Supplementary Material

In this supplementary material, we display in Figures 10–17 the output of our pipeline for 9 datasets (in addition to hc19, whose output is in the introduction, and the GAGE datasets R. sphaeroides, S. Aureus, and hc14). For each dataset we plot

$$\log(1 + a_\ell),$$

the log of one plus the number of repeats of each length $\ell$. From the repeat statistics $a_m, b_{m,n},$ and $c_m$, we produce a feasibility plot. The thick black line denotes the lower bound on feasible $N, L,$ and the green line is the performance achieved by MULTI BRIDGING.
Figure 10: Lactofidus. $G = 2,078,001$, $\ell_{\text{triple}} = 3027$, $\ell_{\text{interleaved}} = 3313$, $\ell_{\text{repeat}} = 5321$.

Figure 11: Buchnera. $G = 642,122$, $\ell_{\text{triple}} = 27$, $\ell_{\text{interleaved}} = 23$, $\ell_{\text{repeat}} = 39$.

Figure 12: Heli51. $G = 1,589,954$, $\ell_{\text{triple}} = 219$, $\ell_{\text{interleaved}} = 2122$, $\ell_{\text{repeat}} = 3478$. 
Figure 13: Salmonella. $G = 2, 215, 568$, $\ell_{\text{triple}} = 112$, $\ell_{\text{interleaved}} = 163$, $\ell_{\text{repeat}} = 1011$.

Figure 14: Perkinsus marinus. $G = 1, 440, 372$, $\ell_{\text{triple}} = 770$, $\ell_{\text{interleaved}} = 92$, $\ell_{\text{repeat}} = 1784$.

Figure 15: Sulfolobus islandicus. $G = 2, 655, 198$, $\ell_{\text{triple}} = 734$, $\ell_{\text{interleaved}} = 761$, $\ell_{\text{repeat}} = 875$. 
Figure 16: Ecoli536. $G = 4,938,920$, $\ell_{\text{triple}} = 2267$, $\ell_{\text{interleaved}} = 3245$, $\ell_{\text{repeat}} = 3353$.

Figure 17: Yasnina. $G = 4,504,254$, $\ell_{\text{triple}} = 3573$, $\ell_{\text{interleaved}} = 3627$, $\ell_{\text{repeat}} = 5358$. 
Lower bounds on coverage depth

The lower bounds are based on a generalization of Ukkonen's condition to shotgun sequencing, as described in Theorem 1. The proof of Theorem 1 follows by a straightforward modification to the argument in [12] and is omitted here.

Theorem 1. Given a DNA sequence $s$ and a set of reads, if there is a pair of interleaved repeats or a triple repeat whose copies are all unbridged, then there is another sequence $s'$ of the same length under which the likelihood of observing the reads is the same.

Lower bound due to interleaved repeats

In this section we derive a necessary condition on $N$ and $L$ in order that the probability of correct reconstruction be at least $1 - \epsilon$.

Recall that a pair of repeats, one at positions $t_1, t_3$ with $t_1 < t_3$ and the second at positions $t_2, t_4$ with $t_2 < t_4$, is interleaved if $t_1 < t_2 < t_3 < t_4$ or $t_2 < t_1 < t_4 < t_3$. From the DNA we may extract a (symmetric) matrix of interleaved repeat statistics $b_{mn}$, the number of pairs of interleaved repeats of lengths $m$ and $n$.

We proceed by fixing both $N$ and $L$ and checking whether or not unbridged interleaved repeats occur with probability higher than $\epsilon$. We will break up repeats into 2 categories: repeats of length at least $L - 1$ (these are always unbridged), and repeats of length less than $L - 1$ (these are sometimes unbridged). We assume that $L > \ell_{\text{interleaved}} + 1$, or equivalently $b_{ij} = 0$ for all $i, j \geq L - 1$, since otherwise there are (with certainty) unbridged interleaved repeats and Ukkonen's condition is violated.

First, we estimate the probability of error due to interleaved repeats of lengths $i < L - 1$ and $j \geq L - 1$. The repeat of length $j$ is too long to be bridged, so an error occurs if the repeat of length $i$ is unbridged. For a repeat, as long as the two copies' locations are not too nearby, each copy is bridged independently and hence the probability that both copies of the repeat of length $i$ are unbridged is $p_i^{\text{unbridged}} = e^{-2N(i-1)}$. (Recall that a repeat is unbridged if both copies are unbridged.)

A union bound estimate gives a probability of error

$$P_{\text{error}} \approx \frac{1}{2} \sum_{m < L - 1} \sum_{n \geq L - 1} b_{mn} e^{-2\lambda(L-m-1)}. \quad (8)$$

Requiring the error probability to be less than $\epsilon$ and solving for $L$ gives the necessary condition

$$L \geq \frac{1}{2\lambda} \log \frac{\gamma_1}{2\epsilon} = \frac{G}{2N} \log \frac{\gamma_1}{2\epsilon}, \quad (9)$$

where $\gamma_1 := \sum_{m < L - 1} \sum_{n \geq L - 1} b_{mn} e^{2(N/G)(m+1)}$ is a simple function of the interleaved repeat statistic $b_{mn}$.

