Sequence analysis

IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth

Yu Peng, Henry C. M. Leung*, S. M. Yiu and Francis Y. L. Chin

Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong

ABSTRACT

Motivation: Next-generation sequencing allows us to sequence reads from a microbial environment using single-cell sequencing or metagenomic sequencing technologies. However, both technologies suffer from the problem that sequencing depth of different regions of a genome or genomes from different species are highly uneven. Most existing genome assemblers usually have an assumption that sequencing depths are even. Such assemblers fail to construct correct long contigs.

Results: We introduce the IDBA-UD algorithm that is based on the de Bruijn graph approach for assembling reads from single-cell sequencing or metagenomic sequencing technologies with uneven sequencing depths. Several non-trivial techniques have been employed to tackle the problems. Instead of using a simple threshold, we use multiple deplethrephthalrephoidal thresholds to remove erroneous k-mers in both low-depth and high-depth regions. The technique of local assembly with paired-end information is used to solve the branch problem of low-depth short repeat regions. To speed up the process, an error correction step is conducted to correct reads of high-depth regions that can be aligned to high-confident contigs. Comparison of the performances of IDBA-UD and existing assemblers (Velvet, Velvet-SC, SOAPdenovo and Meta-IDBA) for different datasets, shows that IDBA-UD can reconstruct longer contigs with higher accuracy.

Availability: The IDBA-UD toolkit is available at our website http://www.cs.hku.hk/~alse/idba_ud

Contact: chin@cs.hku.hk

Received and revised on March 26, 2012; accepted on April 12, 2012

1 INTRODUCTION

Since, over 99% of microbes cannot be cultivated, single-cell sequencing and metagenomic sequencing technologies are used to study these microbes (Chitsaz et al., 2011; Wooley et al., 2010). Single-cell sequencing technology amplifies and sequences genome of an individual cell without cultivation (Chitsaz et al., 2011). Since the amplification bias, the sequencing depths at different regions of the genome can be extremely uneven. Metagenomic sequencing studies a microbe community as a whole (Wooley et al., 2010) and has similar problem of uneven sequencing depths of genomes because different species in a sample have different abundances. Almost all existing de novo assembly tools were designed for single genome with uniform sequencing depth and were used by some recent studies on microbes (Rodrique et al., 2009; Woyke et al., 2009). However, these tools may not be able to produce long contigs when applying to data with highly uneven sequencing depths.

Many existing de novo assembly tools for the next-generation sequencing technologies adopt the de Bruijn graph approach (Butler et al., 2008; Chaisson et al., 2009; Li et al., 2010; Peng et al., 2010; Pevzner et al., 2001; Simpson et al., 2009; Zerbino and Birney, 2008) in which a vertex represents a unique length-k substring called k-mer and an edge connects vertices u and v if and only if u and v appear consecutively in a read. Each read is represented by a path of k-mers in the de Bruijn graph. After error detection and removal, a simple path in the de Bruijn graph represents a contig.

There are three major problems in this approach (Peng et al., 2010):

(a) Incorrect k-mers: Sequencing errors introduce many incorrect k-mers (vertices) that make the de Bruijn graph complicated.

(b) Gap problem: When k is large, especially in regions with lower sequencing depths, some k-mers (i.e. vertices, also edges, in the de Bruijn graph) are missing.

(c) Branching problem: Due to repeat regions or erroneous reads, many branches are introduced in the de Bruijn graph especially when k is small.

For Problem (a), some of these errors can be removed by the topological structure of the graph. For the remaining errors, based on the assumption of uniform sequencing depth and the observation that the multiplicity of an erroneous k-mer is usually smaller than that of a correct k-mer, existing tools use a simple threshold to either prune contigs if the contigs are formed by k-mers of low multiplicity [e.g. Velvet (Zerbino and Birney, 2008) and Abyss (Simpson et al., 2009)] or directly remove k-mers with low multiplicity [IDBA (Peng et al., 2010) and EULER-SR (Chaisson and Pevzner, 2008)]. Note that this also solves some of the branching problems [Problem (c)] due to incorrect k-mers.

For Problems (b) and (c), using a small k will induce more branches whereas using a large k will result in more gaps. Many existing tools [e.g. Velvet (Zerbino and Birney, 2008) and SOAPdenovo (Li et al., 2010)] just pick an appropriate k, some intermediate value, to balance the two problems. On the other hand, the IDBA assembler (Peng et al., 2010) provides a better solution which, instead of using a single k, iterates from $k = k_{min}$ to $k = k_{max}$ at each iteration, the constructed contigs are used as reads for the next iteration. These contigs carry the k-mers of the current iteration, which may be missing in the next iteration, to the next iteration, thus solving some of the gap problems. It then relies on larger k to resolve the branches for the repeat regions.

*To whom correspondence should be addressed.
However, when applying to single cell or metagenomic assembling, highly uneven sequencing depth aggravates these problems further that affect the performance of these tools substantially due to the following issues. Issue (A): erroneous vertices and branches in high-depth regions; Issue (B): gaps in low-depth repeat regions.

Problems (a) and (c) due to Issue (A):
Due to highly uneven sequencing depth, the assumption of an incorrect k-mer having lower multiplicity is not valid. Those incorrect ones in the high-depth regions may even have higher multiplicity than the correct ones in the low-depth regions, thus simply using a single threshold to remove incorrect vertices will not work. Setting the threshold too low induces many incorrect vertices and edges (those in high-depth regions) in the graph. Setting the threshold too high will remove many correct vertices and edges in low-depth regions. We remark that there exist some error correction algorithms for reads/k-mers (Chaisson and Pevzner, 2008; Kelley et al., 2010; Medvedev et al., 2011), but they do not perform very well in datasets with very uneven sequencing depths.

Problems (b) and (c) due to Issue (B):
Recall that most existing assemblers do not have a good method to resolve Problems (b) and (c) probably, except IDBA. Even for IDBA, in low-depth short repeat regions [For very long repeats (longer than the whole span of a paired-end read), it is almost impossible to resolve it.], when k is small, the branching problem makes it difficult to construct a contig to be passed to the next iteration. When k is increased, due to the low-depth issue, we still have the missing k-mer problem (the gap problem).

