A Safety of unitigs: full exposition

Proof of Theorem 1 and Corollary 1

In this subsection, we will prove Theorem 1 and Corollary 1. In the following, we will always have $\mathcal{S}$ be a set of sequenced segments and $w = (x_0, \ldots, x_m)$ be a unitig in $G_{\text{basic}}(\text{sp}^k(\mathcal{S}))$. We start with a lemma that, roughly speaking, says that if a walk corresponding to some $S \in \mathcal{S}$ touches w, it must contain all of w except that it may begin or end somewhere along the way.

Lemma A.1. *Let $\mathcal{S}$ be a set of sequenced segments and let $w = (x_0, \ldots, x_m)$ be a unitig in $G_{basic}(sp^k(\mathcal{S}))$. Let $S \in \mathcal{S}$ and let $g = (g_0, \ldots, g_{|S|})$ be the walk corresponding to S. Suppose there exists i and j such that $x_i = g_j$. Then,*

(i) If $suf_k(S) \notin \{x_i, \ldots, x_{m-1}\}$, then $g_{j+\delta} = x_{i+\delta}$ for all $\delta \in [0, m-i]$.

(ii) If $pre_k(S) \notin \{x_1, \ldots, x_i\}$, then $g_{j-\delta} = x_{i-\delta}$ for all $\delta \in [0, i]$.

Proof. We will only prove (i), since the argument for (ii) is symmetric. We use induction on δ. For $\delta = 0$, we have that the implication of (i) reduces to $g_j = x_i$, which is vacuously true because it is also a condition of the theorem. Now we assume that (i) holds for $\delta - 1$, i.e. $g_{j+\delta-1} = x_{i+\delta-1}$. Since $x_{i+\delta-1} \neq suf_k(G)$, $g_{j+\delta-1}$ is not the last vertex of g. Because $x_{i+\delta-1}$ is a non-last vertex of a unitig, it has only one out-neighbor, which is $x_{i+\delta}$. Therefore, $g_{j+\delta} = x_{i+\delta}$, which shows that that (i) holds for δ. □

Using this lemma, we can now prove some general properties of unsafe unitigs.

Lemma A.2. *Let $\mathcal{S}$ be a set of sequenced segments and let $w = (x_0, \ldots, x_m)$ be a unitig in $G_{basic}(sp^k(\mathcal{S}))$. If w is unsafe then*

(i) $m \geq 1$,

(ii) there exists $S \in \mathcal{S}$ such that $pre_k(S) \in \{x_1, \ldots, x_m\}$,

(iii) there exists $S \in \mathcal{S}$ such that $suf_k(S) \in \{x_0, \ldots, x_{m-1}\}$,

(iv) for all $S \in \mathcal{S}$ and their corresponding walks g, either g and w do not share a vertex, or $pre_k(S) \in \{x_1, \ldots, x_m\}$, or $suf_k(S) \in \{x_0, \ldots, x_{m-1}\}$, and

(v) for all $S \in \mathcal{S}$ and all i, $occ_S(x_i) \leq 2$.

Proof. For (i), consider a unitig that has just one vertex x. Since each k-mer in $G_{\text{basic}}(\text{sp}^k(\mathcal{S}))$, there must be at least one $S \in \mathcal{S}$ whose walk contains x. Hence, the unitig that is composed of only x is safe. For (ii), assume for sake of contradiction that for all $S \in \mathcal{S}$, $pre_k(S) \notin \{x_1, \ldots, x_m\}$. Since every vertex of the graph must be contained in at least one string, let $S' \in \mathcal{S}$ be a string that contains x_m. Applying Lemma A.1(ii) with $i = m$, we get that the walk corresponding to S' must contain w, contradicting that w is unsafe. The case of (iii) is symmetric to (ii), using x_0 instead of x_m and applying Lemma A.1(i) with $i = 0$. For (iv), let $g = (g_0, \ldots, g_{|S|})$ and assume for sake of contradiction that there exists a $S \in \mathcal{S}$ such that g shares a vertex with w and $pre_k(S) \notin \{x_1, \ldots, x_m\}$ and $suf_k(S) \notin \{x_0, \ldots, x_{m-1}\}$. Let x_i and g_j be the vertices of w and g, respectively, that are equivalent. We can apply Lemma A.1 to get that $(g_j, \ldots, g_{j+m-i}) = (x_i, \ldots, x_m)$ and $(g_{j-i}, \ldots, g_j) = (x_0, \ldots, x_i)$. This means that w is a subwalk of g, which is a contradiction. For (v), let $S \in \mathcal{S}$ and let $g = (g_0, \ldots, g_{|S|})$ be its corresponding walk. If g and w do not share any vertices, then $occ_S(x_i) = 0 \leq 2$ for all i and we are done. Otherwise, we can apply (iv) to get that

either (1) $pre_k(S) \in \{x_1, \ldots, x_m\}$ or (2) $suf_k(S) \in \{x_0, \ldots, x_{m-1}\}$. Let us consider (1) — we will omit the argument for (2) since it is symmetrical. Then $g_0 = x_i$ for some $1 \leq i \leq m$. Note that g_0 is the first occurrence of x_i in g. Assume for the sake of contradiction that $occ_S(x_i) > 2$. To get the second occurrence of x_i, g must first visit x_0. After this second visit to x_i, g must continue all the way until x_m if it is to visit x_i for a third time. Therefore, at the second visit to x_i, g must in fact visit $(x_0, \ldots, x_m)$, which contradicts that w is unsafe. □

The case when a sequenced segment contains its first and/or last k-mer more than once puts additional constraints on how it can contain a unitig.

Lemma A.3. *Let $\mathcal{S}$ be a set of sequenced segments and let $w = (x_0, \ldots, x_m)$ be a unitig in $G_{basic}(sp^k(\mathcal{S}))$. Let $S \in \mathcal{S}$ such that at least one of the following holds:*

(i) $occ_S(pre_k(S)) = 2$ and there exists an integer $i \in [1, m]$ such that $x_i = pre_k(S)$, or

(ii) $occ_S(suf_k(S)) = 2$ and there exists an integer $j \in [i, m-1]$ such that $x_j = suf_k(S)$.

Then, $spell(w)$ is not a substring of S iff both (i) and (ii) hold.

Proof. We only prove case (i) since case (ii) is symmetrical. Let g be the walk corresponding to S. In the first phase, g starts from x_i and, since it must visit x_i a second time, continues until x_m. Then at some point it enters w through x_0 and proceeds to visit x_i for the second and last time. We will refer to the time from the end of the first phase to the point it enters x_0 as the second phase, and the rest of the walk as the third phase. Observe that g does not contain w as a subwalk in either the first or second phase.

Now we prove the if direction. During phase 1, g visits x_j exactly once. During phase 2, g does not visit x_j. During phase 3, g proceeds from x_0 forward along the unitig until it hits x_j for the second time. Since x_j is occurs exactly twice and is the last vertex of g, this is the end of g. Since $j < m$, g does not contain w as a subwalk during the third phase.

Now we prove the only if direction. Assume w is not a subwalk of g. Therefore, during the third phase g cannot go until x_m and must stop earlier at some $x_j = suf_k(S)$, for some integer $j \in [i, m-1]$. This x_j was visited once during phase 1 and not visited during phase 2 and now visited a second and final time during phase 3. □

These lemmas are all the pieces we need to prove Theorem 1.

Theorem 1. *Let $\mathcal{S}$ be a set of sequenced segments and let $w = (x_0, \ldots, x_m)$ be a unitig in $G_{basic}(sp^k(\mathcal{S}))$. Then w is unsafe if and only if for all $S \in \mathcal{S}$, one of the following holds:*

(i) S does not contain any k-mer of w,

(ii) $occ_S(pre_k(S)) = 1$ and $pre_k(S) = x_i$ for some $1 \leq i \leq m$,

(iii) $occ_S(suf_k(S)) = 1$ and $suf_k(S) = x_j$ for some $0 \leq j \leq m-1$, or

(iv) $occ_S(pre_k(S)) = occ_S(suf_k(S)) = 2$ and there exists $1 \leq i \leq j \leq m-1$ such that $pre_k(S) = x_i$ and $suf_k(S) = x_j$.

