# Do Read Errors Matter for Genome Assembly?

Ilan Shomorony
UC Berkeley
ilan.shomorony@berkeley.edu

Thomas Courtade
UC Berkeley
courtade@berkeley.edu

David Tse
Stanford University
dntse@stanford.edu

***Abstract*—While most current high-throughput DNA sequencing technologies generate short reads with low error rates, emerging sequencing technologies generate long reads with high error rates. A basic question of interest is the tradeoff between read length and error rate in terms of the information needed for the perfect assembly of the genome. Using an adversarial erasure error model, we make progress on this problem by establishing a critical read length, as a function of the genome and the error rate, above which perfect assembly is guaranteed. For several real genomes, including those from the GAGE dataset, we verify that this critical read length is not significantly greater than the read length required for perfect assembly from reads without errors.**

## I. Introduction

Current DNA sequencing technologies are based on a two-step process. First, tens or hundreds of millions of fragments from random locations on the DNA sequence are read via *shotgun sequencing*. Second, these fragments, called reads, are merged to each other based on regions of overlap, using an *assembly algorithm*.

Roughly speaking, different shotgun sequencing platforms can be distinguished from the point of view of three main metrics: the *read length*, the *read error rate*, and the *read throughput*. In the last decade, the so-called next-generation sequencing platforms have attained considerable success at employing heavy parallelization in order to achieve *high-throughput* shotgun sequencing. This allowed a significant reduction in the cost and time of sequencing, causing an explosion in the number of new sequencing projects and the generation of massive amounts of sequencing data.

In order to guarantee low error rates, most of these next-generation technologies are restricted to *short read lengths*, shifting some of the burden of sequencing to the assembly step. In practice, this results in very fragmented assemblies, with large gaps and little linking information between fragments [1]. On the other hand, recent technologies that generate longer reads suffer from lower throughput and much higher error rates[1].

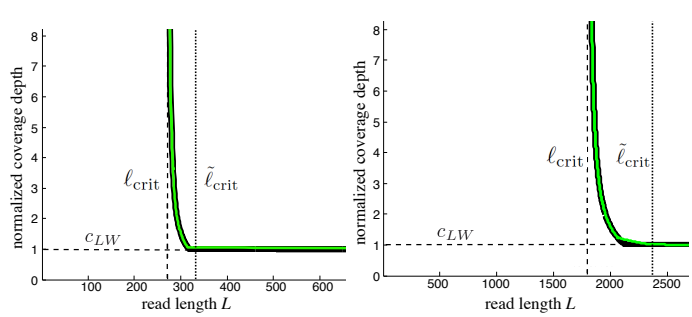
Given this technology trend, the natural questions to ask are: what is the impact of read errors on the performance of assemblers? Is the negative impact of read errors more than offset by the increase in read lengths in long-read technologies? It is well known that read errors have a significant impact on assembly algorithms. For example, in DeBruijn graph based algorithms, read errors create extraneous nodes and edges in the assembly graph, which results in added complexity. However, these observations pertain to *specific* algorithms. A more fundamental question can be asked from an *information-theoretic* point of view: given a read length, an error rate and a coverage depth (number of reads per base), is there enough *information* in the read data to uniquely reconstruct the genome? Do errors significantly increase the read length and/or coverage depth requirements? An answer to these basic feasibility questions can provide an algorithm-independent framework for evaluating different sequencing technologies. It would also settle some speculations in the assembly community on whether read errors have a significant impact in long-read technologies (see for example [2]).

Such a framework was initiated in [3] for *error-free* reads: a feasibility curve relating the read length and coverage depth needed to perfectly assemble a genome was characterized in terms of the repeat complexity of the genome (see examples in Fig. 1). Evaluating this curve on several genomes revealed an interesting threshold phenomenon: if the read length is below a certain critical value $\ell_{\text{crit}}$, reconstruction is impossible; a read length slightly above $\ell_{\text{crit}}$ and a coverage depth close to the Lander-Waterman depth $c_{LW}$ (i.e., just enough reads to cover the whole sequence) is sufficient. The critical read length $\ell_{\text{crit}}$ is given by the length of the longest *interleaved repeat* in the genome, and coincides with the minimum read length $L$ needed to uniquely reconstruct the genome given its $L$-*spectrum*, i.e. the set of reads with one length-$L$ read starting at each position of the sequence, illustrated in Fig. 2. This minimum read length also appeared in earlier works by Ukkonen and Pevzner [4, 5] for reconstruction via *sequencing by hybridization*.

Given this framework, the impact of read errors can be studied by asking how much the critical read length $\ell_{\text{crit}}$ increases when there are errors. In this paper, we investigate this tradeoff for a specific error model: 1) the errors are erasures; 2) the erasures occur at a rate no larger than $D/L$ for each read and for each base in the sequence, but are otherwise arbitrary. Our main result is the characterization of a critical read length $\tilde{\ell}_{\text{crit}}$ above which perfect assembly is always possible. While in the noiseless case $\ell_{\text{crit}}$ is a function of the sequence repeat structure, $\tilde{\ell}_{\text{crit}}$ depends more generally on the error rate and on the *approximate repeats* in the sequence. More concretely, for a sequence $\mathbf{s}$,

$$\tilde{\ell}_{\text{crit}}(\mathbf{s}, D) = \min_{k \geq \ell_{\text{crit}}(\mathbf{s})} k + D \cdot M_{\mathbf{s}}(D, k+1),$$

where $M_{\mathbf{s}}(D, \ell)$ is the maximum number of $D$-approximate




(a) *S. aureus* (b) *R. sphaeroides*

Fig. 1. The thick black curve is a feasibility lower bound for any algorithm, and the green line represents the performance of the Multibridging algorithm [3].

[1]One example of a short-read-length technology is Illumina, with reads of length $\sim 200$ base pairs and error rates of about $1\%$. In contrast, PacBio reads can be several thousand base pairs long, with error rates of about 10-15$\%$.