Velvet-SC (Chitsaz et al., 2011) is the only tool that tries to address the assembling problem of single-cell sequencing data with very uneven sequencing depths. Following Velvet, Velvet-SC picks an appropriate k to balance the gap and the branching problem; and uses variable thresholds to address problems related to Issue (A). Short erroneous contigs are filtered iteratively using different thresholds from low to high sequencing depths based on a global average of the multiplicity of all k-mers. Its performance is already better than existing tools designed for even sequencing depth. However, problems related to Issue (B) are not yet handled. In this article, we propose an assembler called IDBA-UD for de novo assembly of reads with uneven sequencing depths that tackles both issues.

To resolve Issue (A), IDBA-UD extends and enhances the idea of variable thresholds of Velvet-SC (Chitsaz et al., 2011) to filter out erroneous contigs. To cater for very extreme sequencing depths, instead of using a global average of the multiplicity of all k-mers, we adopt variable ‘relative’ thresholds depending on the sequencing depths of their neighboring contigs based on the idea that short contigs with much lower sequencing depths than their neighboring contigs tend to be erroneous. For the gap and branching problems, we follow the approach of IDBA and iterate from a small k to a large k, so that the missing k-mers for large k can be obtained from contigs constructed in the iterations of small k.

Then we tackle Issue (B) as follows. The problem of Issue (B) is due to the low-depth short repeat regions such that using small k, we cannot get the contig out since it is a repeat region and the branches may be complicated due to the ambiguity of using a small value of k. When k increases, however, due to the low sequencing depths some k-mers are missing. Even if we iterate from small k to large k, this problem of missing k-mers cannot be resolved. So, we employ the technique of local assembly with paired-end information to handle these cases. Paired-end reads with one end aligned to some long confident contigs are grouped together. Local assembly is performed on the unaligned ends. Since we consider only the read pairs with one end aligned to the contig, the ambiguity due to small k is removed. If the insert size is longer than the repeat involved, it is likely that we can extend the contig over this repeat region, thus constructing the missing k-mers for large k. Note that this local assembly step can also help to resolve some branching problems in high-depth regions too.

To further reduce the size of the de Bruijn graph and to speed up the assembly process, at every iteration, we conduct an additional error correction step by aligning the erroneous reads from the high-depth regions to confident contigs (i.e., with many supporting reads) which turns out to be very effective.

We compared the performance of IDBA-UD with other assemblers on data in actual situations when the sequencing depths are extremely uneven, e.g., with the ratios larger than 100:1. Experiments on both simulated and real datasets showed that IDBA-UD produces much longer contigs than existing assemblers with higher coverage and precision.

2 METHODS
A flowchart of the major steps of IDBA-UD is shown in Figure 1. IDBA-UD iterates the value of k from \( k_{min} \) to \( k_{max} \). In each iteration, an ‘accumulated de Bruijn graph’ (HD for a fixed k) is constructed from the set of input reads and the contigs (\( C_{k-1} \) and \( C_{k-2} \)) constructed in previous iterations, i.e., these contigs
are treated as input reads for constructing $H_0$. In each iteration, IDBA-UD also progressively increases the value of depth cutoff thresholds for removing some low-depth contigs so as to get longer confident contigs ($C_\delta$) in $H_\delta$.

Error in reads are corrected by aligning the reads to some confident contigs. Some missing k-mers in reads can be recovered from those contigs ($L(C_k)$) reconstructed by local assembling of a small set of paired-end reads with one end aligned to a confident contig. Information of these missing k-mers will be passed on to the next iteration through these contigs ($L(C_k)$) for the construction of $H_{k+1}$. Finally, all outputted contigs are used to form scaffolds using paired-end reads information.

Algorithm 1 shows the pseudocode of IDBA-UD for assembling a set of paired-end reads $R$ with insert distance $\alpha$ and $\delta$. In the first iteration when $k = k_{\text{min}}$, $H_0$ is equivalent to a de Bruijn graph for vertices whose corresponding k-mers have multiplicity at least $\min$ (2 by default) times in all reads. During all the subsequent iterations, some sequencing errors are first removed according to the topological structure of $H_k$, e.g. dead-end contigs and bubbles (Steps (b) and (c)). The dead-end contigs (tangling paths in $H_k$ of lengths shorter than $\alpha$) are likely to be false positives (Li et al., 2010; Simpson et al., 2009; Zerbino and Birney, 2008). Paths (bubbles) representing very similar contigs except at one position and with the same starting vertex and ending vertex are likely to be caused by an error or a single-nucleotide polymorphism (SNP) and they should be merged together into one contig (Hernandez et al., 2008; Simpson et al., 2009; Zerbino and Birney, 2008). When constructing $H_{k+1}$ from $H_k$, each length $s+1$ path in $H_k$ is converted into a vertex $k$-mer and there is an edge from between two vertices if the corresponding $(k+s+1)$-mers appear $\min$ (1 by default) times in reads or once in contigs in $C_k$/$L(C_k)$. In the following subsections, we will describe the other steps of IDBA-UD in detail.

### 2.1 Progressive relative depth

The sequencing depth, ‘depth’ in short, of each simple path (contig) in $H_0$ ($H_k$ which is a copy of $H_0$ is used in Algorithm 1 so as to preserve $H_0$ after the implementation of this step) is used to remove errors. The ‘depth of a contig’ is the average number of reads covering each k-mer in the contig. Note that long contigs are usually correct, because short paths can unlikely be formed by erroneous reads, similarly for high-depth contigs which have supports from many reads. For a contig, whether its length is long or short and whether its depth is high or low cannot be judged by its absolute values as the length of a contig depends on the value of $k$ and the depth of a contig depends on the depths of its neighboring contigs (neighboring contigs can be identified by their adjacency in the de Bruijn graph). Even wrong contigs in high-depth regions may have higher depths than correct contigs from low-depth regions, ‘short’ (< $\alpha$) and ‘relatively low depth’ (< $\min$) contigs are likely to be erroneous and can be removed.