Proof. First we prove the if direction. We will show that for all $S \in \mathcal{S}$ and its corresponding walk g, if one of the four conditions hold, then w is not a subwalk of g. If (i) holds, then w is trivially not a subwalk of g. Now, if $pre_k(S) = x_i$ for some $1 \leq i \leq m$ and x_i is visited only once by g, If (ii) holds, then g starts with x_i but never visits x_i again, therefore $(x_0, \ldots, x_i)$ is not a subwalk of

g. Hence, w is not a subwalk of g. Similarly, if (iii) holds, then $(x_j, \ldots, x_m)$ is not a subwalk of g and hence w is not a subwalk of g. If (iv) holds, then Lemma A.3 implies that w is not a subwalk of g.

Now we prove the only if direction. We will show that for all $S \in \mathcal{S}$ and their corresponding walk g, if w is not a subwalk of g, then one of the four conclusions hold. By Lemma A.2.(iv), either (1) g does not contain any k-mer from w, (2) $pre_k(S) \in \{x_1, \ldots, x_m\}$, or (3) $suf_k(S) \in \{x_0, \ldots, x_{m-1}\}$. In case of (1), g trivially does not contain w, and condition (i) is satisfied. In case of (2), let $i \in [1, m]$ be an integer such that $x_i = pre_k(S)$. By Lemma A.2.(v), $occ_S(x_i)$ is either 1 or 2. If $occ_S(x_i) = 1$, then condition (ii) immediately holds. If $occ_S(x_i) = 2$, then Lemma A.3 implies that there exists an integer $j \in [i, m-1]$ that satisfies condition (iv). In case of (3), let $j \in [0, m-1]$ be an integer such that $x_j = suf_k(S)$. Again, by Lemma A.2.(v), $occ_S(x_j)$ is either 1 or 2. If $occ_S(x_j) = 1$, then condition (iii) immediately holds. If $occ_S(x_j) = 2$, then Lemma A.3 implies that there exists an integer $i \in [i, m-1]$ that satisfies condition (iv). □

Corollary 1. *Let X be a string and let $w = (x_0, \ldots, x_m)$ be a unitig in $G_{basic}(sp^k(X))$. Then $spell(w)$ is not a substring of X iff one of the following holds:*

1. *$occ_X(pre_k(X)) = occ_X(suf_k(X)) = 1$, $pre_k(X) = x_i$, $suf_k(X) = x_{i-1}$ for some $1 \leq i \leq m$.*
2. *$occ_X(pre_k(X)) = occ_X(suf_k(X)) = 2$, $pre_k(X) = x_i$, $suf_k(X) = x_j$ for some $0 < i \leq j < m$.*

Moreover, this can hold for at most one unitig in $G_{basic}(sp^k(X))$.

Proof. We can apply Theorem 1 to set $\mathcal{S} = \{X\}$. Since X must contain all k-mers of w, w is unsafe if and only if condition (ii), (iii) or (iv) from Theorem 1 holds for $S = X$. First, assume Condition (ii) is true for X. Then by Theorem 1, w is unsafe. Consider the walk g corresponding to X. Because g begins at x_i and all vertices in $G_{\text{basic}}(K)$ must be in g at least once, $(x_i, \ldots, x_m)$ is a subwalk of g. This is the one and only occurrence of x_i in g. Since x_i is the first vertex in g occuring only once, x_{i-1} cannot precede x_i. Hence, x_{i-1} must be the end of g, i.e., $suf_k(S) = x_{i-1}$. Note that, this is the one and only occurrence of x_{i-1} in g. Thus, Condition (iii) is also true for X. With a symmetric argument, we can show that if Condition (iii) is satisfied, then Condition (ii) is satisfied. Combining both gives us the first condition of the corollary. Finally, observe that Condition (iv) is identical to Condition 2 in the corollary. The fact that these conditions can hold for at most one unitig follows directly from the fact that there is only one vertex for $pre_k(S)$ in the graph. □

Formal definition of the case of Figure 3

In the Results section, we quantify the number of unsafe unitigs that fall into the case of Figure 3. To make this precise, we give a formal classification for this case. Let X be a genome and let $\mathcal{S}$ be a set of its sequenced segments. We say that an unsafe walk $w = (x_0, \ldots, x_m)$ satisfies the case of Figure 3 if

(i) there exists $0 < i \leq j < m$ such that $\psi = (x_i, \ldots, x_j)$ is a unitig in $G_{\text{basic}}(\text{sp}^k(X))$,
(ii) $spell(\psi)$ occurs at least twice in X,
(iii) in one of the occurrences, the k-mer preceding $spell(\psi)$ is not in $\mathcal{S}$,
(iv) in another of the occurrences, the k-mer following $spell(\psi)$ is not in $\mathcal{S}$,
(v) there exists $S \in \mathcal{S}$ and an integer $i' \in [i, j]$ such that $spell((x_{i'}, \ldots, x_j))$ is a suffix of S, and
(vi) there exists $S \in \mathcal{S}$ and an integer $j' \in [i'-1, j]$ such that $spell((x_i, \ldots, x_{j'}))$ is a suffix of S.

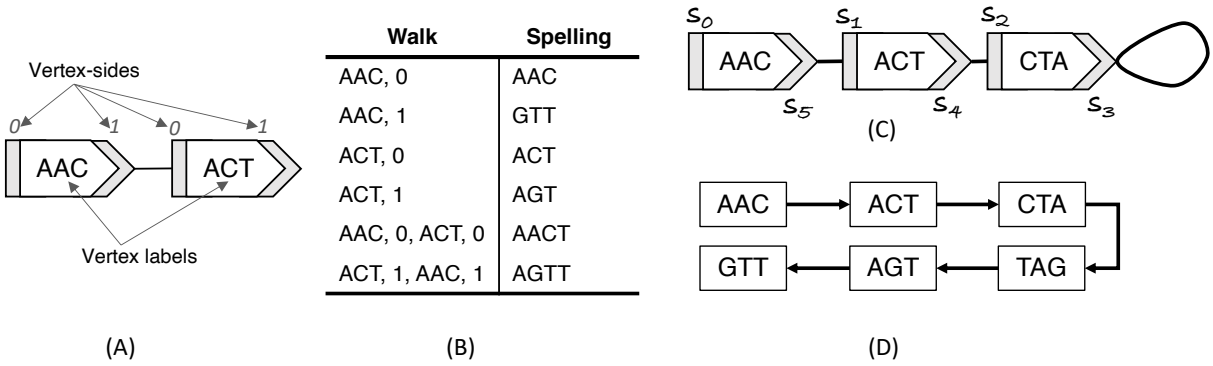

Figure S1: An example illustrating some of the bidirected graph terminology. Panel (A) shows a bidirected graph with two vertices. Panel (B) shows a list of all possible walks in this graph and their spellings. Note that walk $(AAC, 0, ACT, 0)$ and $(ACT, 1, AAC, 1)$ are reverses of each other. The endpoint sides of the walks, in both cases, are $(AAC, 0)$ and $(ACT, 1)$. Panel C and D show an example of a doubled dBG ($G_{\text{dbl}}(K)$) and a bidirected dBG ($G_{\text{bid}}(K)$) using $K = \{AAC, ACT, CTA\}$. Panel C shows $G_{\text{bid}}(K)$ as well as the order of vertex-sides as they appear in a walk $w = (AAC, 0, ACT, 0, CTA, 0, CTA, 1, ACT, 1, AAC, 1) = (s_0, \ldots, s_5)$, with $spell(w) = AACTAGTT$. Note that in this case, since $spell(w)$ is a palindrome, the reverse walk is identical: $rev(w) = w$. Panel D shows $G_{\text{dbl}}(K)$.

B The relationship of $G_{\text{dbl}}(K)$ and $G_{\text{bid}}(K)$: full exposition

In this section, we will prove Theorem 2. We start by providing additional definitions that are necessary to understand the proofs in this section.

Let K be a set of k-mers. A unitig in a directed graph that is not a proper subwalk of another unitig that ends at the same vertex is said to be *prefix-maximal*; a unitig that is not a proper subwalk of another unitig that starts with the same vertex is said to be *suffix-maximal*. Notice that a unitig is maximal iff it is both prefix- and suffix-maximal.