We now estimate the probability of error due to interleaved repeat pairs in which both repeats are shorter than $L - 1$. In this case only one repeat of each interleaved repeat pair must be bridged. Again a union bound estimate gives

$$P_{\text{error}} \approx \frac{1}{2} \sum_{m,n < L - 1} b_{mn} e^{-2\lambda(L-m-1)} e^{-2\lambda(L-n-1)}. \quad (8)$$

\[1\] More precisely, for the two copies of a a repeat of length $t$ to be bridged independently requires that no single read can bridge them both. This means their locations $t$ and $t + d$ must have separation $d \geq L - t - 2$.

\[2\] The union bound on probabilities gives an upper bound, so its use here is only an approximation. To get a rigorous lower bound we can use the inclusion-exclusion principle, but the difference in the two computations is negligible for the data we observed. For ease of exposition we opt to present the simpler union bound estimate.
Requiring the error probability to be less than $\epsilon$ gives the necessary condition

$$L \geq \frac{1}{4\lambda} \log \frac{\gamma_2}{2\epsilon} = \frac{G}{4N} \log \frac{\gamma_2}{2\epsilon}, \quad (10)$$

where $\gamma_2 := \sum_{m,n<L-1} b_{mn} e^{2(N/G)(m+n+2)}$ and similarly to $\gamma_1$ is computed from $b_{mn}$.

### Lower bound due to triple repeats

We translate the generalized Ukkonen’s condition prohibiting unbridged triple repeats into a condition on $L$ and $N$. Let $c_m$ denote the number of triple repeats of length $m$. Then a union bound estimate gives

$$\mathbb{P}(E) \approx \frac{1}{2} \sum_m c_m e^{-3\lambda(L-m-1)}. \quad (11)$$

Requiring $\mathbb{P}(E) \leq \epsilon$ and solving for $L$ gives

$$L \geq \frac{1}{3\lambda} \log \frac{\gamma_3}{2\epsilon} = \frac{G}{3N} \log \frac{\gamma_3}{2\epsilon}, \quad (12)$$

where $\gamma_3 := \sum_m c_m e^{3(N/G)(m+1)}$.

**Remark 7.** As discussed here and in Section , if the DNA sequence is not covered by the reads or there are unbridged interleaved or triple repeats, then reconstruction is not possible. But there is another situation which must be ruled out. Without knowing its length a priori, it is impossible to know how many copies of the DNA sequence are actually present: if the sequence $s$ to be assembled consists of multiple concatenated copies of a shorter sequence, rather than just one copy, the probability of observing any set of reads will be the same. Since it is unlikely that a true DNA sequence will consist of the same sequence repeated multiple times, we assume this is not the case throughout the paper. Equivalently, if $s$ does consist of multiple concatenated copies of a shorter sequence, we are content to reconstruct a single copy. If available, knowledge of the approximate length of $s$ would then allow to reconstruct.

### Proofs for algorithms

#### Proof of Theorem 2 (Greedy)

The greedy algorithm’s underlying data structure is the overlap graph, where each node represents a read and each (directed) edge $(x, y)$ is labeled with the overlap $ov(x, y)$ (defined as the length of the shared prefix/suffix) between the incident nodes’ reads. For a node $v$, the in-degree [out-degree] is the number of edges in the graph directed towards [away from] $v$. The greedy algorithm is described as follows.

**Algorithm 2 Greedy.** Input: reads $R$. Output: sequence $s$.

1. For each read with sequence $x$, form a node with label $x$.
   **Greedy steps 2-3:**
2. Consider all pairs of nodes $x_1, x_2$ in $G$ satisfying $d_{out}(x_1) = d_{in}(x_2) = 0$, and add an edge $(x_1, x_2)$ with largest value $ov(x_1, x_2)$.
3. Repeat Step 2 until no candidate pair of nodes remains.
4. Output the sequence corresponding to the unique cycle in $G$.

**Theorem 2.** Given a sequence $s$ and a set of reads, Greedy returns $s$ if every repeat is bridged.

**Proof.** We prove the contrapositive. Suppose Greedy makes its first error in merging reads $r_i$ and $r_j$ with overlap $ov(r_i, r_j) = \ell$. Now, if $r_j$ is the successor to $r_i$, then the error is due to incorrectly aligning the reads; the other case is that $r_j$ is not the successor of $r_i$. In the first case, the subsequence $s_{t_j}^\ell$ is repeated at location $s_{t_i+L-\ell}^\ell$ and no read bridges either repeat copy. In the second case, there is a repeat $s_{t_j}^\ell = s_{t_i+L-\ell}^\ell$. If $s_{t_i+L-\ell}$ is bridged by some read $r_k$, then $r_i$ has overlap at least $\ell + 1$ with $r_k$, implying that read $r_i$ has already found its successor before step $\ell$ (either $r_k$ or some other read with even higher overlap). A similar argument shows that $s_{t_j}^\ell$ cannot be bridged, hence there is an unbridged repeat. \[\square\]
Proofs for \(K\)-mer algorithms

Background

We give some mathematical background leading to the proof of Theorem 3 (restated below).

Lemma 8. Fix an arbitrary \(K\) and form the \(K\)-mer graph from the \((K+1)\)-spectrum \(S_{K+1}\). The sequence \(s\) corresponds to a unique cycle \(C(s)\) traversing each edge at least once.