There is still a risk of removing short and relatively low-depth correct contigs because some relatively low-depth correct contigs with high-depth neighbors may be broken into short contigs by some wrong contigs (as branches in $H_k$). Based on the observation that these short and relatively low-depth correct contigs usually have higher depths than the short wrong contigs, we can filter these wrong contigs first by increasing the depth cutoff threshold progressively from low to high. After the wrong contigs or branches are removed by a low-depth cutoff threshold, the relative low-depth correct contigs will be linked together to form long confident contigs which will be considered as reads for the next iteration. The key idea to consider the depth of a contig $\epsilon$ is relatively and progressively is shown in Algorithm 2. $T(c)$ represents the depth of contig $c$ and $T_{\text{neighbor}}(c)$ represents the mean depth of $c$’s neighboring contigs. The filtering depth cutoff threshold $t$ is increased by a factor $\alpha$ progressively (as is $\sim10\%$). A geometric increase, instead of absolute increase (as used in Velvet-SC), in the depth cutoff threshold value improves implementation efficiency because the threshold difference is more sensitive at the low-depth values than the high-depth values. In each iteration, short contig $c$ is removed if its depth $T(c)$ is lower than the minimum of cutoff threshold $t$ and the relative threshold $\beta*T_{\text{neighbor}}(c)$ where $\beta$ is in the range of 0.1–0.5.

Algorithm 1 IDBA-UD($R$, $d$, $k$):

1. Pre-Error-Correction (optional)
2. Repeat from $k = k_{\text{min}}$ to $k_{\text{MAX}}$ with step $s$
   a. If $k = k_{\text{max}}$, then construct $H_\delta$ from $R$
   b. else construct $H_k$ from $H_{k-1}$ and $R \cap C_k \cup L(C_k)$
3. Remove dead-ends with length $< \alpha$
4. Merge bubbles
5. $H_k = H_k$
6. Progressive-Relative-Depth($G$, $k$, $t$)
7. until $t > \max(C_k)$
8. return $C_{\text{outs}}$ and scaffolds

Algorithm 2 Progressive-Relative-Depth($G$, $k$, $t$):

1. repeat
2. for each contig $c$ in $G$
3. if $\text{len}(c) < \alpha$ and $T(c) < \min(\beta*T_{\text{neighbor}}(c), \max(C_k))$
4. remove $c$ from $G$
5. $t = t + \min(1, t)$
6. until $t > \max(C_k)$

#### 2.2 Local assembly

IDBA makes use of the contigs (containing the information of some missing k-mers) constructed in each iteration for the construction of the de Bruijn graphs of larger $k$. These missing k-mers may not exist in any of the reads but they might help to fill the gaps in the de Bruijn graphs for larger $k$. This approach still has a limitation that not all the missing k-mers, i.e. contigs containing these k-mers, can be constructed (so not all the gaps can be filled) because of branches. The main contribution of local assembly is to construct these contigs for the missing k-mers, especially in the low-depth regions, based on the information of paired-end reads to eliminate the branches introduced from other parts of the genome.

We shall illustrate this main idea of local assembly through an example (Fig 2). Let us consider the construction of a de Bruijn Graph for $k = 3$, based on two reads, AACT and ACCTG. We have a simple path connecting the 3mers, AAC, ACT and CTG. IDBA can reconstruct the missing 5mer AACTG (not appeared in any reads) by forming a simple path containing it. However, as given in Figure 2, when ACT is a length-3 repeat in the genome (the repeat regions are apart by more than the insert distance) and there are reads covering the region TACCT, containing the other repeat. The 5mer ACT in the de Bruijn graph for $k = 3$ now has two in-branches and two out-branches (refer to the left diagram of Figure 3 where vertex $v$ represents the 3mer ACT, vertices $u$, $w$, $u'$ and $w'$ are for 3mers AAC, CTG, TAC and CCT, respectively).

Under this situation, even when $k$ is increased to 4 and 5 in IDBA (this part of graph will be disconnected in $H_4$ and $H_5$), the missing critical 5mers AACTG cannot be reconstructed from local reads.
Algorithm 3 Local-Assembly(C, R, d, δ):
1: Remove contigs shorter than 2l from C
2: Align reads in R to contigs in C
3: For each (v, δ) ∈ R:
4:  if v uniquely aligned to last δ + 3l bases of c, \( R_c \) := \( R_c \cup \{v\} \)
5:  if v uniquely aligned to last δ + 3l bases of \( c^\prime \), \( R_{c^\prime} \) := \( R_{c^\prime} \cup \{v\} \)
6: For each c ∈ C:
7:  \( LC_c \) := IDBA([last δ + 3l bases of c] \cup \{R_c\} \cup \{R_{c^\prime}\})
8:  \( LC_{c^\prime} \) := IDBA([last δ + 3l bases of \( c^\prime\]) \cup R_{c^\prime})
9: return \( \cup_{c \in C} (LC_c \cup LC_{c^\prime}) \)

Let \( C_b \) be the set of contigs (simple paths) in \( H_2 \). The set of paired-end reads \( R_e \) are those with one read aligned with the ends of each long contig \( c \) (with length at least twice of read length) in \( C_b \) (\( c^\prime \) stands for the reverse complement of \( c \)). The other unaligned ends of these aligned paired-end reads, which would cover the genome regions extended about an insert distance beyond each end of a long contig, are extracted separately. Assume the insert distances of paired-end reads satisfy the normal distribution \( N(\delta, 3) \). IDBA-UD groups the last \( δ + 3l \) bases of \( c \) (\( c^\prime \) together and then locally assembles them into the set of local contigs \( LC_c \) using IDBA [Algorithm 1 without Steps (e), (g), and (h)] as shown in Algorithm 3. Since those reads which are far away from the contig \( c \) will not be mixed up with these unaligned ends, the contig \( c \) and these unaligned ends (reads) of \( R_e \) can be used to construct a smaller and simpler de Bruijn graph whose simple paths (represented by the set of contigs \( LC_c \)) might reconstruct some of the missing \( k \)-mers and be considered as reads for the next iteration. Thus, the contigs can be extended longer and longer at each iteration. The expected number of resolved branches can be computed by Theorem 3 (Appendix).

