\mathrm{pre}}$) from junctions are marked (Supplemental Figure 6). Marked k_{pre}-mers are defined as "junction-neighboring k_{pre}-mers."
(4) Reads that contain junction-neighboring k_{pre}-mers are collected, and k-mers within are counted.
(5) Two tables of k-mers made in (2) and (4) are merged. k-mers whose coverage is less than the threshold value c_i (described below) are discarded.

Using this method, connectivity information in the k_{pre}-mer graph is transferred to the k-mer graph.



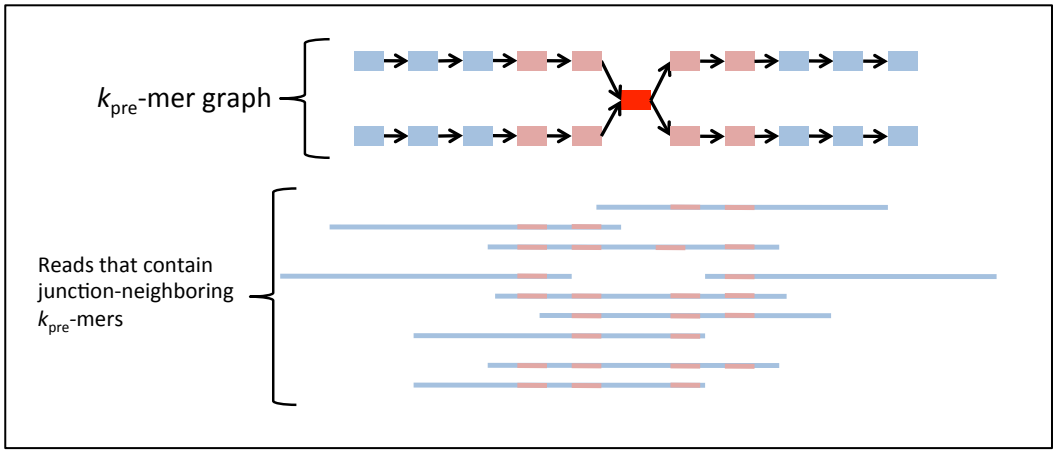
Supplemental Figure 4. *k*-mer extension and solution of repeat sequence

Blue boxes represent *k*-mers. Red boxes represent repetitive sequences.



Supplemental Figure 5. Conversion of a straight node in graph reconstruction

Conversion of node v is displayed ($k_{pre} = 5$, $k = 9$) as an example. The extension of v is continued until paths are forked or the extended length reaches ($k - k_{pre}$). Resulting k-mers are used for the construction of k-mer graph.



Supplemental Figure 6. Extraction of reads in graph reconstruction

$k - k_{pre} = 2$. Blue boxes: k_i-mers in straight nodes. Deep red box: junction node. Light red boxes: k_i-mers within distance of $(k - k_{pre})$ from junction (junction-neighboring k_i-mers).

Reads that contain junction-neighboring k_{pre}-mers are collected, and the k-mers within are counted. Colors in the reads correspond to those of the k_{pre}-mer graph.

Calculation of c_i and k_{max} (Contig-assembly)

Misassemblies are generally caused by the presence of repetitive regions (Supplemental Figure 7). A higher minimum coverage threshold causes a greater number of misassemblies. In addition, an increase in *k* is associated with a decrease in average *k*-mer coverage due to a decline in both the total number of *k*-mers and common *k*-mers among multiple reads. To reduce the rate of misassembly, Platanus calculates the maximum *k* ($k_{\max}$) and minimum coverage values according to the average k_0-mer coverage and average read length. For each graph reconstruction, k_i, c_i, and a_i represent the *k*-mer length, minimum coverage, and estimated average coverage, respectively. The k_0 variable is a constant. Determination of c_0 is described above (Supplemental Methods, *k*-mer coverage distribution). The a_0 variable denotes the average coverage of k_0-mers ($\geq c_0$). The score $s_{\text{split}}(k, k_{\text{pre}}, a, c)$, associated with the probability of splitting correct edges, is defined as follows:

$$s_{\text{split}}(k, k_{\text{pre}}, a, c) = 1 - (1 - e^{-a} \sum_{j=0}^{c-1} \frac{a^j}{j!})^{k-k_{\text{pre}}}$$

where *k* denotes *k*-mer size, k_{pre} denotes *k* of previous graph, *a* denotes average coverage, *c* denotes minimum coverage, and it is assumed that *k*-mer coverage obeys a Poisson distribution. If the score $s_{\text{split}}(k, k_{\text{pre}}, a, c)$ is high, probability of misassembly is high. In addition, *a* is determined as follows:

$$a = a_{\text{pre}} \frac{r - k + 1}{r - k_{\text{pre}} + 1}$$

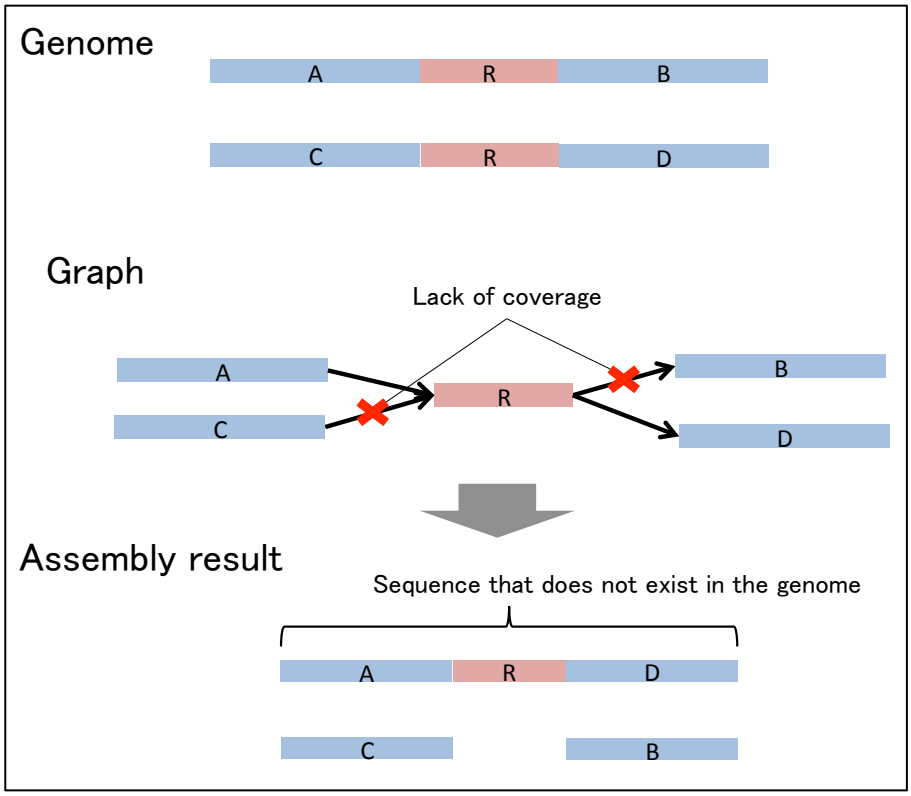
where *r* denotes the average read length.

Two constant values, $c_{\min}$ and $s_{\max}$ (default: $c_{\min} = 2$, $s_{\max} = 10^{-10}$), are introduced herein. $c_{\max}$ is determined as the maximum *c* that satisfies the following conditions:

(1) $c_{\min} \leq c \leq c_{\max}$

(2) $s_{\text{split}}(k, k_{\text{pre}}, a, c) \leq s_{\max}$

where $k = k_{\text{pre}} + k_{\text{step}}$ until *c* satisfies the two existing conditions. If *c* is not found, an acceptable *k* is sought by increasing the k_{pre} value by 1, and $k_{\max} = k$ In short, Platanus determine *k*, thereby satisfying the criteria of the probability of misassembly.



Supplemental Figure 7. Misassembly through repetitive sequence

R (red box) represents a repetitive sequence.

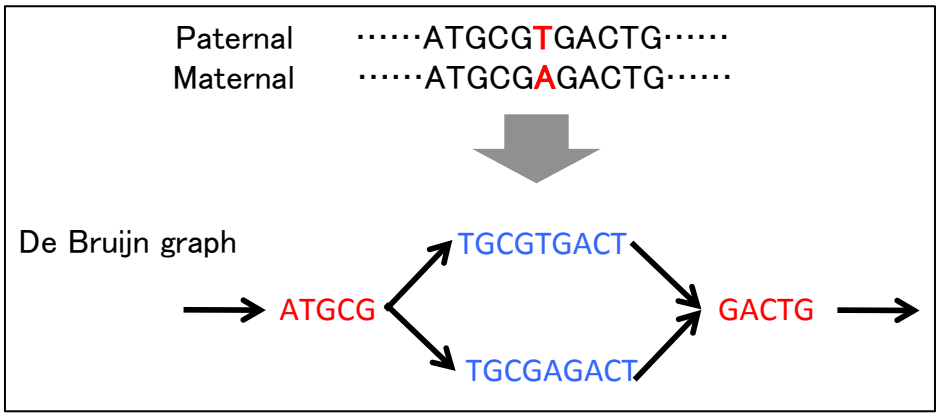
Bubble removal in Contig-assembly

After reconstruction and tip removal of the k_{max}-mer graph, "bubbles" in the graph are removed. Bubble structures are caused by heterozygosity of diploid samples and errors (Supplemental Figure 8). A bubble is defined as a set of two straight nodes and two junction nodes, for which straight nodes are connected to the same junction bidirectionally. In deciding whether to remove a bubble, the coverage and edit distance of two straight nodes are examined. We define edit(v, u) as the edit distance between two straight nodes, v and u, and introduce the constant value C_{bubble} (default, 0.1). A straight node that has lower coverage is deleted if a bubble including straight nodes v and u satisfies the following two conditions:

(1) $\mathrm{cov}(v) + \mathrm{cov}(u) < 1.5a$ (a: average k_{max}-mer coverage)

(2) $\mathrm{edit}(v, u) \leq C_{\mathrm{bubble}} \cdot \max(|v|, |u|)$

Strings of removed nodes are saved for scaffolding and analysis. After bubble removal, strings of the remaining straight nodes are generated as contig outputs.

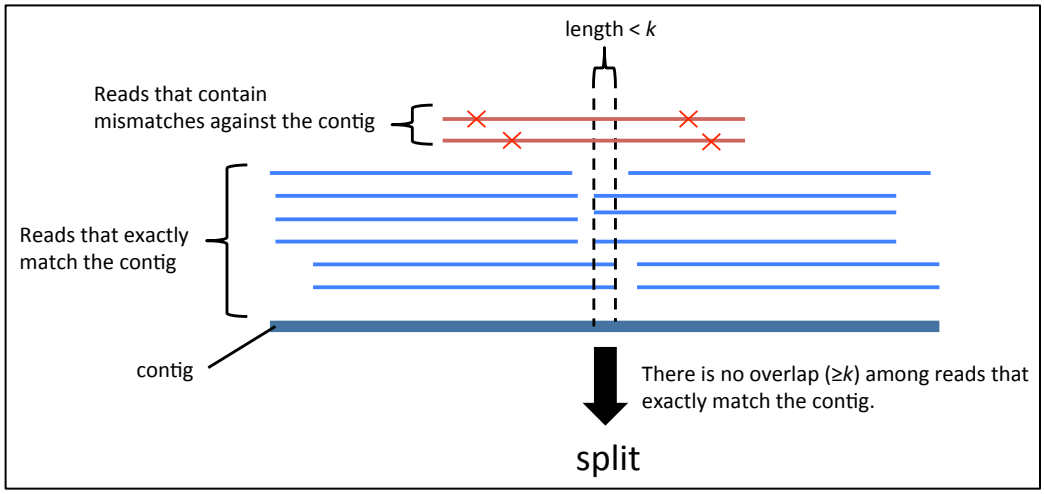


Supplemental Figure 8. Bubble in de Bruijn graph

k = 5. A bubble structure caused by a single nucleotide polymorphism (SNP) is provided as an example.

Check of contigs using exact-match reads (Contig-assembly)

After the removal of bubbles, reads are mapped on the resulting contigs and Platanus attempts to detect misassemblies. Here, reads that exactly match contigs are used. After mapping, a contig is split if it is situated in the inside position where there exists no overlap (length $>k_0$) among mapped reads (Supplemental Figure 9).



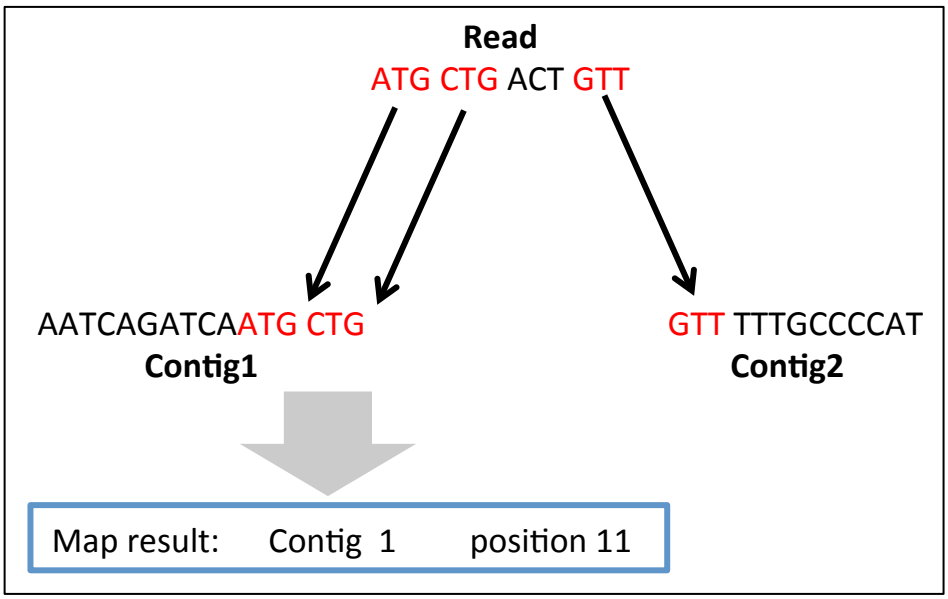
Supplemental Figure 9. Check of contig using exact-match reads

Mapping of reads in scaffolding

In scaffolding, Platanus initially maps all paired reads on contigs. The mapping results are used for the estimation of library insert sizes and for linking contigs. Before mapping, a hash table is made. For this table, keys and values correspond to all *k*-mers (default, $k = 32$) in contigs and positions. A read is mapped in the following manner (Supplemental Figure 10):

(1) *k*-mers in each read are collected without overlaps.
(2) These *k*-mers are queried in the hash table. If a *k*-mer uniquely exists in contigs, it is mapped to the corresponding position.
(3) The contig is determined according to majority rule, using map results of the *k*-mers. If the top number is a tie score, the read is treated as a non-hit.
(4) The position is determined as an average of inferred positions of the mapped *k*-mers.

If one of the paired reads is not mapped, the pair is excluded in the subsequent step.

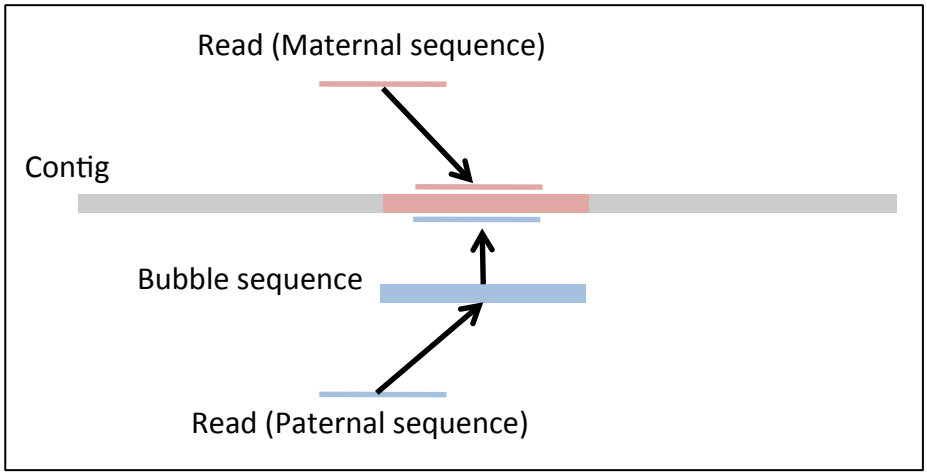


Supplemental Figure 10. Mapping of a read

The *k*-mer size for mapping is 3 in this figure. From the read, 4 *k*-mers are taken. Two are mapped on contig1, 1 is mapped on contig2, and another remains unmapped. According to the majority rule, the read is mapped on contig1.

Mapping of contig-bubbles (Scaffolding)

For diploid samples, the resulting sequences from Contig-assembly are mosaics of homologous chromosomes. Any number of scenarios may occur within one contig, including the observation that some regions represent the paternal chromosome and others represent the maternal chromosome. To map reads from the heterozygous region, Platanus integrates contigs and sequences of bubbles that are deleted in Contig-assembly (Supplemental methods, *Bubble removal in Contig-assembly*). Sequences of bubbles are mapped on contigs using exact matches of edges. The relation of contig and bubble sequences is saved. Reads that are not mapped on contigs are re-queried to bubble sequences. In reads that are mapped on bubble sequences, mapped positions are converted into their corresponding contig positions (Supplemental Figure 11).



Supplemental Figure 11. Mapping of reads from the heterozygous region

Estimation of insert sizes (Scaffolding)

For each library, the insert size is estimated using paired reads that are mapped on the same contig (scaffold). In scaffolding, libraries are used according to the order of input. For longer-insert libraries, insert sizes are estimated using scaffolds constructed by shorter-insert libraries. If both paired reads are mapped onto the same strand or the inferred insert size is a negative value, that pair is not used.

To estimate the insert size, the peak is first detected in the distribution of insert sizes using the window with a size of 101. Means and standard deviations of the insert size are calculated excluding the outliers (<0.5 × peak-insert-size or >1.5 × peak-insert-size).

