Sequence analysis

Assembling short reads from jumping libraries with large insert sizes

Irina Vasilinetc1, Andrey D. Prijbelski1,2,*, Alexey Gurevich1,2, Anton Korobeynikov1,2,3 and Pavel A. Pevzner2,4

1Algorithmic Biology Lab, St. Petersburg Academic University 194021, 2Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St.Petersburg State University, 199004, 3Department of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg, 198504, Russia and 4Department of Computer Science and Engineering, University of California, San Diego, CA 92093-0404, USA

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on February 12, 2015; revised on May 3, 2015; accepted on May 26, 2015

Abstract

Motivation: Advances in Next-Generation Sequencing technologies and sample preparation recently enabled generation of high-quality jumping libraries that have a potential to significantly improve short read assemblies. However, assembly algorithms have to catch up with experimental innovations to benefit from them and to produce high-quality assemblies.

Results: We present a new algorithm that extends recently described EXSPANDER universal repeat resolution approach to enable its applications to several challenging data types, including jumping libraries generated by the recently developed Illumina Nextera Mate Pair protocol. We demonstrate that, with these improvements, bacterial genomes often can be assembled in a few contigs using only a single Nextera Mate Pair library of short reads.

Availability and implementation: Described algorithms are implemented in C++ as a part of SPAdes genome assembler, which is freely available at bioinf.spbau.ru/en/spades.

Contact: ap@bioinf.spbau.ru

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

In an article titled ‘De novo fragment assembly with short mate-paired reads: Does the read length matter?’, Chaisson et al. (2009) argued that availability of paired reads with long and accurate insert sizes (rather than the increase in the read length) is the most important factor for improving the quality of short read assemblies. However, while paired reads with long insert sizes have been extensively used in many assembly projects, robust generation of read-pairs with accurate insert sizes proved to be difficult and have only been achieved recently.

The recently emerged sample preparation technologies open new opportunities for genome assembly from short reads. For example, Illumina Nextera Mate Pair protocol generates long inserts (3 kb and longer) that feature rather tight insert size distribution and small rate of non-circularized fragments that result in read-pairs with abnormal distances. As discussed in Chaisson et al. (2009), such read-pair libraries may enable assemblies approaching the quality of assemblies from long reads of length equal to the insert size. Moreover, they can potentially substitute the existing assembly approaches based on a combination of short paired-end libraries (with insert size less than 1 kb) and long jumping libraries (with insert sizes typically longer than 1 kb) by a pipeline based on a single Nextera Mate Pair library.

However, even though the popular assembly algorithms perform well with the previously proposed approaches to sample preparation, they have not kept up with recent experimental innovations. To catch up, bioinfomaticians either need to design novel tools for every technology improvement or to develop a universal assembler that can be easily modified to support new data types. For example, Ray (Boisvert et al., 2010) and SPAdes (Bankevich et al., 2012)
[with the recently implemented exSPANder (Prjibelski et al., 2014) algorithm] use similar approach to repeat resolution that can be adapted for new types of sequencing data. However, the recently proposed Nextera Mate Pair libraries present new computational challenges that go beyond the algorithmic framework of both Ray and SPAdes. In this work, we describe several algorithmic advances to exSPANder that allows one to efficiently utilize jumping libraries and to perform assembly using only high-quality Nextera Mate Pair libraries.

The exSPANder algorithm (Prjibelski et al., 2014) is based on the path extension framework that was proposed by the authors of the Ray assembler (Boisvert et al., 2010) and later implemented in the Telescopier (Bresler et al., 2012) and PERGA (Zhu et al., 2014) assemblers. Given a path \( P \) in the assembly graph (Bankevich et al., 2012; Prjibelski et al., 2014), exSPANder iteratively attempts to grow it by choosing one of its extension edges (all edges starting at the terminal vertex of the path \( P \)). The assembly graph is defined as simplified de Brujin graph (Pevzner et al., 2001) of \( k \)-mers in reads after removal of bulges, tips and chimeric edges.