2.3 Error correction

To reduce the errors in reads, error correction on some erroneous bases is performed based on the alignment between reads and confident contigs. Errors in reads are corrected only if they can be aligned to contigs with certain similarity, say 95%. The reads which can be multi-aligned to different contigs will not be considered for corrections. This approach of error correction is especially effective for high-depth regions because the confident contigs are well-supported by many reads.

A position of a contig is labeled as ‘confirmed’ if one base type appears over 80% in all reads aligned to that position. Each read, aligned to a contig region with all positions confirmed and the number of different bases no > 3, will be corrected according to the confirmed bases.

A pre-error-correction step for improved efficiency can be used to remove errors in high-depth regions as the first step in IDBA-UD if the sequencing depths are extremely uneven. A medium \( k \)-value and filtering threshold will be used to assemble reads to form contigs and errors in reads are corrected on its alignment with the output contigs.

2.4 Scaffold

The reads are finally aligned to contigs so as to build a scaffold graph in which each vertex \( v \) represents a contig and each edge (\( u, v \)) represents the connection between \( u \) and \( v \) with a support of \( > p \) (3 by default) paired-end reads. After the scaffold graph is built, scaffold algorithm (Li et al., 2010) will be applied to further connect contigs.

3 RESULTS

To evaluate the performance of our experiments, (All experiments were done on a machine with 8-core 2.40 GHz Intel CPU and 144 GB memory. The tested assembler was run with multiple threads, if it supports.) are carried out on several datasets with different properties. Results on existing general purpose assemblers like Velvet (Zerbino and Birney, 2008), SOAPdenovo (Li et al., 2010), IDBA (Peng et al., 2010) and special purpose assemblers like Velvet-SC (Chitsaz et al., 2011), Meta-IDBA (Peng et al., 2011) were compared. Different \( k \)-values were tried for each assembler and the result with best performance are shown and compared.

Two most important statistics, N50 and coverage are calculated to evaluate the contiguity and completeness of assembly results. N50 is the length of the longest contig such that all the contigs longer than this contig cover at least half of the genome being assembled (Earl et al., 2011). Coverage is the proportion of the genome being covered by output contigs. In this article, only correct contigs are considered in the calculation of N50 and coverage. A contig is considered as correct if it can be aligned to the genome reference by BLAT (Kent, 2002) with 95% similarity. For correct contigs, the substitution errors are computed by comparing the alignment between contigs and genome reference. For unaligned contigs, the number of contigs and the number of bases are recorded for comparison.

3.1 Error correction

The performance of our error correction algorithm is assessed by correcting the simulated reads sampled from Lactobacillus delbrueckii genome (∼1.85 Mb). The simulated dataset contains 1.85 million length-100 reads (100×) uniformly sampled from the reference with 1% error rate. The error correction algorithm was executed on this dataset with output contigs of IDBA with \( k = 60 \) (kmin = kmax in this case). The correction result is shown in Table 1. There are 1,586,822 error bases in the dataset. Our algorithm corrected 1,627,727 bases with 1,626,929 (99.95%) being true positive. Note that our target of error correction is to reduce the errors without introducing other errors. The remaining erroneous reads either contain too many errors to be aligned to contigs or are from those regions which cannot be assembled correctly. This high-precision and low-sensitivity error correction algorithm is suitable for high-precision and low-sensitivity error correction algorithm is suitable.

Table 1. Error correction result on simulated 100× length-100 reads of L.delbrueckii (∼1.85 Mb) with 1% error rate

<table>
<thead>
<tr>
<th>No. of errors</th>
<th>No. of corrected bases</th>
<th>No. of TP</th>
<th>No. of FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,586,822</td>
<td>1,627,727</td>
<td>1,626,929 (99.95%)</td>
<td>798 (0.05%)</td>
</tr>
</tbody>
</table>

No. of errors: Total number of error bases. No. of corrected bases: Number of error bases corrected. No. of TP: Number of true positive bases. No. of FP: Number of false positive bases.
for IDBA-UD, because the remaining errors could be handled by other means in the later iterations. Note that error correction which simplifies the graph for next iterations improves efficiency tremendously.

3.2 Low depth assembly

*Lactobacillus plantarum* (∼3.3 Mb) was used as genome reference for simulating low-depth dataset. 10× length-100 paired-end reads were simulated for testing. The assembly results of IDBA-UD, Velvet, SOAPdenovo and IDBA are shown in Table 2.

IDBA-UD has the longest N50 (36,523), which is several folds of the N50 of Velvet, SOAPdenovo and IDBA (13,761 and 8,350). IDBA-UD also has the highest coverage (99.56%), which is higher than the coverage of Velvet, SOAPdenovo and IDBA (98.36, 98.09 and 98.52%). IDBA-UD and IDBA have the least number of erroneous bases. Only 10 contigs (4437 bases) constructed by IDBA-UD and 5 contigs (301 bases) by IDBA cannot be aligned to the genome reference, compared with 1079 contigs (11,250 bases) by SOAPdenovo and 5 contigs (15,921 bases) by Velvet. Among aligned contigs, all assemblers have similar substitution error rate.

IDBA-UD also constructed the longest scaffolds with N50 being 194,322, which is nearly twice of N50 of scaffolds constructed by other assemblers. Although N50 of scaffolds generated by different assemblers are similar, the coverage of scaffolds generated by all assemblers is much lower than that of contigs except IDBA-UD. IDBA-UD made the least misassembly during the scaffolding process, because of having longer and more accurate contigs.