To prove the lemma, note that all \((K+1)\)-mers in \(s\) correspond to edges and adjacent \((K+1)\)-mers in \(s\) are represented by adjacent edges. An induction argument shows that \(s\) corresponds to a cycle. The cycle traverses all the edges, since each edge represents a unique \((K+1)\)-mer in \(s\).

In both SBH and shotgun sequencing the number of times each edge \(e\) is traversed by \(C(s)\) (henceforth called the multiplicity of \(e\)) is unknown a priori, and finding this number is part of the reconstruction task. Repeated \((K+1)\)-mers in \(s\) correspond to edges in the \(K\)-mer graph traversed more than once by \(C(s)\), i.e. having multiplicity greater than one. In order to estimate the multiplicity, previous works seek a solution to the so-called Chinese Postman Problem (CPP), in which the goal is to find a cycle of the shortest total length traversing every edge in the graph (see e.g. [28], [16], [8], [7]). It is not obvious under what conditions the CPP solution correctly assigns multiplicities in agreement with \(C(s)\). For our purposes, as we will see in Theorem 3, the multiplicity estimation problem can be sidestepped (thereby avoiding solving CPP) through a modification to the \(K\)-mer graph.

Ignoring the issue of edge multiplicities for a moment, Pevzner [18] showed for the SBH model that if the edge multiplicities are known with multiple copies of each edge included according to the multiplicities, and moreover Ukkonen’s condition is satisfied, then there is a unique Eulerian cycle in the \(K\)-mer graph and the Eulerian cycle corresponds to the original sequence. (An Eulerian cycle is a cycle traversing each edge exactly once.) Pevzner’s algorithm is thus to find an Eulerian cycle and read off the corresponding sequence. Both steps can be done efficiently.

Lemma 9 (Pevzner [18]). In the SBH setting, if the edge multiplicities are known, then there is a unique Eulerian cycle in the \(K\)-mer graph with \(K = L - 1\) if and only if there are no unbridged interleaved repeats or unbridged triple repeats.

Most practical algorithms (e.g. [16], [22], [23]) condense unambiguous paths (called unitigs by Myers [29] in a slightly different setting) for computational efficiency. The more significant benefit for us, as shown in Theorem 3, is that if Ukkonen’s condition is satisfied then condensing the graph obviates the need to estimate multiplicities. Condensing a \(K\)-mer graph results in a graph of the following type.

Definition 10 (Sequence graph). A sequence graph is a graph in which each node is labeled with a subsequence, and edges \((u, v)\) are labeled with an overlap \(a_{uv}\) such the subsequences \(u\) and \(v\) overlap by \(a_{uv}\) (the overlap is not necessarily maximal). In other words, an edge label \(a_{uv}\) on \(e = (u, v)\) indicates that the \(a_{uv}\)-length suffix of \(u\) is equal to the \(a_{uv}\)-length prefix of \(v\).

The sequence graph generalizes both the overlap graph used by GREEDY in Section (nodes correspond to reads, and edge overlaps are maximal overlaps) as well as the \(K\)-mer algorithms discussed in this section (nodes correspond to \(K\)-mers, and edge overlaps are \(K - 1\)).

In order to speak concisely about concatenated sequences in the sequence graph, we extend the notation \(s^\ell_t\) (denoting the length-\(\ell\) subsequence of the DNA sequence \(s\) starting at position \(t\)) which was introduced in Section ; we abuse notation slightly, and write \(s_{t}^{\text{end}}\) to indicate the subsequence of \(s\) starting at position \(t\) and having length so that its end coincides with the end of \(s\).

We will perform two basic operations on the sequence graph. For an edge \(e = (u, v)\) with overlap \(a_{uv}\), merging \(u\) and \(v\) along \(e\) produces the concatenation \(u_{t}^{\text{end}} v_{t+1}^{\text{end}}\). Contracting an edge \(e = (u, v)\) entails two steps (c.f. Fig. 6): first, merging \(u\) and \(v\) along \(e\) to form a new node \(w = u_{t}^{\text{end}} v_{t+1}^{\text{end}}\), and, second, edges to \(u\) are replaced with edges to \(w\), and edges from \(v\) are replaced by edges from \(w\). We will only contract edges \((u, v)\) with \(d_{\text{out}}(u) = d_{\text{in}}(v) = 1\).
The condensed graph is defined next.

**Definition 11** (Condensed sequence graph). The condensed sequence graph replaces unambiguous paths by single nodes. Concretely, any edge \( e = (u,v) \) with \( d_{\text{out}}(u) = d_{\text{in}}(v) = 1 \) is contracted, and this is repeated until no candidate edges remain.

For a path \( \mathcal{P} = v_1, v_2, \ldots, v_q \) in the original graph, the corresponding path in the condensed graph is obtained by contracting an edge \((v_i, v_{i+1})\) whenever it is contracted in the graph, replacing the node \( v_1 \) by \( w \) whenever an edge \((u, v_1)\) is contracted to form \( w \), and similarly for the final node \( v_q \). It is impossible for an intermediate node \( v_i, 2 \leq i < q \), to be merged with a node outside of \( \mathcal{P} \), as this would violate the condition \( d_{\text{out}}(u) = d_{\text{in}}(v) = 1 \) for edge contraction in Defn. 11.