Let (u, s) be a vertex-side in $G_{\text{bid}}(K)$. We define $d^{il}(u, s)$ to indicate the presence of an inverted loop, i.e. $d^{il}(u, s) = 1$ if there is an inverted loop incident to side (u, s) and $d^{il}(u, s) = 0$ otherwise. A unitig t in $G_{\text{bid}}(K)$ is *prefix-maximal* if it is not a proper subwalk of another unitig that ends at the same vertex-side as t. A unitig is *suffix-maximal* if it is not a proper subwalk of another unitig that starts with the same vertex-side as t. Note that a unitig is maximal iff it is both prefix- and suffix-maximal.

We will prove Theorem 2 by first building a collection of Lemmas. First, we make a simple observation. A palindrome must have an even number of characters, otherwise there is a middle character that would need to be equal to its own reverse complement. Hence, a palindromic walk, in either the doubled or the bidirected graph must have an even number of nucleotides.

Lemma B.1. *Let K be a set of k-mers.*

1. *For all palindromic walks $w = (x_0, \ldots, x_n)$ in $G_{dbl}(K)$, k and n have the same parity.*
2. *For all palindromic walks $t = (u_0, s_0, \ldots, u_n, s_n)$ in $G_{bid}(K)$, k and n have the same property.*

Proof. A palindromic string must have an even number of nucleotides. The number of nucleotides in $spell(w)$ and in $spell(t)$ is $k + n$, Hence the parity of k and n must be the same. □

From here on out, we proceed by first proving Lemmas for the directed de Bruijn graphs (both the regular one and the doubled one) (Appendix B.1), then proving Lemmas for the bidirected graph (Appendix B.2), then proving Lemmas which connect the two types of graphs (Appendix B.3), and, finally, proving Theorem 2 (Appendix B.4).

B.1 Directed graph

First, we make the observation that unitigs cannot repeat vertices unless they are a simple cycle. This is generally stated without proof, but the statement is actually not true when unitigs are allowed to be periodic cycles. In our definition of unitig, we forbid this case, allowing us to prove the observation.

Lemma B.2. *For all unitigs w in a directed graph, either w is a simple cycle or it does not repeat any vertices.*

Proof. Let $w = (x_0, \ldots, x_n)$ be a unitig. Suppose that w repeats a vertex. Let $0 \leq j \leq n$ be the smallest value for which there exists $0 \leq i < j$ such that $x_i = x_j$. If $i > 0$, then x_i has x_{i-1} and x_{j-1} as an in-neighbor. By the minimality of our choice of i, $x_{i-1} \neq x_{j-1}$, and hence $d^-(i) \geq 2$. This contradicts that w is a unitig. If $i = 0$, then let $j + 1 \leq \ell \leq n - 1$ be the largest index greater than j such that $x_\ell = x_{\ell \bmod (j+1)}$. In other words, ℓ is the first place after x_j where the unitig is about to "fall off the cycle". If such an ℓ does not exist, then either $j = n$ and w is a simple cycle, or w is a simple periodic cycle, contradicting the definition of a unitig. Otherwise, the vertex x_ℓ has as out-neighbors both $x_{\ell+1}$ and $x_{\ell+1 \bmod (j+1)}$. By the choice of ℓ, these out-neighbors are distinct and hence $d^+(x_\ell) \geq 2$. This contradicts that w is a unitig. □

A very simple property in the doubled graph is is that the in-degree (respectively, out-degree) of a vertex is equal to the out-degree (respectively, in-degree) of its reverse complement.

Lemma B.3. *Let K be a set of k-mers and let x be a vertex in $G_{dbl}(K)$. Then $d^+(x) = d^-(\overline{x})$ and $d^-(x) = d^+(\overline{x})$.*

Proof. Observe that for all vertices y in the $G_{\text{dbl}}(K)$, there is an edge from x to y in $G_{\text{dbl}}(K)$ iff there is an edge from $\overline{y}$ to $\overline{x}$. This is true even if $x = \overline{y}$ and these two edges are identical. Hence $d^+(x) = d^-(\overline{x})$ and $d^-(x) = d^+(\overline{x})$. □

We defined maximal unitigs as those that are not proper sub-walks of other unitigs. We can give an equivalent definition for directed graphs, in terms of vertex degrees. Since it is widely known, we state it without proof.

Lemma B.4. *Let G be a directed graph and let $w = (x_0, \ldots, x_n)$ be a unitig in G. Then*

(i) w is prefix-maximal if and only if $d^-(x_0) \neq 1$ or there exists a vertex x' that has an edge to x_0 and $d^+(x') > 1$.

(ii) w is suffix-maximal if and only if $d^+(x_n) \neq 1$ or there exists a vertex x' that has an edge from x_n and $d^-(x') > 1$.

Palindromic unitigs play a special role in Theorem 2. We observe that in a palindromic unitig of the doubled graph, the only edge from a k-mer to its reverse complement is the middle one.

Lemma B.5. *Let K be a set of k-mers with odd k. Let $w = (x_0, \ldots, x_n)$ be a palindromic unitig in $G_{dbl}(K)$ that is not a simple cycle. Then for all $0 \leq i \leq n - 1$, we have that $x_i = \overline{x_{i+1}}$ iff $i = (n - 1)/2$.*

Proof. First note that by Lemma B.1, n is odd and $n \geq 1$. Let $m = (n - 1)/2$. Because $spell(w)$ is a palindrome, $x_i = \overline{x_{n-i}}$ for all $0 \leq i \leq n$. The only if direction of the Lemma statement follows immediately by plugging in $i = m$ and getting $x_m = \overline{x_{n-m}} = \overline{x_{m+1}}$. For the if direction, assume that $x_i = \overline{x_{i+1}}$ for all $0 \leq i \leq n - 1$. Then $x_i = \overline{x_{i+1}} = \overline{\overline{x_{n-i-1}}} = x_{n-i-1}$. By the fact that w is not a simple cycle and Lemma B.2, it cannot have any repeated vertices. Hence, $i = n - i - 1$ which only happens when $i = m$. □

We also observe that a maximal unitig that is not a palindrome cannot contain within it a palindrome of length $\geq k$.

Lemma B.6. *A non-palindromic maximal unitig w in $G_{dbl}(K)$ cannot contain a proper sub-unitig that is palindromic.*

Proof. For the sake of contradiction, let z be a proper sub-unitig of w that is a palindrome. First suppose that there exists a k-mer y such that y precedes z in w and $\overline{y}$ follows z in w. In that case, observe that the walk $(y, z, \overline{y})$ is also a sub-unitig of w and also a palindrome. We can then extend z in this way until no longer possible, i.e. there do not exist a k-mer y such that y precedes z in w and $\overline{y}$ follows z in w. Let w' be this maximally extended walk. Note that by construction, w' is a sub-unitig of w and it is proper because w' is palindromic and w is not. Let the first vertex of w' be x, and, hence, the last one is $\overline{x}$.

Consider the case when w starts with x. Because $w \neq w'$, there must exist an out-neighbor u of $\overline{x}$ in w. Its mirror must also exist, i.e. an edge from $\overline{u}$ to x. Lemma B.4 states that x is the first vertex of a maximal unitig, it must either (a) have one other in-neighbor besides $\overline{u}$ or (b) $\overline{u}$ must have at least one other out-neighbor besides x. For case (a), Lemma B.3 implies that $d^+(\overline{x}) = d^-(x) > 1$. For case (b), Lemma B.3 implies that $d^+(\overline{u}) = d^-(u) > 1$. In either case, the degrees of $\overline{x}$ or of u contradict the definition of being part of a unitig. The case when w ends with $\overline{x}$ is symmetric and omitted.

Now consider the case when w does not start with x and does not end with $\overline{x}$. Let a be the vertex preceding x in w, and let b be the vertex following $\overline{x}$ in w. There exist a mirror edge from $\overline{b}$ to x. Since w' was chosen so that it cannot be extended, $\overline{a} \neq b$. Hence x has two distinct in-neighbors, a and $\overline{b}$. Since w contains x as a non-first vertex, this contradicts that w is a unitig. □

B.2 Bidirected graph

As is the case with directed graphs (Lemma B.4), there is a definition of maximality for bidirected unitigs that has to do with degrees rather than sub-unitigs. We are not aware of this equivalence being explicitly proven, so we do so here:

Lemma B.7. *Let K be a set of canonical k-mers. Let $t = (u_0, s_0, \ldots, u_n, s_n)$ be a unitig in $G_{bid}(K)$. Then*

(i) t is prefix-maximal if and only if $d(u_0, s_0) \neq 1$ or there is an edge $\{(u_0, s_0), (u', s')\}$ such that $d(u', s') > 1$, and

(ii) t is suffix-maximal if and only if $d(u_n, 1 - s_n) \neq 1$ or there is an edge $\{(u_n, 1 - s_n), (u', s')\}$ such that $d(u', s') > 1$.