Scaffold graph (Scaffolding)

When paired reads are mapped onto different contigs, we refer to the paired reads as "links." Contigs and links are converted into a graph structure. Contigs in these graphs correspond to nodes and nodes have length values. If the number of links between contigs (nodes) is above the threshold n, an edge connects the nodes. Each edge consists of two values: the number of links and the estimated distance (Supplemental Figure 12). Here, we describe the variables as follows.

r: Average read length
c: Average coverage (Total-length-of-mapped-reads/Total-length-of-contigs)
μ: Average insert size
σ: Standard deviation of insert size
l_1, l_2: Length of two contigs
g: Actual gap size between contigs

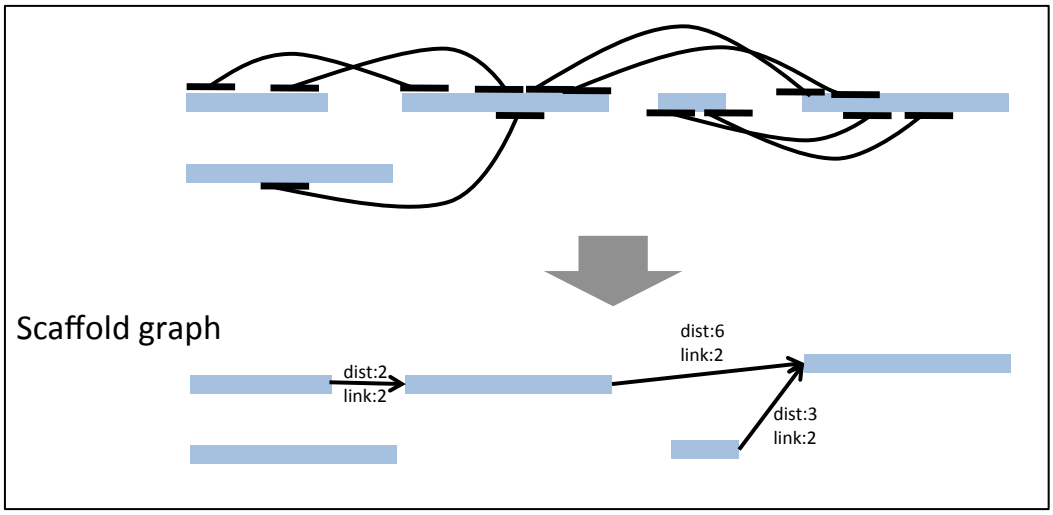
In the estimation of distances, all paired reads that link contigs are treated as their insert sizes equal to the average insert size of the library. For each pair of contigs, the maximum likelihood approach is used for the estimation of distances, assuming that the distribution of insert sizes follows a normal distribution (mean: μ, variance: σ^2). For each pair of reads, $s_i(d)$ (i = 1, 2, …, number-of-pairs) is defined as the insert sizes resulting from the distance (gap size) between contigs, d. Estimated distance is calculated as follows:

$$\underset{d}{\operatorname{argmax}} \prod_{i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(s_i(d)-\mu)^2}{2\sigma^2}\right)$$

Minimum numbers of links, n, are specified in two ways for each library. First, the expected value of the number links between two contigs is calculated. It is assumed that insert sizes obey the normal distribution and that coverage is uniform. The expected number of links is calculated in the following manner:

$$\int_{0}^{l_1-r} \int_{l_1+g+r-y}^{l_1+g+l_2-y} \frac{c}{2r} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dxdy$$

where n_{exp} is determined as the expected number for which $g = \mu - 2r$ and $l_1 = l_2 =$ average contig length. Second, n_{min} is defined as a constant value (default, 3). If $n_{min} < n_{exp}$, scaffolding is initially performed with $n = n_{exp}$ and is subsequently performed with $n = n_{min}$. In other instances, only n_{min} is applied. Note that n_{exp} is calculated for each library.

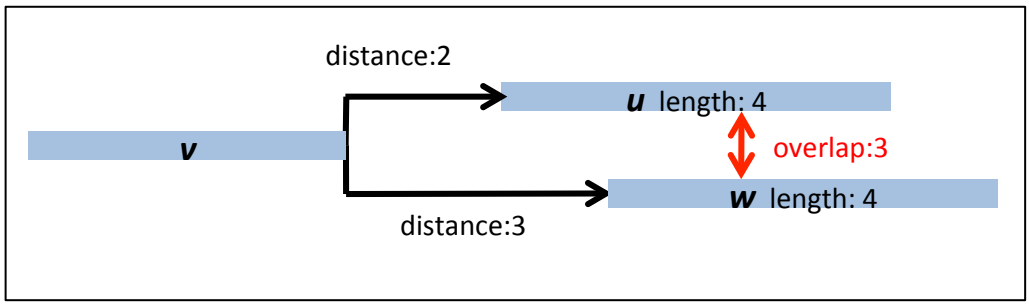


Supplemental Figure 12. Scaffold graph

Blue boxes are contigs (nodes) and black arrows are edges. In this case, the minimum number of links is 2. Each edge possesses 2 values: the number of links (link) and estimated distance (dist).

Conflict of nodes (Scaffolding)

As an example, consider node *v* connected to two nodes, *u* and *w*. Nodes *u* and *w* are positioned based on *v*, considering the distances between the nodes and length of nodes. If *u* and *w* overlap, and the overlap is longer than the threshold *o*, *u* conflicts with *w* (Supplemental Figure 13). *o* denotes a tolerance of overlaps and is applied in one of two ways for each library, as for *n* (denoting the minimum number of links). First, *o* is set to 2σ (σ is the standard deviation of insert size), and then *o* is set to 3σ.



Supplemental Figure 13. Overlap of nodes in the scaffold graph

An example of a conflict in the scaffold graph is shown. *u* and *w* are situated based on *v*. The overlap length is 3.

Construction of scaffolds (Scaffolding)

The variable V_{repeat} represents the set of nodes for which adjacent nodes conflict, and the variable V_{used} represents the set of nodes that are included in scaffolds in this step. The procedure to construct scaffolds from the graph is as follows:

(1) A selection of v_0 is made from a node that satisfies the following conditions:

i. $(v_0 \notin V_{\text{repeat}} \text{ and } v_0 \notin V_{\text{used}})$

ii. V_{scaffold} is initialized as $V_{\text{scaffold}} = \{v_0\}$

(2) An edge (u, w) is searched that satisfies the following conditions:

$$(u \in V_{\text{scaffold}} \text{ and } u \notin V_{\text{repeat}})$$

If w does not conflict with nodes in V_{scaffold}, w is added to V_{scaffold} and V_{used}. If multiple candidates exist for w, candidates whose distance (the number of edges) from v_0 is minimal are added to V_{scaffold} and V_{used}.

(3) Step (2) is repeated until the candidate nodes to be added to V_{scaffold} are lost.

(4) V_{scaffold} is saved as a scaffold.

(5) Steps (1)–(4) are repeated until the candidate nodes of v_0 are lost.

Using this procedure, the risk to connect contigs through repetitive contigs is reduced. In practice, the graph is pre-processed before construction of scaffolds (described below).

Reduction of conflicts (Scaffolding)

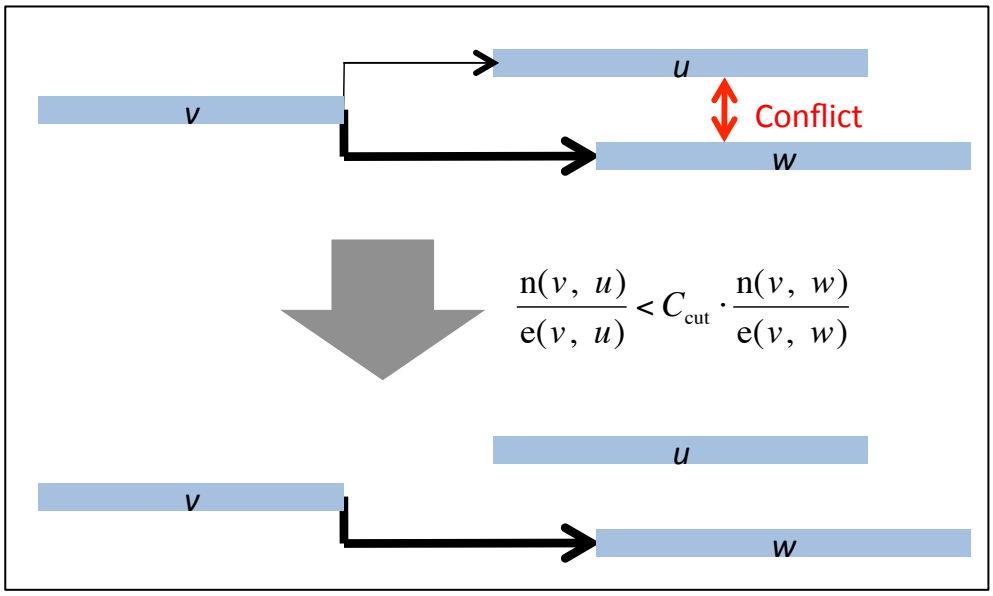
In order to reduce the number of contigs caused by errors or heterozygosity, Platanus begins searching the scaffold graph for pairs of nodes (contigs) that conflict and overlap. The overlaps are represented as negative values of distances between nodes. The node *v* is removed from the graph if the following conditions are satisfied between two nodes, *v* and *u*:

(1) *v* and *u* are conflicting (inferred from another node and edges; Supplemental Figure 13).
(2) Both the length and coverage of *v* are smaller than those of *u*.
(3) Estimated distance between *v* and *u* is smaller than $-o$.
[*o* is the tolerance of overlaps described above in the section *Conflict of nodes (Scaffolding)*]
(4) Coverage of *v* is smaller than 1.5 × average coverage.

Edge cut in scaffold graph (Scaffolding)

In this step, erroneous edges that cause conflicts are deleted. The variables *v*, *u*, and *w* represent the nodes of the scaffold graph where *u* and *w* conflict. We define the number of links in edge (*v*, *u*) as n(*v*, *u*), and the expected number of links in edge (*v*, *u*) as e(*v*, *u*). In the calculation of expected value, gap size (*g*) is substituted with estimated gap size. After introducing a constant value C_{cut} (default, 0.5), edge (*v*, *u*) is deleted if the following condition is satisfied (Supplemental Figure 14):

$$\frac{\mathrm{n}(v,u)}{\mathrm{e}(v,u)} < C_{\text{cut}} \cdot \frac{\mathrm{n}(v,w)}{\mathrm{e}(v,w)}$$



Supplemental Figure 14. Edge cut in the scaffold graph

An example of a conflict in the scaffold graph and the cutting edge is shown.

Bubble removal in scaffold graph (Scaffolding)

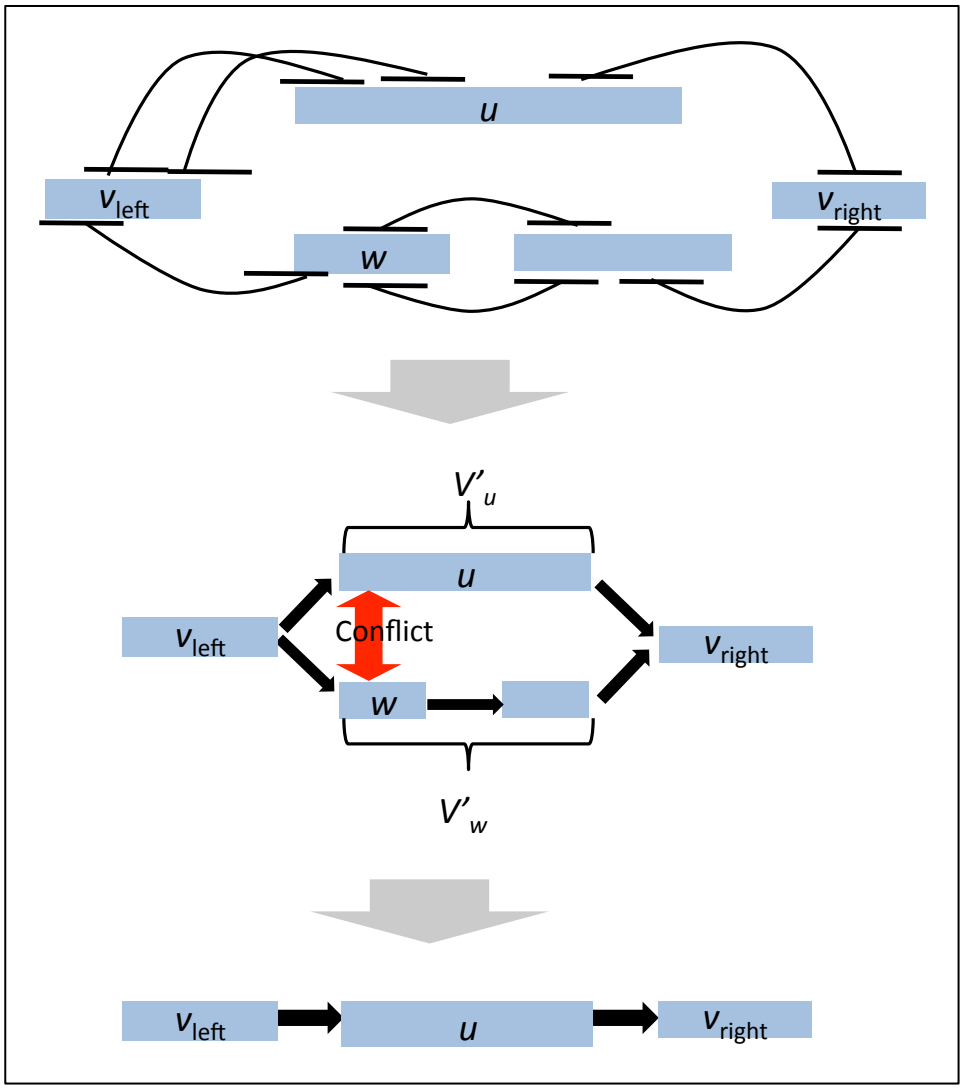
The variables u and w represent nodes that conflict with one another. In this case, scaffolds (set of nodes) V_u and V_w are locally constructed using u and w as starting points (v_0; see Supplemental Methods, *Construction of scaffolds*). If the leftmost and rightmost segments (nodes) of V_u and V_w are common, scaffolds V'_u and V'_w are constructed excluding the common nodes, v_{left} and v_{right}. $|V'|$, cov(V'), bub(V'), and edit(V'_u, V'_w) are defined as length of strings corresponding to V', average coverage of V', the number of contig-bubble mapped on V' (Supplemental Methods, *Mapping of contig-bubbles*) and edit distance between V'_u and V'_w, respectively. For strings corresponding to scaffolds, gaps are represented as 'N' characters. Coverage is estimated using the k-mer coverage of each node. Here, a denotes the average coverage. The variables C_{sim}, C_{homo}, and C_{hetero} are constant values (default, 0.1, 1.5, and 0.75, respectively). Nodes that constitute lower-coverage scaffold (V'_u or V'_w) are deleted if one of the following two conditions is satisfied (Supplemental Figure 15):

(1)

$$\text{cov}(V'_u) + \text{cov}(V'_w) \leq 2a \quad \text{and}$$
$$\text{edit}(V'_u, V'_w) \leq C_{\text{sim}} \cdot \max(|V'_u|, |V'_w|) \quad \text{and}$$
$$(\text{bub}(V'_u) = 0 \ \text{ or } \ \text{bub}(V'_w) = 0)$$

(2)

$$\text{cov}(V'_u) \leq C_{\text{hetero}} \cdot a \quad \text{and}$$
$$\text{cov}(V'_w) \leq C_{\text{hetero}} \cdot a \quad \text{and}$$
$$\text{cov}(v_{\text{right}}) \leq C_{\text{homo}} \cdot a \quad \text{and}$$
$$\text{cov}(v_{\text{left}}) \leq C_{\text{homo}} \cdot a \quad \text{and}$$
$$(\text{bub}(V'_u) = 0 \ \text{ or } \ \text{bub}(V'_w) = 0)$$



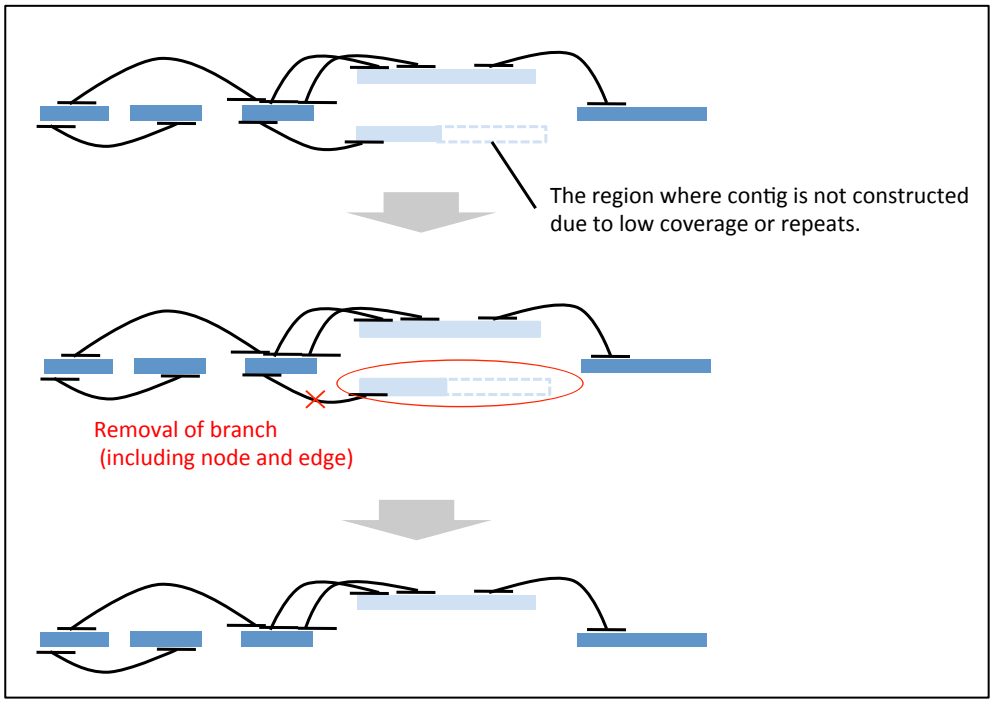
Supplemental Figure 15. Bubble removal in the scaffold graph

Nodes u and w are conflicting. V'_u has a higher coverage than the scaffolds (sets of nodes), and V'_w is deleted.

Branch-cut in the scaffold graph (Scaffolding)

The variables V_u and V_w represent nodes that conflict with one another, and both are connected to the common node (V_{root}). The definitions of a, $|V'|$, cov(V'), bub(V'), and edit(V'_u, V'_w) are the same as defined in the section *Bubble removal in scaffold graph*. Lower-coverage scaffolds (V'_u or V'_w) are deleted if the following condition is satisfied (Supplemental Figure 16):

$\text{cov}(V_{\text{root}}) \leq 1.5a$ and
$\max\left(\text{cov}(V_u), \text{cov}(V_w)\right) < 0.75a$ and
$\text{bub}(V_u) = 0$ and
$\text{bub}(V_w) = 0$



Supplemental Figure 16. Branch-cut in scaffolding

Filling gaps in scaffolding (Scaffolding)

After contigs (nodes) are ordered, Platanus examines whether adjacent contigs (*v* and *u*) possess common sequences.

If $o(v, u) \geq o_{min}$ and $d(v, u) - x(v, u) < o$, the gap between *v* and *u* is filled with the common sequence. o_{min} is the minimum overlap length (default, 32).
[*o*: overlap tolerance described above, in the section *Conflict of nodes (Scaffolding)*].
$x(v, u)$: length of exact match between ends of *v* and *u*.
$d(v, u)$: estimated distance between *v* and *u*.

