Special Issue on Biological Applications of Information Theory in Honor of Claude Shannon's Centennial—Part 2

Fundamental Limits of Genome Assembly Under an Adversarial Erasure Model

Ilan Shomorony, Thomas A. Courtade, *Member, IEEE*, and David Tse, *Fellow, IEEE*

***Abstract*—In contrast to second-generation DNA sequencing technologies, emerging third-generation technologies are capable of delivering reads that are long enough to enable perfect genome assembly. Unfortunately, the benefits of long reads are accompanied by higher rates of read errors. This motivates a question of fundamental import: what read-length and error-rate combinations allow for perfect assembly of the genome? Formal investigation of this tradeoff is complicated by the fact that tractable probabilistic models for the genome sequence and error process fail to capture key features of the problem: real genomes contain long repetitive patterns, and read errors are often bursty and sequence-dependent. In order to circumvent these modeling barriers and take a first step toward the study of this question, we consider a simple setting: the genome sequence is arbitrary, and the read errors are erasures that can occur at adversarially chosen positions, up to a limit in the number of erasures per read and per genome position. In this context, we show that a natural error-correction scheme is optimal in the sense that it recovers the error-free *k-spectrum* of the genome for the largest possible *k*. The worst-case nature of our analysis ensures that the proposed error-correction method is robust and allows us to study its performance under stochastic error models. As a result, we show that, for several real genomes, the impact of read errors on the information-theoretic requirements for perfect assembly is relatively mild.**

***Index Terms*—Genome assembly, DNA sequencing, error correction, erasure model.**

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*.

Manuscript received June 2, 2016; revised September 30, 2016; accepted December 6, 2016. Date of publication December 19, 2016; date of current version February 17, 2017. This work was supported in part by the Center for Science of Information, in part by the NSF Science and Technology Center under Grant CCF-0939370, and in part by the Hellman Fellowship. Part of this work was completed while the authors were visiting the Simons Institute for the Theory of Computing, UC, Berkeley, CA, USA. The associate editor coordinating the review of this paper and approving it for publication was S. M. Moser.

I. Shomorony and T. Courtade are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley CA 94720 USA (e-mail: ilan.shomorony@berkeley.edu; courtade@berkeley.edu).

D. Tse is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: dntse@stanford.edu).

Digital Object Identifier 10.1109/TMBMC.2016.2641440

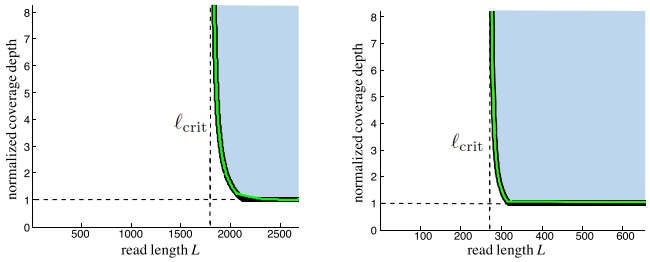
Roughly speaking, different shotgun sequencing platforms can be distinguished from the point of view of three main metrics: the *read length*, the *read error rate*, and the *read throughput*. In the last decade, the so-called next-generation sequencing platforms have attained considerable success at employing heavy parallelization in order to achieve *high-throughput* shotgun sequencing 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 de Bruijn graph based algorithms [2]–[4], read errors create extraneous nodes and edges in the assembly graph, which results in added complexity and poses challenges to the assembly of reliable contigs. In practice, this issue is often dealt with using k-mer filtering [6] and error-correction tools [5], [7]–[11] that attempt to clean up the reads before the assembly algorithm is applied. In fact, it has been shown that even noisy long reads, if properly preprocessed, can be used to obtain finished assemblies of bacterial genomes [12]–[16]. However, such claims must usually be made on a dataset-by-dataset basis. Moreover, error correction tools and assembly pipelines are in general evaluated relative to each other, and the evaluation is contingent on the existence of a reliable reference genome.

A more fundamental question regarding the performance of assembly pipelines 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 a target genome?

[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%.



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

Fig. 1. Feasibility region for a target error probability $\epsilon = 0.01$. The thick black curve is a feasibility lower bound for any algorithm, and the green line represents the performance of the Multibridging algorithm [18]. In the vertical axis, the normalized coverage depth is N/N_{LW}, where $N_{LW} \approx (G/L)\log(G/\epsilon)$ is the Lander-Waterman coverage depth; i.e., the number of reads required to cover the whole genome with a probability $1-\epsilon$.

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 and assembly pipelines.

Such a framework was initiated in [17] and [18] for *error-free* reads. In [18], 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 ℓ_{crit}, reconstruction is impossible; a read length slightly above ℓ_{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 ℓ_{crit} is given by the length of the longest *interleaved repeat* in the genome (Fig. 2). and had appeared in earlier works by Ukkonen [19] and Pevzner [20] in the context of *sequencing by hybridization*.

Given this framework, the impact of read errors on genome assembly can be studied by asking how the information-theoretic requirements captured by these feasibility curves change when there are errors in the reads. In particular, understanding the impact of errors in the critical read length ℓ_{crit} is important from a practical point of view: while increasing the coverage depth of the experiment incurs a (roughly) proportional increase in cost, the read length is usually dictated by the technology and chemistry utilized and cannot be tuned. Hence, designing assembly algorithms which succeed whenever the read length exceeds the information-theoretic requirement is highly desirable.

When reads have errors, a natural conjecture is that one should view *approximate* repeats as exact repeats, and the critical read length required for assembly, instead of ℓ_{crit}, would be $\ell_{\text{crit, app}}$, defined as the length of the longest approximate interleaved repeats, illustrated in Fig. 2(b). This number always exceeds ℓ_{crit} and, when evaluated on real genomes (for an appropriate definition of approximate repeats) can be seen to be substantially larger than ℓ_{crit}.

But is $L > \ell_{\text{crit, app}}$ actually required from a fundamental point-of-view? In this work, we study this question under a simple erasure error model and show that $L > \ell_{\text{crit, app}}$ is

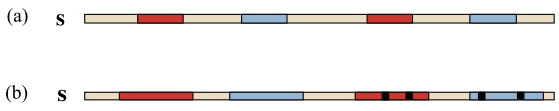


Fig. 2. (a) Interleaved repeats and (b) an approximate interleaved repeats.

not needed for perfect assembly. In fact, under this model, the fundamental read length requirement essentially remains unchanged by the addition of errors, as illustrated in Fig. 3.

The impact of read errors on genome assembly was previously studied in [21] and [22]. Motahari *et al.* [21] proposed an assembly scheme based on first using a *typicality test* to generate a set of cleaned up reads and then assembling them as if they were error free. Based on this approach, they obtained a result in the same spirit of the main result in this paper. More precisely, they showed that, as long as the error rate is below a threshold, the requirements for assembly in terms of read length and coverage depth are the same for noisy and noiseless reads. However, two important modeling assumptions were necessary in order to establish this result.