Proof. We will only prove (i) since the proof of (ii) is symmetric. First, we prove the only if direction. We need to consider three cases. The first case is when $d(u_0, s_0) \neq 1$. If $d(u_0, s_0) = 0$, then t is prefix-maximal because there is no other walk of which it is a subwalk with the same last vertex-side (u_n, s_n). The second case is when $d(u_0, s_0) > 1$. Consider any walk t' that ends in (u_n, s_n) and of which t is a proper subwalk. Observe that (u_0, s_0) would not be the first vertex-side of t'. Therefore, since $d(u_0, s_0) > 1$, t' cannot be a unitig and t must be prefix-maximal. The third case is when (u_0, s_0) has degree one and (u', s') is its only neighbor. Again, consider any walk t' that ends in (u_n, s_n) and of which t is a proper subwalk. Observe that $(u', 1 - s')$ belongs to t' but is not the last vertex-side of t'. Therefore, since we assumed that $d(u', s') > 1$, t' cannot be a unitig and t must be prefix-maximal.

To prove the if direction we prove the contrapositive. In other words, we will show that if the degree of (u_0, s_0) is one and its sole neighbor (u', s') also has degree at most 1, then t is not prefix-maximal. First, observe that $t' = (u', 1 - s', u_0, s_0, \ldots, u_n, s_n)$ is a valid walk, since the edge $\{(u', s'), (u_0, s_0)\}$ exists. Then, observe that the degree of (u', s') is exactly one because it has degree at most one (by our assumption) and also has a neighbor (i.e. (u_0, s_0)). Therefore, the degree requirements for t' being a unitig are fulfilled. Finally, observe that t is a proper subwalk of t' ending in the same vertex-side, (u_n, s_n). Therefore, t is not prefix-maximal. □

In a bidirected graph, a walk and its reverse are either both unitigs or not and, if they are, are either both are maximal or not.

Lemma B.8. *Let K be a set of canonical k-mers and let w be a unitig in $G_{bid}(K)$.*

(i) $rev(w)$ is a unitig in $G_{bid}(K)$.

(ii) w is prefix-maximal iff $rev(w)$ is suffix-maximal.

(iii) w is suffix-maximal iff $rev(w)$ is prefix-maximal.

Proof. Let $(u_0, s_0, \ldots, u_n, s_n) = w$ and $(u'_0, s'_0, \ldots, u'_n, s'_n) = rev(w)$. For (i), by definition of rev, we have that $u'_i = u_{n-i}$ and $s'_n = 1 - s_{n-i}$. Applying the definition of unitig to w, we get that

$$d(u_i, 1 - s_i) \leq 1 \text{ for all } 0 \leq i < n \qquad \text{and} \qquad d(u_i, s_i) \leq 1 \text{ for all } 0 < i \leq n.$$

These can be equivalently stated as

$$d(u'_{n-i}, s'_{n-i}) \leq 1 \text{ for all } 0 \leq i < n \qquad \text{and} \qquad d(u'_{n-i}, 1 - s'_{n-i}) \leq 1 \text{ for all } 0 < i \leq n$$

If we change the index variables, these can be equivalently restated as

$$d(u'_i, s'_i) \leq 1 \text{ for all } 0 < i \leq n \qquad \text{and} \qquad d(u'_i, 1 - s'_i) \leq 1 \text{ for all } 0 \leq i < n.$$

This is precisely the definition of $rev(w)$ being a unitig.

For (ii) and (iii), first observe that Lemma B.7 gives an alternate, equivalent, definition for prefix- and suffix-maximal. For (ii), observe that if apply the alternate definition of suffix-maximal to $rev(w)$ and plug in that $u'_n = u_0$ and $s'_n = 1 - s_0$, we get precisely the alternate definition of w being prefix-maximal. For (iii), observe that if apply the alternate definition of prefix-maximal to $rev(w)$ and plug in that $u'_0 = u_n$ and $s'_0 = 1 - s_n$, we get precisely the alternate definition of w being suffix-maximal. □

While we showed that it is natural for the doubled graph to have a palindromic unitig, this is impossible in a bidirected graph.

Lemma B.9. *Let K be a set of canonical k-mers, with k odd. Then a unitig of $G_{bid}(K)$ cannot be a palindrome.*

Proof. Let $t = (u_0, s_0, \ldots, u_n, s_n)$ be a palindromic walk. By Lemma B.1, n is odd, and so $n \geq 1$. For convenience, let $m = (n - 1)/2$. By definition, $spell(t) = \overline{spell(t)}$. In particular, the two "central" k-mers of $spell(t)$ must be reverse complements of each other. Formally, $orient(lab(u_m), s_m) = orient(lab(u_{m+1}), s_{m+1})$. Since the labels of vertices in a bidirected graph are distinct, $lab(u_m) \neq lab(u_{m+1})$ and hence $s_m = 1 - s_{m+1}$. Applying the definition of a bidirected walk to t, we get that $\{(u_m, 1 - s_m), (u_{m+1}, s_{m+1})\}$ is an edge. The fact that $s_m = 1 - s_{m+1}$ implies that this edge is an inverted loop incident to $(u_m, 1 - s_m)$. Thus $d(u_m, 1 - s_m) \geq 2$, implying that t does not satisfy the definition of being a unitig. □

B.3 Connecting the directed and bidirected graphs

So far, we have proven properties of the doubled graph and of the bidirected graph separately; in this section, we prove lemmas about the relationship between the two graphs, when k is odd. Recall that for a k-mer $x \in K$, we defined $F_V(x) = (u, s)$, where (u, s) is the unique vertex-side in $G_{\text{bid}}(K)$ such that $lab(u) = orient(x, s)$.

Lemma B.10. *Let K be a set of canonical k-mers where k is odd. F_V is a bijection between vertices of $G_{dbl}(K)$ and vertex-sides of $G_{bid}(K)$.*

Proof. To show that F_V is a bijection, we will show that for all vertex-sides (u, s) in $G_{\text{bid}}(K)$, there exists a unique k-mer x in $G_{\text{dbl}}(K)$ such that $F_V(x) = (u, s)$. Consider a value of x such that $F_V(x) = (u, s)$. By definition, $lab(u) = orient(x, s))$. Since k is odd and x is not a palindrome, the value of x satisfying this must be unique. By construction of $G_{\text{dbl}}(K)$ and $G_{\text{bid}}(K)$, k must be a vertex in $G_{\text{dbl}}(K)$. Further, if $x = orient(lab(u), s)$, then $orient(x, s) = orient(orient(lab(u), s), s) = lab(u)$ and so x satisfies the condition that $F_V(x) = (u, s)$. □

We will use F_V^{-1} to denote the inverse of F_V, which was shown in Lemma B.10 is $F_V^{-1}(u, s) = orient(lab(u), s)$. We will use $x \overset{F_V}{\Longleftrightarrow} (u, s)$ to denote that a vertex x of $G_{\text{dbl}}(K)$ and a vertex-side (u, s) in $G_{\text{bid}}(K)$ are associated with each other by F_V.