Masking contig ends (Scaffolding)

Before generating the scaffold sequences, identities and overlaps between ends of adjacent contigs are examined. The variable o_{min} represents the minimum length threshold (default, 32). If the adjacent pair of contigs possesses similar ends (identity ≥ 0.95; length ≥ o_{min}) or the estimated distance has a negative value, both ends are masked with 'N's. The length of 'N's corresponds to the length of similar ends or overlaps. This step is intended to prevent the generation of false tandem repeats.

Check of scaffolds using long-insert libraries (Scaffolding)

After scaffolding is performed using one of the libraries, resulting scaffolds are checked using longer-insert libraries. The check is performed if the sum of physical coverage (insert-size × number-of-mapped-pairs / total-contig-length) of longer-insert libraries is greater than that of previously used shorter-insert libraries. For each of the gaps in scaffolds, Platanus judges whether there exists a misassembly and splits the scaffold if all of following conditions below (1–4) are satisfied (Supplemental Figure 17).

s : The number of paired reads spanning the gap.

s_{exp} : Expected number of paired reads spanning the gap.

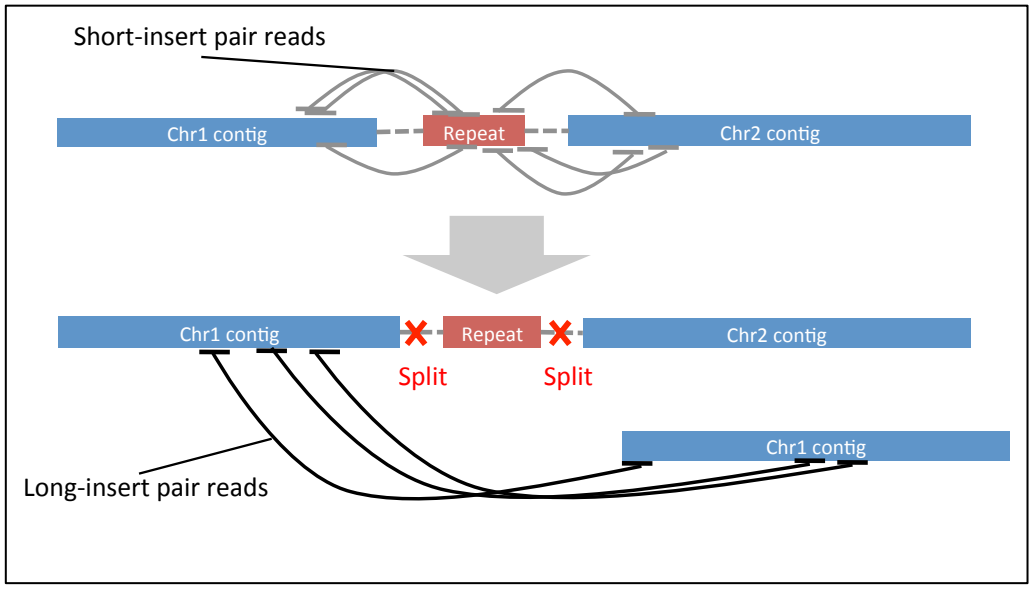
n_{min} : Constant value (default, 3).

(1) $s < n_{\mathrm{min}}$

(2) $s/s_{\mathrm{exp}} < 0.1$

(3) $s_{\mathrm{exp}} > 1$

(4) There are alternative links (number-of-paired-reads $\geq n_{\mathrm{min}}$) relating to contigs that are adjacent to the gap.



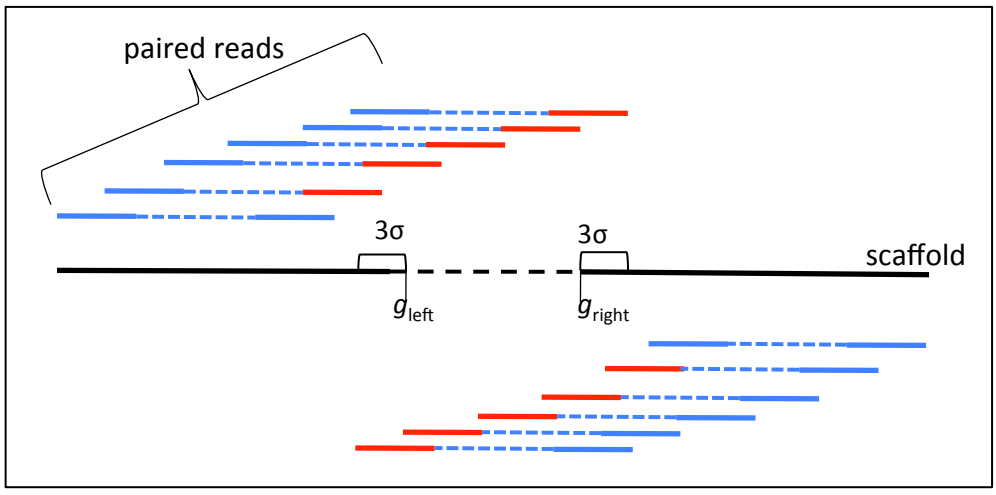
Supplemental Figure 17. Check of scaffolds using long-insert libraries

Collection of reads in gap-close (Gap-close)

The program "Gap-close" closes gaps in scaffolds using paired-end (mate-pair) libraries. First, reads are mapped into scaffolds in the same manner as scaffolding. Means and standard deviations of insert sizes are also calculated. When one of the paired reads is mapped, the position of the counterpart is estimated, assuming that the insert size equals the mean. A read is related to a gap if both of the following two conditions are satisfied (Supplemental Figure 18):

h_{left}: Estimated leftmost position of a read.
h_{right}: Estimated rightmost position of a read.
g_{left}: Leftmost position of a gap.
g_{right}: Rightmost position of a gap.
σ: Standard deviation of insert size of a library.

(1) $g_{\text{left}} - 3\sigma \leq h_{\text{right}}$
(2) $h_{\text{left}} \leq g_{\text{right}}$



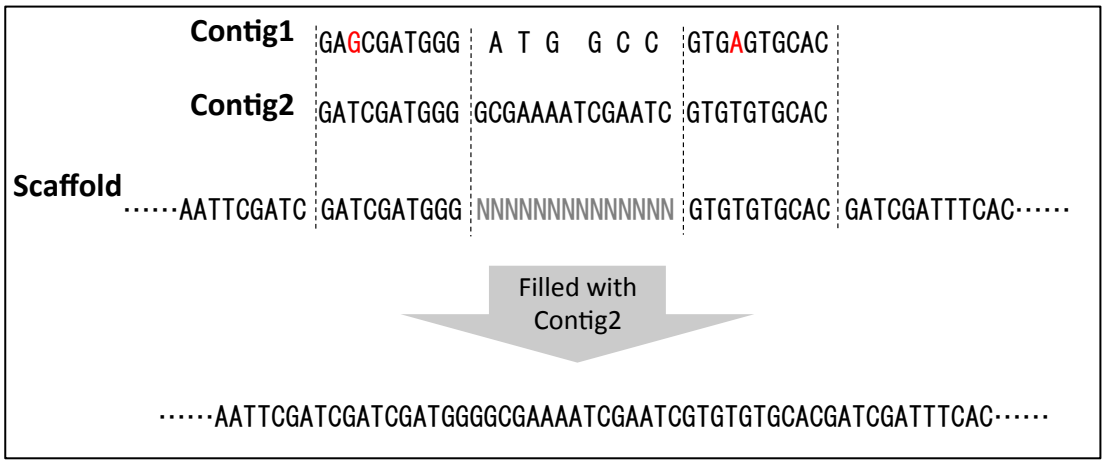
Supplemental Figure 18. Collection of gap-covering reads

Blue and red lines are paired reads. Black line is a scaffold. Dashed lines represent gaps. Red lines are related to a gap of the scaffold.

g_{left}: Leftmost position of a gap in a scaffold.
g_{right}: Rightmost position of a gap in a scaffold.
σ: Standard deviation of a library.

Gap-closing with assembled contigs (Gap-close)

Collected reads are assembled using the de Bruijn graph. The algorithm is similar to contig-assembly, although differences do exist. The strand directions of reads are determined from the paired-end information. Therefore, a *k*-mer and its reverse complement are distinguishable. Several steps are simplified in order to maintain the speed. Assembly is performed using two fixed *k* values (24 and 72). Minimum coverage of *k*-mers is also a fixed value (3). After contigs are generated from de Bruijn graph, overlaps between the end of a contig and the side of a gap are sought. If both ends of a contig have overlaps that are greater than the threshold (default, 32), the gaps are closed using the sequence of the contig. In this case, if a pair of ends has similarity greater than the threshold (default, 0.95), a gap is filled with the contig (Supplemental Figure 19). In addition, after de Bruijn graph-based closing, Platanus applies an overlap-layout-consensus algorithm to the remaining gaps. In this step, overlaps of length ≥ 32 and edit distance ≤ 1 between reads are used. Like scaffolding, libraries are used from short-insert to long-insert libraries. In the final analysis, all libraries are used simultaneously.



Supplemental Figure 19. Gap filling with a contig

In this figure, the minimum overlap length is 8. In the overlap between the contig and gap-neighboring region, mismatches are represented as red characters. A gap in a scaffold is represented as a sequence of N. Overlap of contig2 indicates a higher similarity and the gap is filled with contig2.

Pre-process of reads

We trimmed adaptor sequences and/or low-quality regions in reads. The pre-processed reads were entered as input into all assemblers.

Adaptor sequences are searched using a seed size of 11 bp. If an alignment length of ≥30 bp is discovered, the aligned region and the outside regions are excluded.

To trim low-quality regions, the threshold of quality value, $q_{\min}$, is set to 15. Starting from both 5′ and 3′ ends, the outside regions are trimmed if the quality value is less than 15. Trimming finishes if the run of high-quality bases occurs (quality values ≥ $q_{\min}$), for which length ≥ 11.

Filtering of mate-pair libraries

For mate-pair libraries, PCR duplicates and read pairs that have abnormally short insert sizes were excluded. Mate-pairs were mapped on the sequences (*C. elegans*, reference; others, scaffolds assembled by Platanus using only paired-end libraries), using Bowtie2 (Langmead and Salzberg 2012). For each set of pairs that have common mapping position of 5′ ends, the one that possesses the highest mapping-quality remains, while the remainder are excluded. In addition, pairs whose insert sizes (inferred form mapping results) are less than half the nominal insert size are also excluded. Resulting mate-pair libraries were input into all assemblers.

Pre-process of reads specific for snake of Assemblathon2

For mate-pair libraries with insert-sizes of 2 kb, 4 kb, and 10 kb, there exist internal sequences (CAGTACTG) corresponding to ligation points in circular DNA fragments. We trimmed the region from this 8-bp sequence to the 3′ end before other pre-processing steps.

Pre-process of reads specific for fish of Assemblathon2

For the fosmid-end library (insert size, 40 kb), we extracted the sub-sequences corresponding to the 5–76 position in reads before other pre-process steps. Only these regions represented the genomic sequences, according to the webpage of Assemblathon2.

Parameter optimization for assemblers

For the benchmarks of *C. elegans, S. venezuelensis,* and oyster, we optimized several parameters (e.g., the *k*-mer length and merge level). For each dataset, the settings that generated the largest scaffold NG50 sizes were selected. Values for means and standard deviations of insert sizes are entered as input in the following manner: for *C. elegans*, estimated values using reference sequence; for other species, nominal values.

Note that Platanus automatically estimates the means and standard deviations of insert sizes, except in the case of fish data. For all assemblers, the options related to the number of threads were set to 32.

・Platanus

All parameters were set to the default, except "–m 100" for the Contig-assembly of the oyster data. The value of –m only affects execution time and memory usage. For Contig-assembly, paired-ends were entered as input. For Scaffolding and Gap-close, all libraries were entered as input. The sole exception was the assembly of fish contigs; all libraries entered as input for Contig-assembly and insert-sizes were specified as nominal sizes in Scaffolding.

・SOAPdenovo2

We specified –K values from 21 to 91 (with a step size of 10), and –M option was set as 1 or 3. GapCloser (version: 1.12) was applied to the SOAPdenovo2 outputs ("out.scafSeq") using default parameters. For contig assembly, paired-ends were entered as input. For the other steps, all libraries were entered as input.

・ALLPATHS-LG

The option of PrepareAllPathsInputs.pl was set as “PLOIDY=2” for diploid samples.

・Velvet

The –k option was set from 21 to 91 (with a step size of 10).

・MaSuRCA

All parameters except “JF_SIZE” were set in an identical manner as the example file. JF_SIZE only affects execution time and memory usage, and was set as follows:
C. elegans and *S. venezuelensis*: JF_SIZE=1000000000.
Oyster: JF_SIZE=10000000000.

***De novo* assembly of oyster RNA-seq data**

Paired-end libraries from 13 samples and single-end libraries from 99 samples were preprocessed in the same way as genomic data (Supplemental Methods, Pre-process of reads**).** All preprocessed libraries (paired-end: 44.6 Gbp, single-end: 76.2 Gbp) were input into Trinity assembler (Grabherr et al. 2011) with default setting. Resulting RNA-contigs whose lengths ≥ 500 bp were used for the validation of assemblies. The total size, number, and average length of the RNA-contigs were 56,540,774 bp, 40,503 bp, and 1,396bp, respectively.

Supplemental Information

Details of sequencing data sets

Details of sequencing data sets using in this paper are shown in Supplemental Table 1.

Supplemental Table1. Details of sequencing data sets

Species	*Caenorhabditis elegans* (nematode worm)		
Genome size (bp)	100.3 M		
Insert size (bp)	230	420	4,660
Read length (raw) (bp)	110	110	100
Total size (raw) (bp)	15.1G	7.2G	13.9G
Read length (preprocessed) (bp)	107	106	87
Total size (preprocessed) (bp)	7.2G	6.8G	2.8G

Species	*Strongyloides venezuelensis* (nematode worm)		
Genome size (bp)	57.7M		
Insert size (bp)	200	450	3400
Read length (raw) (bp)	110	100	100
Total size (raw) (bp)	2.9G	5.2G	5.3G
Read length (preprocessed) (bp)	104	96	69
Total size (preprocessed) (bp)	2.7G	5.0G	3.6G

Species	*Crassostrea gigas* (oyster)						
Genome size (bp)	565.7M						
Insert size (bp)	170	500	800	2000	5000	10,000	20,000
Read length (raw) (bp)	90	90	90	90	90	90	90
Total size (raw) (bp)	36.3G	18.7G	18.5G	50.7G	16.6G	18.8G	25.7G
Read length (preprocessed) (bp)	86	84	82	71	82	66	60
Total size (preprocessed) (bp)	34.7G	17.6G	17.0G	29.7G	4.2G	2.0G	2.6G

Species	*Melopsittacus undulates* (bird)							
Genome size (bp)	1085.2M							
Insert size (bp)	220	500	800	2,000	5,000	10,000	20,000	40,000
Read length (raw) (bp)	150	150	150	90	90	90	90	90
Total size (raw) (bp)	48.4G	47.2G	43.1G	47.6G	35.0G	17.0G	16.1G	15.7G
Read length (preprocessed) (bp)	135	128	115	82	81	79	78	65
Total size (preprocessed) (bp)	43.6G	40.5G	33.0G	26.0G	15.7G	7.3G	3.7G	1.7G

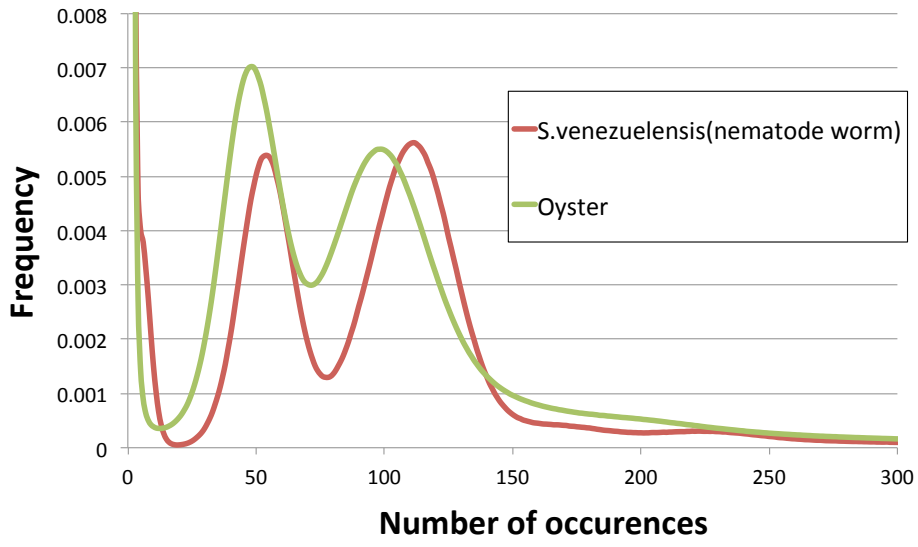
Species	*Boa constrictor constrictor* (snake)			
Genome size (bp)	1431.5M			
Insert size (bp)	400	2,000	4,000	10,000
Read length (raw) (bp)	121	101	101	101
Total size (raw) (bp)	136.0G	25.1G	17.4G	20.5G
Read length (preprocessed) (bp)	118	65	68	75
Total size (preprocessed) (bp)	132.1G	14.3G	1.0G	6.5G

Species	*Maylandia zebra* (fish)						
Genome size (bp)	915.0M						
Insert size (bp)	180	2,500	5,000	7,000	9,000	11,000	40,000
Read length (raw) (bp)	101	101	101	101	101	101	101
Total size (raw) (bp)	60.4G	71.7G	14.5G	16.0G	14.9G	11.6G	3.9G
Read length (preprocessed) (bp)	80	62	62	58	52	58	50
Total size (preprocessed) (bp)	48.0G	27.7G	3.8G	3.8G	5.1G	2.0G	1.0G

For "preprocessed" data, adaptor sequences and low quality regions (threshold of quality value: 15) are trimmed and PCR-duplicate if mate-pairs are excluded.

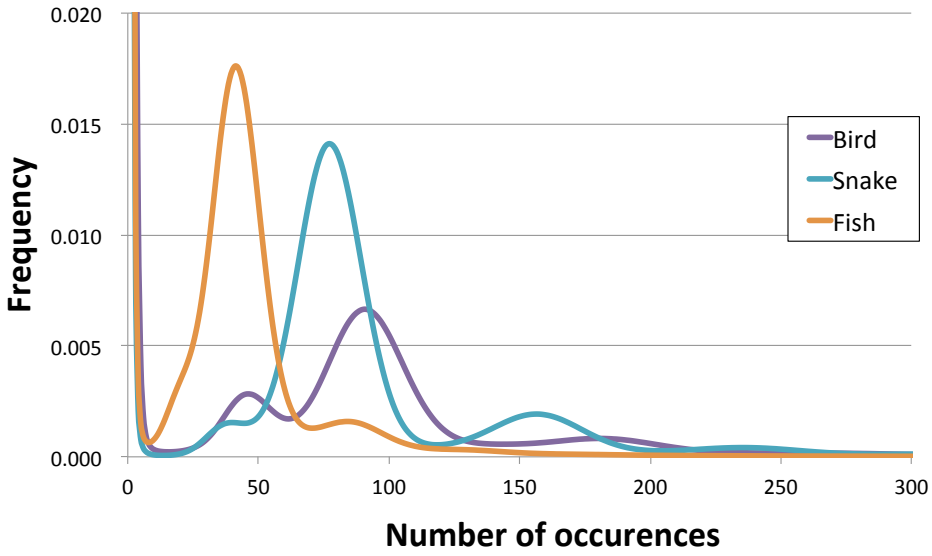
17-mer frequency analysis

Original (pre-normalized) 17-mer frequency distributions of *S. venezuelensis* and oyster are shown in Supplementary Figure 20 and 17-mer frequency distributions of 3 species (bird, snake, fish) are shown in Supplementary Figure 21.