To extend a path \( P \), exSPANder computes the scoring function \( \text{Score}(e) \) for each extension edge \( e \) using read-pairs with one read mapping to \( P \) and another read mapping to \( e \) (referred to as \((P, e)\)-connecting read-pairs). Afterward, exSPANder decides whether to select the top-scoring extension edge or to stop growing \( P \). It iteratively repeats the path extension procedure starting with single-edge paths until every edge in the assembly graph is covered by at least one path and no path can be extended further. To generate equivalent contigs on both strands, exSPANder is implemented as a bidirectional approach that can extend a path in both directions.

While the scoring function \( \text{Score}(e) \) described in Prjibelski et al. (2014) works well with short libraries, it appears to be rather inefficient when using jumping libraries. The key limitation of the previously defined scoring function is that it analyses only \((P, e)\)-connecting read-pairs (where \( e \) is an extension edge of path \( P \)) and ignores read-pairs that connect path \( P \) with other edges. When an edge \( e \) is short and the variations in the insert sizes are large, there is a danger that no \((P, e)\)-connecting read-pairs exist and thus \( \text{Score}(P,e) = 0 \) even if \( e \) is the correct extension edge (Fig. 1). Thus, the decision rule may stop extending path \( P \) even before selecting an incorrect extension edge. Additionally, the approach described in Prjibelski et al. (2014) is inapplicable for scaffolding procedure, since it is unable to jump over coverage gaps. In this article, we extend the exSPANder approach to scaffolding. This extension is important since scaffolding with jumping libraries may dramatically improve the assembly quality.

We describe several algorithms that address these bottlenecks based on the following idea. Consider a set of extension paths \( \mathcal{E} \) [rather than extension edges as in Prjibelski et al. (2014)] that contain all sufficiently long paths (longer than the insert size) starting from the extension edges of the path \( P \). Once the set \( \mathcal{E} \) is constructed, we choose the best-scoring path \( E \) in \( \mathcal{E} \) and extend path \( P \) by the first edge of \( E \). Our analysis has shown that such conservative extension (by the first edge of the best-scoring extension path rather than by the entire path) provides more accurate assemblies. To perform scaffolding procedure, we allow extension paths to ‘jump’ over coverage gaps in the assembly graph.

This intuitive approach, while appealing, is often impractical since the assembly graph is usually tangled, resulting in a prohibitively large number of extension paths. To reduce the running time, we have implemented the new algorithm based on the observation that, instead of the exhaustive search through the set of all extension paths, one can significantly prune this set using single reads and paired-end libraries (if available).

We demonstrate that the new algorithm enables assemblies of nearly complete genomes from a single Nextera Mate Pair library. We also show that SPAdes, coupled with the improved exSPANder algorithm, outperforms other popular assemblers, such as ABySS (Simpson et al., 2009), IDBA-UD (Peng et al., 2012), Ray (Boisvert et al., 2010), SOAPdenovo (Li et al., 2010) and Velvet (Zerbino and Birney, 2008) on various types of datasets.

2 Methods

2.1 The exSPANder framework

2.1.1 Rectangles

exSPANder utilizes the concept of the rectangle graph introduced by Bankevich et al. (2012) and further developed in (Vyahhi et al., 2012). Let \( e \) and \( e' \) be edges in the assembly graph (see (Bankevich et al., 2012) for the construction of the assembly graph) separated by a known distance \( D \) and \((r,r')\) be a read-pair, such that read \( r \) maps to \( e \) at position \( x_0 \) and read \( r' \) maps to \( e' \) at position \( y_0 \) (Fig. 2a). We note that while \( D \) is not known a priori, exSPANder only considers edges that are connected by a previously constructed path, which unambiguously defines distance \( D \).