In the condensed sequence graph \( \mathcal{G} \) obtained from a sequence \( s \), nodes correspond to subsequences via their labels, and paths in \( \mathcal{G} \) correspond to subsequences in \( s \) via merging the constituent nodes along the path. If the subsequence corresponding to a node \( v \) appears twice or more in \( s \), we say that \( v \) corresponds to a repeat. Conversely, subsequences of length \( \ell \geq K \) in \( s \) correspond to paths \( \mathcal{P} \) of length \( \ell - K + 1 \) in the \( K \)-mer graph, and thus by the previous paragraph also to paths in the condensed graph \( \mathcal{G} \).

We record a few simple facts about the condensed sequence graph obtained from a \( K \)-mer graph.

**Lemma 12.** Let \( \mathcal{G}_0 \) be the \( K \)-mer graph constructed from the \((K + 1)\)-spectrum of \( s \) and let \( \mathcal{C}_0 = \mathcal{C}_0(s) \) be the cycle corresponding to \( s \). In the condensed graph \( \mathcal{G} \), let \( \mathcal{C} \) be the cycle obtained from \( \mathcal{C}_0 \) by contracting the same edges as those contracted in \( \mathcal{G}_0 \).

1. Edges in \( \mathcal{G}_0 \) can be contracted in any order, resulting in the same graph \( \mathcal{G} \), so the condensed graph is well-defined. Similarly \( \mathcal{C} \) is well-defined.
2. The cycle \( \mathcal{C} \) in \( \mathcal{G} \) corresponds to \( s \) and is the unique such cycle.
3. The cycle \( \mathcal{C} \) in \( \mathcal{G} \) traverses each edge at least once.

**Theorem 3.** Let \( S_{K+1} \) be the \((K + 1)\)-spectrum of \( s \) and \( \mathcal{G}_0 \) be the \( K \)-mer graph constructed from \( S_{K+1} \), and let \( \mathcal{G} \) be the condensed sequence graph obtained from \( \mathcal{G}_0 \). If Ukkonen’s condition is satisfied, i.e. there are no triple repeats or interleaved repeats of length at least \( K \), then there is a unique Eulerian cycle \( \mathcal{C} \) in \( \mathcal{G} \) and \( \mathcal{C} \) corresponds to \( s \).

**Proof.** We will show that if Ukkonen’s condition is satisfied, the cycle \( \mathcal{C} = \mathcal{C}(s) \) in \( \mathcal{G} \) corresponding to \( s \) (constructed in Lemma 12) traverses each edge exactly once in the condensed \( K \)-mer graph, i.e. \( \mathcal{C} \) is Eulerian. Pevzner’s [18] arguments show that if there are multiple Eulerian cycles then Ukkonen’s condition is violated, so it is sufficient to prove that \( \mathcal{C} \) is Eulerian. As noted in Lemma 12, \( \mathcal{C} \) traverses each edge at least once, and thus it remains only to show that \( \mathcal{C} \) traverses each edge at most once.

To begin, let \( \mathcal{C}_0 \) be the cycle corresponding to \( s \) in the original \( K \)-mer graph \( \mathcal{G}_0 \). We argue that every edge \((u,v)\) traversed twice by \( \mathcal{C}_0 \) in the \( K \)-mer graph \( \mathcal{G}_0 \) has been contracted in the condensed graph \( \mathcal{G} \) and hence in \( \mathcal{C} \). Note that the cycle \( \mathcal{C}_0 \) does not traverse any node three times in \( \mathcal{G}_0 \), for this would imply the existence of a triple repeat of length \( K \), violating the hypothesis of the Lemma. It follows that the node \( u \) cannot have two outgoing edges in \( \mathcal{G}_0 \) as \( u \) would then be traversed three times; similarly, \( v \) cannot have two incoming edges. Thus \( d_{\text{out}}(u) = d_{\text{in}}(v) = 1 \) and, as prescribed in Defn. 11, the edge \((u,v)\) has been contracted. \( \Box \)

**Proofs for SimpleBridging**

Since bridging reads extend one base to either end of a repeat, it will be convenient to use the following notation for extending sequences:

Given an \( X \)-node \( v \) with an incoming edge \((p,v)\) and an outgoing edge \((v,q)\), let

\[
\begin{align*}
    v^\rightarrow q &= v q_{\text{prep}}^1, \quad \text{and} \quad p^\rightarrow v &= p_{\text{prep}}^1 v.
\end{align*}
\]

Here \( v^\rightarrow q \) denotes the subsequence \( v \) appended with the single next base in the merging of \( v \) and \( q \) and \( p^\rightarrow v \) the subsequence \( v \) prepended with the single previous base in the merging of \( p \).
and $v$. For example, if $v = ATTC$, $p = TCAT$, $a_{pv} = 2$, $q = TTCGCC$, and $a_{vq} = 3$, then $v \rightarrow q = ATTCG$, $p \rightarrow v = CATTC$, and $p \rightarrow v \rightarrow q = CATTCG$.