Supplemental Figure 20. Distribution of the number of 17-mer occurrences

Real data from highly heterozygous samples, *S. venezuelensis* and oyster.



Supplemental Figure 21. Distribution of the number of 17-mer occurrences

Data from three species of Assemblathon2.

Details of statistics of *C. elegans* assemblies

Details of statistics of *C.elegans* assemblies are shown in Supplemental Table 2 and Supplemental Figure 22.

Supplemental Table 2. Details of statistics of *C. elegans* assemblies

(A) Corrected Scaffold NG50 (bp)

Heterozygosity (%)	Corrected Scaffold NG50 (bp)				
	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
0.0	361,608	402,062	210,683	394,556	123,438
0.1	388,028	368,779	209,917	403,115	104,756
0.2	411,618	228,033	229,808	403,701	104,352
0.3	392,137	82,457	255,210	372,430	60,253
0.5	366,498	5,994	263,146	288,181	51,097
1.0	368,857	4,928	288,752	234,395	34,491
1.5	355,057	4,999	287,646	202,140	2,099
2.0	341,914	5,178	240,778	150,059	1,932

(B) Scaffold NG50 (bp)

Heterozygosity (%)	Scaffold NG50 (bp)				
	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
0.0	478,744	466,658	420,694	507,513	424,862
0.1	490,975	431,770	482,920	497,363	332,019
0.2	535,328	375,904	430,011	489,092	340,229
0.3	545,914	219,404	460,620	441,950	286,218
0.5	497,387	127,365	475,513	353,955	251,000
1.0	511,190	91,413	466,806	280,050	209,807
1.5	516,958	73,543	472,079	252,105	178,132
2.0	580,832	86,979	351,406	212,590	162,062

(C) Corrected Contig NG50 (bp)

Heterozygosity (%)	Corrected Contig NG50 (bp)				
	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
0.0	48,054	47,111	40,248	45,927	17,739
0.1	48,250	46,131	41,645	43,026	12,263
0.2	47,179	33,243	41,819	34,822	7,831
0.3	43,969	15,142	42,900	25,294	9,474
0.5	43,196	2,746	42,700	23,831	7,652
1.0	39,291	1,316	44,746	16,784	4,231
1.5	35,786	1,057	42,593	9,419	707
2.0	34,030	1,649	38,465	4,013	610

(D) Contig NG50 (bp)

Heterozygosity (%)	Contig NG50 (bp)				
	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
0.0	66,446	67,596	74,686	170,395	19,084
0.1	65,717	66,289	77,050	155,501	12,988
0.2	61,969	45,370	76,439	144,468	8,135
0.3	59,795	19,629	79,194	141,590	10,174
0.5	53,873	2,807	80,813	117,418	8,142
1.0	48,090	1,358	84,827	103,677	4,445
1.5	42,013	1,103	75,918	90,768	766
2.0	39,915	1,810	62,656	60,761	666

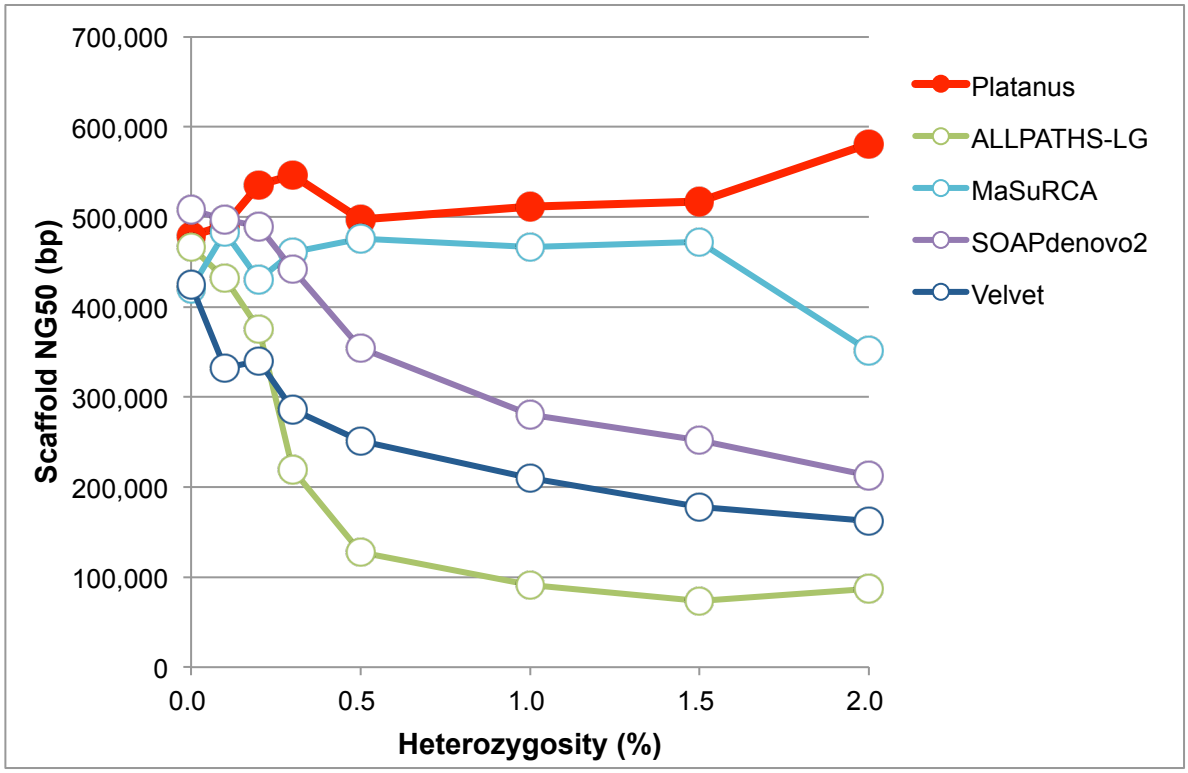
(E) Total size of scaffolds ≥ 500 bp

Heterozygosity (%)	Total size (bp)				
	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
0.0	97,938,690	98,864,571	97,742,565	99,132,377	99,382,943
0.1	98,876,811	99,521,025	99,171,512	99,215,054	99,851,591
0.2	98,876,741	102,601,583	98,996,688	99,504,741	101,204,194
0.3	98,920,320	114,954,930	99,210,505	99,969,847	99,482,862
0.5	98,916,151	153,369,255	99,925,896	99,196,053	100,074,381
1.0	99,223,085	163,953,538	100,913,677	99,086,653	103,773,061
1.5	99,706,135	155,053,323	103,830,661	100,438,152	107,239,890
2.0	100,454,422	170,519,463	105,340,084	99,546,387	110,798,448

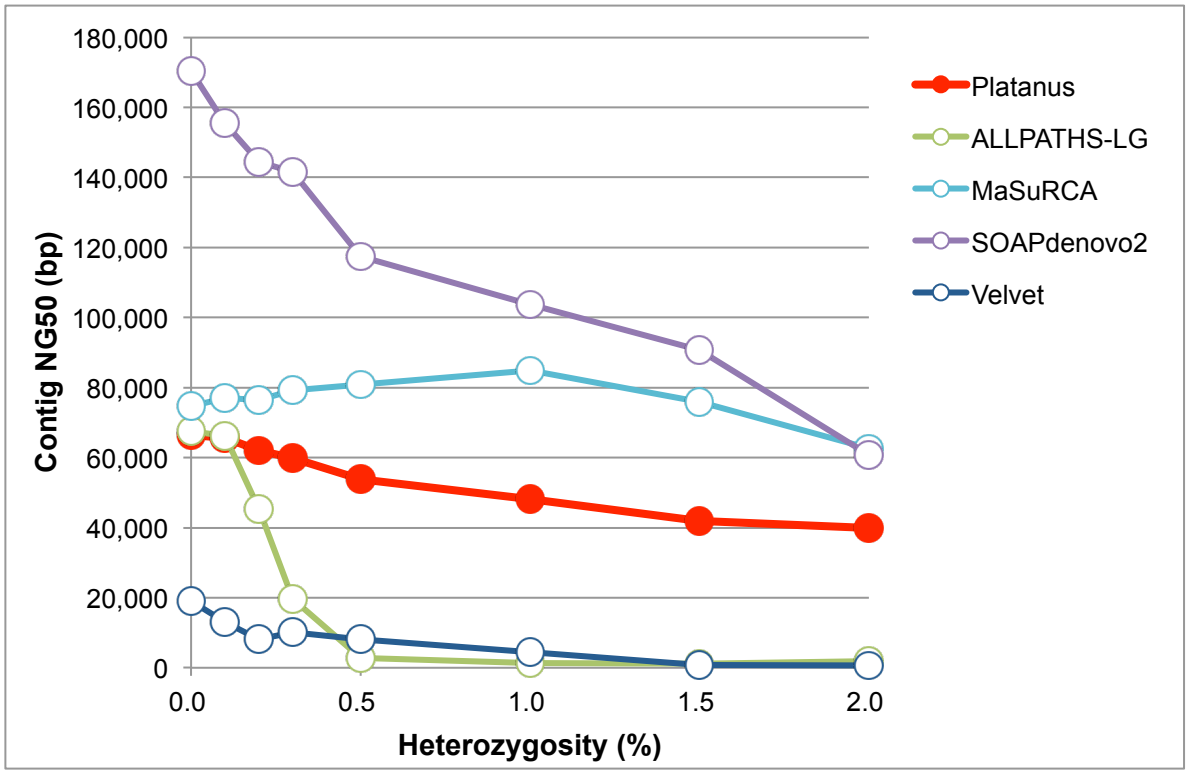
(F) Number of errors (inversion + translocation + relocation)

Heterozygosity (%)	Number of errors (bp)				
	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
0.0	367	272	900	407	1,490
0.1	346	259	878	441	2,044
0.2	333	276	892	450	1,999
0.3	324	335	911	442	3,575
0.5	325	593	881	941	4,206
1.0	290	542	895	1,226	6,285
1.5	300	498	861	1,304	35,440
2.0	256	496	774	1,839	36,840

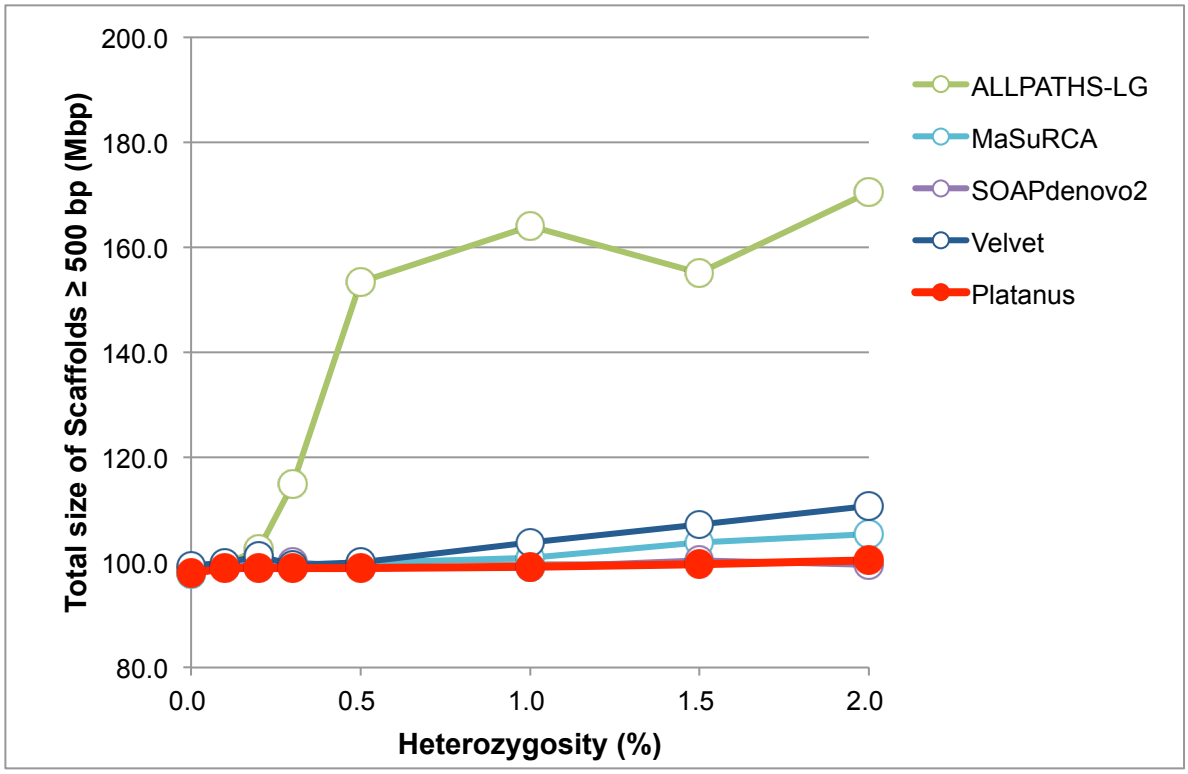
(A)


700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
Scaffold NG50 (bp)
0.0
0.5
1.0
1.5
2.0
Heterozygosity (%)
Platanus
ALLPATHS-LG
MaSuRCA
SOAPdenovo2
Velvet

(B)


180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
Contig NG50 (bp)
0.0
0.5
1.0
1.5
2.0
Heterozygosity (%)
Platanus
ALLPATHS-LG
MaSuRCA
SOAPdenovo2
Velvet

(C)



Supplemental Figure 22. Results of the benchmarks of heterozygosity simulations (*C. elegans*)

(A) Scaffold NG50. (B) Contig NG50. (C) Total sizes of the assemblies (≥500 bp).

Comparison between simulated and estimated heterozygosity

By mapping original and each simulated paired-end reads onto the reference genome (*C. elegans* Sequencing Consortium 1998), the heterozygosity of each simulated data was calculated. These results are shown in Supplemental Table 3.

Supplemental Table 3. Comparison between simulated and estimated heterozygosity (*C. elegans*)

Simulated heterozygosity (%)	Calculated heterozygosity (%)
0.0	1.85×10^{-3}
0.1	0.102
0.2	0.202
0.3	0.302
0.5	0.501
1.0	0.999
1.5	1.49
2.0	1.95

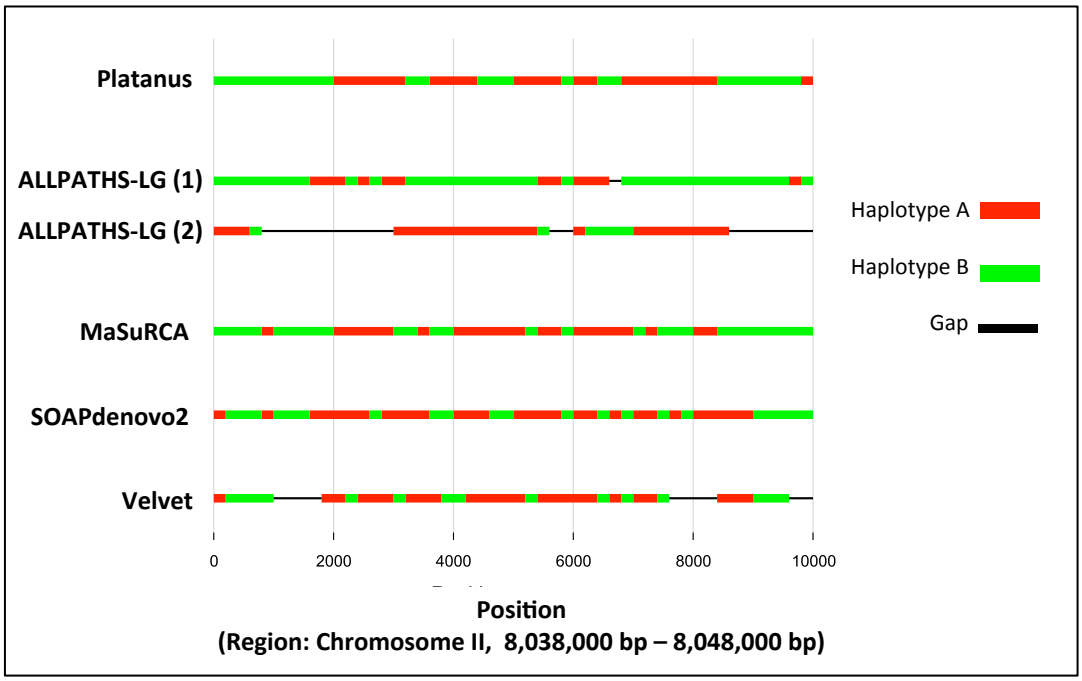
Total size of assembled scaffolds

In the assembly of simulated heterozygous *C.elegans* genomes, the total size of ALLPATHS-LG approaches nearly twice the genome size (2 × 100 Mbp) (Supplemental Figure 22). We were aware of the possibility that ALLPATHS-LG constructed haplotype sequences separately. If so, linkage information can be obtained from the variants, which could be suitable for the post analysis. To investigate the existence of haplotype-sensitive assembly, we first mapped all 200 bp fragments, clipping from each assembler's scaffolds (2.0% heterozygosity) to the simulated diploid reference genome sequences with 2.0% heterozygosity, which were constructed *in silico*. We then detected deviations of each fragment from the haplotypes and finally identified the haplotype junctions (i.e., the borders between different haplotypes). As a result of this process, all assemblies contained such junctions (Supplemental Table 4). Supplemental Figure 23 provides an example of haplotype changes in the 10-Kbp region. In this region, all scaffolds contained multiple junctions, and ALLPATHS-LG provided duplicated scaffolds. Indeed, 85% of the ALLPATHS-LG scaffolds possessed haplotype junctions. This result suggests that the assemblers tested in this study are not designed to construct each haplotype sequence, and that an increase of total assembly sizes is not indicative of the success of haplotype assembly. In the assembly of real highly heterozygous S.venezuelensis genome, the total size of ALLPATHS-LG's scaffold (≥500 bp) was 1.06-fold comparing with the real genome size, in contrast to the 1.63-fold size in the 1.0%-heterozygous *C. elegans* test. The uneven distribution of heterozygosity may have resulted in these observed differences between the real data and the simulated data.