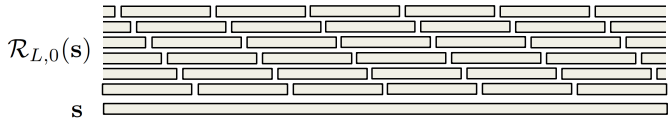



Fig. 2. The sequence $\mathbf{s}$ and its $L$-spectrum, $\mathcal{R}_{L,0}(\mathbf{s})$.

length-$\ell$ repeats in $\mathbf{s}$. Moreover, reminiscent of classical coding theory results, we show that the same read length $\tilde{\ell}_{\text{crit}}$ is sufficient for assembly if instead of erasures we consider substitution errors at half of the rate. In order to characterize $\tilde{\ell}_{\text{crit}}$, we derive a new result about the error correction capability of the $L$-spectrum. More precisely, we show that given a noisy version of the $L$-spectrum of a sequence, it is possible to obtain the noiseless $(k+1)$-spectrum of the same sequence, for any $k$ such that $L > k + D \cdot M_{\mathbf{s}}(D, k+1)$. When $L > \tilde{\ell}_{\text{crit}}$, we can obtain the noiseless $(k+1)$ spectrum for some $k > \ell_{\text{crit}}$, and the noiseless result from [3] implies that perfect assembly is possible.

By evaluating $\tilde{\ell}_{\text{crit}}$ on several real genomes, including those in the GAGE dataset [6], we verify that $\tilde{\ell}_{\text{crit}}$ is not significantly larger than $\ell_{\text{crit}}$. In fact, in most cases, $\tilde{\ell}_{\text{crit}} \approx \ell_{\text{crit}} + 3D$. Hence, if the read length $L$ is chosen above the noiseless requirement $\ell_{\text{crit}}$, perfect assembly is robust to errors up to a threshold (roughly $\frac{1}{3}(L - \ell_{\text{crit}})$ erasures per read).

The impact of read errors on the information theoretic limits of genome assembly has also been studied in the setting of an i.i.d. genome model and asymptotically long genome length [7], building on an earlier work on error-free reads in the same setting [8]. The results are surprising: as long as the error rate is below a threshold (which can be as high as 19% for substitution errors), noisy reads are as good as noiseless reads; i.e., the requirements for assembly in terms of read length and coverage depth are the same in both cases. While this result seems stronger than the result in the present paper, it is proved under the idealistic and unrealistic settings of i.i.d. genome statistics and i.i.d. errors. The present result, on the other hand, is more robust as it applies to arbitrary genome repeat statistics and error statistics.

## II. Problem Setting

In the DNA assembly problem, the goal is to reconstruct a sequence $\mathbf{s} = (s[1], ..., s[G])$ of length $G$ with symbols from the alphabet $\Sigma = \{a, c, g, t\}$. In order to simplify the exposition, we assume a *circular* DNA model; thus, $\{s[i]\}_{i=1}^{\infty}$ is a periodic sequence with (minimum) period $G$. Our results hold in the non-circular case as well under minor modifications. We will let $\mathbf{s}_i^{\ell}$ be the substring of length $\ell$ starting at $s[i]$; i.e., $\mathbf{s}_i^{\ell} = (s[i], s[i+1], ..., s[i+\ell-1])$.

The sequencer provides a multiset of $N$ *reads* $\mathcal{R} = \{\mathbf{r}_1, ..., \mathbf{r}_N\}$ from $\mathbf{s}$, each of length $L$. In the noiseless case, each read is a length-$L$ substring of $\mathbf{s}$ with an unknown starting location. Our focus, however, will be on noisy read models, where each read may be corrupted by noise. The goal is to design an *assembler*, which takes the set of reads $\mathcal{R}$ and attempts to reconstruct the sequence $\mathbf{s}$.

### A. The L-Spectrum Read Model

We will consider a "dense-read" model, in which all the reads in the $L$-spectrum of $\mathbf{s}$ are provided. More precisely, $\mathcal{R}$ will have exactly $G$ reads, one from each possible starting position; i.e., $\mathcal{R} = \{\mathbf{r}_1, ..., \mathbf{r}_G\}$, where $\mathbf{r}_i = \mathbf{s}_i^L$ for $i = 1, ..., G$. We will refer to the error-free $L$-spectrum of $\mathbf{s}$ by $\mathcal{R}_{L,0}(\mathbf{s})$. Notice that the starting position $i$ for each read $\mathbf{r}_i$ is unknown to the assembler.

While such a read model was originally proposed in the context of *sequencing by hybridization* [4, 5, 9], our motivation for using it comes from next-generation sequencing technologies, where the high read throughput can provide large coverage depths at low costs, and a dense read regime is not unrealistic. This way, we can bypass the question of the necessary coverage depth for assembly, and instead focus on the interplay between read length and error rate in the context of assembly feasibility. Moreover, as shown in [3] for noiseless reads, the dense-read model provides valuable insights towards understanding the information-theoretic limits of reconstruction in the more general shotgun read model.

In the $L$-spectrum read model, since we have exactly $G$ reads, an assembly of the reads $\mathcal{R}_{L,0}(\mathbf{s}) = \{\mathbf{r}_1, ..., \mathbf{r}_G\}$ can be thought of as a permutation $\sigma$ of the entries of $(1, ..., G)$. We assume without loss of generality that the identity permutation $\sigma_0 = (1, ..., G)$ yields a correct assembly of $\mathbf{s}$. Notice, however, that the index $i$ of each read $\mathbf{r}_i$ is unknown to the assembler. Notice also that in general, there may be multiple correct assemblies for a sequence $\mathbf{s}$ if $\mathbf{r}_i = \mathbf{r}_j$ for some $i \neq j$.