The idea is that a bridging read is consistent with only one pair $(p, v)$ and $(v, q)$ and thus allows to match up edge $(p, v)$ with $(v, q)$. This is recorded in the following lemma.

**Lemma 13.** Suppose $C$ corresponds to a sequence $s$ in a condensed sequence graph $G$. If a read $r$ bridges an $X$-node $v$, then there are unique edges $(p, v)$ and $(v, q)$ such that $p \rightarrow v$ and $v \rightarrow q$ are adjacent in $r$.

SimpleBridging is described as follows.

**Algorithm 3 SimpleBridging.** Input: reads $R$, parameter $K$. Output: sequence $\hat{s}$.

1. For each subsequence $x$ of length $K$ in a read, form a node with label $x$.
2. For each read, add edges between nodes representing adjacent $K$-mers in the read.
3. Condense the graph as described in Defn. 11.
4. **Bridging step:** See Fig. 7. While there exists an $X$-node $v$ with $d_{in}(v) = d_{out}(v) = 2$ bridged by some read $r$: (i) Remove $v$ and edges incident to it. Add duplicate nodes $v_1, v_2$. (ii) Choose the unique $p_i$ and $q_j$ s.t. $p_i \rightarrow v$ and $v \rightarrow q_j$ are adjacent in $r$ and add edges $(p_i, v_1)$ and $(v_1, q_j)$. Choose the unused $p_i$ and $q_j$, add edges $(p_i, v_2)$ and $(v_2, q_j)$. (iii) Condense the graph.
5. **Finishing step:** Find an Eulerian cycle in the graph and return the corresponding sequence.

**Proofs for MultiBridging**

In this subsection we recall Theorem 6 stating sufficient conditions for correct reconstruction, and derive the corresponding required coverage depth and read length to meet a target probability of correct reconstruction. The subsection concludes with a proof that the sufficient conditions are correct.

**Theorem 6.** The algorithm MultiBridging reconstructs the sequence $s$ if:

(a) all interleaved repeats are bridged
(b) all triple repeats are all-bridged
(c) the sequence is covered by the reads.

Remark 14. Unlike the previous $K$-mer algorithms, DeBruijn and SimpleBridging, it is unnecessary to specify a parameter $K$ for MultiBridging. Implicitly MultiBridging uses $K = 1$, which makes the condition that reads overlap by $K$ equivalent to coverage of the genome.

Figure 2 plots the performance of MultiBridging, obtained by solving for the relationship between $G, N, L$, and $\epsilon$ in order to satisfy the conditions of Lemma 6. We first perform the requisite calculations, and then prove the Lemma.

Condition (a) is already dealt with in (9) and (10), and Condition (c) amounts to the requirement that $\frac{N}{N_{LW}} \geq 1$. 

23
We turn to Condition (b) that all triple repeats are all-bridged. Let $c_m$ denote the number of triple repeats of length $m$. A union bound estimate over triple repeats for the event that one such triple repeat fails to be all-bridged gives

$$P_{\text{error}} \approx \sum_m 3 \cdot c_m e^{-\lambda(L-m-1)^+},$$

and requiring $P_{\text{error}} \leq \epsilon$ and solving for $L$ yields

$$L \geq \frac{1}{\lambda} \log \frac{\gamma_3}{\epsilon} = \frac{G}{N} \log \frac{\gamma_3}{\epsilon},$$

where $\gamma_3 := \sum_m 3c_m e^{(N/G)-(m+1)}$ is computed from the triple repeat statistics $c_m$.

In order to understand the cost of all-bridging triple repeats, compared to simply bridging one copy as required by our lower bound, it is instructive to study the effect of the single longest triple repeat. Setting $c_{\ell_{\text{triple}}} = 1$ and $c_m = 0$ for $m \neq \ell_{\text{triple}}$ makes $\gamma_3 = 3e^{(N/G)-(\ell_{\text{triple}}+1)}$ in (15) and

$$L \geq L^\text{all}_3 := \ell_{\text{triple}} + 1 + \frac{G}{N} \log 3e^{-1}. \quad (16)$$

Bridging the longest triple repeat, as shown in Section 7, requires

$$L \geq L := \ell_{\text{triple}} + 1 + \frac{G}{3N} \log e^{-1}. \quad (17)$$

Solving for $N$ in equations (17) and (16) gives

$$N_3 \geq \frac{G}{3} \cdot \frac{\log e^{-1}}{L - \ell_{\text{triple}} - 1} \quad (18)$$

$$N^\text{all}_3 \geq G \cdot \frac{\log e^{-1} + \log 3}{L - \ell_{\text{triple}} - 1}. \quad (19)$$

The ratio is

$$\frac{N^\text{all}_3}{N_3} = 3 \cdot \frac{\log 3e^{-1}}{\log e^{-1}} \approx 3.72 \quad \text{for } \epsilon = 10^{-2}. \quad (20)$$

This means that if the longest triple repeat is dominant, then for $L$ slightly larger than $\ell_{\text{triple}}$, MultiBridging needs a coverage depth approximately 3.72 times higher than required by our lower bound.

The remainder of this subsection is devoted to the proving Lemma 6.