Figure 2b shows the read-pair \((r,r')\) represented as a point \((x_0, y_0)\) within the rectangle formed by the edges \( e \) and \( e' \). We further refer to this rectangle simply as \((e,e')\). Because of variations in the insert sizes, real read-pairs connecting edges \( e \) and \( e' \) typically correspond to the points that are scattered in the confidence strip—a strip between the 45° lines \( y = x - d_{\text{min}} \) and \( y = x - d_{\text{max}} \) within the rectangle \((e,e')\) (Fig. 2c), where

\[
d_{\text{min}} = D - l_{\text{min}} + \text{ReadLength},
\]

\[
d_{\text{max}} = D - l_{\text{max}} + \text{ReadLength}.
\]

Here, \([l_{\text{min}}, l_{\text{max}}]\) is the confidence interval of the insert size distribution defined as the smallest insert size interval that contains at least 80% of read-pairs. Since reads may have variable lengths (e.g. after quality trimming or error correction), we set \text{ReadLength} as the maximal read length across all reads in the current library.

2.1.2 The decision rule

The decision rule relies on the scoring function that will be described in the next section. To extend a path \( P \), we score all extension edges \( e_1, \ldots, e_n \) and select \( e_i \) as a correct extension if the following conditions are met:

1. \( \text{Score}_r(e_i) > C \cdot \text{Score}_r(e_j) \) for all \( j \neq i \)
2. \( \text{Score}_r(e_i) > \Theta \)

Here, \( C \) and \( \Theta \) are the exSPANder parameters described in (Prjibelski et al., 2014). If no extension edge meets these conditions, the algorithm stops growing path \( P \).
Observed on 27 December 2017

The points outside the confidence strip appear in eight out of nine simple rectangles and represent read-pairs with abnormal insert sizes. The confidence strip is a horizontal range of values bounding the observed statistics. Points outside this range indicate possible errors in the library construction or sequencing process. The confidence strip is determined by the statistical analysis of the observed data and the expected distribution under a null hypothesis.

2.2 Scoring function

Given a path \( P = (p_1, p_2, \ldots, p_n) \) and its extension path \( E = (e_1, e_2, \ldots, e_m) \), the composite rectangle for paths \( P \) and \( E \) is formed by \( n \times m \) simple rectangles \((p_i, e_j)\) (Fig. 2d). The distance \( D_{ij} \) between a pair of edges \( p_i \) and \( e_j \) according to the paths \( P \) and \( E \) is equal to

\[ D_{ij} = \sum_{k=1}^{n} \text{Length}(p_k) + \sum_{l=1}^{j-1} \text{Length}(e_l). \]

We define the following variables [see Prjibelski et al. (2014) for more details]:

- \( \text{Expected}_{ij}(p_i, e_j) \): the expected number of \((p_i, e_j)\)-connecting read-pairs;
- \( \text{Observed}_{ij}(p_i, e_j) \): the observed number of \((p_i, e_j)\)-connecting read-pairs.

To filter out rectangles filled by chimeric read-pairs, we use the support function

\[ \text{Support}_{ij}(p_i, e_j) = \begin{cases} 1, & \text{Density}_{ij}(p_i, e_j) > \Psi \\ 0, & \text{otherwise} \end{cases} \]

where the density value is defined as

\[ \text{Density}_{ij}(p_i, e_j) = \frac{\text{Observed}_{ij}(p_i, e_j)}{\text{Expected}_{ij}(p_i, e_j)}. \]

Thus, exSPAnier ignores rectangles with a density lower than a pre-defined threshold \( \Psi \).

While this approach worked well for short libraries, our benchmarking revealed that it deteriorates for long libraries that typically have a higher rate of read-pairs with abnormal insert sizes (referred to as chimeric read-pairs).

We thus define a new support function that evaluates whether a simple rectangle \((p_i, e_j)\) within the composite rectangle \((P, E)\) can be considered as trusted and used for calculating the scoring function:

\[ \text{Support}_{ij}(p_i, e_j) = \begin{cases} 1, & \text{Observed}_{ij}(p_i, e_j) > \eta \\ 0, & \text{otherwise} \end{cases} \]

where \( \eta \) is a user-controlled parameter with very conservative default value \( \eta = 30 \).