### B. Adversarial Erasure Model

As in the classical coding theory literature, we will study the problem of DNA assembly with noisy reads from the perspective of an *adversarial* noise model. Given that actual sequencing noise profiles are complex (non-i.i.d., asymmetric across bases) and technology-dependent, this approach avoids the need for a probabilistic noise model by instead focusing on a worst-case scenario. Moreover, under this model we can hope to obtain deterministic and non-asymptotic conditions for perfect assembly, which can be more easily analyzed in terms of real genome data.

Motivated by the fact that sequencing technologies usually provide a quality score for each base that is read (which could be thresholded into "good" and "bad" bases), and in order to simplify the problem, we will consider an *erasure* model. The reads in $\mathcal{R}$ will be length-$L$ sequences from the alphabet $\Sigma' = \{a, c, g, t, \varepsilon\}$, where $\varepsilon$ corresponds to an erasure. Thus, a read starting at position $i$ from $\mathbf{s}$ can be written as $\mathbf{r}_i = (r_i[0], ..., r_i[L-1])$, where either $r_i[j] = s[i+j]$ or $r_i[j] = \varepsilon$, for $1 \leq i \leq G$ and $0 \leq j \leq L-1$.

For a fixed parameter $D$, the adversarial erasure model will be constrained by a maximum error rate of $D/L$ within each read, and for each base. Since in our read model each base $s[i]$ is read $L$ times $(r_{i-(L-1)}[L-1], r_{i-(L-2)}[L-2], ..., r_i[0])$, these constraints can be written as follows:

a) There are at most $D$ erasures per read.
b) Each base $s[i]$ is erased at most $D$ times across all reads.

We will use $\mathcal{R}_{L,D}(\mathbf{s})$ to refer to the $L$-spectrum of $\mathbf{s}$, $\mathcal{R}_{L,0}(\mathbf{s})$, after being corrupted by erasures satisfying (a) and (b).

In the context of an adversarial noise model with deterministic constraints, it makes sense to restrict our attention to potential sequences $\hat{\mathbf{s}}$ that are *consistent* with the reads $\mathcal{R}_{L,D}(\mathbf{s})$. A sequence $\hat{\mathbf{s}}$ is said to be consistent with $\mathcal{R}_{L,D}(\mathbf{s})$ if it could have generated the set of reads $\mathcal{R}_{L,D}(\mathbf{s})$ according to the erasure model in (a) and (b). By extension, we will say that an assembly $\sigma$ of $\mathcal{R}_{L,D}(\mathbf{s})$ is consistent if there exists a sequence $\hat{\mathbf{s}}$, consistent with $\mathcal{R}_{L,D}(\mathbf{s})$, that could have generated the reads in $\mathcal{R}_{L,D}(\mathbf{s})$ according to the positions determined by $\sigma$. As illustrated in Fig. 3, we notice that (b) guarantees that a consistent assembly $\sigma$

Fig. 3. Part of a consistent assembly for $L = 5$ and $D = 2$. Notice that there can be at most $D$ erasures per read and per "column" of the assembly $\sigma$. Moreover, all non-erased bases in a column must agree.

defines, up to cyclic shifts, a unique consistent sequence in $\Sigma^G$, which we will refer to as $\hat{\mathbf{s}}(\sigma)$.

The fundamental feasibility question corresponds to asking which values of $L$ allow unambiguous reconstruction. Formally, it corresponds to the following algorithm-independent question.

**Question 1.** *Consider a fixed circular sequence* $\mathbf{s} \in \Sigma^G$. *What values of* $L$ *guarantee that, for an arbitrary set of erased reads* $\mathcal{R}_{L,D}(\mathbf{s})$, $\mathbf{s}$ *is the unique sequence consistent with* $\mathcal{R}_{L,D}(\mathbf{s})$*?*

## III. Assembly in the Noiseless Case

The assembly problem in Question 1 was first studied in [4] in the noiseless setting $D = 0$. Notice that when $L = 1$, $\mathcal{R}_{1,0}(\mathbf{s})$ is simply the multi-set $\{s[1], ..., s[G]\}$ and any permutation $\sigma$ of $(1, ..., G)$ is a consistent assembly. Hence, $\mathbf{s}$ cannot be reconstructed unambiguously, unless all of its symbols are the same. On the other hand, when $L = G$, there is a unique assembly of $\mathcal{R}_{G,0}(\mathbf{s}) = \{\mathbf{s}\}$, and $\mathbf{s}$ can always be reconstructed unambiguously. Question 1 is thus equivalent to asking for the threshold $\ell_{\text{th}}$ for which $\mathbf{s}$ can be reconstructed if and only if $L > \ell_{\text{th}}$. In [4], this threshold is established as a function of the *repeat* structure of the sequence $\mathbf{s}$, as we explain next.

A *repeat* of length $\ell$ in $\mathbf{s}$ is a subsequence appearing twice at some positions $t_1$ and $t_2$ (so $\mathbf{s}_{t_1}^{\ell}$ and $\mathbf{s}_{t_2}^{\ell}$) that is maximal; i.e., $s[t_1 - 1] \neq s[t_2 - 1]$ and $s[t_1 + \ell] \neq s[t_2 + \ell]$. Two pairs of repeats $\mathbf{s}_{a_1}^{\ell}, \mathbf{s}_{a_2}^{\ell}$ and $\mathbf{s}_{b_1}^{k}, \mathbf{s}_{b_2}^{k}$ are *interleaved* if $a_1 < b_1 \leq a_2 < b_2$. Due to the circular DNA model, since a subsequence $\mathbf{s}_t^{\ell}$ can also be written as $\mathbf{s}_{t+mG}^{\ell}$ for any integer $m$, we additionally require that $b_2 - a_1 < G$. The length of a pair of interleaved repeats $\mathbf{s}_{a_1}^{\ell}, \mathbf{s}_{a_2}^{\ell}$ and $\mathbf{s}_{b_1}^{k}, \mathbf{s}_{b_2}^{k}$ is defined to be $\min(\ell, k)$. We let $\ell_{\text{inter}}(\mathbf{s})$ be the length of the longest pair of interleaved repeats in $\mathbf{s}$ and set $\ell_{\text{crit}}(\mathbf{s}) = \ell_{\text{inter}}(\mathbf{s}) + 1$. The results from [4, 5] imply the following:

**Theorem 1.** *If* $L > \ell_{\text{crit}}(\mathbf{s})$*, then* $\mathbf{s}$ *is the unique sequence that is consistent with* $\mathcal{R}_{L,0}(\mathbf{s})$*. Conversely, if* $L \leq \ell_{\text{crit}}(\mathbf{s})$*, there exists a sequence* $\mathbf{s}' \neq \mathbf{s}$ *that is also consistent with* $\mathcal{R}_{L,0}(\mathbf{s})$*.*

In other words, Theorem 1 characterizes the threshold on $L$ that fully answers Question 1. We point out that, in the previous literature [3, 4], $\ell_{\text{crit}}$ was defined in terms of the length of pairs of interleaved repeats (defined in a more restrictive way) and the length of *triple repeats*. However, one can verify that by considering the more general definition of interleaved repeats above, triple repeats are included as a special case.

Notice that, while Theorem 1 characterizes the minimum $L$ that guarantees perfect reconstruction, $\ell_{\text{crit}}(\mathbf{s})$ is a function of the ground truth $\mathbf{s}$, and is not known a priori. However, the following corollary of Theorem 1 readily follows:

**Corollary 1.** *If a sequence* $\hat{\mathbf{s}}$ *is consistent with* $\mathcal{R}_{L,0}(\mathbf{s})$ *and* $L > \ell_{\text{crit}}(\hat{\mathbf{s}})$*, then* $\hat{\mathbf{s}} = \mathbf{s}$*.*

Since $\ell_{\text{crit}}(\hat{\mathbf{s}})$ can be computed from the assembled sequence $\hat{\mathbf{s}}$, this result means that $L > \ell_{\text{crit}}(\hat{\mathbf{s}})$ provides a certificate that $\hat{\mathbf{s}} = \mathbf{s}$, even without previous knowledge of $\ell_{\text{crit}}(\mathbf{s})$.

## IV. Main Results

In the previous section, we described how Theorem 1 fully characterizes when assembly is possible given the noiseless $L$-spectrum. In this section, we seek a similar characterization in the case where reads are noisy.

Notice that for the erasure setting described in Section II, one possible erasure pattern is to have the last $D$ bases from each read erased, which effectively results in noiseless reads of length $L - D$. Therefore, the converse part of Theorem 1 implies that, if $L \leq \ell_{\text{crit}}(\mathbf{s}) + D$, there is a read set $\mathcal{R}_{L,D}(\mathbf{s})$ and a sequence $\hat{\mathbf{s}} \neq \mathbf{s}$ that is consistent with $\mathcal{R}_{L,D}(\mathbf{s})$. But how much larger than $\ell_{\text{crit}}(\mathbf{s}) + D$ does the read length $L$ have to be in order to guarantee unambiguous correct reconstruction? In other words, how do erasures degrade the fundamental limit characterized by Theorem 1?

Our main result is the introduction of a new sequence-dependent quantity, $\tilde{\ell}_{\text{crit}}(D, \mathbf{s})$, such that, if $L > \tilde{\ell}_{\text{crit}}(\mathbf{s}, D)$, $\mathbf{s}$ is the unique sequence consistent with $\mathcal{R}_{L,D}(\mathbf{s})$. In general, $\ell_{\text{crit}}(\mathbf{s}) + D < \tilde{\ell}_{\text{crit}}(\mathbf{s}, D)$ for $D > 0$, and one can construct an arbitrary sequence $\mathbf{s} \in \Sigma^G$ for which the gap between the two quantities is significant. However, by computing $\ell_{\text{crit}} + D$ and $\tilde{\ell}_{\text{crit}}$ for actual genomes, we verify that they are often close, as shown in Table I.

Rather than being defined in terms of exact repeats, as is the case of $\ell_{\text{crit}}(\mathbf{s})$, $\tilde{\ell}_{\text{crit}}(\mathbf{s})$ depends more generally on *approximate repeats*. For a set of segments $\mathcal{S}$ of a given length $\ell$; i.e., $\mathcal{S} \subset \Sigma^{\ell}$, we will first define the radius of $\mathcal{S}$ to be

$$\rho(\mathcal{S}) = \min_{\mathbf{x} \in \Sigma^{\ell}} \max_{\mathbf{y} \in \mathcal{S}} d_H(\mathbf{y}, \mathbf{x}), \tag{1}$$

where $d_H(\mathbf{y}, \mathbf{x})$ is the Hamming distance between $\mathbf{y}$ and $\mathbf{x}$. We will say that the segments in $\mathcal{S}$ are $d$-approximate copies if $\rho(\mathcal{S}) \leq d$. Intuitively, a sequence $\mathbf{s}$ that contains a large set $\mathcal{S}$ of length-$\ell$ segments with a small radius $\rho(\mathcal{S})$ has more ambiguity in terms of assembly. To capture that, we will let $M(d, \ell)$ correspond to the maximum number of $d$-approximate length-$\ell$ segments in $\mathbf{s}$; i.e.,

$$M_{\mathbf{s}}(d, \ell) = \max \left\{ |\mathcal{S}| : \mathcal{S} \subset \mathcal{R}_{\ell,0}(\mathbf{s}), \rho(\mathcal{S}) \leq d \right\}. \tag{2}$$

Notice that $M_{\mathbf{s}}(d, \ell)$ is monotonically decreasing in $\ell$. We let

$$\tilde{\ell}_{\text{crit}}(\mathbf{s}, D) = \min_{k \geq \ell_{\text{crit}}(\mathbf{s})} k + D \cdot M_{\mathbf{s}}(D, k + 1). \tag{3}$$

Notice that $\tilde{\ell}_{\text{crit}}(\mathbf{s}, D) \geq \tilde{\ell}_{\text{crit}}(\mathbf{s}, 0) = \ell_{\text{crit}}(\mathbf{s})$. Our main result is the following.