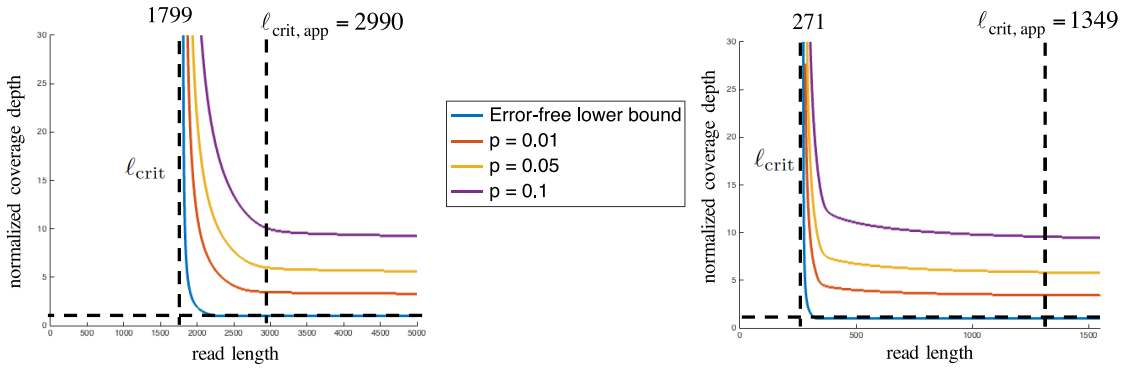
1) The genome was assumed to be an i.i.d. sequence of length G, and the asymptotic regime $G \to \infty$ was considered.
2) Errors were assumed to be i.i.d. for some fixed and known probability p.

It is well known that in practice the main obstacle to assembly are long repeats in the genome, which are not well modeled by i.i.d. sequences. Therefore, due to (1), the asymptotic result in [21] cannot be used to characterize sequence-specific bounds such as those in Fig. 3. While (2) is motivated by nominal error rates that are usually provided for each sequencing technology, read errors often occur in bursts (generating so-called *chimeric segments*) and in a sequence-specific fashion (e.g., in homopolymers). Typicality-based approaches are usually sensitive to deviations from the probabilistic model assumed, even if the true error process is more tractable than the one assumed (i.e., it has a lower error rate, or a higher capacity). Hence, an error-correction scheme that does not rely heavily on the error model is desirable.

In this work, we take an initial step in studying this problem without assumptions (1) and (2), and consider the following setting: the genome sequence is deterministic and comprises a single chromosome, and the read errors are erasures[2] that can occur at adversarially chosen positions, but under constraints in the number of erasures per read and per base in the genome.

Under this framework, we ask a fundamental question: when is there enough *information* in the set of reads to allow error correction to be performed in an unambiguous way? We formally pose this question by defining the *k-spectrum* of the genome (i.e., the set of all error-free length-k substrings) as the goal of the error correction problem. As we argue, reconstructing the k-spectrum for a large k is as difficult as the perfect reconstruction problem, and we can focus on the problem of

[2]When a base is erased, it is replaced by an erasure symbol ε. Erasure models are easier to analyze than models with insertions, deletions and substitutions, and hence are are commonly used in information theory as a starting point in the study of fundamental performance limitations.



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

Fig. 3. Lower bound for assembly feasibility region with error-free reads and upper bound when reads have erasures with probability p.

characterizing the largest value of k for which the set of reads unambiguously determines the k-spectrum.

In order to answer this question, we develop a notion of *worst-case typicality*, which serves as a test to identify true k-mers based on the read data. Under the adversarial sequence and erasure model considered, this typicality approach has strong theoretical guarantees: it can reconstruct the k-spectrum for the largest possible k. Furthermore, the techniques developed under this adversarial model, once carried over to a probabilistic erasure model, exhibit nice robustness properties. This allows us to combine this typicality test with assembly algorithms developed for error-free reads to gain insight into the fundamental performance limitations of genome assembly from reads with errors. In particular, we can explicitly compute the requirements for perfect assembly in terms of read length and coverage depth for small bacterial genomes, and show that they do not differ significantly from the error-free case considered in [18], as illustrated in Fig. 3. This result is relevant as it suggests that the higher error rates of current long-read sequencing technologies may not fundamentally impact their ability to yield perfect assemblies.

The rest of the paper is organized as follows. In Section II, we motivate the spectrum reconstruction problem as a way to study the error correction problem, describe the erasure model, and state our main result. In Section III, we describe the worst-case typicality test, which we use to prove our main result. Section IV is dedicated to show that, although our main results are derived under an adversarial erasure model, they can be used to compute feasibility curves for a probabilistic erasure model. Finally, Section V discusses practical limitations of the approach considered in this paper, and Section VI concludes the paper.

II. Assembly via Spectrum Reconstruction

In the DNA assembly problem, the goal is to reconstruct a sequence $\mathbf{s} = (s[1], \ldots, s[G])$ of length G with symbols from the alphabet $\Sigma = \{\mathsf{A}, \mathsf{G}, \mathsf{C}, \mathsf{T}\}$. In order to simplify the exposition and avoid edge effects, we will 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 ℓ starting at $s[i]$; i.e., $\mathbf{s}_i^{\ell} = (s[i], s[i+1], \ldots, s[i+\ell-1])$.

The sequencer provides a set of N reads $\mathcal{R} = \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$ 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 in $\mathcal{R}$ may be corrupted by noise. We define the k-spectrum of $\mathbf{s}$ as the multiset $\mathcal{S}_k(\mathbf{s}) = \{\mathbf{s}_t^k : t = 1, \ldots, G\}$, and we let $\text{supp}(\mathcal{S}_k(\mathbf{s})) \subset \Sigma^k$ be the support of such multiset. Whenever the sequence $\mathbf{s}$ is clear from the context, we will simply write $\mathcal{S}_k$ and $\text{supp}(\mathcal{S}_k)$. Following the convention in the field, we will often refer to a length-k sequence in Σ^k as a k-mer. In this work, we consider three related assembly problems. In decreasing order of difficulty, they are

A1. Reconstruct the entire sequence $\mathbf{s}$ from $\mathcal{R}$.
A2. Reconstruct the k-spectrum of $\mathbf{s}$, $\mathcal{S}_k(\mathbf{s})$, from $\mathcal{R}$, for some $k \leq L$.
A3. Reconstruct the support of the k-spectrum of $\mathbf{s}$, $\text{supp}(\mathcal{S}_k(\mathbf{s}))$, from $\mathcal{R}$, for some $k \leq L$.

While reconstructing the spectrum or just its support rather than the complete sequence $\mathbf{s}$ may seem significantly more modest goals than (A1), we claim they are almost as difficult as perfect assembly, and can be seen as intermediate steps towards (A1).

Consider the problem (A2) of reconstructing the spectrum $\mathcal{S}_k$. We know from results in [19] and [20] that for every sequence $\mathbf{s}$, there exists a critical read length $\ell_{\text{crit}}(\mathbf{s})$ as a function of the repeat statistics of $\mathbf{s}$ (see Appendix A for a formal definition), for which we have the following.

Theorem 1: If $k > \ell_{\text{crit}}(\mathbf{s})$, then $\mathbf{s}$ is the unique sequence with k-spectrum $\mathcal{S}_k(\mathbf{s})$. Conversely, if $k \leq \ell_{\text{crit}}(\mathbf{s})$, there exists a sequence $\mathbf{s}' \neq \mathbf{s}$ for which $\mathcal{S}_k(\mathbf{s}) = \mathcal{S}_k(\mathbf{s}')$.

Hence by solving the assembly problem (A2) for sufficiently large k, we also solve the (standard) assembly problem (A1). As it turns out, similar guarantees can be given for the assembly problem (A3). As we describe in Section VI, previous results from [18] and [23] imply that there exists another critical read length $\bar{\ell}_{\text{crit}}(\mathbf{s}) \geq \ell_{\text{crit}}(\mathbf{s})$ such that, if $k > \bar{\ell}_{\text{crit}}(\mathbf{s})$, $\mathbf{s}$ can be assembled from $\text{supp}(\mathcal{S}_k)$. More precisely, we have the following.