We will use $m_C(v)$ to denote the multiplicity (traversal count) a cycle $C$ assigns a node $v$. The multiplicity $m_C(v)$ is also equal to the number of times the subsequence $v$ appears in the sequence corresponding to $C$. For an edge $e$, we can similarly let $m_c(e)$ be the number of times $C$ traverses the edge. The following key lemma relates node multiplicities with the existence of X-nodes.

**Lemma 15.** Let $C$ be a cycle in a condensed sequence graph $G$, where $G$ itself is not a cycle, traversing every edge at least once. If $v$ is a node with maximum multiplicity at least 2, i.e. $m(v) = \max_{u \in C} m_C(u) \geq 2$, then $v$ is an X-node. As a consequence, if $m_C(v) \geq 3$ for some $v$, i.e. $C$ traverses some node at least three times, then $m_C(u) \geq 3$ for some X-node $u$.

**Proof.** Let $v$ be a node with maximum multiplicity $m_C(v) = \max_{u \in G} m_C(u)$. We will show that $v$ is an X-node, i.e. $d_{\text{out}}(v) \geq 2$ and $d_{\text{in}}(v) \geq 2$.

We prove that $d_{\text{out}}(v) \geq 2$ by supposing that $d_{\text{out}}(v) = 1$ and deriving a contradiction. Denote the outgoing edge from $v$ by $e = (v, u)$, where $u$ is distinct from $v$ since otherwise $G$ is a cycle. If $d_{\text{in}}(u) \geq 2$, then $u$ must be traversed more times than $v$, contradicting the maximality of $m_C(v)$, and if $d_{\text{in}}(u) = 1$, then the existence of the edge $e$ contradicts the fact that $G$ is condensed. The argument showing that $d_{\text{in}}(v) \geq 2$ is symmetric to the case $d_{\text{in}}(v) \geq 2$.

**Proof of Lemma 6.** We assume that all triple repeats are all-bridged, that there are no unbridged interleaved repeats, and that all reads overlap their successors by at least 1 base pair. We wish to show that MultiBridging returns the original sequence.

Consider the condensed sequence graph $G_0$ constructed in steps 1-3 of MultiBridging. Suppose all X-nodes that are either all-bridged or correspond to bridged 2-repeats have been resolved according to repeated application of the procedure in step 4 of MultiBridging, resulting in a condensed sequence graph $G$. We claim that 1) $s$ corresponds to a cycle $C$ in $G$ traversing every edge at least once, 2) $C$ is Eulerian, and 3) $C$ is the unique Eulerian cycle in $G$.
Proof of Claim 1. Let $G_n$ be the graph after $n$ resolution steps, and suppose that $C_n$ is a cycle in $G_n$ corresponding to the sequence $s$ and traversing all edges. We will show that there exists a cycle $C_{n+1}$ in $G_{n+1}$ corresponding to $s$ and traversing all edges, and that $G_t = G$ for a finite $t$, so by induction, there exists a cycle $C$ in $G$ corresponding to $s$ and traversing all edges. The base case $n = 0$ was shown in Lemma 8. Moving on to arbitrary $n > 0$, let $v$ be an X-node in $G_n$ labeled as in Fig. 6. The X-node resolution step is constructed precisely to preserve the existence of a cycle corresponding to $s$. Each traversal of $v$ by the cycle $C_n$ assigns an incoming edge $(p,v)$ to an outgoing edge $(v,q)$, and the resolution step correctly determines this pairing by the assumption on bridging reads.

Note that all X-nodes in the graph $G_{n+1}$ continue to correspond to repeats in $s$. The process terminates: let $L(G_t) = \sum_{v \in G_t} m_C(v)1_{m_C(v)>1}$ and observe that $L(G_t)$ is strictly decreasing in $i$. Thus $s$ corresponds to a cycle $C$ in $G$ traversing each edge at least once.

Proof of Claim 2. We next show that $C$ is an Eulerian cycle. If $G$ is itself a cycle, and $s$ is not formed by concatenating multiple copies of a shorter subsequence (assumed not to be the case, see discussion at end of Section 5), then $C$ traverses $G$ exactly once and is an Eulerian cycle. Otherwise, if $G$ is not a cycle, then we may apply Lemma 15 to see that any node with $m_C(v) \geq 3$ implies the existence of an X-node $u$ with $m_C(u) \geq 3$. Node $u$ must be all-bridged, by hypothesis, which means that an additional X-node resolution step can be applied to $G$, a contradiction. Thus each node $v$ in $G$ has multiplicity $m_C(v) \leq 2.$

We can now argue that no edge $e = (u,v)$ is traversed twice by $C$ in the condensed sequence graph $G$, as it would have been contracted. Suppose $m_C(e) \geq 2$. The node $u$ cannot have two outgoing edges as this implies $m_C(u) \geq 3$; similarly, $v$ cannot have two incoming edges. Thus $d_{out}(u) = d_{in}(v) = 1$, but by Defn. 11 the edge $e = (u,v)$ would have been contracted.

Proof of Claim 3. It remains to show that there is a unique Eulerian cycle in $G$. All X-nodes in $G$ must be unbridged 2-X-nodes (correspond to 2-repeats in $s$), as all other X-nodes were assumed to be bridged and have thus been resolved in $G$.