In general, IDBA-UD achieved its best performance by iterating *k* from 20 to 100, whereas Velvet, SOAPdenovo and IDBA had best performance when *k* is set to a small value (21, 31 and 20–40, respectively). Since, the local assembly procedure can reconstruct missing *k*-mers, IDBA-UD can iterate *k* to a large value to construct very long contigs. The other assemblers are not able to reconstruct missing *k*-mers so that a reasonably small *k* is used to balance the gaps and branches problem. The running time and memory cost are more or less the same among all assemblers.

3.3 Local assembly

The expected number of resolved branches (Theorem 3 in Appendix) by IDBA-UD, IDBA and the actual numbers for all assemblers for different repeat lengths *k* are shown in Table 3. We ran all assemblers with a specific *k* (*k*max) and measured repeats with length (*k*−10, *k*). Since SOAPdenovo and Velvet cannot handle even values of *k* due to the palindrome problem, 29, 39 etc. are considered in Table 3. The number of resolved branches by IDBA-UD is slightly smaller than the expected number because some of the (*k* + 2)-mers are removed as dead-end. As the *k*-value increases, the number of resolved branches drops and the number of wrong contigs also increases because of the missing (*k* + 2)-mers. Comparing with other assemblers, IDBA-UD can resolve more repeats by increasing *k* and gets longer contigs.

Table 3. Expected numbers of resolved branches by IDBA-UD, IDBA and the real number of all assembly on the same dataset in Table 2

<table>
<thead>
<tr>
<th><em>k</em></th>
<th>All branches</th>
<th>IDBA-UD (expected)</th>
<th>IDBA-UD</th>
<th>IDBA (expected)</th>
<th>IDBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>20–40</td>
<td>5509</td>
<td>2117</td>
<td>1019</td>
<td>618</td>
<td>375</td>
</tr>
<tr>
<td>21</td>
<td>5507</td>
<td>2115</td>
<td>1018</td>
<td>616</td>
<td>373</td>
</tr>
<tr>
<td>22</td>
<td>5497</td>
<td>2103</td>
<td>1000</td>
<td>584</td>
<td>329</td>
</tr>
<tr>
<td>23</td>
<td>5298</td>
<td>1986</td>
<td>933</td>
<td>513</td>
<td>280</td>
</tr>
<tr>
<td>24</td>
<td>5259</td>
<td>1861</td>
<td>795</td>
<td>512</td>
<td>282</td>
</tr>
<tr>
<td>25</td>
<td>5151</td>
<td>1536</td>
<td>617</td>
<td>419</td>
<td>219</td>
</tr>
</tbody>
</table>

3.4 Single cell assembly

Two single-cell short read sequencing datasets; *Escherichia coli* (lane 1) and *Staphylococcus aureus* (Chitsaz et al., 2011; http://bix.ucsd.edu/projects/singlecell/) were used to test the performance of IDBA-UD, SOAPdenovo, Velvet and Velvet-SC. Velvet-SC was run after EULER error correction as the authors suggested. The assembly result we presented is slightly different from that in Velvet-SC paper, because we calculated the N50 for aligned contigs rather than all contigs. Genomes sequences of *E.coli* str. K-12 substr. MG1655 and *S.aureus* subsp. aureus USA300_FPR3757 were downloaded from NCBI and were used as reference for validation. The statistics of the assembly results of different assemblers are summarized in Tables 4 and 5.

3.4.1 De novo assembly of *E.coli*

According to (Chitsaz et al., 2011), the average sequencing depth of single-cell sequencing data of *E.coli* is ∼650X and the reads are sampled very unevenly. For contigs, SOAPdenovo and Velvet have similar N50 (6,428 and 7,679); Velvet-SC has the second best N50 (34,454), as it considers the property of uneven depth to remove errors; IDBA-UD has the longest N50 (82,007), as it considers uneven relative depth to remove errors and uses local assembly to reconstruct missing *k*-mers in low-depth regions. The contigs constructed by IDBA-UD also have the highest coverage. IDBA-UD and Velvet-SC have the least number of substitution errors. All assemblers constructed some contigs cannot be aligned to the reference. Some of them are really misassembled contigs, but the alignment of reads against contigs
The assembly results on real single-cell sequencing data of *E. coli* (~6.46 Mb)

<table>
<thead>
<tr>
<th>k</th>
<th>No.</th>
<th>N50</th>
<th>Max len</th>
<th>Cov (%)</th>
<th>Sub.err (%)</th>
<th>Err no. (len)</th>
<th>Time (s)</th>
<th>Mem (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDBA-UD</td>
<td>40–100</td>
<td>187</td>
<td>82 007</td>
<td>224 018</td>
<td>95.01</td>
<td>0.0017</td>
<td>87 347</td>
<td>148</td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>75</td>
<td>60 008</td>
<td>64 248</td>
<td>50 965</td>
<td>92.42</td>
<td>0.0042</td>
<td>693 335</td>
<td>4419</td>
</tr>
<tr>
<td>Velvet</td>
<td>45</td>
<td>76</td>
<td>18 597</td>
<td>67 117</td>
<td>92.69</td>
<td>0.0151</td>
<td>160 266</td>
<td>1707</td>
</tr>
<tr>
<td>Velvet-SC</td>
<td>55</td>
<td>372</td>
<td>34 454</td>
<td>157 931</td>
<td>92.74</td>
<td>0.0019</td>
<td>78 279</td>
<td>516</td>
</tr>
</tbody>
</table>

The read length is 100, insert distance is ~215 and the average depth is ~600×.

<table>
<thead>
<tr>
<th>No.</th>
<th>N50</th>
<th>Max len</th>
<th>Cov (%)</th>
<th>Sub.err (%)</th>
<th>Err no. (len)</th>
<th>Time (s)</th>
<th>Mem (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDBA-UD</td>
<td>60–100</td>
<td>122</td>
<td>87 502</td>
<td>175 236</td>
<td>94.46</td>
<td>0.0027</td>
<td>14 247</td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>95</td>
<td>1005</td>
<td>12 214</td>
<td>92 978</td>
<td>96.61</td>
<td>0.0067</td>
<td>791 220</td>
</tr>
<tr>
<td>Velvet</td>
<td>55</td>
<td>520</td>
<td>15 800</td>
<td>67 677</td>
<td>94.14</td>
<td>0.0043</td>
<td>7999</td>
</tr>
<tr>
<td>Velvet-SC</td>
<td>55</td>
<td>322</td>
<td>29 719</td>
<td>221 773</td>
<td>93.12</td>
<td>0.0042</td>
<td>43 312</td>
</tr>
</tbody>
</table>

The read length is 100, insert distance is ~265, and the average depth is ~2300×.