While it is not clear how to select optimal \( \eta \), our analysis of the distribution of the number of chimeric points within all rectangles across a wide range of bacterial genomes revealed that hardly any rectangles contain more than 30 chimeric points. We note that by setting the parameter \( \eta \), one essentially ignores the contribution from small rectangles containing less than \( \eta \) points. While it may appear that we unfairly ignore small rectangles, our benchmarking revealed that this approach actually improves the assembly quality. Moreover, the users can change the parameter \( \eta \) in the case of assembly projects with unusually low or high coverage.

When the extension path contains a single edge \( e_j \), exSPAnier uses the following path-edge scoring function:

\[ \text{Score}_{ij}(e_j) = \frac{\sum_{i=1}^{n} \text{Support}_{ij}(p_i, e_j) \cdot \text{Expected}_{ij}(p_i, e_j)}{\sum_{i=1}^{n} \text{Expected}_{ij}(p_i, e_j)}. \]

In this article, we generalize this path-edge scoring function to a path–path scoring function as follows:

\[ \text{Score}(E) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \text{Support}_{ij}(p_i, e_j) \cdot \text{Expected}_{ij}(p_i, e_j)}{\sum_{i=1}^{n} \sum_{j=1}^{m} \text{Expected}_{ij}(p_i, e_j)}. \]

Given a confidence interval \([l_{\text{min}}, l_{\text{max}}]\) of the insert size distribution, we call a path long if its length exceeds \( l_{\text{max}} \) and short otherwise. The algorithm that we describe below limits attention to the long extension paths \((e_1, e_2, \ldots, e_m)\) whose prefix \((e_1, e_2, \ldots, e_{m-1})\) is short. Given a path \( P \), for each extension edge \( e_j \) of \( P \), we construct a set of extension paths \( E \). A top-scored path \( E \) in \( E \) is called a representative path for edge \( e \) (Fig. 3a). The score of the extension edge \( e \) is now defined as the score of its representative path, i.e.

\[ \text{Score}(e) = \text{Score}(E). \]

The new approach to define \( \text{Score}(e) \) can be applied to the libraries with both short and long insert sizes. While the resulting improvements are significant for long insert sizes, they appear to be minor for the libraries with small insert sizes. Thus, in the default exSPAnier mode, the new scoring function is used only for the jumping libraries.
Extend only paths that are shorter than

\( E \) of the assembly graph (shown in blue) and its extension edges \( e_1 \) and \( e_2 \). The representative path for \( e_1 \) is shown in red above the graph, whereas another extension path for \( e_1 \) (with lower score) is shown in orange. (b-d) A step-by-step example of the work of the ExtensionSearch algorithm for constructing extension paths (marked orange) for edge \( e \), using the pre-constructed paths from the set \( S \) (shown as five green paths in the top figure). At each step, the ExtensionSearch algorithm grows an extension path \( E \) (that is shorter than \( I_{\text{max}} \)) using only \( E \)-consistent extension edges (with respect to \( S \)).

2.3 Constructing extension paths

In the case of long libraries, the Depth-First Search (DFS) algorithm for exploring all paths and selecting representative paths becomes rather slow, e.g. the number of extension paths can reach millions even for a bacterial genome. To decrease the number of extension paths, we developed the ExtensionSearch algorithm that uses a set of pre-constructed paths \( S \) that are generated by the previous EXSPANDER algorithm [Priebelski et al., 2014] using single reads and short paired-end libraries if they are available.

We use the notation \( \text{Suffix}(E, l) \) [and \( \text{Prefix}(E, l) \)] to refer to a path formed by the last \((l)\) edges of a path \( E \). For a path \( E \) and a set of pre-constructed paths \( S \), we say that an extension edge \( e \) of path \( E \) is \( E \)-consistent (with respect to \( S \)) if \( S \) contains a path \( E' \), such that \( \text{Suffix}(E + e, l) = \text{Prefix}(E', l) \) for some \( l > 0 \). Instead of extending a path \( E \) by all its extension edges (like in the DFS algorithm), ExtensionSearch uses the following heuristics to construct extension paths:

- Extend a path \( E \) only by \( E \)-consistent extension edges;
- Extend only paths that are shorter than \( I_{\text{max}} \).