**Theorem 2.** *If* $L > \tilde{\ell}_{\text{crit}}(\mathbf{s}, D)$*, then* $\mathbf{s}$ *is the unique sequence that is consistent with* $\mathcal{R}_{L,D}(\mathbf{s})$*.*

The main tool used to prove Theorem 2 is a result about spectrum error correction. More precisely, we show that from a noisy version of the $L$-spectrum of $\mathbf{s}$ $\mathcal{R}_{L,D}(\mathbf{s})$, it is possible to obtain $\mathcal{R}_{L',0}(\mathbf{s})$, for some effective read length $L' < L$. This result and the proof of Theorem 2 are presented in Section V.

As in the noiseless case, we point out that $\tilde{\ell}_{\text{crit}}(\mathbf{s}, D)$ cannot be computed a priori, since it is a function of the ground truth sequence $\mathbf{s}$. However, Theorem 2 can in fact be used to obtain a certificate result analogous to Corollary 1, allowing one to certify

whether an assembly $\hat{s}$ is correct, even without prior knowledge of $\tilde{\ell}_{\text{crit}}(\mathbf{s})$ and $M_{\mathbf{s}}(D,\cdot)$.

**Corollary 2.** *If a sequence* $\hat{\mathbf{s}}$ *is consistent with* $\mathcal{R}_{L,D}(\mathbf{s})$ *and* $L > \tilde{\ell}_{\text{crit}}(\hat{\mathbf{s}})$*, then* $\hat{\mathbf{s}} = \mathbf{s}$.

*Proof:* If $\hat{\mathbf{s}}$ is consistent with $\mathcal{R}_{L,D}(\mathbf{s})$, by the definition of consistency, $\mathcal{R}_{L,D}(\mathbf{s})$ can be viewed as a set of reads $\mathcal{R}_{L,D}(\hat{\mathbf{s}})$ from $\hat{\mathbf{s}}$, with an erasure pattern satisfying (a) and (b). But from Theorem 2, if $L > \tilde{\ell}_{\text{crit}}(\hat{\mathbf{s}})$, $\hat{\mathbf{s}}$ is the unique sequence that is consistent with $\mathcal{R}_{L,D}(\hat{\mathbf{s}}) = \mathcal{R}_{L,D}(\mathbf{s})$. Since $\mathbf{s}$ must also be consistent with $\mathcal{R}_{L,D}(\hat{\mathbf{s}})$, we must have $\hat{\mathbf{s}} = \mathbf{s}$. ■

In Table I, we show the value of $\tilde{\ell}_{\text{crit}}(\mathbf{s}, D)$ computed for several real genomes. Computing $\tilde{\ell}_{\text{crit}}(\mathbf{s}, D)$ is generally impractical from a computational standpoint, so the values in Table I are based on heuristics implemented by a sequence alignment tool called Nucmer [10]. We choose the value of $D$ such that $D/\ell_{\text{crit}} \approx 15\%$. We point out that the first two genomes, *R. sphaeroides* and *S. aureus* are from the GAGE dataset [6], which is used as a benchmark for assemblers. Notice that, with the exception of *E. coli 536*, in all cases $\tilde{\ell}_{\text{crit}}(\mathbf{s}, D) = \ell_{\text{crit}}(\mathbf{s}) + mD$, for $m \in \{2, 3, 4\}$. This occurs because, for the genomes considered, $\ell_{\text{crit}}(\mathbf{s})$ is already long enough so that there aren't many approximate repeats of that length.

| Genome ($\mathbf{s}$) | $\ell_{\text{crit}}(\mathbf{s})$ | $\tilde{\ell}_{\text{crit}}(\mathbf{s}, D)$ | $D$ |
|---|---|---|---|
| *R. sphaeroides* | 271 | 331 | 30 |
| *S. aureus* | 1799 | 2399 | 200 |
| *A. ferrooxidans* | 2628 | 3228 | 300 |
| *E. coli 536* | 3245 | 4462 | 450 |
| *E. coli K-12* | 1744 | 2544 | 200 |

TABLE I
COMPUTED $\tilde{\ell}_{\text{crit}}(\mathbf{s}, D)$ FOR $D/\ell_{\text{crit}} \approx 15\%$

While the results in this section were presented for an erasure model, they can be extended to a substitution error model. In fact, if instead of $D$ erasures per read and per base, we have $D/2$ substitution errors, the proofs of Theorems 2 and 3 can be modified accordingly, and the statements still hold. We will restrict the discussion to the erasure case for simplicity.

## V. SPECTRUM ERROR CORRECTION

The main result we use to prove Theorem 2 is a statement about when it is possible to take a noisy $L$-spectrum of $\mathbf{s}$ and unambiguously construct its noiseless $L'$-spectrum, for $L' < L$.

**Theorem 3.** *Suppose that, for some* $k$*, we have*

$$L > k + D \cdot M_{\mathbf{s}}(D, k+1). \tag{4}$$

*Then, for any sequence* $\hat{\mathbf{s}}$ *that is consistent with* $\mathcal{R}_{L,D}(\mathbf{s})$*,* $\mathcal{R}_{k+1,0}(\hat{\mathbf{s}}) = \mathcal{R}_{k+1,0}(\mathbf{s})$.