We will map the sequence $s$ to another sequence $s'$, allowing us to use the characterization of Lemma 9 for SBH with known multiplicities. Denote by $G'$ the graph obtained by relabeling each node in $G$ by a single unique symbol (no matter the original node label length), and setting all edge overlaps to 0. Through the relabeling, $C$ corresponds to a cycle $C'$ in $G'$, and let $s'$ be the sequence corresponding to $C'$. Writing $S'_2$ for the 2-spectrum of $s'$, the graph $G'$ is by construction precisely the 1-mer graph created from $S'_2$, and there is a one-to-one correspondence between X-nodes in $G'$ and unbridged repeats in $s'$. Through the described mapping, every unbridged repeat in $s'$ maps to an unbridged repeat in $s$, with the order of repeats preserved.

There are multiple Eulerian cycles in $G$ only if there are multiple Eulerian cycles in $G'$ since the graphs have the same topology, and by Lemma 9 the latter occurs only if there are unbridged interleaved repeats in $s'$, which by the correspondence in the previous paragraph implies the existence of unbridged interleaved repeats in $s$.

Truncation estimate for bridging repeats (Greedy and MultiBridging)

The repeat statistics $a_m$ and $c_m$ used in the algorithm performance curves are potentially overestimates. This is because a large repeat family—one with a large number of copies $f$—will result in a contribution $\binom{f}{2} \approx f^2/2$ to $a_m$ and $\binom{f}{3} \approx f^3/6$ to $c_m$.

We focus here on deriving an estimate for the required $N, L$ for bridging all repeats with probability $1 - \epsilon$. This upper bound reduces the sensitivity to large families of short repeats. The analogous derivation for all-bridging all triple-repeats is very similar and is omitted.

Suppose there are $a_m$ repeats of length $m$. The probability that some repeat is unbridged is ap-
proximately, by the union bound estimate,
\[ \mathbb{P}(E) \approx \sum_{m} a_m e^{-2\lambda (L-m)}. \]  
(21)

Requiring \( \mathbb{P}(E) \leq \epsilon \) and solving for \( L \) gives
\[ L \geq L_{K-cov} \left( x, \frac{\epsilon}{2} \right) := x + \frac{1}{\lambda} \log \frac{2N}{\epsilon}. \]  
(23)

Thus, for any \( x \leq L \), we may replace (22) by
\[ L \geq \max_{x} \left\{ \frac{1}{2\lambda} \log \frac{2\gamma(x)}{\epsilon}, L_{K-cov} \left( x, \frac{\epsilon}{2} \right) \right\}, \]  
where \( \gamma(x) = \sum_{m>x} a_m e^{2(N/G)m} \), and obtain a looser bound.

**Critical window calculations**

**Window size if \( \ell_{\text{interleaved}} \gg \ell_{\text{triple}} \)**

We focus here on the bound due to interleaved repeats (rather than triple repeats, treated subsequently), and furthermore assume that the effect of the single largest interleaved repeat is dominant. In this case \( \ell_{\text{interleaved}} = L_{\text{crit}} \) is the length of the shorter of the pair of interleaved repeats, and let \( \ell_1 \) be the length of the longer of the two. For \( L_{\text{crit}} < L \leq \ell_1 + 1 \), we are in the setting of (9) but with a redefined \( \gamma_1 = e^{2(N/G)(L_{\text{crit}}-1)} \). Thus,
\[ L \geq L_{\text{crit}} + \frac{G}{2N} \log \epsilon^{-1}, \]  
and solving for \( N \) gives
\[ N_{\text{repeat}} = \frac{G}{2} \frac{\log \epsilon^{-1}}{L - \ell_2 - 1}. \]  
(26)

Let \( L^* \) be the value of \( L \) at which the curve described by constraint (26) intersects the Lander-Waterman coverage value, i.e. \( N_{\text{repeat}}(L^*) = N_{\text{LW}}(L^*) := N^* \). This is the minimum read length for which coverage of the sequence suffices for reconstruction.

We now solve for \( L^*/L_{\text{crit}} \). First, the Lander-Waterman equation (2) at \( N = N^* \) is
\[ N^* = \frac{G}{L^*} \log \frac{N^*}{\epsilon}, \]  
(27)

and setting equal the right-hand sides of (27) and (26) at \( L = L^* \) gives
\[ \frac{G}{L^*} \log \frac{N^*}{\epsilon} = \frac{G}{2} \log \epsilon^{-1} \]
\[ \log \log \frac{N^*}{\epsilon} - \frac{G}{L_{\text{crit}}} \leq x \]  
where
\[ x := \frac{\log \epsilon^{-1} - \log N^* + \log \epsilon^{-1}}{2}. \]  
(29)

Since \( x \leq \frac{1}{2} \), equation (28) implies \( L^* \geq 2L_{\text{crit}} \), and combined with the obvious inequality \( L^* \geq L_{\text{crit}} \), we have \( L_{\text{crit}} \leq L^* \leq 2L_{\text{crit}} \). Thus
\[ N_{\text{LW}}(2L_{\text{crit}}) \leq N^* \leq N_{\text{LW}}(L_{\text{crit}}), \]  
(30)

and applying the Lander-Waterman fixed-point equation (2) yet again gives
\[ \frac{G}{2L_{\text{crit}}} \log \frac{N_{\text{LW}}(2L_{\text{crit}})}{\epsilon} \leq N^* \leq \frac{G}{L_{\text{crit}}} \log \frac{N_{\text{LW}}(L_{\text{crit}})}{\epsilon}. \]  
(31)