As the result, we obtain a set of extension paths that ‘comply’ with the set of pre-constructed paths \( S \) with length at least \( I_{\text{max}} \) (Fig. 3b-d).

2.4 Scaffolding

2.4.1 Jumping over the coverage gaps

Existing genome assemblers are often complemented by scaffolding tools such as Bambus (Pop et al., 2004), Opera (Gao et al., 2011), SCARPA (Donmez and Brudno, 2013), SOPRA (Dayarian et al., 2010) and SSPACE (Boetzer et al., 2011) to improve the contiguity of the resulting assemblies. However, most of these scaffolding tools work with contigs rather than the assembly graphs and none of them uses the rectangle framework that EXSPANDER utilizes. We thus added a rectangle-based scaffolder to EXSPANDER and evaluated its performance.

**Jumping from an out-tip to an in-tip.** A coverage gap typically creates a pair of tips in the assembly graph [see Bankevich et al. (2012) for details] formed by an out-tip (an edge that ends in a vertex with out-degree 0) and an in-tip (an edge that starts in a vertex with in-degree 0). We further refer to these tips as paired tips. Since coverage gaps are usually short, read-pairs from the jumping libraries often span them. Below we describe how ExtensionSearch finds pairs of tips and uses them for scaffolding.

Let \( E \) be an extension path for a path \( P \), such that \( E \) ends with an out-tip \( e \) and \( \text{Length}(E) < I_{\text{max}} \) (Fig. 5a). To further extend path \( E \), we search for an in-tip paired with \( e \) by considering a set of \( P \)-supported edges, i.e. all edges \( e' \) that have \((P,e')\)-connecting
JUMPING OVER COVERAGE GAPS IN THE ASSEMBLY GRAPH
FROM AN OUT-TIP TO AN IN-TIP

(a)

(b)

(c)

JUMPING OVER COVERAGE GAPS IN THE ASSEMBLY GRAPH
FROM AN OUT-TIP TO AN INTERNAL EDGE

(d)

(e)

Fig. 5. Jumping over coverage gaps in the assembly graph from an out-tip to an in-tip: (a) an extension path \( E \) (marked red) ending with an out-tip \( e \); (b) the only \( P \)-supported edge \( e' \) is an in-tip (marked green); (c) the path \( E \) is extended by an edge \( e' \) with a gap (red line). Jumping over coverage gaps in the assembly graph from an out-tip to an internal edge: (d) the only \( P \)-supported edge \( e' \) is not an in-tip (marked green) and (e) paths \( E_1 \) and \( E_2 \) that start with in-tips and contain \( P \)-supported edge \( e' \), both paths are considered as possible extensions for \( E \) (shown by red lines)

read-pairs. Since the set of \( P \)-supported edges typically contains many false short edges (due to high rate of chimeric read-pairs in the jumping libraries), we filter out all \( P \)-supported edges shorter than a certain threshold (default value is 500 bp).

Figure 5b presents a simple example when there is only one \( P \)-supported edge \( e' \) and this edge is an in-tip. We then consider \( e \) and \( e' \) as paired tips and extend the path \( E \) by edge \( e' \) with a gap (Fig. 5c). Since the size of a gap is difficult to estimate using a jumping library with large insert size variation, we assign a fixed length to each gap (100 bp by default). Once we extend the path \( E \) with an edge, we continue growing \( E \) using ExtensionSearch until its length exceeds \( l_{\text{max}} \).

Jumping from an out-tip to an internal edge. A more difficult (and more common) case is illustrated in Figure 5d, where there exists only one \( P \)-supported edge \( e' \) and this edge is not an in-tip. Such situations arise when the in-tip paired with \( e \) is either not \( P \)-supported or is shorter than 500 bp (and thus ignored). To grow the path \( E \) in such cases, we consider all in-tips in the assembly graph and extend them further using ExtensionSearch limiting the path length to \( l_{\text{max}} - \text{Length}(E) \). Afterward, we select only paths that contain edge \( e' \) and use them as potential extensions for \( E \) (Fig. 5e).