Theorem 3 says that, by finding a consistent assembly of $\mathcal{R}_{L,D}(\mathbf{s})$, we can obtain the (noiseless) $(k+1)$-spectrum of $\mathbf{s}$, as long as $k$ satisfies (4). Therefore, when $L > \tilde{\ell}_{\text{crit}}(\mathbf{s}, D)$, if we let $k^\star$ be the minimizer in (3), we have that $L > k^\star + D \cdot M_{\mathbf{s}}(D, k^\star + 1)$ and, by Theorem 3, any $\hat{\mathbf{s}}$ that is consistent with $\mathcal{R}_{L,D}(\mathbf{s})$ has the same $(k^\star+1)$-spectrum $\mathcal{R}_{k^\star+1,0}(\mathbf{s})$. But since, $k^\star + 1 > \ell_{\text{crit}}(\mathbf{s})$, Theorem 1 implies that there is only one sequence that is consistent with $\mathcal{R}_{k^\star+1,0}(\mathbf{s})$, and we must have $\hat{\mathbf{s}} = \mathbf{s}$. This proves Theorem 2.

Next, we turn to the proof of Theorem 3. Suppose that we pick some $k$ satisfying (4) and that $\sigma$ is a consistent assembly for the set of reads $\mathcal{R}_{L,D}(\mathbf{s})$ with assembled sequence $\hat{\mathbf{s}} = \hat{\mathbf{s}}(\sigma)$. The main idea of the proof is to show that $(k+1)$-blocks in $\mathbf{s}$ and $\hat{\mathbf{s}}$ are in one-to-one correspondence; i.e., $\hat{\mathbf{s}}_t^{k+1} = \mathbf{s}_{\tau(t)}^{k+1}$ for a bijective mapping $\tau : \{1, ..., G\} \to \{1, ..., G\}$, which implies $\mathcal{R}_{k+1,0}(\mathbf{s}) = \mathcal{R}_{k+1,0}(\hat{\mathbf{s}})$.

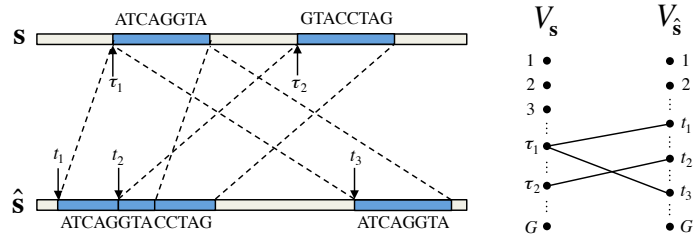



Fig. 4. We place an edge $(u, v)$ in $(V_{\mathbf{s}}, V_{\hat{\mathbf{s}}}, E)$ if $\mathbf{s}_u^{k+1} = \hat{\mathbf{s}}_v^{k+1}$. In this example, $(\tau_1, t_1)$, $(\tau_2, t_2)$ and $(\tau_1, t_3)$ are some of the edges in $(V_{\mathbf{s}}, V_{\hat{\mathbf{s}}}, E)$.

In order to show the existence of this bijection $\tau$, we consider a bipartite graph $(V_{\mathbf{s}}, V_{\hat{\mathbf{s}}}, E_{k+1})$, where $V_{\mathbf{s}} = V_{\hat{\mathbf{s}}} = \{1, ..., G\}$ and $E = \{(u, v) \in V_{\mathbf{s}} \times V_{\hat{\mathbf{s}}} : \mathbf{s}_u^{k+1} = \hat{\mathbf{s}}_v^{k+1}\}$, as illustrated in Fig. 4. The existence of the bijective mapping $\tau$ is equivalent to the existence of a perfect matching in $(V_{\mathbf{s}}, V_{\hat{\mathbf{s}}}, E)$. Hence, Theorem 3 is equivalent to the following:

**Claim 1.** *There exists a perfect matching in* $(V_{\mathbf{s}}, V_{\hat{\mathbf{s}}}, E)$.

For a set of nodes $U \subset V_{\hat{\mathbf{s}}}$, we let $\delta(U) = \{v \in V_{\mathbf{s}} : (v, u) \in E \text{ for } u \in U\}$ be the set of neighbors of $U$. We will show that, for any $U \subset V_{\hat{\mathbf{s}}}$, $|\delta(U)| \geq |U|$, and by Hall's marriage theorem, Claim 1 will follow. We will first state the following lemma, which establishes $|\delta(U)| \geq |U|$ for the special case of sets $U$ of the form $U_{\mathbf{x}} = \{u \in V_{\hat{\mathbf{s}}} : \hat{\mathbf{s}}_u^{k+1} = \mathbf{x}\}$ for some $\mathbf{x} \in \Sigma^{k+1}$.

**Lemma 1.** *For the bipartite graph* $(V_{\mathbf{s}}, V_{\hat{\mathbf{s}}}, E)$*,* $|\delta(U_{\mathbf{x}})| \geq |U_{\mathbf{x}}|$*, for any* $\mathbf{x} \in \Sigma^{k+1}$.

The proof of Lemma 1 is at the end of this section. Now consider a general set $U \in V_{\hat{\mathbf{s}}}$. Let $\mathcal{S}_U^{k+1} = \{\mathbf{s}_u^{k+1} \in \Sigma^{k+1} : u \in U\}$. Since two nodes $u, u' \in U$ with $\mathbf{s}_u^{k+1} \neq \mathbf{s}_{u'}^{k+1}$ cannot be connected to the same node $v \in V_{\mathbf{s}}$, we have

$$\begin{aligned} |\delta(U)| &= \sum_{\mathbf{x} \in \mathcal{S}_U^{k+1}} |\delta(U_{\mathbf{x}} \cap U)| = \sum_{\mathbf{x} \in \mathcal{S}_U^{k+1}} |\delta(U_{\mathbf{x}})| \\ &\geq \sum_{\mathbf{x} \in \mathcal{S}_U^{k+1}} |U_{\mathbf{x}}| \geq \sum_{\mathbf{x} \in \mathcal{S}_U^{k+1}} |U_{\mathbf{x}} \cap U| = |U|, \end{aligned}$$

where the first inequality follows from Lemma 1. By applying Hall's theorem, Claim 1 follows, implying that, $\mathcal{R}_{k+1,0}(\mathbf{s}) = \mathcal{R}_{k+1,0}(\hat{\mathbf{s}})$. Therefore, to conclude the proof of Theorem 3, we just need to prove Lemma 1.