Theorem 2 [18]: If $k > \bar{\ell}_{\text{crit}}(\mathbf{s})$, the Multibridging algorithm correctly assembles $\mathbf{s}$ from $\text{supp}(\mathcal{S}_k)$.

Therefore, by solving (A3) for sufficiently large k, we also solve (A1). As we will discuss in the subsequent sections, aiming to reconstruct the support of the spectrum of $\mathbf{s}$ has several advantages with respect to reconstructing the complete spectrum. Moreover, as we describe in Appendix A, the difference between $\ell_{\text{crit}}(\mathbf{s})$ and $\bar{\ell}_{\text{crit}}(\mathbf{s})$ is a technicality, and in many practical settings $\ell_{\text{crit}}(\mathbf{s}) = \bar{\ell}_{\text{crit}}(\mathbf{s})$. Therefore, in these cases, all three problems are equivalent.

A. Adversarial Erasure Model

In this work, we will study the problem of error correction of sequencing data by viewing it as the spectrum reconstruction problem. We will build upon the ideas from [24], and consider this problem under an adversarial erasure model. Given that actual sequencing noise profiles are complex (non-i.i.d., sequence-dependent, bursty) and technology-dependent, such an adversarial model is intended to prevent the development of techniques that are tied to a specific probabilistic model. In Section IV, we will evaluate the techniques developed under the adversarial erasure model on a probabilistic erasure setting, and show that they still provide powerful error correction techniques.

The adversarial erasure model herein proposed can be seen as a generalization of the model considered in [24]. Erasures can be practically motivated by the fact that sequencing technologies usually provide a quality score for each base that is sequenced, which could be thresholded into "good" and "bad" bases (although in practice the "good" bases are not guaranteed to be correct). The reads in $\mathcal{R}$ will be length-L sequences from the alphabet $\Sigma' = \{\mathsf{A}, \mathsf{G}, \mathsf{C}, \mathsf{T}, \varepsilon\}$, where ε corresponds to an erasure. Thus, a read starting at position i from $\mathbf{s}$ can be written as $\mathbf{r}_i = (r_i[0], \ldots, r_i[L-1])$, where either $r_i[j] = s[i+j]$ or $r_i[j] = \varepsilon$, for $1 \leq i \leq G$ and $0 \leq j \leq L-1$. Notice that an erasure is distinct from a deletion (which is more commonly studied in the context of sequencing data) since the location of an erasure is known due to the symbol ε.

In [24], we focused on the L-spectrum read model, or the *dense-read* model. More precisely, $\mathcal{R}$ contained exactly one length-L read from every position in $\mathbf{s}$, and these reads were corrupted by erasures in an adversarial manner. For a fixed parameter D, the adversarial erasure model in [24] was constrained as follows:

(a') There are at most D erasures per read.

(b') Each base $s[t]$ is erased at most D times across all reads.

While the simplicity of this model made it analytically tractable, the assumption of an even coverage depth across the whole sequence is unrealistic. Hence in this work we move away from this dense-read model and instead allow $\mathcal{R}$ to be a set of N reads with arbitrary starting positions in $\mathbf{s}$, with the only constraint being that there is at least one read starting in every W-length window.[3] Thus, if we let $\mathcal{R}_\tau^W$ be the set of reads with starting positions in $s[\tau], s[\tau+1], \ldots, s[\tau+W-1]$ ($\mathcal{R}_\tau^W$ is not known by the assembler), we have $|\mathcal{R}_\tau^W| \geq 1$ for $\tau = 1, 2, \ldots, G$. While constraint (a') above translates to this general model in a straightforward manner, we need a new way to redefine constraint (b'). Hence we will define an additional parameter $p \in (0, 1)$ (the erasure rate) and require that the erasures satisfy the following constraints:

(a) There are at most pL erasures per read.

(b) Each base $s[t]$ is erased in at most a fraction p of the reads in $\mathcal{R}_\tau^W$, for $t - L < \tau \leq t - W + 1$.

Notice that if $t - L < \tau \leq t - W + 1$, all reads in $\mathcal{R}_\tau^W$ cover $s[t]$ and (b) is well defined. For the L-spectrum read model from [24], the constraints in (a) and (b) reduce to (a') and (b') for $W = L$ and $p = D/L$.

We point out that although the erasure model is adversarial, the constraints imposed mimic a probabilistic erasure process, where bases are erased with some probability p. Therefore, if one assumes a standard probabilistic model where reads are sampled independently and uniformly at random from the genome and each base is erased independently with probability p, this adversarial model can be shown to hold with high probability as $N \to \infty$ for a fixed W.

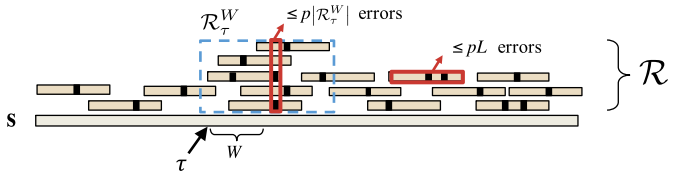


Fig. 4. Illustration of adversarial model constraints given by (a) and (b).

B. Maximum Reconstructible Support

We will focus on studying the support reconstruction problem under the adversarial model given by (a) and (b). Thus we are interested in assembling the support of the k-spectrum from the set of erased reads $\mathcal{R}$, for k as large as possible. As DNA sequences tend to be highly repetitive and present patterns that are difficult to model, we will once again take a worst-case approach by considering a minimax formulation of (A3). Let $\mathbf{R}(\mathbf{s}, p, W)$ be the set of all possible sets of reads $\mathcal{R}$ from $\mathbf{s}$ with erasures satisfying the adversarial constraints (a) and (b). We will write $\mathcal{R} \Rightarrow \text{supp}(\mathcal{S}_k)$ if the set of reads $\mathcal{R}$ implies $\text{supp}(\mathcal{S}_k)$, in the sense that $\text{supp}(\mathcal{S}_k)$ is the only possible support of the k-spectrum that is consistent with $\mathcal{R}$ under the erasure model in (a) and (b). We are then interested in characterizing the *maximum reconstructible support*

$$k^\star(p, W) = \min_{\mathbf{s},\ \mathcal{R} \subset \mathbf{R}(\mathbf{s}, p, W)} \max\{k : \mathcal{R} \Rightarrow \text{supp}(\mathcal{S}_k)\}. \quad (1)$$

In words, $k^\star(p, W)$ is the largest k whose spectrum can be unambiguously reconstructed for any sequence $\mathbf{s}$ from adversarially corrupted set of reads. Intuitively, characterizing $k^\star(p, W)$ corresponds to devising an error-correction scheme that takes the set of noisy reads $\mathcal{R}$ and attempts to construct a set of "clean" reads of length k, one from each position in the genome, for k as large as possible. The worst-case nature of (1) guarantees that the devised error-correction scheme does not exploit potentially unrealistic assumptions of probabilistic models.

[3] This is natural in the context of the standard Lander-Waterman probabilistic model for sequencing, where a number of reads $N = \alpha N_{LW} \approx \alpha(G/L)\log(G/\epsilon)$ implies the existence of at least one read starting in every window of length L/α with probability $1-\epsilon$. As practical values of α may be in the range 5 to 10, W can be thought of as a relatively small fraction of L.