Supplemental Table 4. Statistics of haplotype junctions in C. elegans data (heterozygosity: 2.0%)

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Number of junctions of haplotypes	143,359	92,940	170,172	184,989	149,030
Rate of scaffolds including junctions (%)	17.88	85.84	11.66	43.67	32.96
Average interval between junctions (bp)	1398.9	2157.7	1178.5	1084.1	1345.7

Every 200 bp sequences in the scaffolds were mapped to the simulated diploid genome using Bowtie2 and assigned to one of the haplotypes. Junctions were detected if adjacent fragments had different haplotypes. Gaps or variant-free regions were skipped.



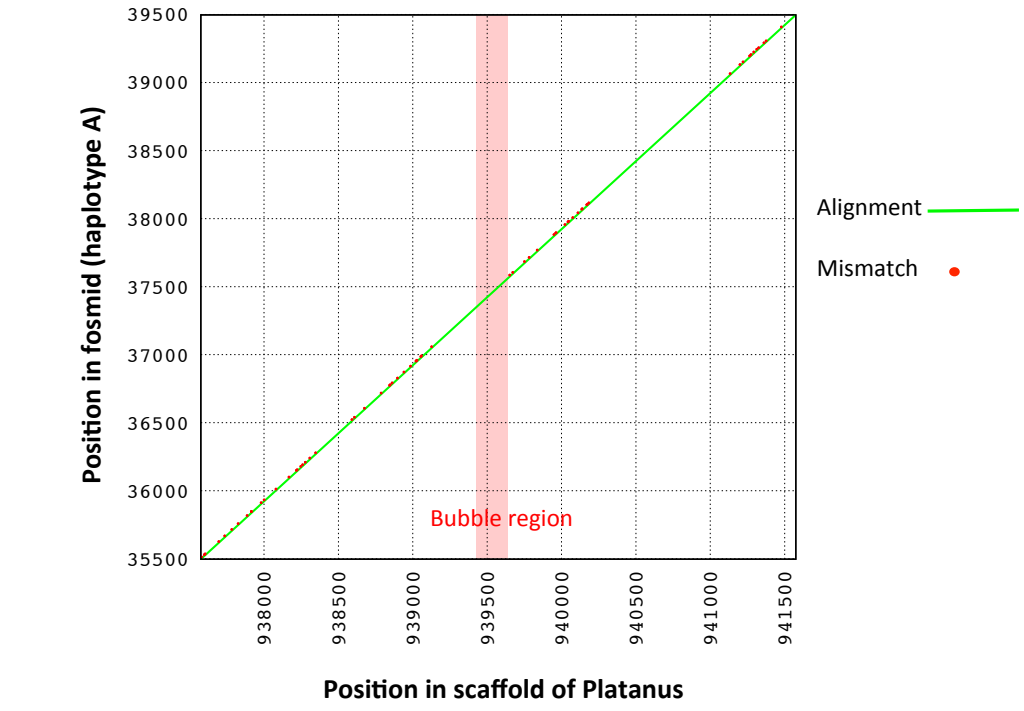
Supplemental Figure 23. Example of the junctions of haplotypes in *C. elegans* assemblies (heterozygosity: 2.0%)

For the benchmark of 2.0%-heterozygous *C. elegans* data, scaffolds corresponding to a certain region (size: 10 kbp) are displayed. All 200 bp of the scaffolds are mapped on the simulated diploid genome, assigned to one of the haplotypes, and the scaffold regions are differentiated (red or green) according to haplotypes. For this region, ALLPATHS-LG produces duplicated scaffolds, and both are shown.

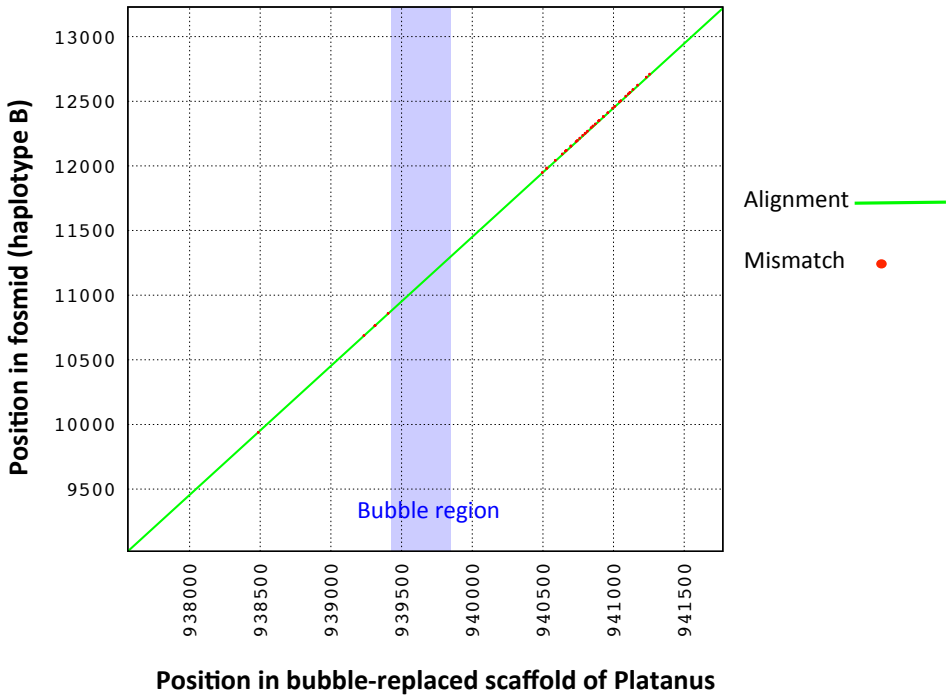
Example of "bubble removal" in scaffolding (*S. venezuerensis*)

we provide the example of "bubble removal" in scaffolding (Fig. 4A, B) using a dot plot analysis by 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 Figure 24B). These results indicate that Platanus correctly resolved the region containing a relatively large indel, many SNVs, and several small indels existed simultaneously using the bubble removal routine.

(A)



(B)



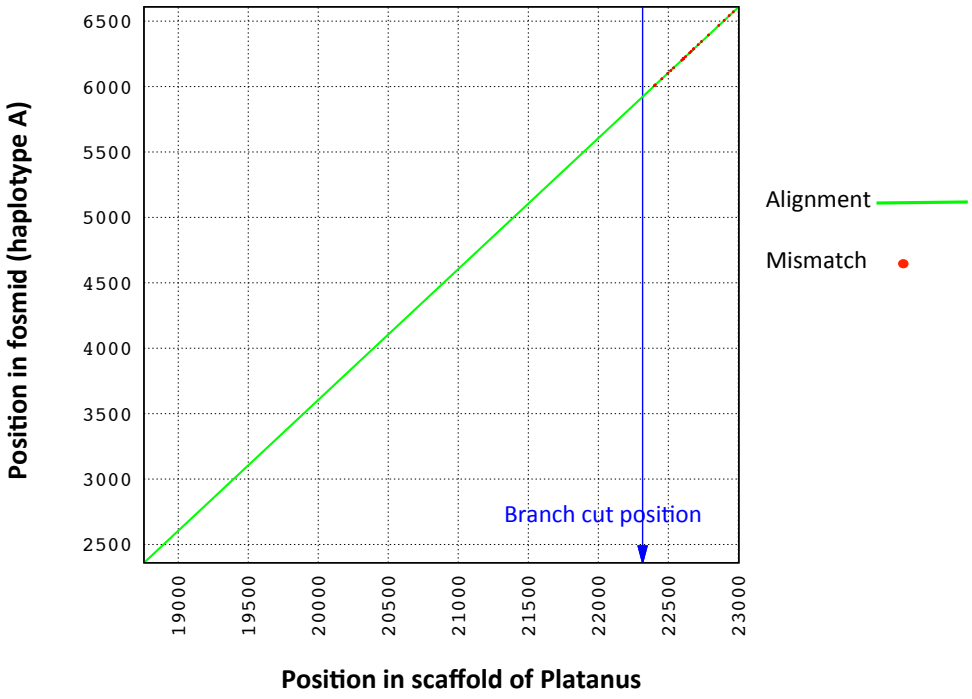
Supplemental Figure 24. Example of a heterozygous region resolved by "bubble removal"

(A) Alignment dot plot between a fosmid (haplotype A) and a scaffold of Platanus. The red box indicates the region on the scaffold corresponding to the bubble (removed contig).

(B) Alignment dot plot between a fosmid (haplotype B) and a scaffold, in which the region corresponding to the bubble is replaced with the removed sequence (removed contig).

Example of "branch cut" in scaffolding (*S. venezuerensis*)

We provide the 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, low coverage depth. Three indels were apparent whose sizes were 126 bp, 715 bp, and 1,206 bp, with high heterozygosity (1.93%). The scaffold sequence (ContigP2 – ContigA2 – ContigN2) could be aligned to one fosmid of the pair (Supplemental Figure 25), and the removed branch (size: 1,217 bp; ContigB2) matched the other fosmid, confirming the correctness of Platanus' resolution.

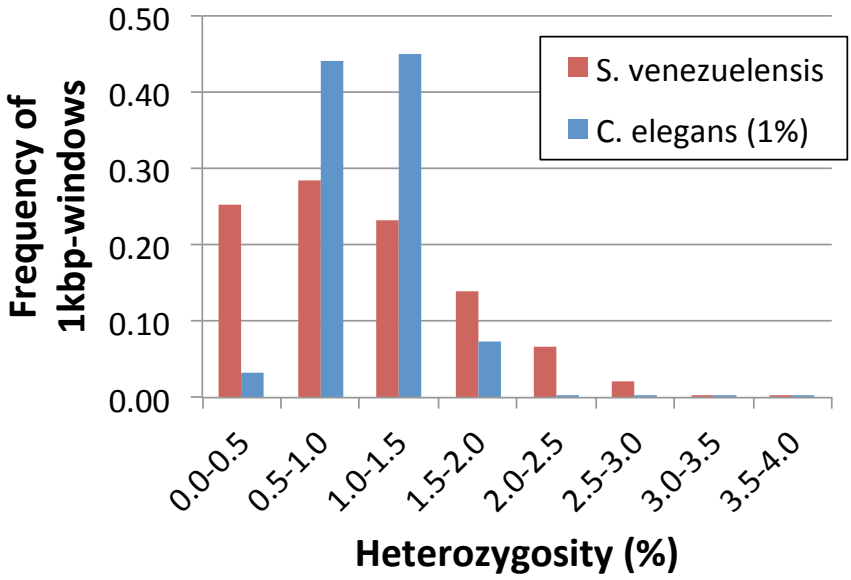


Supplemental Figure 25. Example of a heterozygous region resolved by "branch cut"

Alignment dot plot between a fosmid (haplotype A) and a scaffold of Platanus.

Distribution of heterozygosity across the entire *S.venezuerensis* genome

SNVs and small indels on the scaffolds were detected by mapping paired-end reads (see Methods), and heterozygosity was calculated for every 1 kbp non-overlapping 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.

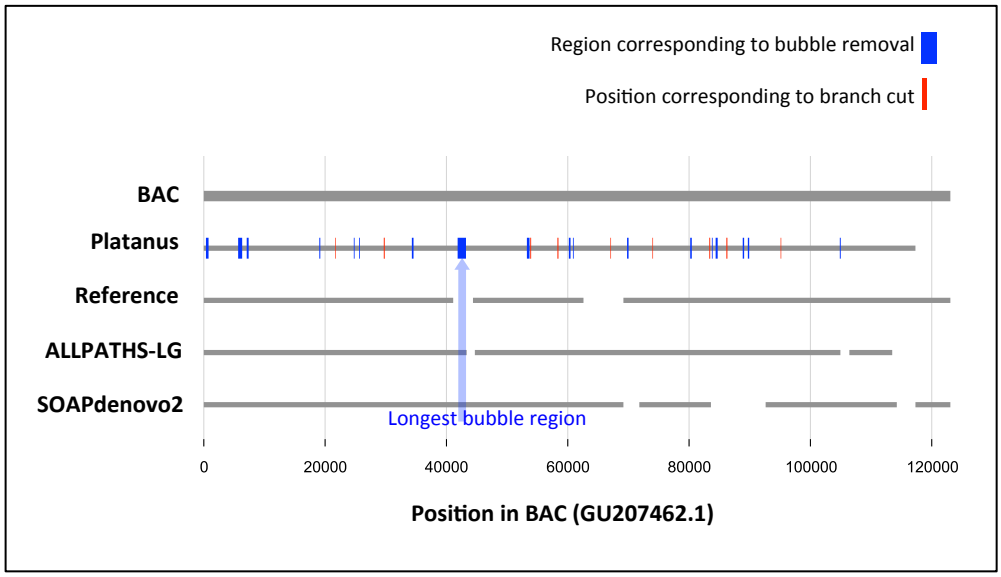


Supplemental Figure 26. Distribution of heterozygosity on scaffolds

SNVs and small indels on Platanus' scaffolds were detected by mapping paired-ends (see Methods). Heterozygosity ((SNVs + indels) / base) was calculated for each 1 kbp window. *S. venezuelensis* data were real data, where as *C. elegans* data consisted of simulated heterozygosity (1.0%).

Example of alignments between BAC and scaffolds in oyster genome assembling

We provide an example of the region that Platanus was able to resolve but the fosmid-based references were divided (denoted by the blue arrow in Supplemental Figure 27). In this region, Platanus removed a bubble of greater than 1 Kbp in length, corresponding to a large indel (approximately 800 bp in length). In the fosmid-based methods, each fosmid was first assembled as a haplotype sequence. Next the resulting fosmids were connected according to the overlap-layout-consensus algorithm, merging haplotypes into consensus sequences. In the first step, fosmid assembly should succeed even for highly heterozygous regions because each haplotype is assembled separately. However, complex variants can interfere with the assembly process in the next merging step. This example demonstrates that even fosmid-based methods can fail to handle highly heterozygous regions including structural variations. Platanus is designed to be able to resolve such regions in its Scaffolding step.



Supplemental Figure 27. Example of alignments between BAC and scaffolds.

Blue boxes indicate the regions corresponding to “bubble removal”. Red lines indicate the position of branch-cut. The blue arrow indicate the longest bubble-removal region in the the BAC, (see Supplemental Information). The accession code of BAC is GU207462.1.

Comparison of scaffold-NG50 length between each assembler and Platanus

In Supplemental Table 5, comparison results between each assembler's scaffold-NG50 and Platanus' are shown.

Supplemental Table 5. NG50 / Platanus-NG50 (Scaffold)

	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
C. elegans (heterozygosity: 0.0%)	0.975	0.879	1.060	0.887
C. elegans (heterozygosity: 0.1%)	0.879	0.984	1.013	0.676
C. elegans (heterozygosity: 0.2%)	0.702	0.803	0.914	0.636
C. elegans (heterozygosity: 0.3%)	0.402	0.844	0.810	0.524
C. elegans (heterozygosity: 0.5%)	0.256	0.956	0.712	0.505
C. elegans (heterozygosity: 1.0%)	0.179	0.913	0.548	0.410
C. elegans (heterozygosity: 1.5%)	0.142	0.913	0.488	0.345
C. elegans (heterozygosity: 2.0%)	0.150	0.605	0.366	0.279
S. venezuelensis	0.061	0.642	0.318	0.062
Oyster	0.404	N/A	0.305	N/A

Comparison between Platanus scaffolds and fosmid-based reference in oyster genome assembling using RNA-seq mapping

For Platanus' scaffolds and the reference, there were 1,357 Platanus-specific RNA-contigs (Platanus: mapped; reference: no hits with coverage ≥50%; Supplemental Table 6). In the RNA-contigs, 928 have significant hits with BLASTX (e-value ≤ 10^{-5}) in the nr database. Conversely, there were only 366 reference-specific RNA-contigs.

Supplemental Table 6. Comparison between Platanus and RNA-mapping reference (oyster)

	Total (bp)	Number	Average length (bp)	Number of significant-hit contigs (nr)
Platanus-specific (Reference: nohit)	1,680,407	1,357	1238.3	928
Reference-specific (Platanus: nohit)	540,263	366	1476.1	271

RNA-contigs (number: 40,503, total: 56,540,774 bp) were aligned to scaffolds using blat. If a RNA-contig was mapped to Platanus' scaffold (with 90% covered by a single scaffold) and no hit against the reference had a coverage ≥50%, that contig is considered to be "Platanus-specific" and vice versa for "Reference-specific." A "significant-hit contig" is an RNA-contig with hits of BLASTX (e-value ≤10-5) on the nr database.

Assembly of the Assemblathon2 data

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) for Assemblathon2. The summaries of results for this section are shown in Table 1 and detail results are shown in Supplemental Table 7. Platanus recorded the highest values for both scaffold NG50 (bird: 21,684,294 bp, snake: 17,165,953 bp). For the snake assembly in particular, the scaffold NG50 of Platanus was unexpectedly large at more than three times the second largest value. 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 the ALLPATHS_LG. One important feature of the fish data is the low coverage depth (52.5×) of its paired-end reads, which probably reduced Platanus' scaffold NG50 value.