*Proof of Lemma 1:* Let $U_{\mathbf{x}} = \{t_1, ..., t_q\} \subset V_{\hat{\mathbf{s}}}$, where $t_1, ..., t_q$ are distinct and $\hat{\mathbf{s}}_{t_1}^{k+1} = ... = \hat{\mathbf{s}}_{t_q}^{k+1} = \mathbf{x}$. Consider one such block $\hat{\mathbf{s}}_t^{k+1}$, for $t \in \{t_1, ..., t_q\}$. There are $L - k$ reads that cover $\hat{\mathbf{s}}_t^{k+1}$ in $\hat{\mathbf{s}}$, as illustrated in Fig. 5. These are the reads given by $\mathbf{r}_{\sigma^{-1}(t-n)}$, for $n = 0, 1, ..., L - k - 1$. Notice that read $\mathbf{r}_{\sigma^{-1}(t-n)}$ was originally obtained from the segment $\mathbf{s}_{\sigma^{-1}(t-n)}^L$ from the true sequence $\mathbf{s}$. The consistency requirement on $\sigma$ thus implies that $d_H(\mathbf{s}_{\sigma^{-1}(t-n)}^L, \mathbf{s}_{t-n}^L) \leq D$. Moreover, if we just focus on the $(k+1)$-block corresponding to $\hat{\mathbf{s}}_t^{k+1}$, we have $d_H(\mathbf{s}_{\sigma^{-1}(t-n)+n}^{k+1}, \mathbf{s}_t^{k+1}) = d_H(\mathbf{s}_{\sigma^{-1}(t-n)+n}^{k+1}, \mathbf{x}) \leq D$, which holds for each $t \in \{t_1, ..., t_q\}$ and $n = 0, ..., L - k - 1$.

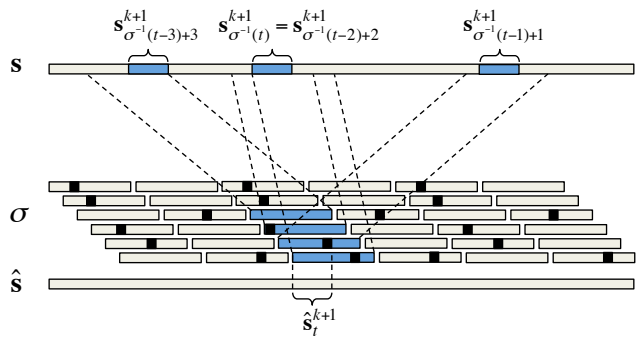



Fig. 5. For an arbitrary length-$(k+1)$ block $B$ in $\hat{s}$, the $L-k$ reads that completely cover $B$ according to the assembly $\sigma$ are shaded (in this example, $L=6$ and $k=2$). By mapping these $L-k$ reads back to $\mathbf{s}$, we find the corresponding $(k+1)$-blocks in $\mathbf{s}$ given by $\mathbf{s}^{k+1}_{\sigma^{-1}(t-j)+j}$ for $j=0,1,2,3$. Notice that, in this example, $\sigma^{-1}(t)=\sigma^{-1}(t-2)+2$, because reads $\mathbf{r}_{\sigma^{-1}(t)}$ and $\mathbf{r}_{\sigma^{-1}(t-2)+2}$ are aligned to each other in the same way in $\mathbf{s}$ and $\hat{\mathbf{s}}$.

If we now consider the set of all such $(k+1)$-blocks in $\mathbf{s}$

$$\mathcal{S} = \left\{ \mathbf{s}^{k+1}_{\sigma^{-1}(t_i-n)+n} : n = 0, ..., L-k-1, i = 1, ..., q \right\}, \quad (5)$$

since $d_H(\mathbf{y}, \mathbf{x}) \leq D$ for each $\mathbf{y} \in \mathcal{S}$, we have that $\rho(\mathcal{S}) \leq D$. Hence, if we let

$$\mathcal{T} = \left\{ \sigma^{-1}(t_i - n) + n : 0 \leq n \leq L-k-1, 1 \leq i \leq q \right\}$$

be the starting positions of these blocks in $\mathbf{s}$, $\mathcal{T}$ must satisfy $|\mathcal{T}| \leq M_{\mathbf{s}}(D, k+1)$. Now consider the set of $(n,i)$ pairs

$$\mathcal{B} = \{(n,i) : 0 \leq n \leq L-k-1, 1 \leq i \leq m\}.$$

We will define a partition on $\mathcal{B}$ according to the value of $\sigma^{-1}(t_i - n) + n$. More precisely, we will let

$$\mathcal{B}_\tau = \left\{ (n,i) \in \mathcal{B} : \sigma^{-1}(t_i - n) + n = \tau \right\},$$

for $\tau \in \mathcal{T}$. It is clear that $\{\mathcal{B}_\tau\}_{\tau \in \mathcal{T}}$ is a partition of $\mathcal{B}$. We claim that there exist distinct $\tau_1, ..., \tau_q \in \mathcal{T}$ such that $|\mathcal{B}_{\tau_j}| \geq D+1$, for $j = 1, ..., q$. Suppose by contradiction that this is not the case, and we have at most $q-1$ parts $\mathcal{B}_\tau$ with $|\mathcal{B}_\tau| \geq D+1$. Notice that, since $\sigma : (1, ..., G) \to (1, ..., G)$ is one-to-one, $\sigma^{-1}(t_i - n) + n \neq \sigma^{-1}(t_j - n) + n$ if $t_i \neq t_j$, and, for any $\tau$, we must have $|\mathcal{B}_\tau| \leq L-k$. Therefore, since (4) implies $L-k-1 \geq D \cdot M_{\mathbf{s}}(D, k+1)$,

$$\begin{aligned} \sum_{\tau \in \mathcal{T}} |\mathcal{B}_\tau| &\leq (q-1)(L-k) + (|\mathcal{T}| - q + 1)D \\ &\leq (q-1)(L-k) + D \cdot M_{\mathbf{s}}(D, k+1) \\ &= q(L-k) - 1. \end{aligned}$$

But since $\sum_{\tau \in \mathcal{T}} |\mathcal{B}_\tau| = |\mathcal{B}| = q(L-k)$, we have a contradiction.