When assembling real datasets, we often encounter several \( P \)-supported edges longer than 500 bp. In this case, we apply the strategy described above, i.e. extend the paths starting from every in-tip and choose paths containing at least one \( P \)-supported edge. If some of the extension paths remain shorter than \( l_{\text{max}} \), we continue growing them with the ExtensionSearch algorithm.

However, growing paths from every in-tip can be extremely time-consuming, especially for single-cell datasets that often contain thousands of coverage gaps. To address this issue, we implement the described approach the other way around; instead of starting a path from every in-tip, we start growing paths from \( P \)-supported edges in the reverse direction and select only those that reach an in-tip. Since the number of long \( P \)-supported edges is typically much smaller than the number of all in-tips in the assembly graph in the case of single-cell assemblies, this implementation greatly reduces the running time of ExSPANder.

2.4.2 Jumping over complex subgraphs

Even after applying the described heuristics for run-time reduction, there are cases when the number of extension paths remains huge. Such cases are often associated with complex subgraphs of the assembly graph (typically formed by short repeats with high multiplicities). To keep the running time under control, ExSPANder stops growing a path if the number of its extension paths exceeds a threshold. The threshold is automatically calculated based on the number of edges in the assembly graph (e.g. the threshold value is set to 50 for graphs with > 10 000 edges).

A better approach is to use jumping libraries and to ‘jump over’ complex subgraphs in the assembly graph (instead of generating all extension paths traversing these complex subgraphs). If the ‘size’ of a complex subgraph is smaller than the insert size, a path \( P \) may start on ‘one side’ of this subgraph and be connected with edges on the ‘other side’ of this subgraph, thus allowing us to jump over the entire subgraph and to continue growing \( P \). The question is how to identify such complex subgraphs and jump over them.

To jump over complex subgraphs, we first identify a path \( P \) with unusually large number of extension paths. Afterward, we find all \( P \)-supported edges and order them using the rule: an edge \( e \) follows an edge \( e' \) if \( e' \) can be reached by ExtensionSearch within a certain distance starting from \( e \). If all \( P \)-supported edges can be ordered into a single list, we extend path \( P \) by the first edge in this list (with a gap).

3 Results

To evaluate how availability of Nextera Mate Pair libraries affects the quality of assembly, we assembled several bacterial genomes using multicell datasets provided by Illumina (Nextera Mate Pair libraries only). In this article, we describe benchmarking (using seven different assemblers and three scaffolders) on *Meiothermus ruber* str. 21T (Tindall et al., 2010) Nextera Mate Pair library and additionally, on *Escherichia coli* str. K12 subsp. MG1655 (Blattner et al., 1997) single-cell dataset that contains both a short paired-end library and a long jumping library. Results for other Nextera Mate Pair libraries are presented in the Supplementary Material.

Since the assemblers we used for comparison were not designed for assembling Nextera Mate Pair libraries, we have conducted
Assembling jumping libraries

Table 1. Comparison of the assemblies for the M. ruber str. 21T (Tindall et al., 2010) (genome size 3.1 Mb) Nextera Mate Pair library