Supplemental Table 7A. Statistics of the Assemblathon2 assembly (Bird)

Description of assemblers			Assembly statistics			Fosmid validation		
Team name	Principal software used	Type of data used	Total (≥500bp)	Scaffold NG50 (bp)	Contig NG50 (bp)	Top-hits-length (bp)	Identity (%)	Number of contained fosmids
Platanus	**Platanus**	**Illumina**	**1,129,507,736**	**21,684,294**	**56,973**	**1,026,396**	**99.40**	**84**
ABL	HyDA	Illumina, Roche-454	962,501,883	3,003	3,003	415,799	99.63	18
Allpaths	ALLPATHS-LG	Illumina	1,148,183,934	17,716,398	57,617	1,001,521	99.22	77
BCM-HGSC	SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLPATHS-LG, Atlas-Link, Atlas-GapFill, Phrap	Illumina, Roche-454, PacBio	1,322,103,957	17,484,024	184,968	995,482	99.40	80
BCM-HGSC*	SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLPATHS-LG, Atlas-Link, Atlas-GapFill, Phrap	Illumina, Roche-454, PacBio	1,323,179,685	17,488,428	115,689	992,609	99.44	80
CBCB	Celera assembler and PacBio Corrected Reads (PBcR)	Illumina, Roche-454, PacBio	1,219,131,251	2,030,397	116,375	1,021,740	99.34	81
CoBig2	4Pipe4 pipeline, Seqclean, Mira, Bambus2	Roche-454	11,169,988	0	0	30,036	89.25	0
MLK-Group	ABySS	Illumina	1,871,173,099	150,983	36,470	876,624	99.45	58
Meraculous	meraculous	Illumina	1,081,601,988	9,834,474	37,061	975,863	99.43	71
Newbler-454	Newbler	Roche-454	1,117,368,293	11,495,203	66,030	995,738	99.38	74
Phusion	Phusion2, SOAPdenovo, SSPACE	Illumina	1,334,459,170	1,479,065	74,321	931,879	99.26	66
Ray	Ray	Illumina	1,266,700,501	673,924	44,663	987,974	99.41	78
SGA	SGA	Illumina	1,152,568,402	3,584,181	17,785	912,595	99.49	59
SOAPdenovo	SOAPdenovo	Illumina	1,150,787,380	14,644,723	40,684	1,009,988	99.24	81
SOAPdenovo*	SOAPdenovo	Illumina	1,147,593,642	14,617,523	54,932	1,010,089	99.29	81
SOAPdenovo**	SOAPdenovo	Illumina	1,147,651,023	14,617,693	53,557	1,010,137	99.29	81

Supplemental Table 7B. Statistics of the Assemblathon2 assembly (Snake)

Description of assemblers			Assembly statistics			Fosmid validation		
Team name	Principal software used	Type of data used	Total (≥500bp)	Scaffold NG50 (bp)	Contig NG50 (bp)	Top-hits-length (bp)	Identity (%)	Number of contained fosmids
Platanus	**Platanus**	**Illumina**	**1,441,357,474**	**17,165,953**	**48,614**	**368,203**	**99.80**	**53**
ABySS	ABySS	Illumina	1,513,411,700	508,539	28,651	356,343	99.80	47
BCM-HGSC	SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLP A THS-LG, Atlas-Link, Atlas-GapFill, Phrap	Illumina	1,439,656,907	1,563,800	12,727	333,109	99.51	39
CRACS	ABySS, SSPACE, Bowtie, and FASTX	Illumina	1,514,104,169	738,101	16,852	344,776	99.76	39
Curtain	SOAPdenovo, fastx_toolkit, bwa, samtools, velvet, curtain	Illumina	1,497,321,002	58,954	4,476	279,969	99.34	25
GAM	GAM, CLC and ABySS	Illumina	1,367,480,854	19,350	4,499	306,517	99.22	29
Meraculous	meraculous	Illumina	1,426,938,655	1,247,790	34,344	356,103	99.85	47
PRICE	PRICE	Illumina	312,669,723	0	0	60,537	94.92	2
Phusion	Phusion2, SOAPdenovo, SSPACE	Illumina	1,522,735,963	4,494,848	77,302	365,061	99.71	53
Ray	Ray	Illumina	1,527,313,311	143,603	18,158	340,844	99.76	42
SGA	SGA	Illumina	1,442,930,293	4,536,273	28,397	357,395	99.71	51
SOAPdenovo	SOAPdenovo	Illumina	1,608,796,330	2,004,523	18,848	365,687	99.77	51
Symbiose	Monument, SSPACE, SuperScaffolder, GapCloser	Illumina	1,989,598,057	1,794,214	82,841	339,360	99.58	45

Supplemental Table 7C. Statistics of the Assemblathon2 assembly (Fish)

Description of assemblers			Assembly statistics		
Team name	Principal software used	Type of data used	Total (≥500bp)	Scaffold NG50 (bp)	Contig NG50 (bp)
Platanus	**Platanus**	**Illumina**	**825,154,699**	**2,371,946**	**6,587**
ABySS	ABySS	Illumina	849,812,973	1,013,854	4,030
Allpaths	ALLPATHS-LG	Illumina	845,883,785	3,677,909	14,105
BCM-HGSC	SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLPATHS-LG, Atlas-Link, Atlas-GapFill, Phrap	Illumina	868,269,785	4,850,564	17,895
CSHL	Metassembler, ALLPATHS, SOAPdenovo	Illumina	845,089,232	3,418,986	16,740
CSHL*	Metassembler, ALLPATHS, SOAPdenovo	Illumina	844,657,792	3,418,986	16,517
CSHL**	Metassembler, ALLPATHS, SOAPdenovo	Illumina	545,734,294	1,560	1,550
CTD	Unspecified	Illumina	1,325,939,605	3,139	935
CTD*	Unspecified	Illumina	933,447,632	935	3,139
CTD**	Unspecified	Illumina	989,896,657	1,285	1,285
IOBUGA	ALLPATHS-LG, SOAPdenovo	Illumina	825,949,698	261,156	1,472
IOBUGA*	ALLPATHS-LG, SOAPdenovo	Illumina	2,193,012,109	51,717	556
Meraculous	meraculous	Illumina	801,481,081	764,900	3,962
Ray	Ray	Illumina	787,647,067	37,280	6,829
SGA	SGA	Illumina	812,239,698	88,883	4,739
SOAPdenovo	SOAPdenovo	Illumina	1,063,706,964	1,665,791	7,001
Symbiose	Monument, SSPACE, SuperScaffolder, GapCloser	Illumina	1,101,461,395	1,731,822	30,444

Team names, principal software used, and type of data used without Platanus are from the Assemblathon2 paper (Bradnam et al. 2013). *and ** indicates the alternative assembling results submitted by same team.

For the bird and snake genomes, the procedures of fosmid validation are the same as those for *S. venezuelensis* (Table 4). For the fish genome, fosmid data are not available. The number of fosmids in the bird genome: 86. Total length of fosmids in the bird genome: 1,035,129 bp. The number of fosmids in the snake genome: 56. Total length of fosmids in the snake genome: 378,186 bp

Scaffold NG10-90 length (bp)

Supplemental Table 8. Scaffold NG10-90 length(bp)

(A) *C.elegans* (0.0%)

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
NG10	1,299,912	1,381,440	1,034,395	1,289,828	1,430,751
NG20	1,057,372	1,012,348	781,669	1,011,737	1,031,386
NG30	804,064	802,422	674,619	808,239	718,216
NG40	578,582	600,891	485,530	627,891	552,498
NG50	478,744	466,658	420,694	511,116	436,119
NG60	382,364	376,538	325,678	403,044	321,258
NG70	304,681	290,317	234,815	308,761	212,774
NG80	203,659	194,126	169,609	214,374	145,031
NG90	112,112	118,336	80,754	135,279	77,580

(B) *C.elegans* (0.1%)

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
NG10	1,507,739	1,280,286	1,260,211	1,291,095	1,080,467
NG20	1,118,163	870,208	1,073,314	1,012,332	874,181
NG30	833,592	707,788	751,741	803,869	626,309
NG40	638,510	541,384	591,818	622,297	479,630
NG50	490,975	431,770	482,920	503,183	361,849
NG60	410,194	349,491	347,389	404,715	250,384
NG70	339,508	272,641	266,002	305,329	194,614
NG80	226,408	184,507	194,975	211,058	138,162
NG90	137,759	117,282	100,689	136,400	77,130

(C)*C.elegans (*0.2%)

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
NG10	1,618,829	1,114,597	1,194,022	1,500,581	1,134,779
NG20	1,180,520	804,196	938,808	940,430	745,422
NG30	888,969	634,922	761,774	805,146	596,877
NG40	689,059	482,045	597,472	602,921	428,963
NG50	535,328	375,904	430,011	489,092	340,229
NG60	428,773	294,284	349,906	379,622	262,353
NG70	341,982	209,437	277,283	293,423	197,949
NG80	224,998	153,506	202,553	202,237	141,198
NG90	136,349	96,208	110,772	123,884	84,078

(D) *C.elegans* (0.3%)

	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	1,706,653	626,135	1,160,198	1,261,449	789,635
NG20	1,096,294	456,730	975,497	879,218	590,193
NG30	852,632	360,603	773,873	741,682	459,920
NG40	697,718	277,924	587,611	556,075	379,644
NG50	579,390	219,404	460,620	441,950	299,631
NG60	448,504	172,465	383,935	347,540	217,336
NG70	359,483	133,539	309,512	266,033	147,297
NG80	257,254	103,567	208,497	183,515	93,487
NG90	148,904	75,929	120,869	116,841	47,291

(E) *C.elegans* (0.5%)

	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	1,238,095	236,448	1,272,404	950,483	939,107
NG20	948,144	203,086	1,020,232	800,138	655,305
NG30	729,332	175,721	779,038	568,611	463,144
NG40	600,900	152,779	594,781	433,522	337,070
NG50	497,387	127,365	475,513	363,065	259,666
NG60	382,616	109,375	390,032	290,747	196,779
NG70	319,057	94,762	310,998	222,863	142,912
NG80	224,518	83,175	220,529	164,889	91,462
NG90	134,593	72,075	133,248	94,426	49,101

(F) *C.elegans* (1.0%)

	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	1,297,465	179,692	1,498,834	735,944	810,795
NG20	1,149,556	143,512	1,131,825	541,413	573,389
NG30	833,362	123,052	747,427	444,954	453,838
NG40	651,970	106,752	604,221	363,226	317,914
NG50	511,190	91,413	466,806	283,097	243,330
NG60	425,517	80,341	378,265	235,133	179,858
NG70	344,783	72,753	302,525	186,200	136,810
NG80	223,772	63,420	212,100	135,227	89,253
NG90	139,858	54,176	125,733	78,550	51,144

(G) *C.elegans* (1.5%)

	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	1,827,873	158,231	1,214,376	736,944	668,279
NG20	1,082,509	120,379	930,268	483,186	514,278
NG30	829,962	99,275	728,678	390,611	354,197
NG40	648,996	85,505	595,175	312,447	254,601
NG50	516,958	73,543	472,079	252,105	188,027
NG60	429,160	63,924	373,247	202,087	141,319
NG70	343,322	55,792	294,906	153,400	95,869
NG80	213,721	47,683	208,561	107,145	60,445
NG90	134,986	41,908	135,272	63,670	28,289

(H) *C.elegans* (2.0%)

	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	1,284,850	182,554	967,957	648,847	604,533
NG20	1,140,197	138,882	758,245	449,541	433,420
NG30	804,731	117,845	603,073	331,717	326,488
NG40	695,866	99,421	463,321	261,688	243,770
NG50	580,832	86,979	351,406	212,590	179,009
NG60	476,594	76,750	306,697	166,087	139,660
NG70	351,908	67,216	230,394	132,960	94,833
NG80	221,209	59,951	169,994	96,257	64,100
NG90	139,114	52,800	109,939	54,218	30,207

(I) *S.venezuelensis*

	Platanus	**Allpaths-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	2,102,022	58,107	463,944	431,670	54,008
NG20	925,806	42,052	343,565	249,005	36,783
NG30	791,096	30,161	268,016	176,099	28,115
NG40	375,176	22,351	219,753	137,042	21,369
NG50	274,622	16,765	176,206	87,219	17,006
NG60	196,403	12,696	133,437	60,314	13,331
NG70	122,790	9,346	104,351	31,881	10,115
NG80	65,702	6,467	66,847	9,104	7,073
NG90	26,004	4,245	33,740	1,042	4,566

(J) Oyster

	Platanus	ALLPATHS-LG	SOAPdenovo2	Reference
NG10	1,006,309	601,021	445,119	1,116,604
NG20	703,142	368,683	274,279	809,743
NG30	560,689	268,463	185,387	635,477
NG40	463,742	199,001	144,420	495,744
NG50	381,943	154,144	116,321	392,835
NG60	304,883	120,995	95,555	290,497
NG70	241,514	93,217	78,905	210,443
NG80	174,316	67,456	65,604	129,849
NG90	114,291	43,004	53,864	59,925

(K) Bird

	Platanus	ABL	Allpaths	BCM-HGSC	BCM-HGSC*	CBCB
NG10	62,880,713	11,895	51,850,585	50,744,786	50,778,477	6,821,342
NG20	46,789,031	7,252	43,673,219	36,361,565	36,382,184	4,485,048
NG30	32,634,265	5,160	32,321,052	28,710,576	28,729,177	3,403,870
NG40	26,995,412	3,903	24,784,735	21,262,213	21,272,772	2,526,179
NG50	21,684,294	3,003	17,716,398	17,484,024	17,488,428	2,030,397
NG60	17,827,529	2,293	10,697,592	13,216,098	13,231,830	1,578,131
NG70	14,282,982	1,682	5,910,946	11,271,271	11,278,255	1,164,744
NG80	9,727,504	1,114	2,593,459	7,546,938	7,554,023	754,764
NG90	5,968,307	386	1,158,564	4,996,169	4,997,906	358,342

	CoBig2	MLK Group	Meraculous	Newbler-454	Phusion	Ray
NG10	0	1,209,031	21,971,264	44,441,969	8,393,225	1,768,581
NG20	0	699,238	19,817,668	38,027,496	5,554,260	1,324,848
NG30	0	435,911	16,720,271	21,998,984	3,194,816	1,032,789
NG40	0	257,153	12,212,215	15,965,151	2,075,433	823,189
NG50	0	150,983	9,834,474	11,495,203	1,479,065	673,924
NG60	0	90,999	6,380,895	7,590,506	995,317	547,665
NG70	0	61,478	4,920,731	5,453,759	633,921	431,830
NG80	0	46,449	3,252,565	2,975,273	377,375	309,595
NG90	0	36,667	1,242,094	1,498,047	160,198	210,806

	SGA	SOAPdenovo	SOAPdenovo*	SOAPdenovo**
NG10	12,778,529	38,231,362	38,200,684	38,202,411
NG20	9,415,776	32,052,133	31,998,732	32,003,806
NG30	6,515,882	24,070,917	23,999,486	23,999,488
NG40	4,771,230	20,086,533	20,006,866	20,016,674
NG50	3,584,181	14,644,723	14,617,523	14,617,693
NG60	1,729,186	11,905,535	11,852,284	11,851,058
NG70	960,227	9,598,542	9,590,173	9,589,746
NG80	375,337	7,180,134	6,941,360	6,942,997
NG90	95,638	4,921,711	4,962,297	4,966,586

(L) Snake

	Platanus	ABySS	BCM-HGSC	CRACS	Curtain	GAM
NG10	47,094,403	1,248,993	3,583,202	1,768,259	127,593	51,871
NG20	29,240,389	949,273	2,713,362	1,349,350	98,649	37,782
NG30	23,496,602	759,330	2,274,804	1,084,842	81,607	29,958
NG40	20,967,559	614,516	1,867,836	903,848	68,639	24,064
NG50	17,165,953	508,539	1,563,800	738,101	58,954	19,350
NG60	12,697,854	408,619	1,267,990	587,518	49,980	15,294
NG70	8,170,376	324,081	978,833	458,950	41,983	11,395
NG80	5,314,452	241,034	717,580	336,507	33,797	7,585
NG90	2,412,727	152,731	429,094	201,554	24,781	3,005

	Meraculous	PRICE	Phusion	Ray	SGA	SOAPdenovo
NG10	3,502,579	6,983	11,845,940	352,064	12,993,518	5,622,710
NG20	2,576,484	2,859	9,156,505	269,788	9,425,130	3,931,095
NG30	1,984,478	0	6,903,134	214,371	7,605,643	2,950,295
NG40	1,549,323	0	5,405,344	176,354	5,740,483	2,475,891
NG50	1,247,790	0	4,494,848	143,603	4,536,273	2,004,523
NG60	972,173	0	3,555,306	116,626	3,540,879	1,652,388
NG70	749,529	0	2,719,556	90,597	2,795,617	1,300,009
NG80	518,879	0	1,954,421	68,068	1,837,486	1,009,391
NG90	283,470	0	1,116,595	43,902	965,325	738,510

	Symbiose
NG10	4,418,467
NG20	3,256,016
NG30	2,555,656
NG40	2,169,012
NG50	1,820,636
NG60	1,511,774
NG70	1,162,919
NG80	832,722
NG90	533,219

(M) Fish

	Platanus	ABySS	Allpaths	BCM-HGSC	CSHL	CSHL*
NG10	8,489,002	3,298,405	12,336,369	17,740,094	14,896,045	14,897,187
NG20	6,392,478	2,539,529	8,670,083	10,852,127	8,654,806	8,655,503
NG30	4,083,009	1,939,393	6,709,969	7,965,905	6,635,705	6,635,838
NG40	3,123,730	1,365,921	4,702,643	6,282,426	4,471,403	4,471,803
NG50	2,371,946	1,013,854	3,677,909	4,850,564	3,418,986	3,418,986
NG60	1,770,377	668,305	2,691,423	3,466,513	2,548,971	2,549,735
NG70	1,089,089	401,184	1,883,411	2,448,799	1,818,535	1,818,535
NG80	430,819	89,403	1,017,865	1,296,719	1,016,930	1,024,270
NG90	529	1,309	148,548	232,649	140,914	143,581

	CSHL**	CTD	CTD*	CTD**	IOBUGA	IOBUGA*
NG10	5,536	1,833	6,501	2,670	830,833	96,292
NG20	3,815	1,452	5,041	2,087	591,655	80,082
NG30	2,803	1,225	4,205	1,741	443,165	66,866
NG40	2,089	1,063	3,606	1,489	342,555	58,018
NG50	1,560	935	3,139	1,285	261,156	51,802
NG60	0	828	2,758	1,114	193,297	47,703
NG70	0	736	2,424	964	127,067	43,018
NG80	0	654	2,126	829	57,655	40,571
NG90	0	580	1,856	706	1,199	40,148

	Meraculous	Ray	SGA	SOAPdenovo	Symbiose
NG10	3,016,460	146,640	432,108	4,529,614	5,522,939
NG20	2,056,696	100,589	274,234	3,602,707	4,048,514
NG30	1,522,853	72,924	190,973	2,705,482	3,058,302
NG40	1,089,609	53,011	131,107	2,079,923	2,229,139
NG50	764,900	37,280	88,883	1,665,791	1,731,822
NG60	478,564	23,116	56,307	1,326,539	1,378,765
NG70	247,526	9,975	36,875	1,010,856	973,031
NG80	20,407	2,011	3,671	739,727	575,050
NG90	188	0	0	461,692	129,181

NG(x) length indicates the length for which the collection of all sequences of that length or longer contains x% of the genome size

NG10-NG90 number of Platanus' scaffold

Supplemental Table 9. NG10-NG90 number of Platanus' scaffold

	NG10 (#)	NG20 (#)	NG30 (#)	NG40 (#)	NG50 (#)	NG60 (#)	NG70 (#)	NG80 (#)	NG90 (#)
C. elegans (heterozygosity: 0.0%)	5	14	25	40	59	83	112	153	218
C. elegans (heterozygosity: 0.1%)	6	14	25	39	57	79	106	142	196
C. elegans (heterozygosity: 0.2%)	5	13	23	35	52	73	99	135	192
C. elegans (heterozygosity: 0.3%)	5	13	24	38	54	75	101	135	192
C. elegans (heterozygosity: 0.5%)	6	15	27	42	60	83	111	148	207
C. elegans (heterozygosity: 1.0%)	6	14	24	38	55	77	103	138	195
C. elegans (heterozygosity: 1.5%)	5	13	24	37	54	75	102	138	197
C. elegans (heterozygosity: 2.0%)	5	13	24	37	53	72	97	133	190
S. venezuelensis	3	7	14	26	44	69	107	173	315
Oyster	45	113	204	316	450	616	823	1,097	1,500
Bird	2	4	7	10	15	20	27	36	51
Snake	3	7	13	19	27	37	51	73	113
Fish	9	22	40	65	99	144	208	337	36,017

NG(x) number indicates the number of sequences of which length are larger than NG(x) length.