Now consider the segments $\mathbf{s}^{k+1}_{\tau_j}$ with $|\mathcal{B}_{\tau_j}| \geq D+1$, for $j = 1, ..., q$. Since $\tau_1, ..., \tau_q$ are all distinct, these segments start at different points in $\mathbf{s}$. Moreover, since $|\mathcal{B}_{\tau_j}| \geq D+1$, each $\mathbf{s}^{k+1}_{\tau_j}$ is covered by $D+1$ reads from the reads that cover $\hat{\mathbf{s}}^{k+1}_{t_i}$, $i = 1, ...q$. Notice that these must be distinct reads from the multiset $\mathcal{R}_{L,D}(\mathbf{s})$. This is because two distinct pairs $(n,i)$ and $(m,j)$ in $\mathcal{B}_\tau$ must have $n \neq m$, and the corresponding reads are $\mathbf{r}_{\sigma^{-1}(t_i-n)} = \mathbf{r}_{\tau-n}$ and $\mathbf{r}_{\sigma^{-1}(t_j-m)} = \mathbf{r}_{\tau-m}$, which are distinct reads (not necessarily different sequences from $\Sigma^L$). Finally, as illustrated in Fig. 6, we note that, since there are at most $D$ erasures per base in $\mathbf{s}$, we have that $\mathbf{s}^{k+1}_{\tau_j} = \mathbf{x}$, for $j = 1, ..., q$. We conclude that $|\delta(U)| \geq q$. ■

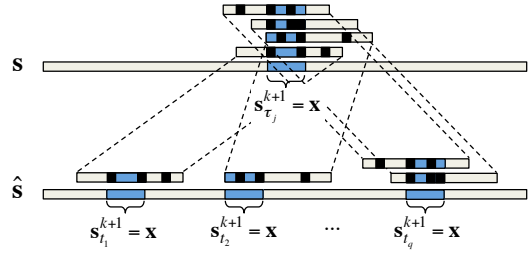



Fig. 6. If $|\mathcal{B}_{\tau_j}| \geq D+1$, at least $D+1$ of the reads that cover one of $\hat{\mathbf{s}}^{k+1}_{t_1}, ..., \hat{\mathbf{s}}^{k+1}_{t_q}$ in $\hat{\mathbf{s}}$ also cover $\mathbf{s}^{k+1}_{\tau_j}$ in $\mathbf{s}$ (in this example, $D=3$). Since there are at most $D$ erasures per base in $\mathbf{s}$, we must have $\mathbf{s}^{k+1}_{\tau_j} = \mathbf{x}$.

## VI. Concluding Remarks

Our results show that for several actual genomes, if we are in a dense-read model with reads 20-40% longer than the noiseless requirement $\ell_{\text{crit}}(\mathbf{s})$, perfect assembly feasibility is robust to erasures at a rate of about 10%. While this is not as optimistic as the message from [7], we emphasize that we consider an adversarial error model. When errors instead occur at random locations, it is natural to expect less stringent requirements.

Another message provided by our results deals with error correction. Most current sequencing technologies employ error correction algorithms based on aligning reads to form clusters and outputing a cleaned-up read for each cluster. However, the spectrum error correction result from Theorem 3 suggests that a "global" approach to generating cleaned-up reads (based on finding a consistent assembly and looking at its spectrum) may perform better than cluster-based, or local, error correction.

A direction for future work is to replace the dense-read model with a shotgun read model. While the $L$-spectrum approach is motivated by the high-throughput of current technologies, it bypasses the question of the actual coverage depth required for assembly. As was the case in [3], we expect the read length requirements from the dense-read model to translate into *bridging* conditions in the shotgun model, allowing one to compute the coverage required for perfect reconstruction with high probability.

## Acknowledgment

This work is partially supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

## References

[1] S. L. Salzberg, "Mind the gaps," *Nature methods*, vol. 7, no. 2, pp. 105–106, 2010.
[2] E. Myers. (2014, Feb.). [Online]. Available: https://twitter.com/thegenemyers/status/437349388676263937
[3] G. Bresler, M. Bresler, and D. Tse, "Optimal assembly for high throughput shotgun sequencing," *BMC Bioinformatics*, 2013.
[4] E. Ukkonen, "Approximate string matching with q-grams and maximal matches," *Theoretical Computer Science*, vol. 92, no. 1, 1992.
[5] P. Pevzner, "DNA physical mapping and alternating Eulerian cycles in colored graphs," *Algorithmica*, vol. 13, pp. 77–105, 1995.
[6] S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marcais, M. Pop, and J. A. Yorke, "GAGE: a critical evaluation of genome assemblies and assembly algorithms," *Genome Research*, vol. 22, no. 3, pp. 557–567, 2012.
[7] A. Motahari, K. Ramchandran, D. Tse, and N. Ma, "Optimal DNA shotgun sequencing: Noisy reads are as good as noiseless reads," *Proc. of IEEE International Symposium on Information Theory*, pp. 1640–1644, 2013.
[8] A. Motahari, G. Bresler, and D. Tse, "Information theory of DNA shotgun sequencing," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6273–6289, Oct. 2013.
[9] P. E. C. Compeau, P. Pevzner, and G. Tesler, "How to apply de Bruijn graphs to genome assembly," *Nature Biotechnology*, vol. 29, 2011.
[10] [Online]. Available: http://mummer.sourceforge.net/