<table>
<thead>
<tr>
<th>Assembler</th>
<th>K</th>
<th>NG50</th>
<th>No. Largest</th>
<th>No. GF</th>
<th>No. N’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABySS</td>
<td>59</td>
<td>61</td>
<td>100 162 4 99.4</td>
<td>599</td>
<td></td>
</tr>
<tr>
<td>IDBA-UD</td>
<td>55</td>
<td>139</td>
<td>168 7 99.1</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Opera</td>
<td>111</td>
<td>44</td>
<td>272 18 99.2</td>
<td>115</td>
<td></td>
</tr>
<tr>
<td>Ray</td>
<td>57</td>
<td>7</td>
<td>373 45 94.2</td>
<td>1834</td>
<td></td>
</tr>
<tr>
<td>SCARPA</td>
<td>-</td>
<td>129</td>
<td>61 389 1 99.2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>71</td>
<td>1</td>
<td>1449 411 244 92.8</td>
<td>36360</td>
<td></td>
</tr>
<tr>
<td>SSPACE</td>
<td>-</td>
<td>3011</td>
<td>-</td>
<td>782 17</td>
<td></td>
</tr>
<tr>
<td>Velvet-SC</td>
<td>105</td>
<td>1099</td>
<td>21 18 98.3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>SPAdes 3.0</td>
<td>-</td>
<td>1701</td>
<td>6 2879 3 99.9</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>SPAdes 3.5 ctg</td>
<td>-</td>
<td>1716</td>
<td>6 1716 2 99.9</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>SPAdes 3.5</td>
<td>-</td>
<td>2892</td>
<td>5 2893 2 99.9</td>
<td>23</td>
<td></td>
</tr>
</tbody>
</table>

Each assembly is represented by scaffolds if not indicated otherwise (marked with ctg). K, the k-mer size; NG50, in kb; no., the total number of contigs/scaffolds longer than 500 bp, largest for the length of the longest contig/scaffold assembled (in kb); no. mis is the number of misassemblies; GF, the percentage of genome mapped; No. N’s, the number of unspecified nucleotides per 100 kb. In each column, the best values are indicated in bold.

Table 2. Comparison of the assemblies for the E.coli st. K12 subsp. MG1655 (Blattner et al., 1997) (genome size 4.7 Mb) single-cell dataset with paired-end and jumping libraries

<table>
<thead>
<tr>
<th>Assembler</th>
<th>K</th>
<th>NG50</th>
<th>No. Largest</th>
<th>No. mis</th>
<th>GF</th>
<th>No. N’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only paired-end library</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABySS</td>
<td>47</td>
<td>87</td>
<td>186 199 9 90.5</td>
<td>111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDBA-UD</td>
<td>105</td>
<td>482</td>
<td>264 7 95.0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opera</td>
<td>-</td>
<td>513</td>
<td>202 4 94.6</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ray</td>
<td>55</td>
<td>422</td>
<td>198 3 94.8</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>71</td>
<td>16</td>
<td>728 150 7 78.2</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPACE</td>
<td>-</td>
<td>534</td>
<td>209 2 94.6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velvet</td>
<td>81</td>
<td>122</td>
<td>178 3 62.3</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velvet-SC</td>
<td>55</td>
<td>634</td>
<td>134 4 92.7</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPAdes 3.0</td>
<td>-</td>
<td>453</td>
<td>283 4 94.9</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPAdes 3.5 ctg</td>
<td>-</td>
<td>459</td>
<td>284 4 94.9</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPAdes 3.5</td>
<td>-</td>
<td>452</td>
<td>284 4 94.9</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both paired-end and jumping libraries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABySS</td>
<td>47</td>
<td>87</td>
<td>186 199 9 90.5</td>
<td>111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALLPATHS-LG</td>
<td>4</td>
<td>483</td>
<td>132 23 50.7</td>
<td>3801</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDBA-UD</td>
<td>-</td>
<td>507</td>
<td>416 16 95.0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opera</td>
<td>-</td>
<td>388</td>
<td>202 28 94.6</td>
<td>359</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ray</td>
<td>55</td>
<td>361</td>
<td>281 28 92.7</td>
<td>427</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCARPA</td>
<td>-</td>
<td>413</td>
<td>630 29 94.8</td>
<td>505</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>71</td>
<td>16</td>
<td>673 63 73.8</td>
<td>743</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPACE</td>
<td>-</td>
<td>402</td>
<td>488 21 94.7</td>
<td>478</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velvet</td>
<td>81</td>
<td>24</td>
<td>175 17 67.5</td>
<td>330</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velvet-SC</td>
<td>55</td>
<td>-</td>
<td>1617 1 0 29.4</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPAdes 3.0</td>
<td>-</td>
<td>404</td>
<td>1551 11 95.6</td>
<td>108</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPAdes 3.5 ctg</td>
<td>-</td>
<td>414</td>
<td>653 10 95.6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPAdes 3.5</td>
<td>-</td>
<td>392</td>
<td>693 10 95.6</td>
<td>36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each assembly is represented by scaffolds if not indicated otherwise (marked with ctg). K, the k-mer size; NG50, in kb; no., the total number of contigs/scaffolds longer than 500 bp, largest for the length of the longest contig/scaffold assembled (in kb); no. mis is the number of misassemblies; GF, the percentage of genome mapped; No. N’s, the number of unspecified nucleotides per 100 kb. In each column, the best values are indicated in bold.