As we describe in Appendix B, the following upper and lower bounds hold:

$$\min(L-W+1, \lceil 1/p \rceil - 1) \leq k^*(p, W) \leq L - W + 1. \quad (2)$$

In particular, we point out that the lower bound $\lceil 1/p \rceil - 1$ can be understood as representing the k-mer count approach to error correction. Since $1/p$ is the average length of an error-free segment, one can reconstruct the $(\lceil 1/p \rceil - 1)$-mer spectrum of $\mathbf{s}$ by simply extracting all $(\lceil 1/p \rceil - 1)$-mers of all reads, and keeping those with no errors.

While characterizing $k^*(p, W)$ exactly in general is challenging, it is unclear how relevant this quantity is to the assembly problem of real genomes, since considering the worst-case sequence $\mathbf{s}$ can be too pessimistic. Ideally, we would like to characterize (1) after we restrict $\mathbf{s}$ to be in a large set that contains most real genomes, but is still amenable to an analytical solution. We will show that by constraining the set of sequences $\mathbf{s}$ to sequences that do not have too many long approximate repeats, the upper bound of $L - W + 1$ is in fact achievable.

C. Main Result

We derive a method for estimating the k-spectrum $\mathcal{S}_k$ based on a test that decides, given the set of reads $\mathcal{R}$, whether a given k-mer should be included in the reconstruction $\widehat{\mathcal{S}}_k$. This allows us to achieve the upper bound in (2) under a mild restriction on the possible sequences $\mathbf{s}$. Hence, under this restriction, the proposed method is worst-case optimal from the point of view of obtaining the maximum reconstructible support (1).

Motivated by the results in [24], we will use the approximate repeat statistics in order to define this set of "allowed" sequences in the maximum reconstructible support formulation from Section II-B. For a set of segments $\mathcal{U}$ of a given length ℓ; i.e., $\mathcal{U} \subset \Sigma^\ell$, we define the radius of $\mathcal{U}$ to be

$$\rho(\mathcal{U}) = \min_{\mathbf{x} \in \Sigma^\ell} \max_{\mathbf{y} \in \mathcal{U}} d_H(\mathbf{y}, \mathbf{x}), \quad (3)$$

where $d_H(\mathbf{y}, \mathbf{x})$ is the Hamming distance between $\mathbf{y}$ and $\mathbf{x}$. We will say that the segments in $\mathcal{U}$ are q-approximate copies if $\rho(\mathcal{U}) \leq q\ell$. As illustrated in Fig. 5, if one were to plot $\mathcal{S}_\ell(\mathbf{s})$ as points in the metric space Σ^ℓ, the existence of several points in close proximity, i.e., a large set $\mathcal{U} \subset \mathcal{S}_\ell$ with a small radius $\rho(\mathcal{U})$ implies that $\mathbf{s}$ has more ambiguity in terms of assembly. To capture that, we let $V_{\mathbf{s}}(r, \ell)$ be the maximum number of r-approximate length-ℓ segments in $\mathbf{s}$; i.e.,

$$V_{\mathbf{s}}(r, \ell) \triangleq \max\{|\mathcal{U}| : \mathcal{U} \subset \mathcal{S}_\ell(\mathbf{s}), \rho(\mathcal{U}) \leq r\}. \quad (4)$$

This quantity was used in [24] to characterize when the set of reads $\mathcal{R}$ uniquely determines $\mathcal{S}_k$, but under the assumption of dense reads; i.e., $\mathcal{R}$ contains one read starting at each position of $\mathbf{s}$. More precisely, it is shown that if

$$L > k + D \cdot V_{\mathbf{s}}(D, k+1), \quad (5)$$

where D is the error parameter in the erasure model in (a') and (b'), then $\mathcal{R}$ uniquely determines the full spectrum $\mathcal{S}_k$ (thus solving problem (A2) for this setting). In this work, we depart from the assumption of dense reads and use (4) instead to define our set of genomes of interest as

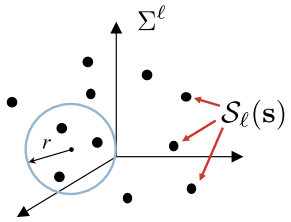


Fig. 5. If we consider plotting S_ℓ as points in Σ^ℓ, $V_{\mathbf{s}}(r, \ell)$ is the maximum number of points that can be enclosed by a ball of radius r.

TABLE I
$L = \ell_{\text{CRIT}}$ AND $W = 0.15L$

Genome ($\mathbf{s}$)	$V_{\mathbf{s}}(pL, L - W + 1)$
R. sphaeroides	2
S. aureus	3
A. ferrooxidans	3
E. coli 536	3
E. coli K-12	4

$$\mathcal{G}(p, L, W) = \{\mathbf{s} : V_{\mathbf{s}}(pL, L - W + 1) < 1/p\}. \quad (6)$$

In words, a sequence $\mathbf{s}$ is in $\mathcal{G}(p, L, W)$ if it contains less than $(1/p)$ pL-approximate repeats of length $L - W + 1$. As we remarked in Section II-B, W should be thought of as a small fraction of L, in which case requiring $\mathbf{s} \in \mathcal{G}(p, L, W)$ corresponds to requiring the number of approximate repeats at a length close to L to not exceed $1/p$. As shown in Table I, by computing $V_{\mathbf{s}}(pL, L - W + 1)$ for several real genomes when $L = \ell_{\text{crit}}$ (the required read length for assembly from noiseless reads) and $W = 0.15L$, we have $V_{\mathbf{s}}(pL, L - W + 1) \leq 4$, which guarantees that $\mathbf{s} \in \mathcal{G}(p, L, W)$ as long as the worst-case erasure rate satisfies $p < 1/4$.

Our main result is the characterization of the maximum recoverable support under this restriction to "reasonable" genomes.

Theorem 3: For the adversarial erasure model in (a) and (b) with $p < 1/2$, we have

$$k^\star(p, W, \mathcal{G}) \triangleq \min_{\mathbf{s} \in \mathcal{G}(p, L, W), \mathcal{R} \subset \mathbf{R}(\mathbf{s}, p, W)} \max\{k : \mathcal{R} \Rightarrow \text{supp}(\mathcal{S}_k)\} = L - W + 1. \quad (7)$$

The result in Theorem 3 states that, as long as $\mathbf{s} \in \mathcal{G}$, a set of noisy reads satisfying (a) and (b) always allows the reconstruction of the $(L - W + 1)$-spectrum. In other words, even if the noise on the reads is adversarial and the sequence is the worst case in the set $\mathcal{G}$, one can still unambiguously obtain a set of cleaned up reads of length $L - W + 1$. Therefore, the effective loss in read length incurred by read errors is W in the worst-case and, when W is a relatively small fraction of L (which is a natural assumption when sequencing at a reasonable coverage, as we argued before), this result supports the message that noisy reads are essentially as good as noiseless reads, first observed in [21].

To prove this result we will describe a technique to construct an estimate of the k-spectrum $\widehat{\mathcal{S}}_k$ from the set of noisy reads $\mathcal{R}$. In the spirit of the approach in [21], this construction can

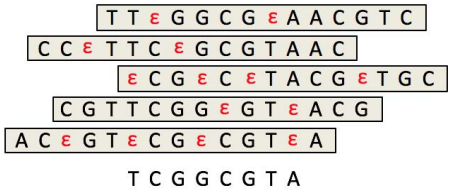


Fig. 6. The 7-mer TCGGCGTA is $(3, 2, 5)$-typical.

be seen as a typicality-like test, which, nonetheless, does not assume a specific probability distribution for the erasures or for the underlying sequence $\mathbf{s}$.