Recall that for two $G_{\text{dbl}}(K)$ k-mers x_1 and x_2, we define the mapping $F_E(x_1, x_2) = \{(u_1, 1 - s_1), (u_2, s_2)\}$, where $(u_1, s_1) = F_V(x_1)$ and $(u_2, s_2) = F_V(x_2)$. Though the mapping is not a bijection, it preserves the property of being an edge in the respective graph[2]:

Lemma B.11. *Let K be a set of canonical k-mers where k is odd. Let x_1 and x_2 be vertices in $G_{dbl}(K)$. We have that (x_1, x_2) is an edge in $G_{dbl}(K)$ if and only if $F_E(x_1, x_2)$ is an edge in $G_{bid}(K)$.*

Proof. By the definition of bidirected edges, $F_E(x_1, x_2) = \{(u_1, 1 - s_1), (u_2, s_2)\}$ is an edge iff

$$suf(orient(lab(u_1), s_1)) = pre(orient(lab(u_2), s_2)). \tag{1}$$

Recall that by the definition of F_E, $lab(u_1) = orient(x_1, s_1)$ and $lab(u_2) = orient(x_2, s_2)$. We can therefore rewrite Equation (1) equivalently as

$$suf(orient(orient(x_1, s_1), s_1)) = pre(orient(orient(x_2, s_2), s_2)). \tag{2}$$

Now, using the fact that $orient(orient(y, s), s) = y$, for all y and s, we can rewrite Equation (2) as

$$suf(x_1) = pre(x_2) \tag{3}$$

Since we obtained Equation (3) from Equation (1) using equivalent transformations, it shows that the two statements are equivalent and completes the proof. □

[2]As an aside, we mention how one would obtain a bijection. This is not necessary for the proofs of this paper, but may be a useful observation in its own right. Let E be the set of edges in $G_{\text{dbl}}(K)$, let $\alpha \subseteq E$ be all the self-mirror edges, and let β be the partition of $E \setminus \alpha$ into mirror edge-pairs. For example, if $E = \{(AGG, GGA), (TCC, CCT), (TTA, TAA)\}$, then $\alpha = \{(TTA, TAA)\}$ and $\beta = \{\{(AGG, GGA), (TCC, CCT)\}\}$. For an element $\{(x, y), (\overline{y}, \overline{x})\} \in \beta$, we define $F_{EG}(\{(x, y), (\overline{y}, \overline{x})\}) = F_E(x, y)$. For a self-mirror edge $(x, y) \in \alpha$, we define $F_{EG}(\{(x, y)\}) = F_E(x, y)$. One can then show that F_{EG} is a bijection between $\alpha \cup \beta$ and edges in $G_{\text{bid}}(K)$.

One particular case of Lemma B.11 that we will often invoke is that there is an edge from x to $\overline{x}$ in $G_{\text{dbl}}(K)$ if and only if there is an inverted loop incident to $(u, 1-s)$ in $G_{\text{bid}}(K)$.

Now recall that F_W is defined as a function that maps a walk $w = (x_0, \ldots, x_n)$ in $G_{\text{dbl}}(K)$ to a sequence $F_W(w) = (u_0, s_0, \ldots, u_n, s_n)$, with $(u_i, s_i) = F_V(x_i)$ for all $0 \leq i \leq n$. We show that $F_W(w)$ is in fact a walk in $G_{\text{bid}}(K)$ and, moreover, F_W is a bijection from the set of walks in $G_{\text{dbl}}(K)$ to the set of walks in $G_{\text{bid}}(K)$.

Lemma B.12. *Let K be a set of canonical k-mers where k is odd. F_W is a spell-preserving bijection from the set of walks in $G_{dbl}(K)$ to the set of walks in $G_{bid}(K)$.*

Proof. Let $w = (x_0, \ldots, x_n)$ be a walk in $G_{\text{dbl}}(K)$ and let $(u_i, s_i) = F_V(x_i)$ for all $0 \leq i \leq n$. We will first show that $F_W(w) = (u_0, s_0, \ldots, u_n, s_n)$ is a walk in $G_{\text{bid}}(K)$. By definition of F_V, $F_W(w)$ is a sequence of vertex-sides. Consider the edge from x_i to x_{i-1}, for all $1 \leq i \leq n$. By Lemma B.11, there is an edge $\{(u_{i-1}, 1-s_{i-1}), (u_i, s_i)\}$ in $G_{\text{bid}}(K)$. This shows that every two consecutive vertex-sides in $F_W(w)$ are connected by an edge, thus completing the proof that $F_W(w)$ is a walk. The fact that it is spell preserving follows from its definition.

To show that F_W is a bijection, we need to show that for all walks $t = (u_0, s_0, \ldots, u_n, s_n)$ in $G_{\text{bid}}(K)$, there exists a unique walk w in $G_{\text{dbl}}(K)$ such that $t = F_W(w)$. Let $w = (x_0, \ldots, x_n)$ be an arbitrary walk in $G_{\text{dbl}}(K)$. In order for $F_W(w) = t$, we need that $F_V(x_i) = (u_i, s_i)$ for all $0 \leq i \leq n$. Because F_V is bijection (Lemma B.10), there is exactly one value of x_i to satisfy this, and that is $x_i = F^{-1}(u_i, s_i) = \mathit{orient}(\mathit{lab}(u_i), s_i)$. Therefore, $w = (\mathit{orient}(\mathit{lab}(u_0), s_0), \ldots, \mathit{orient}(\mathit{lab}(u_n), s_n))$ is the unique walk in $G_{\text{dbl}}(K)$ to satisfy $F_W(w) = t$. □

Given the above proof, we can write the inverse of F_W as $F_W^{-1}(u_0, s_0, \ldots, u_n, s_n) = (\mathit{orient}(\mathit{lab}(u_0), s_0), \ldots, \mathit{orient}(\mathit{lab}(u_n), s_n))$. We will use $w \overset{F_W}{\Longleftrightarrow} t$ to denote that a walk w in $G_{\text{dbl}}(K)$ and a walk t in $G_{\text{bid}}(K)$ are associated with each other by F_W.

Notice that if k were to be even, then Lemma B.12 would not hold. In particular, Let $x \in K$ be a palindrome k-mer and let u be the vertex in $G_{\text{bid}}(K)$ such that $\mathit{lab}(u) = x$. Then both of the walks $(u, 0)$ and $(u, 1)$ would spell x, while in the $G_{\text{dbl}}(K)$ there would only be one walk that spells x.

Since unitigs are defined in terms of degrees, it is useful to first understand how the degrees of vertices in $G_{\text{dbl}}(K)$ relate to the degrees of vertex sides in $G_{\text{bid}}(K)$.

Lemma B.13. *Let K be a set of canonical k-mers where k is odd. Let x be a vertex in $G_{dbl}(K)$ and let (u, s) be a vertex-side is $G_{bid}(K)$ such that $x \overset{F_V}{\Longleftrightarrow} (u, s)$. Then,*

(i) $d^+(x) = d(u, 1-s) - d^{il}(u, 1-s)$

(ii) $d^-(x) = d(u, s) - d^{il}(u, s)$.

Proof. For proving part (i), we will first prove an upper bound and then a matching lower bound. We start with the upper bound. Let Y be the set of all out-neighbors of x which are not equal to $\overline{x}$. Note that Y may contain x. Let $Y' = \{F_V(y) \mid y \in Y\}$ and observe that since F_V is injective (Lemma B.10), $|Y'| = |Y|$. By Lemma B.11, for each vertex-side $(u', s') \in Y'$, there is an edge $\{(u, 1-s), (u', s')\}$ and so $d(u, 1-s) \geq |Y'|$.

We show that $d^+(x) = d(u, 1-s) - d^{il}(u, 1-s)$ by considering two cases. In the first case, assume that there does not exist an edge $(x, \overline{x})$. Then $d^+(x) = |Y|$. Moreover, by Lemma B.11, the edge $\{(u, 1-s), (u, 1-s)\}$ does not exist, so $d^{il}(u, 1-s) = 0$. Putting these facts together, $d^+(x) = |Y| = |Y'| \leq d(u, 1-s) = d(u, 1-s) - d^{il}(u, 1-s)$.

In the second case, assume that there exists an edge $(x, \overline{x})$. Lemma B.11 says that there is an inverted loop incident to side $(u, 1-s)$, so $d^{il}(u, 1-s) = 1$. An inverted loop adds 2 to the degree of $(u, 1-s)$, i.e. $d(u, 1-s) \geq |Y'| + 2$; it also contributes 1 to out-degree of x, i.e. $d^+(x) = |Y| + 1$. Putting these together, we get $d^(x) = |Y| + 1 = |Y'| + 1 \leq d(u, 1-s) - 1 = d(u, 1-s) - d^{il}(u, 1-s)$.