- **Table 4.** The assembly results on real single-cell sequencing data of *E. coli* (~6.46 Mb)
- **Table 5.** The assembly result on real single-cell sequencing data of *S. aureus* (~2.87 Mb)

reference showed that some non-aligned contigs are from regions with structure variations.

After scaffolding, all assemblers produced longer scaffolds and lower coverage, but the difference between contigs and scaffolds is not much except SOAPdenovo. SOAPdenovo increased the N50 from 6428 to 25 244, but the coverage dropped from 92.42% to 86.49%. This means that the uneven depth of single cell assembly makes the scaffolding very difficult so that assemblers either cannot construct long scaffolds or make many mistakes in scaffolding procedure. Since IDBA-UD produced very long contigs, although scaffolding did not connect many contigs, the scaffolds generated by IDBA-UD have the longest N50 and highest coverage.

### 3.4.2 De novo assembly of *S. aureus*

The sequencing depth of single-cell sequencing data of *S. aureus* is ~2300×, much higher than that of *E. coli*. SOAPdenovo and Velvet performed better for dataset with higher sequencing depth. The contig N50 of SOAPdenovo and Velvet became 12 214 and 15 800, about twice of that of *E. coli*. IDBA-UD and Velvet-SC had similar performance as before, and handle reads with uneven depth quite well. Generally, higher sequencing depth does not affect the quality of assembly result much. The scaffolding also increases N50 and reduces coverage. The substitution error rates are very low for all assemblers for this high sequencing depth.

SOAPdenovo is the fastest assemblers among four assemblers IDBA-UD took about twice the time of SOAPdenovo to perform de novo single cell assembly. The memory cost of IDBA-UD is much less than the others, as it did the filtering on k-mers. Depending on the nature of assemblers, different assemblers achieved its best performance with different k-values. SOAPdenovo got its best performance by using relatively large k-values (45 and 55) to reduce errors in high-depth regions and introducing more gaps problem at the same time. It then relies on paired-end information to connect contigs to form scaffolds. Therefore, it produced more and shorter contigs and more gaps in its scaffolds than the others. Velvet preferred a relatively small k-values (45 and 55), probably because it contains a very sophisticated algorithm to remove the errors in de Bruijn graph. Velvet-SC had best performance with a moderate k-value (55) and relies on iteratively removing low-depth erroneous contigs to form long contigs. IDBA-UD is able to iterate k-values from small to large to build a de Bruijn graph with less gaps and less branches and obtain the best performance, through local assembly producing missing k-mers and iterate depth for reducing the errors.

3.5 Metagenomic assembly

#### 3.5.1 De novo assembly of simulated metagenomes data

To evaluate the performance of IDBA-UD on metagenomic data, we considered a simulated dataset with extremely uneven depth. This dataset was synthesized by combining simulated reads of three species *L. plantarum* (~3.3 Mb), *L. delbrueckii* (~1.85 Mb) and *L. reuteri reuteri* F275 *K. iisato* (~2 Mb) from the same genus. Length-100 reads were sampled from these three species with sequencing depth 10× (low depth), 100× (moderate depth) and 1000× (high depth), respectively, with 1% error rate. The simulated paired-end reads have an insert distance following normal distribution N(500, 50). IDBA-UD, SOAPdenovo, Velvet and Meta-IDBA were executed on this simulated metagenomic sequencing dataset. Since the depth is highly uneven, the Pre-Error-Correction of IDBA-UD was activated to remove errors. The experiment results are showed in Table 6. In addition to the statistics we presented before, the sequencing depth of each species was considered for comparison.

SOAPdenovo and Velvet did not perform well on this simulated metagenomic dataset. N50 of SOAPdenovo and Velvet are only 461 and 418, respectively. The contigs constructed by SOAPdenovo and Velvet covered most regions of moderate-depth species (95.39% and 95.40%, respectively).
At each iteration, missing assembly result on simulated metagenomic dataset of Table 6. The assembly results on human gut microbial short read data (SRR041654 and SRR041655) from NCBI Velvet 55 33 165 3476 171 410 65 630 428 30 419 5065 358 885 65 887 174 40 914 176 35

The sequencing depth of these species are 10×, 100× and 1000×, respectively. The read depth is 100, error rate is set to 1% and the insert distance follows normal distribution N(500, 50).

<table>
<thead>
<tr>
<th>k</th>
<th>Contigs</th>
<th>Scaffolds</th>
<th>Time (min)</th>
<th>Mem (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. N50</td>
<td>Cov (%)</td>
<td>10× 1000×</td>
<td>1000×</td>
</tr>
<tr>
<td>IDBA-UD</td>
<td>20–100</td>
<td>546 444879 99.15 99.31 98.48 95.92</td>
<td>94</td>
<td>335 104696 97.80 99.23 95.90 97.21</td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>45</td>
<td>16266 461 79.57 95.45 95.49 94.39</td>
<td>1374 166</td>
<td>10056 1356 74.01 83.43 93.37 93.28</td>
</tr>
<tr>
<td>Velvet</td>
<td>45</td>
<td>12018 418 72.73 94.46 95.24 16.66</td>
<td>6200 710</td>
<td>11976 41 72.56 94.34 95.24 16.66</td>
</tr>
<tr>
<td>Meta-IDBA</td>
<td>20–100</td>
<td>2636 4588 89.62 85.29 96.96 99.09</td>
<td>127 292</td>
<td>46</td>
</tr>
</tbody>
</table>

The read length is 100 and the insert distance is 260.