III. A Worst-Case Typicality Test

In order to prove Theorem 3, we will introduce a test that decides whether a k-mer is a true k-mer (i.e., a k-mer that appears in $\mathbf{s}$) by clustering similar reads satisfying certain properties. In the flavor of [21], we can view this procedure as a typicality test, where we check for a given k-mer $\mathbf{x}$, whether there are reads in $\mathcal{R}$ that look like typical outputs of passing $\mathbf{x}$ through the erasure channel. However, since we are dealing with the adversarial erasure model described in Section II-A, the test can be thought of as a worst-case typicality test.

Definition 1: Given the set of reads $\mathcal{R}$, a k-mer $\mathbf{x} \in \Sigma^k$ is (D_h, D_v, m)-typical if we can find k-mers $\mathbf{x}_1, \ldots, \mathbf{x}_m$ from distinct reads in $\mathcal{R}$ satisfying

1) *consistency:* $\mathbf{x}_i[t] = \varepsilon$ or $\mathbf{x}_i[t] = \mathbf{x}[t]$, for $1 \le i \le m$ and $1 \le t \le k$
2) *horizontal constraint:* $|\{t : \mathbf{x}_i[t] = \varepsilon\}| \le D_h$, for $1 \le i \le m$
3) *vertical constraint:* $|\{i : \mathbf{x}_i[t] = \varepsilon\}| \le D_v$, for $1 \le t \le k$

Notice that given the consistency property, the horizontal constraint can also be written as $d_H(\mathbf{x}, \mathbf{x}_i) \le D_h$, where the Hamming distance is defined over the extended alphabet Σ'.

In Fig. 6, we provide an example of a typical k-mer. Intuitively, we would like to choose the parameters D_h (horizontal error rate), D_v (vertical error rate), and m to generate a test that is guaranteed to work under the worst-case model given by (a) and (b). The main property that makes this notion of typicality powerful is that, depending on the choice of parameters and the approximate repeat statistics of $\mathbf{s}$ given by $V_{\mathbf{s}}(\cdot, \cdot)$, it has a no-false-positive guarantee. This is expressed in the following theorem.

Theorem 4: Given a set of reads $\mathcal{R}$, if $\mathbf{x}$ is a (D_h, D_v, m)-typical k-mer and $D_v V_{\mathbf{s}}(D_h, k) < m$, then $\mathbf{x} \in \mathcal{S}_k$.

Proof: Consider a (D_h, D_v, m)-typical k-mer $\mathbf{x}$ and the k-mers $\mathbf{x}_1, \ldots, \mathbf{x}_m$ satisfying the properties in Definition 1. Each of these k-mers (with erasures) must have originated from some k-mer in $\mathbf{s}$. Let $\mathcal{S} = \{\mathbf{s}^k_{t_1}, \ldots, \mathbf{s}^k_{t_M}\} \subset \mathcal{S}_k$ be the length-k segments in $\mathbf{s}$ from which at least one of $\mathbf{x}_1, \ldots, \mathbf{x}_m$ originated. Since $d_H(\mathbf{x}, \mathbf{x}_i) \le D_h$ for $i = 1, \ldots, m$, we have $d_H(\mathbf{x}, \mathbf{s}^k_j) \le D_h$ for $i = 1, \ldots, M$. This implies that $\rho(\{\mathbf{s}^k_{t_1}, \ldots, \mathbf{s}^k_{t_M}\}) \le D_h$ and hence $M \le V_{\mathbf{s}}(D_h, k)$. We will show that we must have $\mathbf{x} = \mathbf{s}^k_{t_i}$ for some $i \in \{1, \ldots, M\}$.

Suppose by contradiction that $\mathbf{x} \ne \mathbf{s}^k_{t_i}$ for $i = 1, \ldots, m$. Then each $\mathbf{s}^k_{t_i}$ must differ from $\mathbf{x}$ in at least one position. Now partition the k-mers $\mathcal{U} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ into $\mathcal{U}_1, \ldots, \mathcal{U}_M$ according to which length-k block $\mathbf{s}^k_{t_i}$ they originated from. All reads from $\mathcal{U}_i$ must have an erasure in the position where $\mathbf{x}$ and $\mathbf{s}^k_{t_i}$ differ. Since $\mathbf{x}$ is a (D_h, D_v, m)-typical k-mer, from the third property in Definition 1, we must have $|\mathcal{U}_i| \le D_v$. Since this holds for $i = 1, \ldots, M$, we have

$$|\mathcal{U}| = \sum_{i=1}^{M} |\mathcal{U}_i| \le M D_v \le D_v V_{\mathbf{s}}(D_h, k) < m,$$

which is a contradiction as $|\mathcal{U}| = m$. ∎

The requirement that $D_v V_{\mathbf{s}}(D_h, k) < m$ may be intuitively understood if we rewrite it as $D_v/m < 1/V_{\mathbf{s}}(D_h, k)$. For many real genomes, as shown in Table I, provided that $k \approx \ell_{\text{crit}}$, $V_{\mathbf{s}}(D_h, k)$ is a small number often no larger than 4. Since D_v/m can be thought of as the erasure rate, we are essentially requiring the erasure rate to be the reciprocal of the maximum number of approximate repeats. We say that Theorem 4 implies a no-false-positive property because of the following direct consequence.

Corollary 1: If we have $D_v V_{\mathbf{s}}(D_h, k) < m$, then the k-spectrum assembler given by

$$\widehat{\mathcal{S}}_k(D_h, D_v, m) \triangleq \left\{\mathbf{x} \in \Sigma^k : \mathbf{x} \text{ is } (D_h, D_v, m)\text{-typical}\right\} \tag{8}$$

satisfies $\widehat{\mathcal{S}}_k(D_h, D_v, n) \subset \mathcal{S}_k$.

Furthermore, we point out that Theorem 4 and Corollary 1 are completely independent of the erasure model used. The only constraint, namely that $D_v V_{\mathbf{s}}(D_h, k) < m$, is just a function of the test parameters and of the sequence repeat statistics. Hence, this result still holds in a probabilistic erasure model, for instance. This fact will be explored later on, in Section IV.

Notice that through this typicality approach we cannot characterize the multiplicity of each element in the multiset $\mathcal{S}_k$ (i.e., $\widehat{\mathcal{S}}_k$ is a set, not a multiset). Hence, we can only hope for $\widehat{\mathcal{S}}_k = \text{supp}(\mathcal{S}_k)$.

Theorem 5: For any sequence $\mathbf{s} \in \mathcal{G}(p, L, W)$ and set of reads $\mathcal{R}$ satisfying the adversarial model in Section II-A,

$$\bigcup_{m \ge 1} \widehat{\mathcal{S}}_k(pL, pm, m) = \text{supp}(\mathcal{S}_k). \tag{9}$$

for any $k \le L - W + 1$.