Contig NG10-90 length (bp)

Supplemental Table 10. Contig NG10-90 length(bp)

(A) *C.elegans* (0.0%)

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
NG10	192,938	232,734	226,404	493,427	64,326
NG20	132,773	148,008	157,954	353,545	44,993
NG30	105,754	113,873	121,324	273,396	33,345
NG40	83,454	89,207	92,067	220,498	25,367
NG50	66,446	67,596	74,686	171,976	19,084
NG60	50,307	51,346	58,570	135,291	13,928
NG70	37,192	37,916	43,099	109,298	9,609
NG80	25,825	25,363	28,241	77,355	5,967
NG90	13,107	11,950	14,042	44,139	2,462

(B) *C.elegans* (0.1%)

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
NG10	198,400	241,348	228,369	481,772	45,061
NG20	133,450	152,706	165,161	332,567	31,571
NG30	105,276	117,106	125,254	269,881	22,984
NG40	83,660	89,493	94,564	195,019	17,228
NG50	65,717	66,289	77,050	155,501	13,031
NG60	49,697	48,082	61,106	123,786	9,679
NG70	37,148	34,836	44,061	99,441	6,774
NG80	25,322	22,406	30,146	72,090	4,284
NG90	13,699	8,954	16,078	41,678	1,870

(C)*C.elegans (*0.2%)

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
NG10	182,449	154,212	235,712	456,205	24,347
NG20	126,745	111,128	161,994	349,586	17,023
NG30	98,839	81,180	124,031	237,129	13,235
NG40	78,218	61,579	97,066	185,863	10,272
NG50	61,969	45,370	76,439	144,468	8,138
NG60	47,772	33,384	60,345	115,548	6,219
NG70	35,843	21,974	44,884	90,855	4,541
NG80	24,176	12,085	30,342	67,132	2,977
NG90	12,884	3,169	16,678	38,499	1,403

(D) *C.elegans* (0.3%)

	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	164,485	76,478	242,359	415,911	38,745
NG20	116,656	52,399	165,335	283,366	26,970
NG30	93,880	37,795	127,075	219,515	19,690
NG40	75,046	27,826	97,766	173,145	14,288
NG50	59,795	19,629	79,194	141,590	10,182
NG60	44,404	12,685	61,631	114,440	7,031
NG70	33,057	6,249	45,941	89,237	4,541
NG80	22,817	3,191	32,603	64,772	2,521
NG90	12,144	2,265	17,687	36,775	844

(E) *C.elegans* (0.5%)

	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	164,394	15,825	253,792	334,037	31,569
NG20	111,998	7,947	180,389	226,597	21,433
NG30	87,618	3,712	132,925	178,201	15,507
NG40	67,621	3,144	103,556	147,712	11,295
NG50	53,873	2,807	80,813	117,418	8,149
NG60	42,268	2,527	63,948	96,025	5,642
NG70	30,808	2,250	47,699	74,562	3,632
NG80	20,950	1,930	33,728	49,603	2,056
NG90	10,716	1,379	18,236	28,462	676

(F) *C.elegans* (1.0%)

	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	143,669	2,518	238,294	310,724	14,339
NG20	105,778	2,111	184,274	210,488	10,267
NG30	77,587	1,830	135,419	159,713	7,712
NG40	60,117	1,594	106,132	123,655	5,914
NG50	48,090	1,358	84,827	103,677	4,449
NG60	36,611	1,060	66,862	81,752	3,185
NG70	26,364	0	49,690	62,064	2,172
NG80	18,213	0	34,309	44,650	1,291
NG90	9,081	0	19,292	22,672	466

(G) *C.elegans* (1.5%)

	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	126,966	2,414	212,609	239,276	2,954
NG20	90,723	1,958	158,619	182,872	1,973
NG30	69,584	1,650	124,361	143,873	1,433
NG40	54,402	1,384	95,359	114,745	1,049
NG50	42,013	1,103	75,918	90,768	766
NG60	32,391	0	60,465	71,544	540
NG70	24,134	0	46,913	53,645	344
NG80	16,318	0	33,222	38,212	169
NG90	7,793	0	18,797	20,637	54

(H) *C.elegans* (2.0%)

	Platanus	**ALLPATHS-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	115,537	3,373	167,527	207,610	2,490
NG20	83,739	2,725	124,880	139,001	1,669
NG30	63,810	2,365	96,613	101,961	1,215
NG40	50,834	2,070	78,570	78,465	901
NG50	39,915	1,810	62,656	60,761	666
NG60	29,911	1,542	50,683	46,279	475
NG70	21,698	1,244	38,755	34,863	310
NG80	14,438	594	27,688	24,035	163
NG90	6,725	0	15,584	12,959	60

(I) *S.venezuelensis*

	Platanus	**Allpaths-LG**	**MaSuRCA**	**SOAPdenovo2**	**Velvet**
NG10	294,286	17,901	249,158	159,770	6,594
NG20	180,007	10,929	186,992	118,350	4,519
NG30	132,006	5,766	143,163	82,821	3,322
NG40	101,406	2,824	110,833	63,257	2,530
NG50	71,357	2,008	84,739	48,010	1,946
NG60	48,673	1,632	61,140	31,658	1,497
NG70	27,520	1,343	42,012	13,806	1,132
NG80	10,176	1,030	21,664	4,721	825
NG90	3,177	0	8,590	338	570

(J) Oyster

	Platanus	ALLPATHS-LG	SOAPdenovo2	Reference
NG10	25,770	34,313	28,293	76,689
NG20	18,536	24,195	20,944	56,394
NG30	14,326	18,763	16,863	43,642
NG40	11,341	15,068	13,958	34,329
NG50	9,011	12,025	11,719	26,430
NG60	7,051	9,519	9,785	19,740
NG70	5,363	7,221	8,180	13,642
NG80	3,849	5,046	6,754	7,461
NG90	2,453	2,785	5,424	0

(K) Bird

	Platanus	ABL	Allpaths	BCM-HGSC	BCM-HGSC*	CBCB
NG10	164,476	11,895	208,498	728,856	378,150	324,259
NG20	116,174	7,252	141,346	482,370	259,905	232,760
NG30	90,439	5,160	103,105	348,475	192,960	183,004
NG40	71,602	3,903	76,412	256,506	149,767	146,787
NG50	56,973	3,003	57,617	184,968	115,689	116,375
NG60	44,679	2,293	42,074	133,777	86,391	90,165
NG70	33,585	1,682	29,668	92,909	61,424	67,182
NG80	23,385	1,114	19,832	55,761	40,829	47,373
NG90	13,069	386	10,847	26,309	20,980	27,380

	CoBig2	MLK Group	Meraculous	Newbler-454	Phusion	Ray
NG10	0	82,626	124,439	201,391	310,885	114,381
NG20	0	61,784	84,465	144,150	198,285	84,074
NG30	0	50,156	62,527	110,233	141,283	67,642
NG40	0	42,339	48,067	85,730	101,039	54,888
NG50	0	36,470	37,061	66,030	74,321	44,663
NG60	0	31,498	28,037	49,785	51,814	35,808
NG70	0	27,592	20,287	36,342	34,878	28,062
NG80	0	24,240	13,122	23,139	20,881	20,963
NG90	0	21,156	5,674	10,991	9,719	13,846

	SGA	SOAPdenovo	SOAPdenovo*	SOAPdenovo**
NG10	54,640	129,237	151,153	148,108
NG20	38,313	89,472	108,749	106,806
NG30	29,152	68,217	85,454	83,834
NG40	22,770	52,723	68,607	67,108
NG50	17,785	40,684	54,932	53,557
NG60	13,382	30,264	43,288	42,163
NG70	9,398	22,075	32,955	31,916
NG80	5,665	14,691	23,558	22,632
NG90	1,769	8,072	13,987	13,417

(L) Snake

	Platanus	ABySS	BCM-HGSC	CRACS	Curtain	GAM
NG10	124,221	70,553	35,293	43,137	13,224	12,521
NG20	93,609	53,301	25,883	32,275	9,430	9,154
NG30	74,263	42,690	20,282	25,678	7,241	7,149
NG40	60,205	34,985	16,096	20,834	5,694	5,683
NG50	48,614	28,651	12,727	16,852	4,476	4,499
NG60	38,536	23,262	9,776	13,385	3,435	3,477
NG70	29,548	18,314	7,129	10,245	2,487	2,538
NG80	20,730	13,587	4,479	7,221	1,582	1,609
NG90	11,715	8,657	1,373	3,818	603	552

	Meraculous	PRICE	Phusion	Ray	SGA	SOAPdenovo
NG10	92,490	6,983	198,026	44,540	73,992	44,916
NG20	68,539	2,859	147,670	33,899	54,907	34,102
NG30	54,112	0	118,241	27,138	43,581	27,555
NG40	43,035	0	95,908	22,241	35,273	22,716
NG50	34,344	0	77,302	18,158	28,397	18,848
NG60	27,073	0	61,719	14,597	22,468	15,525
NG70	20,347	0	47,472	11,312	16,972	12,519
NG80	13,883	0	33,533	8,120	11,663	9,687
NG90	6,786	0	18,187	4,542	5,857	7,016

	Symbiose
NG10	219,233
NG20	164,409
NG30	129,320
NG40	102,520
NG50	82,841
NG60	65,963
NG70	50,398
NG80	35,826
NG90	20,613

(M) Fish

	Platanus	ABySS	Allpaths	BCM-HGSC	CSHL	CSHL*
NG10	28,255	15,141	51,557	64,735	63,310	62,337
NG20	19,083	10,516	35,844	45,188	43,750	43,295
NG30	13,706	7,734	26,479	33,550	32,143	31,694
NG40	9,780	5,688	19,630	25,064	23,616	23,354
NG50	6,587	4,030	14,105	17,895	16,740	16,517
NG60	3,771	2,550	9,378	11,918	10,972	10,858
NG70	1,110	1,074	4,955	6,482	5,822	5,770
NG80	0	0	0	927	0	0
NG90	0	0	0	0	0	0

	CSHL**	CTD	CTD*	CTD**	IOBUGA	IOBUGA*
NG10	5,529	1,833	6,501	2,670	6,073	2,015
NG20	3,807	1,452	5,041	2,087	4,149	1,397
NG30	2,796	1,225	4,205	1,741	2,978	1,033
NG40	2,081	1,063	3,606	1,489	2,144	764
NG50	1,550	935	3,139	1,285	1,472	556
NG60	0	828	2,758	1,114	0	388
NG70	0	736	2,424	964	0	242
NG80	0	654	2,126	829	0	127
NG90	0	580	1,856	706	0	0

	Meraculous	Ray	SGA	SOAPdenovo	Symbiose
NG10	17,596	22,802	20,520	22,950	117,549
NG20	11,878	16,008	13,929	16,346	81,516
NG30	8,429	12,075	9,977	12,309	60,014
NG40	5,984	9,193	7,088	9,378	43,292
NG50	3,962	6,829	4,739	7,001	30,444
NG60	2,115	4,774	2,690	4,932	19,197
NG70	526	2,670	980	3,050	9,196
NG80	0	451	0	1,173	2,444
NG90	0	0	0	0	981

NG(x) length indicates the length for which the collection of all sequences of that length or longer contains x% of the genome size

Statistics of assemblies of purely simulated data (*C. elegans*)

The pure simulation tests have been performed. Using the simulator of Illumina data (pIRS, Hu et al. 2012), we generated the data from *C. elegans* genome with the same coverage depth and insert sizes as the data in the manuscript (Supplemental Table 1). Heterozygosities were simulated in two ways, 0% and 1%. The results are displayed in Supplemental Table11.

Supplemental Table 11. Statistics of assemblies of purely simulated data (*C. elegans*)

(A) Heterozygosity: 0%

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Total (≥500 bp)	99,095,271	99,244,554	99,368,655	99,316,698	100,242,764
Number of scaffolds (≥500 bp)	623	446	1,930	486	696
Scaffold NG50 (bp)	1,662,234	954,649	494,503	1,233,697	1,827,862
Corrected scaffold NG50 (bp)	673,078	741,372	164,950	774,682	834,325
Contig NG50 (bp)	267,789	77,347	122,258	298,574	50,874
Corrected contig NG50 (bp)	134,027	67,354	65,769	84,248	47,498
Number of errors	241	145	858	481	296

(B) Heterozygosity: 1%

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Total (≥500 bp)	100,720,046	63,315,575	105,861,126	99,542,464	105,804,436
Number of scaffolds (≥500 bp)	666	5,075	7,082	1,341	1,917
Scaffold NG50 (bp)	6,157,496	10,165	1,746,798	417,346	306,871
Corrected scaffold NG50 (bp)	1,229,293	1,234	600,124	319,602	35,926
Contig NG50 (bp)	143,138	1,006	266,923	136,170	6,404
Corrected contig NG50 (bp)	72,476	925	75,980	10,825	5,879
Number of errors	187	359	483	1,160	5,814

The trend is similar to HiSeq2000-based simulation, that is, only Platanus and MaSuRCA do not decrease (corrected) NG50 according to the increase of heterozygosity, and the numbers of errors of Platanus is less than those of MaSuRCA. However, all cases but one (ALLPATHS-LG in heterozygosity 1%) indicate larger scaffold NG50 compared to the real-data tests (HiSeq2000-simulation-free C. elegans data and S. venezuelensis data, see Supplemental Table 2 and Table 4). The differences are significant especially for Platanus and MaSuRCA, though two paired-ends and one

mate-pair libraries were prepared for all tests of nematodes.

Additionally, to investigate whether ALLPATHS-LG successfully accepted 230bp-paired-end of HiSeq2000-C.elegans data (Supplemental Table 1), we generated another data sets (insert-sizes of 180 bp, 430 bp and 4.7 kbp), and compared with abovementioned data sets (insert-sizes of 230bp, 430bp and 4.7kbp). Both data had the same total size as HiSeq2000 data, and heterozygosities were set as 0% and 1%. The assembly result is shown in the Supplemental Table 12.

Supplemental Table 12. Test of ALLPATHS-LG with overlapping paired-ends (simulation data, *C.elegans*)

	Heterozygosity: 0%		**Heterozygosity: 1%**	
	Shortest-insert: 180	**Shortest-insert: 230**	**Shortest-insert: 180**	**Shortest-insert: 230**
Total (≥500 bp)	99,359,985	99,244,554	60,777,760	63,315,575
Number of scaffolds (≥500 bp)	419	446	7,584	5,075
Scaffold NG50 (bp)	1,020,326	954,649	5,586	10,165
Corrected scaffold NG50 (bp)	851,821	741,372	1,183	1,234
Contig NG50 (bp)	101,353	77,347	880	1,006
Corrected contig NG50 (bp)	79,230	67,354	851	925
Number of errors	161	145	343	359

Replacing 230bp-paired-end with 180bp-paired-end, corrected scaffold NG50 increased by 13% when heterozygosity was 0%. Unexpectedly, that replacement decreased corrected scaffold NG50 when heterozygosity was 1%. These results infer that ALLPATHS-LG does not miss the performance significantly and can accept 230bp-paired-end as overlapping paired-end.

Turning-off tests for Platanus' features

We executed benchmark test turning off Platanus' features (1) *k*-mer extension, (2) bubble-removal in contig-assembly, (3) removal of hetero-regions in scaffolding. Note that "hetero removal of hetero-regions in scaffolding" includes both bubble-removal and branch-cut. Inputs were *C. elegans* (heterozygosity: 0%, 1%, 2%)) and *S. venezuelensis* data described in the manuscript. The results are shown in Supplemental Table 13.

Supplemental Table 13. Results of turning-off tests for Platanus' features

(A) C.elegans (heterozygosity: 0%)

	Default	**− (*k*-mer-extension)**	**− (bubble-removal in contig-assembly)**	**− (hetero-removal in scaffolding)**
Total (≥500 bp)	97,938,690	96,528,882	98,022,430	98,715,700
Number of scaffolds (≥500 bp)	942	1,953	852	1,214
Scaffold NG50 (bp)	478,744	306,733	481,025	415,878
Corrected scaffold NG50 (bp)	361,608	226,959	375,389	334,052
Contig NG50 (bp)	66,446	35,712	67,574	66,447
Corrected contig NG50 (bp)	48,054	30,899	49,046	48,800
Number of errors	367	257	342	274

(B) *C.elegans* (heterozygosity: 1%)

	Default	**− (*k*-mer-extension)**	**− (bubble-removal in contig-assembly)**	**− (hetero-removal in scaffolding)**
Total (≥500 bp)	99,223,085	96,211,233	140,267,578	100,641,997
Number of scaffolds (≥500 bp)	1,289	2,193	24,301	4,579
Scaffold NG50 (bp)	511,190	291,355	45,673	381,587
Corrected scaffold NG50 (bp)	368,857	204,518	20,728	195,751
Contig NG50 (bp)	48,090	8,887	6,268	14,472
Corrected contig NG50 (bp)	39,291	8,547	5,970	13,665
Number of errors	290	160	203	228

(C) *C.elegans* (heterozygosity: 2%)

	Default	**− (*k*-mer-extension)**	**− (bubble-removal in contig-assembly)**	**− (hetero-removal in scaffolding)**
Total (≥500 bp)	100,454,422	96,157,873	183,642,064	108,879,145
Number of scaffolds (≥500 bp)	2,264	2,434	10,902	17,986
Scaffold NG50 (bp)	580,832	270,290	81,927	207,088
Corrected scaffold NG50 (bp)	341,914	204,425	26,451	100,787
Contig NG50 (bp)	39,915	6,704	22,094	7,545
Corrected contig NG50 (bp)	34,030	6,517	16,994	7,316
Number of errors	256	129	146	170

(D) *S. venezuelensis*

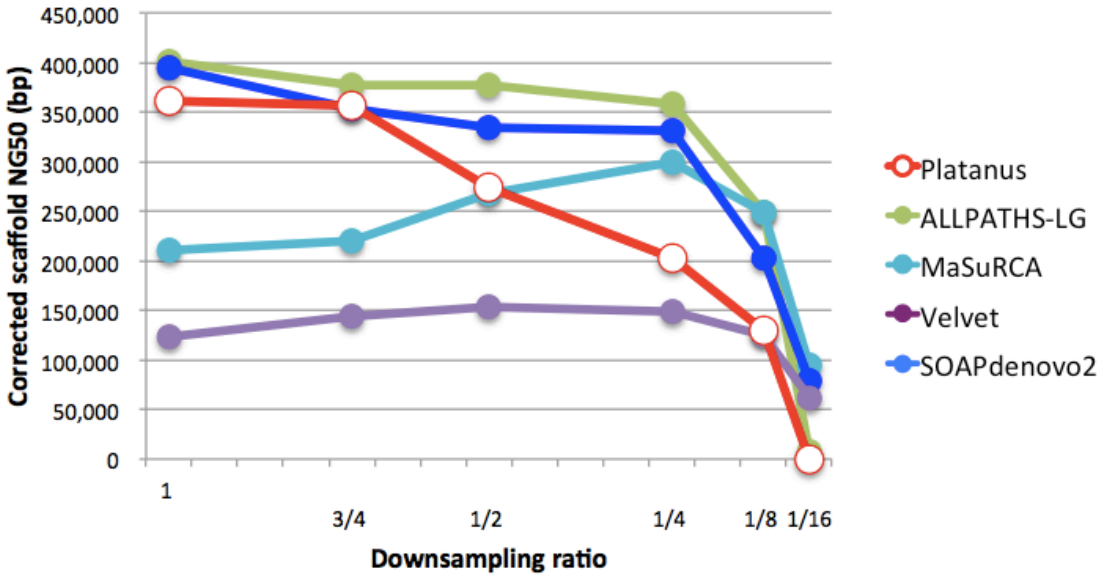
	Default	**− (*k*-mer-extension)**	**− (bubble-removal in contig-assembly)**	**− (hetero-removal in scaffolding)**
Total (≥500 bp)	58,503,689	45,493,271	74,553,871	60,793,871
Number of scaffolds (≥500 bp)	2,560	1,660	9,304	7,026
Scaffold NG50 (bp)	274,622	127,238	52,920	108,844
Contig NG50 (bp)	71,357	15,708	7,754	15,983

When k-mer extension was turned off, scaffold (corrected) NG50 decreased in all cases, inferring that k-mer extension is fundamentally important function. As expected, deactivations of bubble-removal in contig-assembly and hetero-removal in scaffolding reduced scaffold (corrected) NG50 for heterozygous input data. Therefore, it is indicated that these features are effective for heterozygous samples.