For the lower bound, let Z' be the set of all vertex-sides (u', s') such that $(u', s') \neq (u, 1-s)$ and there is an edge $\{(u, 1-s), (u', s')\}$. Let $Z = \{z \mid F_V(z) \in Z'\}$. By Lemma B.10, $|Z| = |Z'|$. By Lemma B.11, for every $z \in Z$, there is an edge from x to z in $G_{\text{dbl}}(K)$ and therefore $d^+(x) \geq |Z| = |Z'|$.

Now we show that $d(u, 1-s) \leq d^+(x) + d^{il}(u, 1-s)$ by considering two cases. In the first case, assume that there is no inverted loop touching $(u, 1-s)$. Then, $d(u, 1-s) = |Z'|$ and $d^{il}(u, 1-s) = 0$. We can therefore write $d(u, 1-s) = |Z'| + d^{il}(u, 1-s) \leq d^+(x) + d^{il}(u, 1-s)$. In the second case, assume there exists an inverted loop touching $(u, 1-s)$. In this case, $d(u, 1-s) = |Z'| + 2$. By Lemma B.11, there is an edge from x to $\overline{x}$ and $\overline{x} \notin Z$. Thus, $d + (x) \geq |Z| + 1$. Putting this together, $d(u, 1-s) = |Z'| + 2 = |Z| + 2 \leq d^+(x) + 1 = d^+(x) + d^{il}(u, 1-s)$.

For part (ii), observe that $F_V(\overline{x}) = (u, 1-s)$. We can then apply part (i) of this theorem to $\overline{x}$, u, and $1-s$, and get that $d^+(\overline{x}) = d(u, s) - d^{il}(u, s)$. By Lemma B.3, $d^-(x) = d^+(\overline{x})$, and hence $d^-(x) = d^+(\overline{x}) = d(u, s) - d^{il}(u, s)$. □

An immediate consequence of the degree-preserving lemma is that if $F(w)$ is a unitig, then so is w. The converse is not always true however.

Lemma B.14. *Let K be a set of canonical k-mers where k is odd. Let $w = (x_0, \ldots, x_n)$ and $t = (u_0, s_0, \ldots, u_n, s_n)$ be two walks related by $w \overset{F_W}{\Longleftrightarrow} t$.*

(i) If t is a unitig, then w is a unitig.

(ii) If w is a unitig and for all $1 \leq i \leq n$, $x_{i-1} \neq \overline{x_i}$, then t is a unitig.

Proof. For (i), when $n = 0$, w is trivially a unitig because it has only one vertex. For $n > 0$, since t is a unitig, $d(u_i, s_i) = 1$ for $0 < i \leq n$. Moreover, since an inverted loop would make a degree ≥ 2, we have $d^{il}(u_i, s_i) = 0$. Using Lemma B.13, $d^-(x_i) = 1$. Similarly, for all $0 \leq i < n$, $d(u_i, 1-s_i) = 1$, $d^{il}(u_i, 1-s_i) = 0$, and Lemma B.13 gives that $d^+(x_i) = 1$. Hence w is a unitig.

For (ii), first observe that there is no inverted loop incident to (u_i, s_i), for $1 \leq i \leq n$. If that were the case, then Lemma B.11 implies that there is an edge from $\overline{x_i}$ to x_i. Since w is a unitig, the only in-neighbor of x_i is x_{i-1}. Hence, $x_{i-1} = \overline{x_i}$, which contradicts the conditions of the Lemma. Now, since $d^{il}(u_i, s_i) = 0$, Lemma B.13 implies that $d(u_i, s_i) = d^-(x_i) + d^{il}(u_i, s_i) = d^-(x_i) = 1$. Using a symmetrical argument (omitted), $d(u_j, 1-s_j) = 1$ for all $0 \leq j < n$. Therefore, t is a unitig. □

Similarly, we can relate the maximality of unitigs in $G_{\text{dbl}}(K)$ and $G_{\text{bid}}(K)$. A maximal unitig in $G_{\text{dbl}}(K)$ is maximal in $G_{\text{bid}}(K)$, on the condition that is a unitig in $G_{\text{bid}}(K)$; however, the other direction only holds with a restrictive condition.

Lemma B.15. *Let K be a set of canonical k-mers where k is odd. Let $w = (x_0, \ldots, x_n)$ and $t = (u_0, s_0, \ldots, u_n, s_n)$ be two walks related by $w \overset{F_W}{\Longleftrightarrow} t$. Suppose that both w and t are unitigs.*

(i) If t is prefix-maximal and has no lonely inverted loop at the first endpoint side, then w is prefix-maximal.

(ii) If w is prefix-maximal, then t is prefix-maximal.

(iii) If t is suffix-maximal and has no lonely inverted loop at the last endpoint side, then w is suffix-maximal.

(iv) If w is suffix-maximal, then t is suffix-maximal.

Proof. We will prove (i) and (ii) only, since the proofs of (iii) and (iv) are symmetric. For (i), if there is more than one edge incident to (u_0, s_0), then $d(u_0, s_0) \geq 2$. If there are no edges incident to (u_0, s_0), then $d(u_0, s_0) = 0$. In both cases, Lemma B.13 implies that $d^-(x_0) = d(u_0, s_0) \neq 1$ and Lemma B.4 implies that w is prefix-maximal.

Now consider the case that $d(u_0, s_0) = 1$. By the conditions of the Lemma, there is no inverted loop incident at (u_0, s_0), and Lemma B.13 implies $d^-(x_0) = 1$. Since t is prefix-maximal, by Lemma B.7, there is a vertex side (u', s') and an edge $e = \{(u', s'), (u_0, s_0)\}$ such that $d(u', s') > 1$. Let $x' = F_V^{-1}(u', 1 - s')$ and Lemma B.11 implies that there is an edge from x' to x_0 in $G_{\text{dbl}}(K)$. Observe that because $d(u_0, s_0) < 2$, e is not an inverted loop. Therefore, (u', s') has at least one incident edge that is not an inverted loop. Because an inverted loop adds at least two to the degree, $d(u', s') - d^{il}(u', s') > 1$. Thus, Lemma B.13 implies that $d^+(x') > 1$. By Lemma B.4, w is a prefix-maximal unitig.

For (ii), suppose for the sake of contradiction that t is not prefix-maximal. Then Lemma B.7 implies that $d(u_0, s_0) = 1$ and there exists a vertex-side (u', s') with $d(u', s') = 1$ and an edge $e = \{(u', s'), (u_0, s_0)\}$. Let $x' = F_V^{-1}(u', 1 - s')$. Note that $d^{il}(u_0, s_0) = d^{il}(u', s') = 0$ because vertex-sides with degree 1 cannot have an inverted loop incident to them. Lemma B.13 then implies that $d^-(x_0) = d(u_0, s_0) = 1$ and $d^+(x') = d(u', s') = 1$. In addition, Lemma B.11 applied to e says that there is an edge from x' to x. By Lemma B.4, these facts imply that w is not prefix-maximal, which is a contradiction. □

Theorem 2 has a condition that there are no circular unitigs. We now show that this implies that a unitig in $G_{\text{bid}}(K)$ cannot have lonely inverted loops incident to both of the endpoint sides.

Lemma B.16. *Let K be a set of canonical k-mers where k is odd. Let $w = (x_0, \ldots, x_n)$ be a walk in $G_{dbl}(K)$ such that $F_W(w)$ is a unitig. If the two endpoint sides of $F_W(w)$ have lonely inverted loops incident on them, then $w' = (x_0, \ldots, x_n, \overline{x_n}, \ldots, \overline{x_0}, x_0)$ is a circular unitig in $G_{dbl}(K)$.*

Proof. First, to show that w' is a walk in $G_{\text{dbl}}(K)$, we need to show that there exist edges $(\overline{x_0}, x_0)$ and $(x_n, \overline{x_n})$. This follows by applying Lemma B.11 to the inverted loop edges at the endpoints of $F(w)$, i.e. to $\{(u_0, s_0), (u_0, s_0)\}$ and $\{(u_n, 1 - s_n), (u_n, 1 - s_n)\}$.