Proof: First we notice that $\mathbf{s} \in \mathcal{G}(p, L, W)$ implies $pm \cdot V_{\mathbf{s}}(pL, k) < m$ for every $m \ge 1$. Hence, from Corollary 1, $\widehat{\mathcal{S}}_k(pL, pm, m) \subset \text{supp}(\mathcal{S}_k)$ for $m \ge 1$. Conversely, consider an arbitrary k-mer $\mathbf{s}^k_i$ from the sequence $\mathbf{s}$. Now if we set $\tau = i - W + 1$, we will have $t - L < \tau \le t - W + 1$ for every $t \in [i : i + k - 1]$. Hence if we let $m = |\mathcal{R}^W_\tau|$, among the reads in $\mathcal{R}^W_\tau$ we have at most pm erasures for each base in $\mathbf{s}^k_i$ (see Fig. 7), implying that $\mathbf{s}^k_i$ is (pL, pm, m)-typical for $m \ge 1$, and $\mathbf{s}^k_i \in \widehat{\mathcal{S}}_k(pL, pm, m)$. We conclude that $\text{supp}(\mathcal{S}_k) \subset \cup_{m \ge 1} \widehat{\mathcal{S}}_k(pL, pm, m)$. ∎

Theorem 3 now follows straightforwardly.

Proof of Theorem 3: The lower bound (achievability) to $k^*(p, W, \mathcal{G})$ follows from Theorem 5. Since for every sequence $\mathbf{s} \in \mathcal{G}(p, L, W)$ and $\mathcal{R} \in \mathbf{R}$, (9) provides a way to unambiguously reconstruct $\text{supp}(\mathcal{S}_k)$ for $k \le L - W + 1$, we must have $L - W + 1 \le k^\star(p, W, \mathcal{G})$. By noticing that the sequence used to derive the upper bound $k^\star(p, W, \mathcal{G}) \le L - W + 1$ (see Appendix B) is in $\mathcal{G}(p, L, W)$ when $p < 1/2$, Theorem 3 follows. ∎

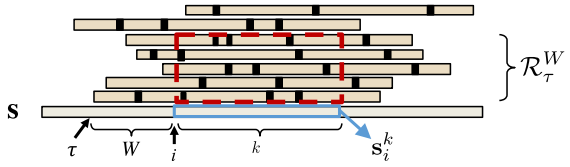


Fig. 7. Due to the k-mers in the dashed rectangle, the k-mer $\mathbf{s}_i^k$ must be (pL, pm, m)-typical for $m = |\mathcal{R}_\tau^W|$.

Intuitively, reducing W corresponds to making the adversarial model in (a) and (b) closer to a probabilistic i.i.d. model. Hence, Theorem 5 suggests that in a probabilistic model one should be able to reconstruct $\text{supp}(\mathcal{S}_{L-W+1})$ where $L-W+1 \approx L$, as long as $p < 1/V_{\mathbf{s}}(pL, L-W+1)$. Since we expect $V_{\mathbf{s}}(pL, L-W+1)$ to be a small number for real genomes, this result has a similar message to the one in [21]: up to a certain value of the error rate p, a typicality-type test can convert noisy reads into noiseless reads that are almost as long as the original reads.

IV. Analysis Under a Probabilistic Model

While the techniques in Section III were introduced with the goal of characterizing the maximum reconstructible support of the spectrum in the worst-case setting described in Section II-B, it can be seen as a general test to identify true k-mers from a set of noisy reads $\mathcal{R}$ under any erasure model. In particular, an important observation is that the no-false-positive property stated in Corollary 1 is independent of the erasure model described in Section II-A and holds under any arbitrary erasure model. When the conditions (a) and (b) in Section II-A are not satisfied everywhere, the spectrum reconstruction in (9) may fail to contain the entire support but will still satisfy $\cup_{m\geq 1}\widehat{\mathcal{S}}_k(pL, pm, m) \subset \text{supp}(\mathcal{S}_k(\mathbf{s}))$. Hence it can be seen as a technique to generate an (error-free) subset of the k-spectrum. It is then natural to consider the performance of the approach introduced in Section III when applied to a probabilistic setting where both the read locations and the erasures are random.

For concreteness, let us consider the standard model where N reads are sampled independently and uniformly at random from the genome. Suppose the read errors are erasures that occur independently with probability p. Due to Theorem 4, and motivated by the spectrum assembler of Corollary 1, a natural approach is to attempt to reconstruct the k-spectrum by considering

$$\bigcup_{m\geq 1} \widehat{\mathcal{S}}_k(qL, qm, m) \subset \text{supp}(\mathcal{S}_k(\mathbf{s})). \tag{10}$$

The parameter q can be thought of as a fraction $q > p$ that we do not expect the rate of erasures to exceed. As long as $q \cdot V_{\mathbf{s}}(qk, k) < 1$ (which is independent of the error model), the above spectrum reconstruction approach is guaranteed by Corollary 1 to generate a subset of $\text{supp}(\mathcal{S}_k)$. In fact, one can do better by noticing that, for $k' > k$, $V_{\mathbf{s}}(qk, k) \leq V_{\mathbf{s}}(qk', k')$, and $q \cdot V_{\mathbf{s}}(qk', k') < 1$. Therefore, one can consider the set

$$\widetilde{\mathcal{R}}_{k,q} \triangleq \bigcup_{k'\geq k} \bigcup_{m\geq 1} \widehat{\mathcal{S}}_k\big(qk', qm, m\big). \tag{11}$$

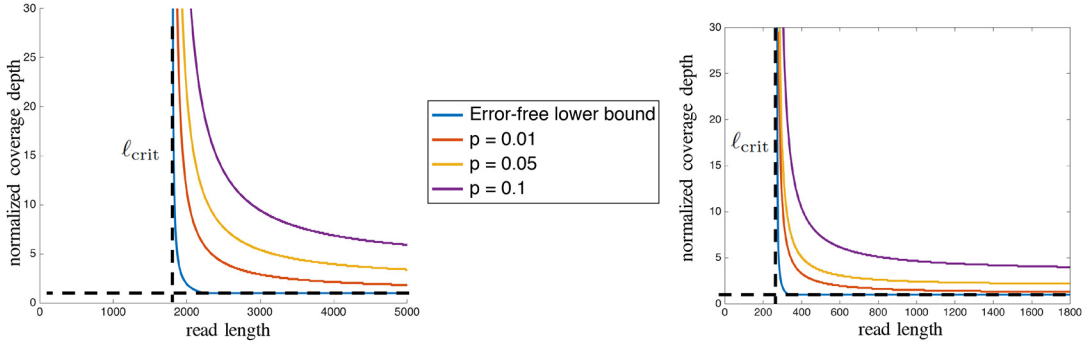
of cleaned up k'-mers for different values of $k' \geq k$. As our goal is to assemble $\mathbf{s}$, we can view the (random) set $\widetilde{\mathcal{R}}_{k,q}$ as generating a set of error-free reads of variable lengths. The problem of genome assembly from variable-length error-free reads was studied in [18] and [25]. Sufficient conditions for these algorithms to succeed were derived in [18] in terms of "bridging" of repeats. Given the probabilistic model considered, one can compute the probability that $\widetilde{\mathcal{R}}_{k,q}$ satisfies these conditions, which would in turn guarantee that perfect assembly can be achieved. By repeating this process for different values of the erasure probability p, we obtain the curves in Fig. 3.