Additional optimization and selected the optimal k-mer sizes (with respect to N50) to ensure fair benchmarking with these assemblers. We ran ABySS 1.3.6 (Simpson et al., 2009), Ray 2.0.0 (Boisvert et al., 2010), SOAPdenovo 2.0.4 (Li et al., 2010), Velvet 1.2.10 (Zerbino and Birney, 2008) and Velvet-SC (Chitsaz et al., 2011) (based on Velvet 0.7.62, released on March 11, 2011) with the k-mer sizes 59, 71, 99 and 105, respectively. Iterative de Bruijn graph assemblers IDBA-UD 1.1.1 (Peng et al., 2012), SPAdes 3.0 (with the previous version of exSPAdes (Prijibelski et al., 2014)) and SPAdes 3.5.0 (that implements the new algorithm described in this article) were run with the default parameters.

The resulting assemblies were evaluated with QUAST 2.3 (Gurevich et al., 2013) using standard metrics: NG50 (NG50 corrected for assembly errors), the total number of scaffolds in the assembly, the size of the largest scaffold, the number of misassemblies, the fraction of genome covered and the number of uncalled bases (N) in the assembly.

Table 1 benchmarks various assemblers on M. ruber Nextera Mate Pair library (mean insert size 3.6 kb). Some of the assemblers used in the comparison produce rather inaccurate assemblies (e.g. 95 misassemblies for Ray and 244 misassemblies for SOAPdenovo). Also, some assemblers generate very large number of unspecified symbols N (ABySS, Ray, SOAPdenovo and Velvet). Interestingly, most assemblers showed rather unstable behavior with Nextera Mate Pair libraries with exception of IDBA-UD and SPAdes (originally developed as single-cell assemblers). SPAdes + exSPAdes assembles an almost complete M. ruber genome with less than 0.03% of unspecified nucleotides and the largest scaffold capturing more than 93% of the genome.

SPAdes generates similar high-quality assemblies on all Nextera Mate Pair libraries (see Supplementary Material) and often results in assemblies of very few contigs with the quality that approaches the assemblies from long Pacific Biosciences reads (Chin et al., 2013). Thus, Nextera Mate Pair libraries provide a valuable low-cost trade-off when compared with the assemblies that use Pacific Biosciences reads.

In Table 2, we also present a comparison between selected tools on E.coli single-cell dataset with paired-end and jumping libraries. In addition, we ran ALLPATHS-LG (Gnerre et al., 2011) (build 47561, released on September 15, 2013) with the default parameters using both libraries together. As Table 2 indicates SPAdes (both versions 3.0 and 3.5) and IDBA-UD outperform other tools on this dataset. However, when using both libraries together, SPAdes generates more accurate and continuous assemblies comparing to IDBA-UD and at the same time captures the largest genome fraction.
Acknowledgement

We would like to thank Ilya Chorny from Illumina for providing various datasets for Nextera Mate Pair libraries, which are now available at Illumina BaseSpace public repository (https://basespace.illumina.com).

Funding

This study was supported by the Russian Science Foundation [14-50-00069].

Conflict of Interest

: none declared.

References