Writing this out gives
\[ \frac{\log \epsilon^{-1} - \log G}{L_{\text{crit}}} + \log \log \frac{N_{\text{LW}}(L_{\text{crit}})}{\epsilon} + \log \epsilon^{-1} \leq x \]
\[ \leq \frac{\log \epsilon^{-1} - \log G}{L_{\text{crit}}} - 1 + \log \log \frac{N_{\text{LW}}(2L_{\text{crit}})}{\epsilon} + \log \epsilon^{-1}, \]
and this can be relaxed to
\[ \frac{\log \epsilon^{-1} - \log G}{L_{\text{crit}}} + \log \epsilon^{-1} + \log \log \frac{G}{L_{\text{crit}}} \leq x \]  
(32)

Letting
\[ r := \frac{\log G}{\log \epsilon^{-1}}, \]  
(33)
we have to a very good approximation

\[ \frac{L^*}{L_{\text{crit}}} \approx \frac{2(r + 1)}{2(r + 1) - 1}. \]  

(34)

For \( G \sim 10^8, L_{\text{crit}} \sim 1000, \) and \( \epsilon = 5\%, \) we get \( \log \frac{G}{L_{\text{crit}}} \approx 13.8 \) and \( \log \epsilon^{-1} \approx 3.0, \) so \( r \approx 4.6 \) and

\[ \frac{L^*}{L_{\text{crit}}} = \frac{2(r + 1)}{2(r + 1) - 1} \approx 1.1. \]

From (33) we see that the relative size of \( \log \epsilon^{-1} \) and \( \log \frac{G}{L_{\text{crit}}} \) determines the size of the critical window. If in the previous example \( \epsilon = 10^{-5}, \) say, then \( \frac{L^*}{L_{\text{crit}}} \) increases to 1.3. As \( \epsilon \) tends to zero, \( r \) approaches zero as well and \( \frac{L^*}{L_{\text{crit}}} \to 2. \)

**Window size if \( \ell_{\text{triple}} \gg \ell_{\text{interleaved}} **

We now suppose the single longest triple repeat dominates the lower bound and estimate the size of the critical window. In this case \( \ell_{\text{triple}} = L_{\text{crit}} - 1 \) is the length of the longest triple repeat. Since we don’t have matching lower and upper bounds for triple repeats, we separately compute the critical window size for each.

We start with the lower bound. For \( L > L_{\text{crit}}, \) the minimum value of \( N \) required in order to bridge the longest triple repeat is given by (18) and repeated here:

\[ N_{\text{triples}} = \frac{G}{3} \cdot \frac{\log \epsilon^{-1}}{L - L_{\text{crit}}}. \]  

(35)

As for the interleaved repeats case considered earlier, we let \( L^* \) be the value of \( L \) at which the curve described by constraint (35) intersects the Lander-Waterman coverage value, i.e. \( N_{\text{triple}}(L^*) = N_{\text{LW}}(L^*) := N^*. \) This is the minimum read length for which coverage of the sequence suffices for reconstruction.

A similar procedure as leading to (28) gives \( L^*/L_{\text{crit}} = 3/(3 - x) \) with \( x \) defined in (29). One can check that the estimates on \( x \) in (32) continue to hold, and we therefore get

\[ \frac{L^*}{L_{\text{crit}}} \approx \frac{3(r + 1)}{3(r + 1) - 1}. \]  

(36)

For the same example as before, \( G \sim 10^8, L_{\text{crit}} \sim 1000, \) and \( \epsilon = 5\%, \) we get \( r \approx 4.6 \) and

\[ \frac{L^*}{L_{\text{crit}}} \approx \frac{3(r + 1)}{3(r + 1) - 1} \approx 1.06. \]

Changing \( \epsilon \) to \( 10^{-5} \) makes \( \frac{L^*}{L_{\text{crit}}} \approx 1.17, \) and as \( \epsilon \) (and hence also \( r \)) tends to zero, \( \frac{L^*}{L_{\text{crit}}} \to \frac{3}{2}. \)

The analogous computation for \( L^*/L_{\text{crit}} \) for the upper bound, as given by \( N^\text{all} \) in (18), yields

\[ \frac{L^*}{L_{\text{crit}}} = \frac{r + 1}{r + \frac{\log 3}{\log \epsilon^{-1}}} \approx 1.12, \]  

(37)

for the example with \( G \sim 10^8, L_{\text{crit}} \sim 1000, \) and \( \epsilon = 5\%. \) The critical window size of the upper bound is about twice as large as that of the lower bound for typical values of \( G \) and \( L_{\text{crit}}, \) with \( \epsilon \) moderate. But as \( \epsilon \to 0, \) we see from (37) that \( L^*/L_{\text{crit}} \to \infty, \) markedly different to the \( L^*/L_{\text{crit}} \to \frac{3}{2} \) observed for the lower bound.