We notice that for both genomes considered, these sufficiency curves are not very far from the error-free lower bound, particularly in terms of the read length required. The coverage depth requirement for this scheme, however, is larger than the Lander-Waterman coverage that is sufficient in the error-free case. Part of the reason for this discrepancy is our strict objective of perfect assembly. This requires the set of error-corrected reads $\widetilde{\mathcal{R}}_{k,q}$ to cover the entire genome, which in turn requires the original set of reads $\mathcal{R}$ to cover the entire genome multiple times. Alternatively, one can consider relaxing this requirement, and allowing the final assembly to contain erasures at a rate no larger then p. One can then modify the scheme to utilize, in addition to the error-free reads in $\widetilde{\mathcal{R}}_{k,q}$, the original reads in $\mathcal{R}$ that did not contribute to any of the typical k'-mers included in $\widetilde{\mathcal{R}}_{k,q}$ to "fill in" the gaps of coverage. Such an approach, although not technically achieving the original goal of perfect assembly, has milder coverage depth requirements, as shown in Fig. 8. In fact, one can show that the sufficiency curves obtained under this modified setting have a horizontal asymptote at 1 (i.e., the Lander-Waterman coverage) matching the lower bound.

V. Practical Considerations

The typicality-based error-correction scheme described in Section III is theoretical in nature, and its main goal is to allow us to study fundamental limits of error correction, illustrated by the curves in Fig. 8. As the error correction of sequencing data is a problem of significant practical importance, several remarks on the limitations of the model and the connections with practical approaches are in order.

Real sequencing platforms generate reads that are susceptible to substitution errors, insertions and deletions. Therefore, the erasure model considered in this paper is restrictive and should be understood as a first step towards considering the more challenging case of general error models. We point out that the typicality-based nature of the test considered makes it straightforward to generalize our error correction scheme. In particular, Definition 1 would regard a k-mer $\mathbf{x}$ as (D_h, D_v, m)-typical if we can find segments $\mathbf{x}_1, \ldots, \mathbf{x}_m$ (not necessarily of length k) such that $d_E(\mathbf{x}, \mathbf{x}_i) \leq D_h$, where $d_E(\cdot,\cdot)$ refers to the edit distance, and these segments can be *aligned* to $\mathbf{x}$ so that each base in $\mathbf{x}$ is supported by at least $m - D_v$. However, besides the computational obstacles imposed by such a definition, the no-false-positive property of 1 does not hold under this generalization, and new techniques must be developed



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

Fig. 8. Sufficiency curves when erasures at a rate p are allowed on the assembled sequence. Reads are assumed to be sampled independently and uniformly at random, and erasures occur independently with probability p.

in order to carry out an analysis similar to the one in this paper.

The error correction scheme proposed in Section III does not lend itself easily to a computationally tractable approach. In fact, a naive implementation would require one to consider all 4^k possible k-mers and, for each one, attempt to align reads to it to verify the typicality conditions, leading to a $O(4^k N(L-k)k)$ running time. However, the idea of considering one k-mer at a time is clearly not practical and one should use the k-mers in the reads themselves as potential typical k-mers. Notice that if there are m k-mers $\mathbf{x}_1, \ldots, \mathbf{x}_m$ satisfying $d_H(\mathbf{x}, \mathbf{x}_i) \leq D_h$ for $i = 1, \ldots, m$, it follows that $d_H(\mathbf{x}_i, \mathbf{x}_j) \leq 2D_h$ for $i, j \in \{1, \ldots, m\}$. Therefore, a practical proxy for identifying typical k-mers is to first cluster k-mers extracted from the reads to identify sets of k-mers with pairwise distance at most $2D_h$ and then only test the typicality of k-mers that are well supported by the cluster (or near the cluster center). The computational bottleneck to such an approach would lie in the identification of pairs of k-mers within a fixed distance of each other. This would in principle require a $O(N^2(L-k)^2k)$ running time, which is still impractical, but hashing-based techniques (such as the locality-sensitive hash used in [15]) can be used to speed up this process at the expense of a small loss in accuracy. We also point out that, when viewed in this way, the approach proposed in this paper resembles cluster-based error correction schemes such as those studied in [7] and [8], except that it uses the worst-case typicality test in Section III instead of a test based on a statistical model for the sequence and the error process to infer the set of cleaned-up k-mers.

VI. Conclusion

In this work, we investigate the problem of genome assembly without making probabilistic assumptions about the underlying sequence and the noise process. Adopting an adversarial erasure model, we propose a typicality-based algorithm for correcting read errors which clusters k-mers with a "worst-case typical" erasure pattern and then uses these clusters to infer error-free reads of (potentially) shorter length. Under the worst-case formulation considered, this approach is proved to be optimal in the sense that it recovers the maximum reconstructible support of the spectrum for the target sequence.

By leveraging our worst-case analysis for fixed maximum erasure rates, we compute conditions on read length and coverage depth that are sufficient for perfect assembly of several real genomes under a stochastic erasure model. When evaluated for real genomes, we observe that the information-theoretic requirements for perfect assembly do not vary significantly due to the introduction of erasures. More specifically, the critical read length required for perfect assembly in the presence of erasures, a parameter dictated by available technology, is approximately the same as that required for perfect assembly from error-free reads.

While this paper focused on reads with erasures, a direction for future work is to extend the techniques to the more general case of substitution errors and indels. We point out that the typicality-based nature of the test considered makes it straightforward to generalize our error correction scheme. However, when errors are no longer erasures, the no-false-positive property of Corollary 1 does not hold, and thus the error correction scheme can produce false reads. As a result, a performance analysis of standard assembly algorithms such as [18] and [23], which are designed for error-free reads, on the resulting set of "almost" error-corrected reads is highly non-trivial. Nevertheless, we conjecture that similar results will hold for more general error modes.

Appendix A
From Spectrum Reconstruction to Perfect Assembly

In this section, we describe in detail the critical lengths $\ell_{\text{crit}}(\mathbf{s})$ and $\bar{\ell}_{\text{crit}}(\mathbf{s})$, which guarantee that, from $\mathcal{S}_k$ and $\text{supp}(\mathcal{S}_k)$ respectively, one can reconstruct the complete sequence $\mathbf{s}$.

A *repeat* of length ℓ in $\mathbf{s}$ is a subsequence appearing twice at some positions t_1 and t_2 (so $\mathbf{s}_{t_1}^{\ell} = \mathbf{s}_{t_2}^{\ell}$) 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 $a_1 < b_1 < a_2 < b_2$. 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 $\mathbf{s}$.

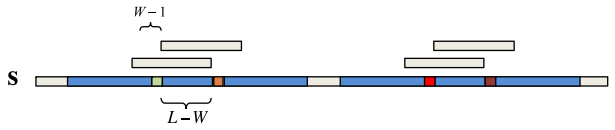


Fig. 9. Sequence $\mathbf{s}$ for which $\text{supp}(S_{L-W+2})$ cannot be reconstructed unambiguously.

A *triple repeat* of length ℓ in $\mathbf{s}$ is a subsequence appearing three times at some positions $t_1 < t_2 < t_3$ (so $\mathbf{s}_{t_1}^{\ell} = \mathbf{s}_{t_2}^{\ell} = \mathbf{s}_{t_3}^{\ell}$) that is maximal; i.e., $s[t_1 - 1]$, $s[t_2 - 1]$ and $s[t_3 - 1]$ are not all equal and $s[t_1 + \ell]$, $s[t_2 + \ell]$ and $s[t_3 + \ell]$ are not all equal. Notice that, for the circular DNA model we consider, we can define three segments $A = \mathbf{s}_{t_1}^{t_2 - t_1}$, $B = \mathbf{s}_{t_2}^{t_3 - t_2}$ and $C = \mathbf{s}_{t_3}^{G + t_3 - t_1}$. We will say that a triple repeat is transpose-invariant if A, B and C are not all distinct. The reason for this terminology is that, if $A = B$ for instance, the circular sequences defined by ABC and ACB are the same. We will let $\ell_{\text{triple}}(\mathbf{s})$ be the length of the longest triple repeat in $\mathbf{s}$ that is not transpose-invariant, and $\bar{\ell}_{\text{triple}}(\mathbf{s})$ be the length of the longest triple repeat in $\mathbf{s}$. Clearly, $\bar{\ell}_{\text{triple}}(\mathbf{s}) \geq \ell_{\text{triple}}(\mathbf{s})$.