95.24%) and low-depth species (95.45% and 94.66%) but a small portion of high-depth regions (39.40% and 16.66%). It is because they make a tradeoff between low-depth and high-depth regions and cannot handle them together. SOAPdenovo and Velvet have the best performance in N50 and coverage when k = 45. Small k is chosen because the genome size of low-depth species (∼3.3 Mb) is larger than high-depth species (∼2.3 Mb). In fact, their performance for k = 55 are similar except that the coverage of low-depth species decreases, the coverage of moderate-depth species remains the same and the coverage of high-depth species increases.

Meta-IDBA designed for metagenomic assembly iterates k from small to large to capture both low-depth and high-depth regions. At each iteration, missing k-mers in low-depth and moderate-depth regions introduce fragmentation in the assembly result. Thus, Meta-IDBA outperforms SOAPdenovo and Velvet, but performs worse than IDBA-UD. The contigs constructed by Meta-IDBA covered >99% of the regions in low-depth species, slightly <90% of the regions in low-depth species and moderate-depth species.

As expected, IDBA-UD outperforms SOAPdenovo, Velvet and Meta-IDBA in all aspects. N50 (44 879) of contigs constructed by IDBA-UD is ∼10 times the second best N50 (4588) of contigs constructed by Meta-IDBA. IDBA-UD also has the best coverage (99.15%), >10% higher than the second highest by Meta-IDBA. The contigs constructed by IDBA-UD covered almost all the region of three species, the uneven depth does not affect the assembly quality of IDBA-UD. Similar to single-cell assembly, scaffolding in all assemblers produced longer contigs but lower coverage. As for substitution error rate, IDBA-UD and Meta-IDBA have much higher accuracy and IDBA-UD constructed the least number of misassembled contigs. The running time of all assemblers are similar, but IDBA-UD and Meta-IDBA used about half of memory as SOAPdenovo and Velvet.

### Table 6. The assembly result on simulated metagenomic dataset of L.plantarum (∼3.3 Mb), L.delbrueckii (∼1.85 Mb) and L.reuteri F275 Kitasato (∼2 Mb)

<table>
<thead>
<tr>
<th>k</th>
<th>Contigs</th>
<th>Scaffolds</th>
<th>Time (min)</th>
<th>Mem (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. N50</td>
<td>Cov (%)</td>
<td>10× 1000×</td>
<td>1000×</td>
</tr>
<tr>
<td>IDBA-UD</td>
<td>20–100</td>
<td>546 444879 99.15 99.31 98.48 95.92</td>
<td>94</td>
<td>335 104696 97.80 99.23 95.90 97.21</td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>45</td>
<td>16266 461 79.57 95.45 95.49 94.39</td>
<td>1374 166</td>
<td>10056 1356 74.01 83.43 93.37 93.28</td>
</tr>
<tr>
<td>Velvet</td>
<td>45</td>
<td>12018 418 72.73 94.46 95.24 16.66</td>
<td>6200 710</td>
<td>11976 41 72.56 94.34 95.24 16.66</td>
</tr>
<tr>
<td>Meta-IDBA</td>
<td>20–100</td>
<td>2636 4588 89.62 85.29 96.96 99.09</td>
<td>127 292</td>
<td>46</td>
</tr>
</tbody>
</table>

The read length is 100 and the insert distance is 260.

3.5.2 De novo assembly of human gut microbial short read data Real human gut microbial sequencing data were used to assess the performance of IDBA-UD. The datasets (SRR041654 and SRR041655) were downloaded from NCBI for assembly. The reads were generated by Illumina Genome Analyzer II with read length 100 and insert distance 260. IDBA-UD, SOAPdenovo, Velvet and Meta-IDBA were compared with this dataset (The Pre-Error-Correction of IDBA-UD was activated.) Since there is no reference, we used the largest total contig size of all assemblers as the estimated genome size (98 407 199) for N50 calculation, and did not analyze the completeness of assembly by comparing genome coverage. MetaGeneAnnotator (Noguchi et al., 2008; only complete genes predicted by MetaGeneAnnotator are considered as recovered) was applied to the output of each assembler to predict the number of genes recovered. The statistics of assembly results are summarized in Table 7.

The contigs of SOAPdenovo and Velvet have similar N50, whereas SOAPdenovo produced much more contigs than Velvet. The total contig size of SOAPdenovo is 92 059 177. Meta-IDBA produced the smallest number of contigs, but N50 of contigs constructed by Meta-IDBA is larger than SOAPdenovo and Velvet. According to the analysis of the assembly result of simulated data, it is probably because Meta-IDBA reconstructed most of the high-depth regions but missed some low-depth regions. IDBA-UD has the largest total contig size and the highest N50 (18 658). The contigs constructed by IDBA-UD contain the largest number of predicted genes (66 298), which is 50% more than that of SOAPdenovo and Velvet. The results reveal that IDBA-UD can assemble metagenomic data better than all the other assemblers. The running time of IDBA-UD is between SOAPdenovo and Velvet. The memory cost of Meta-IDBA and Meta-IDBA is also about half of SOAPdenovo and Velvet.