Second, to show that w' is a unitig, we will show that all the necessary vertex degrees are 1. By Lemma B.14, w is a unitig, and hence $d^+(x_i) = 1$ for all $0 \leq i < n$ and $d^-(x_i) = 1$ for all $0 < i \leq n$. Let $(u_i, s_i) = F_V(x_i)$ for all $0 \leq i \leq n$. Because the endpoint sides of $F(w)$ each have a lonely inverted loop, $d(u_0, s_0) = 2$ and $d(u_n, 1 - s_n) = 2$. Applying Lemma B.13, $d^-(x_0) = d(u_0, s_0) - d^{il}(u_0, s_0) = 2 - 1 = 1$ and $d^+(x_n) = d(u_n, 1 - s_n) - d^{il}(u_n, 1 - s_n) = 1$. Applying Lemma B.3 to all these, we get that $d^-(\overline{x_i}) = 1$ for all $0 \leq i \leq n$ and $d^+(\overline{x_i}) = 1$ for all $0 \leq i \leq n$. □

B.4 Proof of Theorem 2

Theorem 2. *Let K be a set of canonical k-mers where k is odd and $G_{dbl}(K)$ does not contain a circular unitig.*

(i) The function F_W is a bijection from $D_{non\text{-}pal}$ to $B_{no\text{-}loop}$.

(ii) The function rev is a bijection between $B_{last\text{-}loop}$ and $B_{first\text{-}loop}$.

(iii) HEAD *is a bijection from* D_{pal} *and* $B_{last\text{-}loop}$

Proof.

(i) We already know from Lemma B.12 that F_W is a bijection between walks in $G_{\text{dbl}}(K)$ and $G_{\text{bid}}(K)$. It remains to show that

(1) For a unitig w that is maximal and non-palindromic in $G_{\text{dbl}}(K)$, $F_W(w) \in B_{\text{no-loop}}$.

(2) For a unitig $t \in B_{\text{no-loop}}$, $F^{-1}(t)$ is a maximal and non-palindromic unitig in $G_{\text{dbl}}(K)$.

First, we prove (1). Because w is a non-palindromic maximal unitig, by Lemma B.6, there is no edge $0 \leq i < n$ such that $x_i = \overline{x_{i+1}}$, because then (x_i, x_{i+1}) would be a palindromic sub-unitig of w. Hence we can apply Lemma B.14 to say that $F_W(w)$ is a unitig and we can apply Lemma B.15 to say that $F_W(w)$ is maximal. Hence $F_W(w) \in B$. To show that $F_W(w) \notin B_2 \cap B_3$, first assume for the sake of contradiction that there is a lonely inverted loop at the last endpoint side of $F_W(w)$. Then by Lemma B.11 there is an edge from x_n to $\overline{x_n}$. By Lemma B.13, $d^+(x_n) = 2 - 1 = 1$. By Lemma B.3, $d^-(\overline{x_n}) = d^+(x_n) = 1$. Because w is maximal, if $d^+(x_n) = 1$, then $d^-(\overline{x_n}) > 1$. This is a contradiction. The argument that there is no lonely inverted loop at the first endpoint side of $F_W(w)$ is symmetric and omitted.

Now, we prove (2). Let $w = F^{-1}(t)$. Since t is a unitig, Lemma B.14 implies that w is a unitig also. Moreover, Lemma B.9 implies that t is non-palindromic; since F_W is spelling preserving (Lemma B.12), w is also non-palindromic. Since the Theorem assumes that $G_{\text{dbl}}(K)$ does not have circular unitigs, Lemma B.16 implies that t cannot have a lonely inverted loop at both endpoints. Since $t \notin B_2 \cup B_3$, it also cannot have an inverted loop at exactly one endpoint. We can therefore apply Lemma B.15 to get that w is maximal.

(ii) Observe that rev is by definition a function that is its own inverse and is a bijection on the set of walks in $G_{\text{bid}}(K)$. Furthermore, Lemma B.8 implies that rev remains a bijection when restricted to maximal unitigs in $G_{\text{bid}}(K)$. Finally, observe that for a walk t, the first (respectively, last) endpoint side of t is the last (respectively, first) endpoint side of $rev(t)$. These facts together imply that rev is a bijection between $B_{\text{first-loop}}$ and $B_{\text{last-loop}}$.

(iii) To show that HEAD is a bijection we show

(1) for all $w \in D_{\text{pal}}$, $\text{HEAD}(w) \in B_{\text{last-loop}}$,

(2) for all $t \in B_{\text{last-loop}}$, there exists a $w \in D_{\text{pal}}$ such that $\text{HEAD}(w) \in B_{\text{last-loop}}$.

(3) the above w is unique.

First, we prove (1). Let $w = (x_0, \ldots, x_n)$. By Lemma B.1, n is odd and at least 1. Let $m = (n-1)/2$ and let $h \triangleq (x_0, \ldots, x_m)$. Since w is a palindromic unitig and, by the conditions of the Theorem, non-circular, Lemma B.5 implies that for all $0 \leq i < n$, $x_i \neq \overline{x_{i+1}}$. Then by Lemma B.14, $\text{HEAD}(w) = F_W(h)$ is a unitig. Simultaneously, because w is a maximal unitig, h is a prefix-maximal unitig. Lemma B.15 then implies that $F_W(h)$ is prefix-maximal.

Now we show that $F_W(h)$ is suffix-maximal and has a lonely inverted loop at the last endpoint. Let $(u_0, s_0, \ldots, u_m, s_m) \triangleq F_W(h)$. Since w is palindromic, Lemma B.5 implies that $x_m = \overline{x_{m+1}}$, and, hence, $u_m = u_{m+1}$. By Lemma B.11, there is an inverted loop incident to $(u_m, 1 - s_m)$, i.e. the last endpoint of $F_W(h)$. Because w is a unitig, $d^+(x_m) = d^-(x_{m+1}) = 1$, Lemma B.13 then implies that $d(u_m, 1 - s_m) = d^+(x_m) + d^{il}(u_m, 1 - s_m) = 2$. By Lemma B.7, $F_W(h)$ is suffix-maximal and therefore we have shown that $F_W(h) \in B_{\text{last-loop}}$.

Next we prove (2). Let $(u_0, s_0, \ldots, u_n, s_n) = t$ and let $x_i = F_V^{-1}(u_i, s_i)$. Let $w = (x_0, \ldots, x_n, \overline{x_n}, \ldots, \overline{x_0})$ be a sequence of vertices in $G_{\text{dbl}}(K)$. We will first show that w is a walk, then that it is palindromic, then that it is a unitig, and finally that it is maximal. Note that w is equivalently defined to be the concatenation of $F_W^{-1}(t)$ with $F_W^{-1}(rev(t))$. Applying Lemma B.12, the sequences $(x_0, \ldots, x_n)$ and $(\overline{x_n}, \ldots, \overline{x_0})$ are walks. Since t is in $B_{\text{last-loop}}$, there is an inverted loop incident to $(u_n, 1 - s_n)$. By Lemma B.11, this implies there is an edge from x_n to $\overline{x_n}$ in $G_{\text{dbl}}(K)$. Therefore, w is a walk. It is palindromic by its definition. Since t is a unitig, by Lemma B.8, $rev(t)$ is a unitig. Now applying Lemma B.14, w and $rev(w)$ are both unitigs. Because the inverted loop is lonely, $d(u_n, 1 - s_n) = 2$, and by Lemma B.13, $d^+(x_n) = 1$. Applying Lemma B.3, $d^-(\overline{x_n}) = 1$. Hence w is a unitig.

As t is in $B_{\text{last-loop}}$, this implies that no lonely inverted loop is incident to (u_0, s_0). We can apply Lemma B.15 to get that $F^{-1}(t)$ is prefix-maximal. Because w starts with $F^{-1}(t)$, w is also prefix-maximal. By Lemma B.8, $F^{-1}(rev(w))$ is suffix-maximal. Because w ends with $F^{-1}(rev(w))$, w is also suffix-maximal. Hence, w is maximal.