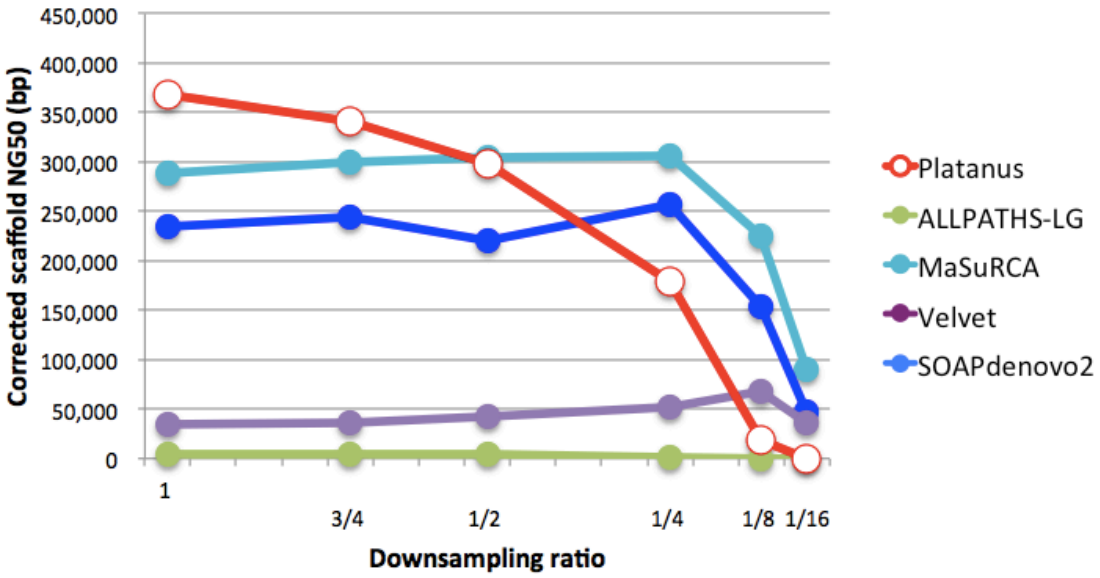
Downsampling benchmark test (*C. elegans*)

We performed the benchmark test using reduced amount of sequence data for *C.elegans* (heterozygosity: 0%, 1%, 2%) to measure the impact for the assembling result of sequence amounts (Downsampling benchmark test). Downsampled datum sizes are 3/4, 1/2, 1/4, 1/8, and 1/16. Results are displayed in the Supplemental Figure 28 and the Supplemental Table 14.

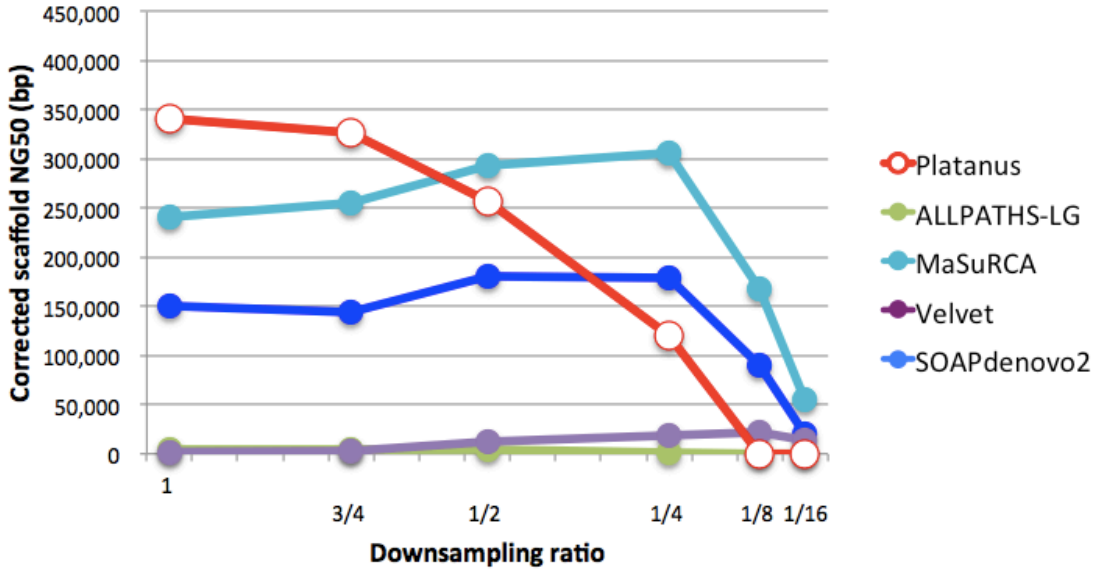
(A) *C. elegans* (heterozygosity: 0%)



(B) *C. elegans* (heterozygosity: 1%)



(C) *C. elegans* (heterozygosity: 2%)



Supplemental Figure 28. Corrected scaffold NG50 in downsampling tests

Supplemental Table 14. Detail statistics of downsampling benchmark tests

(A) *C. elegans* (heterozygosity: 0%)

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Downsampling ratio: 1/1					
Total (≥500 bp)	97,938,690	98,864,571	97,742,565	99,132,377	99,382,943
Number of scaffolds (≥500 bp)	942	619	1,369	794	1,375
Scaffold NG50 (bp)	478,744	466,658	420,694	507,513	436,119
Corrected scaffold NG50 (bp)	361,608	402,062	210,683	394,556	123,438
Contig NG50 (bp)	66,446	67,596	74,686	170,395	19,084
Corrected contig NG50 (bp)	48,054	46,361	40,248	45,927	19,938
Number of errors	367	272	900	407	1,490
Downsampling ratio: 3/4					
Total (≥500 bp)	97,880,813	98,763,216	98,479,896	99,071,997	99,497,441
Number of scaffolds (≥500 bp)	1,012	602	1,264	865	1,329
Scaffold NG50 (bp)	455,094	480,286	506,779	476,394	430,832
Corrected scaffold NG50 (bp)	356,829	378,418	221,125	353,562	144,941
Contig NG50 (bp)	65,256	68,458	66,167	155,175	21,808
Corrected contig NG50 (bp)	47,219	45,967	39,721	43,981	20,267
Number of errors	375	325	854	536	1,189
Downsampling ratio: 1/2					
Total (≥500 bp)	98,130,065	98,726,036	99,257,018	98,983,521	99,261,376
Number of scaffolds (≥500 bp)	1,143	642	1,114	974	1,401
Scaffold NG50 (bp)	417,069	428,223	461,620	421,981	387,445
Corrected scaffold NG50 (bp)	273,728	377,702	268,513	334,428	154,132
Contig NG50 (bp)	56,056	65,793	58,087	153,213	20,702
Corrected contig NG50 (bp)	40,595	42,947	38,256	43,263	19,351
Number of errors	400	304	855	515	1,113
Downsampling ratio: 1/4					
Total (≥500 bp)	96,416,108	98,696,419	98,945,932	99,012,707	98,893,576
Number of scaffolds (≥500 bp)	2,577	736	1,016	971	1,774
Scaffold NG50 (bp)	257,681	428,259	480,121	421,984	343,426
Corrected scaffold NG50 (bp)	203,642	358,152	300,653	332,270	148,787
Contig NG50 (bp)	27,924	52,614	42,143	131,448	14,409
Corrected contig NG50 (bp)	24,731	35,783	30,748	40,472	13,745
Number of errors	244	285	861	555	989

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Downsampling ratio: 1/8					
Total (≥500 bp)	93,770,232	98,474,312	99,623,297	98,677,012	99,049,859
Number of scaffolds (≥500 bp)	6,758	1,397	1,335	1,954	2,445
Scaffold NG50 (bp)	170,745	304,508	437,750	254,468	264,900
Corrected scaffold NG50 (bp)	129,402	248,726	248,776	203,303	125,794
Contig NG50 (bp)	15,087	27,920	24,069	64,621	11,379
Corrected contig NG50 (bp)	13,596	20,237	19,554	28,438	10,701
Number of errors	230	464	1,038	757	1,410
Downsampling ratio: 1/16					
Total (≥500 bp)	10,838	75,698,629	101,746,165	98,510,063	100,385,334
Number of scaffolds (≥500 bp)	6	7,440	4,962	6,963	6,887
Scaffold NG50 (bp)	0	21,120	173,469	121,000	135,774
Corrected scaffold NG50 (bp)	0	6,899	94,498	79,674	62,353
Contig NG50 (bp)	0	1,016	3,489	10,522	3,673
Corrected contig NG50 (bp)	0	480	3,255	5,226	3,461
Number of errors	8	812	1,795	1,358	1,629

(B) *C. elegans* (heterozygosity: 1%)

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Downsampling ratio: 1/1					
Total (≥500 bp)	99,223,085	163,953,538	100,913,677	99,086,653	103,773,061
Number of scaffolds (≥500 bp)	1,289	5,061	2,803	1,575	2,642
Scaffold NG50 (bp)	511,190	91,413	466,806	280,050	243,330
Corrected scaffold NG50 (bp)	368,857	4,928	288,752	234,395	34,491
Contig NG50 (bp)	48,090	1,358	84,827	103,677	4,445
Corrected contig NG50 (bp)	39,291	1,316	44,746	16,784	4,360
Number of errors	290	542	895	1,226	6,285
Downsampling ratio: 3/4					
Total (≥500 bp)	99,163,809	150,279,673	100,225,734	98,818,960	103,129,963
Number of scaffolds (≥500 bp)	1,341	5,758	2,509	1,410	2,569
Scaffold NG50 (bp)	504,880	69,518	499,699	298,021	217,593
Corrected scaffold NG50 (bp)	340,818	4,860	299,378	244,472	36,432
Contig NG50 (bp)	42,275	1,107	71,502	106,094	4,329
Corrected contig NG50 (bp)	35,778	1,077	43,120	18,000	4,135
Number of errors	315	657	829	1,189	5,531

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Downsampling ratio: 1/2					
Total (≥500 bp)	98,595,467	126,737,588	100,137,899	100,939,856	102,641,002
Number of scaffolds (≥500 bp)	1,330	6,657	2,001	2,474	2,544
Scaffold NG50 (bp)	479,535	42,406	532,649	310,326	221,497
Corrected scaffold NG50 (bp)	297,513	4,837	305,051	221,003	42,094
Contig NG50 (bp)	28,493	0	61,245	98,439	4,209
Corrected contig NG50 (bp)	25,136	0	38,721	5,178	4,044
Number of errors	353	519	887	859	4,875
Downsampling ratio: 1/4					
Total (≥500 bp)	95,754,933	69,626,022	99,518,376	98,797,170	101,657,122
Number of scaffolds (≥500 bp)	3,567	7,934	1,172	1,460	2,777
Scaffold NG50 (bp)	224,151	9,946	482,535	339,915	198,489
Corrected scaffold NG50 (bp)	179,349	1,158	306,847	257,520	52,811
Contig NG50 (bp)	7,358	0	40,093	83,253	3,973
Corrected contig NG50 (bp)	7,080	0	29,217	17,432	3,821
Number of errors	143	399	922	1,011	3,804
Downsampling ratio: 1/8					
Total (≥500 bp)	83,767,296	57,815,377	99,374,458	97,971,268	101,381,085
Number of scaffolds (≥500 bp)	16,206	9,328	1,357	2,710	3,650
Scaffold NG50 (bp)	34,922	4,468	418,564	196,971	200,010
Corrected scaffold NG50 (bp)	18,716	0	225,193	153,697	67,885
Contig NG50 (bp)	1,069	0	15,973	42,112	4,126
Corrected contig NG50 (bp)	1,047	0	13,417	15,796	3,971
Number of errors	214	473	1,254	1,080	2,401
Downsampling ratio: 1/16					
Total (≥500 bp)	7,978	13,377,870	101,657,657	98,102,179	102,069,954
Number of scaffolds (≥500 bp)	2	5,795	6,085	9,417	9,303
Scaffold NG50 (bp)	0	0	200,674	81,641	109,050
Corrected scaffold NG50 (bp)	0	0	90,557	47,639	35,800
Contig NG50 (bp)	0	0	2,327	6,257	2,307
Corrected contig NG50 (bp)	0	0	2,197	3,344	2,216
Number of errors	11	249	1,854	1,327	2,135

(C) *C. elegans* (heterozygosity: 2%)

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Downsampling ratio: 1/1					
Total (≥500 bp)	100,454,422	170,519,463	105,340,084	99,546,387	110,798,448
Number of scaffolds (≥500 bp)	2,264	6,467	4,763	1,727	10,692
Scaffold NG50 (bp)	580,832	86,979	351,406	212,590	179,009
Corrected scaffold NG50 (bp)	341,914	5,178	240,778	150,059	1,932
Contig NG50 (bp)	39,915	1,810	62,656	60,761	666
Corrected contig NG50 (bp)	34,030	1,648	38,465	4,013	1,721
Number of errors	256	496	774	1,839	36,840
Downsampling ratio: 3/4					
Total (≥500 bp)	100,146,918	160,635,826	104,758,302	99,482,324	110,747,609
Number of scaffolds (≥500 bp)	2,075	7,164	4,781	1,797	8,891
Scaffold NG50 (bp)	563,178	72,054	429,998	215,803	169,528
Corrected scaffold NG50 (bp)	326,921	5,071	255,751	143,567	2,139
Contig NG50 (bp)	35,293	1,667	62,279	61,553	689
Corrected contig NG50 (bp)	30,515	1,504	38,095	4,219	636
Number of errors	294	645	763	1,808	37,489
Downsampling ratio: 1/2					
Total (≥500 bp)	98,449,434	144,734,796	103,810,835	99,595,608	112,970,331
Number of scaffolds (≥500 bp)	1,838	8,459	4,690	2,263	5,851
Scaffold NG50 (bp)	395,336	50,013	473,524	232,986	152,357
Corrected scaffold NG50 (bp)	256,464	5,056	294,279	181,402	12,114
Contig NG50 (bp)	18,024	1,353	58,528	61,235	1,807
Corrected contig NG50 (bp)	16,441	1,207	36,779	4,562	1,727
Number of errors	228	613	776	1,100	10,950
Downsampling ratio: 1/4					
Total (≥500 bp)	94,120,074	75,874,771	101,653,790	99,570,849	110,355,220
Number of scaffolds (≥500 bp)	5,784	10,058	3,271	2,265	5,600
Scaffold NG50 (bp)	180,087	10,725	512,796	232,827	145,335
Corrected scaffold NG50 (bp)	121,055	1,565	305,790	179,223	19,312
Contig NG50 (bp)	3,967	0	38,113	47,175	1,765
Corrected contig NG50 (bp)	3,832	0	27,471	4,372	1,683
Number of errors	143	462	895	1,198	7,470

	Platanus	ALLPATHS-LG	MaSuRCA	SOAPdenovo2	Velvet
Downsampling ratio: 1/8					
Total (≥500 bp)	43,974,267	26,734,177	101,263,467	98,331,567	109,333,015
Number of scaffolds (≥500 bp)	34,186	8,090	2,974	6,820	7,489
Scaffold NG50 (bp)	301	0	361,688	126,053	135,460
Corrected scaffold NG50 (bp)	260	0	168,031	89,918	22,461
Contig NG50 (bp)	229	0	10,001	23,852	1,639
Corrected contig NG50 (bp)	226	0	8,682	6,293	1,587
Number of errors	103	284	1,482	863	4,533
Downsampling ratio: 1/16					
Total (≥500 bp)	12,470	3,682,752	102,854,429	99,920,871	108,539,408
Number of scaffolds (≥500 bp)	5	2,262	8,985	16,041	15,327
Scaffold NG50 (bp)	0	0	144,551	37,972	64,141
Corrected scaffold NG50 (bp)	0	0	55,633	20,166	13,202
Contig NG50 (bp)	0	0	1,575	3,616	1,272
Corrected contig NG50 (bp)	0	0	1,506	2,017	1,236
Number of errors	4	116	2,020	1,234	2,589

As shown in Supplemental Figure 28, for corrected scaffold NG50, we revealed that Platanus was sensitive to the downsampling effect. The fewer amount of input data, the shorter correct scaffold NG50 were obtained. This tendency was applicable for every test sets. For the other assemblers, these tendencies were not observed. Especially for MaSuRCA and SOAPdenovo2, the results from greater than 1/4 (about x35 sequence coverage) amount of input datum, almost same length of corrected NG50 were gained.

These observations imply that for MaSuRCA and SOAPdenovo2, about x35 sequence redundancy were sufficient to assemble and reached plateaus, so even much more amount of sequences were given the corrected scaffold NG50 were not grown. In contrast, Platanus' optimum coverage might be at 100% (x140 sequence redundancy) or much more amount of sequences. But in the test of 1% and 2% heterozygosity, even comparing the best corrected scaffold NG50 of each assembler marked among all the redundancy, Platanus marked the best statistics within de Bruijn based assemblers and slightly better results than overlap-layout based assembler, MaSuRCA.

References

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.

Hu X, Yuan J, Shi Y, Lu J, Liu B, Li Z, Chen Y, Mu D, Zhang H, Li N, Yue Z, Bai F, Li H, Fan W; pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics* **28:** 1533-1535.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biology* **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. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078-2079.