### Table 7. The assembly results on human gut microbial short read data (SRR041654 and SRR041655) from NCBI

<table>
<thead>
<tr>
<th>k</th>
<th>Contigs</th>
<th>Scaffolds</th>
<th>Time (min)</th>
<th>Mem (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. N50</td>
<td>Cov (%)</td>
<td>10× 1000×</td>
<td>1000×</td>
</tr>
<tr>
<td>IDBA-UD</td>
<td>20–100</td>
<td>39221 18658 39221 18658</td>
<td>37512</td>
<td>20737 673651 98243412</td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>55</td>
<td>99213 2666 113678 92059177 93961</td>
<td>4575 319705 97788489</td>
<td>41 993</td>
</tr>
<tr>
<td>Velvet</td>
<td>55</td>
<td>33165 3476 171410 65630428</td>
<td>50419</td>
<td>5065 358885 65887174</td>
</tr>
<tr>
<td>Meta-IDBA</td>
<td>20–100</td>
<td>19978 7710 325768</td>
<td>48693</td>
<td>49</td>
</tr>
</tbody>
</table>

The read length is 100 and the insert distance is 260.
Table 8. The assembly results on simulated metagenomic datasets in different taxonomic levels

<table>
<thead>
<tr>
<th>k</th>
<th>Contigs</th>
<th>Scaffolds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Genus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IDBA-UD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20–100</td>
<td>713</td>
</tr>
<tr>
<td></td>
<td>Velvet</td>
<td>20,427</td>
</tr>
<tr>
<td></td>
<td>Meta-IDBA</td>
<td>20–100</td>
</tr>
<tr>
<td></td>
<td>Family</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>IDBA-UD</td>
<td>20–100</td>
</tr>
<tr>
<td></td>
<td>SOAPdenovo</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Meta-IDBA</td>
<td>20–100</td>
</tr>
<tr>
<td></td>
<td>Class</td>
<td>45</td>
</tr>
</tbody>
</table>

For each level, five test cases with three randomly selected species are generated for testing. The depths of three species are set to 100, 500, and 2000, respectively. The real length is 100, error rate is set to 1% and the insert distance follows normal distribution N(500, 50).

In general, all assemblers have the best performance on dataset in class level and the worst performance on dataset in genus level. Contigs and scaffolds generated by IDBA-UD have the largest N50 and highest coverage in all three kinds of datasets. Velvet generated the second highest coverage, but it produced the shortest N50. The scaffolds generated by SOAPdenovo have the second largest N50, but they covered the least part of genomes. Meta-IDBA is somehow in the middle, because it is designed to handle similar subspecies problem rather than different expression levels. In the assembly results of IDBA-UD, all species got similar coverage. All the other assemblers generated high coverage for middle-depth (50×) species but lower coverage for low-depth and high-depth species, because they balanced the gap problems in low-depth species and high-depth species and benefited the middle-depth species. Consistent with previous experiments, IDBA-UD outperformed all the existing assemblers in sequencing data with highly uneven depth. For metagenomic data, there are more common k-mers between genomes from subspecies of the same species than genomes from different species. This information is used in Meta-IDBA for assembling metagenomic data. As a future work, we should study how to integrate this information in IDBA-UD for better performance.

Funding: This work was supported in part by HKGRF funding (HKU 7116/08E, HKU 71909/E) and HKU Genomics SRT funding.

Conflict of Interest: none declared.

REFERENCES


Assume Theorems 2 and 3 indicates the expected number of repeats resolved by local assembly. Woyke, T. or the (k + 2)-mers can be used to convert branches to simple paths. If some of these (k + 2)-mers are missing in reads due to low depth or errors, then local assembly may be used to reconstruct them as shown in Figure 2.

In this section, we try to calculate the expected number of branches caused by repeats that can be resolved by (k + 2)-mers which already exist in reads or reconstructed by locally assembly. Theorem 2 gives the expected number of (k + 2)-mers covering a repeat being sampled f times (applied by IDBA). Theorem 3 gives the expected number of (k + 2)-mers covering a repeat being sampled f times or reconstructed by locally assembly (applied by IDBA-UD). Thus, the difference between these two expected numbers as given in Theorems 2 and 3 indicates the expected number of repeats resolved by local assembly.

THEOREM 1. Assume t length-l reads are uniformly sampled from a length-g genome with error rate e, the probability that a k-mer v appearing in the genome being sampled at least m times is at least

\[
P_{k,m,v} = \sum_{i=0}^{m-1} \binom{t}{i} p^i (1-p)^{t-i}
\]

where

\[
p = \frac{(t-k+1)\lambda}{g-t+1} \times (1-e)^k
\]

PROOF. (v is sampled in a read) \[ P(v) \leq \text{Pr}(\text{a read containing } v \text{ is sampled}) \]

\[ = \frac{(t-k+1)x}{g-t+1} \times (1-e)^k \]

The probability that a correct k-mer v appears m times is at most \[
\sum_{i=0}^{m-1} \binom{t}{i} p^i (1-p)^{t-i}
\]

so the result follows.

THEOREM 2. Assume t length-l reads are uniformly sampled from a length-g genome with error rate e and R_k is the set of repeats with length k. If the support requirement for resolving a branch is f, then the expected number of resolved branches \(R_k\) from k to k+1 is at least

\[
\sum_{x \in R_k} \sum_{b \in Y(x)} P_{k+2,f,x}(b)
\]

where \(Y(x)\) is the set of the (k + 2)-mers covering repeat r in the genome and \(x(b)\) is the number of times that b appear in the genome.

PROOF. To resolve a repeat of length k, a (k + 2)-mer appearing at least f times in the reads is needed and the probability of such (k + 2)-mer is \(P_{k+2,f,x}(b)\) (Theorem 1).

If reads are localized to a specific region of genome, then the branches caused by distant repeats may disappear and convert to simple paths. In this way, local assembly may generate local contigs which help to resolve branches.

THEOREM 3. Assume t length-l reads are uniformly sampled from a length-g genome with error rate e, \(R_k\) is the set of repeats with length k and \(L_{R_k}\) is the set of length-k repeats occurring at least twice in the genome within insert distance. If the support requirement for resolving a branch is f and IDBA-UD is applied on the data from k to k+1, then the expected number of resolved branches \(L_{R_k}\) is at least

\[
|L_{R_k}| + \sum_{b \in L_{R_k}} \sum_{x \in Y(b)} P_{k+2,f,x}(b)
\]

where \(Y(b)\) is the set of the (k + 2)-mers covering repeat r in the genome and \(x(b)\) is the number of times that b appear in the genome.

PROOF. If a repeat does not appear twice within insert distance, then it can be resolved by IDBA-UD, i.e. \(|L_{R_k} - L_{R_k}|\) and those repeats occurring more than once within the insert distance can only be resolved by the (k + 2)-mers appearing at least f times in reads (Theorem 2).