For (3), let $(u_0, s_0, \ldots, u_n, s_n) = t$ and let $x_i = F_V^{-1}(u_i, s_i)$. Let w' be a walk in D_{pal} such that $\text{HEAD}(w') \in B_{\text{last-loop}}$. We will show that $w' = (x_0, \ldots, x_n, \overline{x_n}, \ldots, \overline{x_0})$. Since $\text{HEAD}(w')$ has $n+1$ vertices, w must have $2n+2$ vertices. Hence we can write $w = (x'_0, \ldots, x'_{2n+1})$. Since w is a palindrome, we have that $x'_i = \overline{x'_{2n+1-i}}$ for all $0 \leq i \leq 2n+1$. We can therefore rewrite w as $w = (x'_0, \ldots, x'_n, \overline{x'_n}, \ldots, \overline{x'_0})$. Next, observe that $\text{HEAD}(w') = F_W((x'_0, \ldots, x'_n))$. Since this must be equal to t and F_W is a bijection (Lemma B.12), we get that $(x'_0, \ldots, x'_n) = (x_0, \ldots, x_n)$. We can therefore rewrite w as $w = (x_0, \ldots, x_n, \overline{x_n}, \ldots, \overline{x_0})$, which is the same as w.

□

C Experimental details

Choice of k parameter for the assemblers: To ensure that the results across the assemblers are comparable, we set the k parameter in a way so that the set of unitigs constructed are as close as possible. The ideal way is to set k such that the underlying k-mer sets K used for all assemblers are same. However, there was a practical limitation for that. We note that both SPAdes and MEGAHIT are a multi-k assemblers, so the k parameter is just the maximum allowed k-mer size. When we pass the value k to the assemblers, both SPAdes and MEGAHIT use k-mer set and $(k+1)$-mer set to construct unitigs, whereas bcalm, ABySS, and minia uses a node-centric de Bruijn graph with only k-mer sets as vertices. As such, we found that the output unitigs of SPAdes and MEGAHIT with a value of k are more similar to unitigs of bcalm and ABySS created with $k+1$. We also note that SPAdes and MEGAHIT only allow odd k, which is why we needed to use an even k for G_{dbl}.

In Table 3, we therefore passed $k = 74$ to bcalm and $k = 73$ to SPAdes and MEGAHIT. Since Theorem 1 is valid for all k, this was not an issue for Table 3. We used the default parameter for minimum k-mer coverage for both assemblers.

For Table 6, we passed $k = 31$ to all assemblers, since Theorem 2 only applies when the vertex lengths are of odd k. Since SPAdes and MEGAHIT by default use both k-mer and $(k+1)$-mer set to construct unitigs, the number of palindromic unitigs (433) differs from the number in minia and ABySS (440). However, this is not a problem because we are not comparing the numbers between assemblers but only within assemblers.

Detection of palindrome splitting artifact In this section, we use the notation $S[i : j]$ to denote substring of string S starting at index i and ending at index j. Let $w = (x_0, \ldots, x_n)$ be a palindromic unitig in D_{pal} and let p be its spelling. We say a unitig in D_{pal} is *fully-covered* if there exists some contig that aligns to an interval which contains p's interval in the reference. Let $k' \triangleq (k-1)/2$. We say w is *split* if there exists at least one contig c such that either

1. c aligns to an interval that starts before p's interval and ends exactly at position $|p|/2 + k'$ of p's interval and there are no other contigs with alignments intersecting $p[|p|/2 + k' + 1 : |p|]$, or

2. c aligns to an interval that ends after p's interval and starts exactly at location $|p|/2 - k' + 1$ of p's interval and there are no other contigs with alignments intersection $p[1 : |p|/2 + k']$.

We say w is *ambiguous* if it does not fall into either category.

To motivate these cases, observe that the length of p is $n + k$ and, because p is a palindrome and k is odd, n must be odd. Let $w' = (x_0, \ldots, x_{\frac{n-1}{2}})$ be the first half of the walk w and let p' be its spelling. By Theorem 2, $\text{HEAD}(w) \in B_{\text{last-loop}}$ and $rev(\text{HEAD}(w)) \in B_{\text{first-loop}}$. Then, $p' = p[1 : \frac{n-1}{2} + k] = p[1 : |p|/2 + k']$. Then,

1. $spell(\text{HEAD}(w)) = spell(F_W(w')) = spell(w') = p[1 : |p|/2 + k']$, and

2. $spell(rev(\text{HEAD}(w))) = spell(rev(F_W(w'))) = \overline{spell(F_W(w'))} = \overline{p[p/2 - k' + 1 : |p|]}$.

The cases we describe therefore correspond to observing the alignments of $\text{HEAD}(w)$ and $rev(\text{HEAD}(w))$ to the corresponding places of p and not observing any other bidirected unitigs aligning across the middle boundaries.

CAMI dataset: We used the benchmark called "low complexity dataset" in (Sczyrba et al 2017). Since our analysis requires error-free reads, we re-simulated the reads using identical genomes and abundances (as detailed in supplementary materials of (Nurk, Meleshko, et al 2017)). Table S1 shows the properties and relative abundances of the genomes. We used CAMISIM (Fritz et al 2019) for the simulations, with read length of 150nt and insert size 150.

Table S1: Characteristics of all 30 genomes consituting the CAMI "low complexity" dataset. The coverage refers to the depth-of-coverage for each genome, in both the benchmark ((Nurk, Meleshko, et al 2017)) and our simulations.

Accession	Species name	n. contigs in reference	n. bases in reference	Coverage	Abundance
AEGL00000000	Gamma proteobacterium IMCC2047	815	2,234,019	873.3	76.21%
AAVV01000001	Marine gamma proteobacterium HTCC2080	25	3,576,081	53	7.41%
AGFI01000001	Paenibacillus sp. Aloe-11	334	5,792,040	22	4.98%
ACXM01000000	Thermoplasmatales archaeon I-plasma	6	1,684,836	21	1.38%
ACZW02000000	Erysipelotrichaceae bacterium 5_2_54FAA	10	3,137,098	16	1.96%
ARQX01000000	Gamma proteobacterium SCGC AAA076-D13	81	1,663,375	14	0.91%
ATUD01000000	Patulibacter americanus DSM 16676	32	4,470,560	9	1.57%
GCF_000236585.1	Thermus sp. CCB_US3_UF1	2	2,263,488	8	0.71%
ARCO01000000	Chloroflexi bacterium SCGC AB-629-P13	64	842,066	8	0.26%
PRJNA586334	Marinimicrobia bacterium SCGC AB-629-J13	103	1,123,146	8	0.35%
AGUD01000000	Patulibacter medicamentivorans	353	5,092,500	7	1.39%
CAXW010000000	Firmicutes bacterium CAG:114	292	2,332,166	4	0.36%
ANLA01000000	Formosa sp. AK20	47	3,055,484	3	0.36%
GCF_000445995.2	Geobacillus sp. JF8	2	3,486,308	2	0.27%
GCA_000496235.1	Uncultured archaeon A07HR60	14	2,876,249	1.9	0.21%
AMFN01000000	Enterobacteriaceae bacterium LSJC7	34	4,616,889	1.8	0.32%
AMYX01000000	Alpha proteobacterium LLX12A	289	5,961,098	1.4	0.33%
AMSP01000000	Brevibacterium casei S18	43	3,664,641	1.2	0.17%
GCA_000403475.2	Lachnospiraceae bacterium 3-2	4	4,455,623	1	0.17%
AANX02000000	Burkholderia mallei 2002721280	208	5,690,468	0.9	0.20%
GCA_000209385.2	Lachnospiraceae bacterium 2_1_46FAA	1	2,219,029	0.9	0.08%
GCF_000219815.1	Weissella koreensis KACC 15510	2	1,441,470	0.7	0.04%
ARQU01000000	Alpha proteobacterium SCGC AAA536-G10	148	2,161,697	0.6	0.05%
AOUN01000000	Sphingopyxis sp. MC1	24	3,653,464	0.4	0.06%
GCF_000015985.1	Rhodobacter sphaeroides ATCC 17029	3	4,489,380	0.4	0.07%
AMFB01000000	Bradyrhizobium sp. DFCI-1	98	7,645,871	0.3	0.09%
NC_021024.1	Butyrate-producing bacterium SM4/1	1	3,108,859	0.3	0.04%
ARSS01000000	Alpha proteobacterium SCGC AAA015-O19	159	1,742,143	0.2	0.01%
CBEH010000000	Firmicutes bacterium CAG:170	375	2,449,192	0.2	0.02%
NC_023004.1	Candidatus Saccharibacteria bacterium RAAC3_TM7_1	1	845,464	0.1	0.00%