Finally, we define the two critical read lengths as

$$\ell_{\text{crit}}(\mathbf{s}) = \max\left[\ell_{\text{inter}}(\mathbf{s}), \ell_{\text{triple}}(\mathbf{s})\right]$$
$$\bar{\ell}_{\text{crit}}(\mathbf{s}) = \max\left[\ell_{\text{inter}}(\mathbf{s}), \bar{\ell}_{\text{triple}}(\mathbf{s})\right].$$

The first critical read length provides the guarantee for when $\mathbf{s}$ can be unambiguously assembled from the complete k-spectrum $\mathcal{S}_k$. More precisely, results in [19] and [20] imply the following:

Theorem 6: If $k > \ell_{\text{crit}}(\mathbf{s})$, then $\mathbf{s}$ is the unique sequence with k-spectrum $\mathcal{S}_k(\mathbf{s})$. Conversely, if $k \leq \ell_{\text{crit}}(\mathbf{s})$, there exists a sequence $\mathbf{s}' \neq \mathbf{s}$ for which $\mathcal{S}_k(\mathbf{s}) = \mathcal{S}_k(\mathbf{s}')$.

Similarly, the second critical read length provides a guarantee of reconstruction of $\mathbf{s}$ when we have the support of the k-spectrum.

Theorem 7 [18]: If $k > \bar{\ell}_{\text{crit}}(\mathbf{s})$, the Multibridging algorithm correctly assembles $\mathbf{s}$ from $\text{supp}(\mathcal{S}_k)$.

We point out that the distinction between these two notions of the critical read length is often not very explicit in the literature.

APPENDIX B
SIMPLE BOUNDS ON $k^*(p, W)$

A simple lower bound for $k^\star(p, W)$ can be obtained by noticing that if we look at $\lceil 1/p \rceil - 1$ consecutive positions in $\mathbf{s}$ and $\lceil 1/p \rceil - 1 \leq L - W + 1$, (b) guarantees that there will be one read where none of these positions is erased. To see this, consider some segment $\mathbf{s}_i^{\lceil 1/p \rceil - 1}$ and the length-W window to its left, and suppose by contradiction that all reads (say m) starting in this window have at least one erasure in $\mathbf{s}_i^{\lceil 1/p \rceil - 1}$. By the pigeonhole principle, at least one of the symbols in $\mathbf{s}_i^{\lceil 1/p \rceil - 1}$ is erased at least $m/(\lceil 1/p \rceil - 1) > pm$ times, which contradicts (b). Hence, we always have $\mathcal{R} \Rightarrow \text{supp}(\mathcal{S}_{\lceil 1/p \rceil - 1})$, and

$$\min(L - W + 1, \lceil 1/p \rceil - 1) \leq k^*(p, W).$$

On the other hand, the example in Fig. 9 shows that $L - W + 1$ is an upper bound.

In the sequence $\mathbf{s}$ in Fig. 9, there are two segments (in blue) which are identical except at two locations. If the gap between these two locations has length $L - W$, it is not difficult to see that in order to unambiguously reconstruct $\text{supp}(\mathcal{S}_{L-W+2})$, we would need at least one read that covers both of the distinguishing bases. But this is equivalent to requiring a read to have a starting position in a window of length $W - 1$. Since in our model we just require that $\mathcal{R}_\tau^W \neq \emptyset$ it is possible that no read will cover both of these positions. This implies that

$$k^*(p, W) \leq L - W + 1.$$

We point out that while the example above may seem contrived at first, long nearly-exact repeats that only differ in a few spread out positions are common genomic patterns, and indeed represent a bottleneck for error correction in practice.

ACKNOWLEDGEMENT

Part of this work was completed while the authors were visiting the Simons Institute for the Theory of Computing, UC, Berkeley, USA.

REFERENCES

[1] S. L. Salzberg, "Mind the gaps," *Nat. Methods*, vol. 7, no. 2, pp. 105–106, 2010.
[2] P. A. Pevzner, H. Tang, and M. S. Waterman, "An Eulerian path approach to DNA fragment assembly," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 17, pp. 9748–9753, 2001.
[3] Y. Peng, H. C. M. Leung, S.-M. Yiu, and F. Y. L. Chin, "IDBA-A practical iterative de Bruijn graph de novo assembler." in *Proc. Annu. Int. Conf. Res. Comput. Mol. Biol.*, Lisbon, Portugal, 2010, pp. 426–440.
[4] D. R. Zerbino and E. Birney, "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs," *Genome Res.*, vol. 18, no. 5, pp. 821–829, 2008.
[5] D. R. Kelley, M. C. Schatz, and S. L. Salzberg, "Quake: Quality-aware detection and correction of sequencing errors," *Genome Biol.*, vol. 11, no. 11, p. 1, 2010.
[6] M. R. Crusoe *et al.*, "The Khmer software package: Enabling efficient nucleotide sequence analysis," *F1000Research*, vol. 4, p. 900, Sep. 2015.
[7] P. Medvedev, E. Scott, B. Kakaradov, and P. Pevzner, "Error correction of high-throughput sequencing datasets with non-uniform coverage," *Bioinformatics*, vol. 27, no. 13, pp. i137–i141, 2011.
[8] S. I. Nikolenko, A. I. Korobeynikov, and M. A. Alekseyev, "BayesHammer: Bayesian clustering for error correction in single-cell sequencing," *BMC Genomics*, vol. 14, no. 1, p. 1, 2013.
[9] L. Ilie, F. Fazayeli, and S. Ilie, "HiTEC: Accurate error correction in high-throughput sequencing data," *Bioinformatics*, vol. 27, no. 3, pp. 295–302, 2011.
[10] S. Koren *et al.*, "Hybrid error correction and de novo assembly of single-molecule sequencing reads," *Nat. Biotechnol.*, vol. 30, no. 7, pp. 693–700, 2012.
[11] Y. Heo, X.-L. Wu, D. Chen, J. Ma, and W.-M. Hwu, "BLESS: Bloom filter-based error correction solution for high-throughput sequencing reads," *Bioinformatics*, vol. 30, no. 10, pp. 1354–1362, 2014.
[12] S. Koren *et al.*, "Reducing assembly complexity of microbial genomes with single-molecule sequencing," *Genome Biol.*, vol. 14, no. 9, p. 1, 2013.
[13] S. Koren and A. M. Phillippy, "One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly," *Current Opin. Microbiol.*, vol. 23, pp. 110–120, Feb. 2015.
[14] C.-S. Chin *et al.*, "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data," *Nat. Methods*, vol. 10, no. 6, pp. 563–569, 2013.
[15] K. Berlin *et al.*, "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing," *Nat. Biotechnol.*, vol. 33, no. 6, pp. 623–630, 2015.

[16] G. M. Kamath, I. Shomorony, F. Xia, T. A. Courtade, and D. Tse, "Hinge: Long-read assembly achieves optimal repeat resolution," *bioRxiv*, Aug. 2016, Art. no. 062117.
[17] A. S. Motahari, G. Bresler, and D. Tse, "Information theory of DNA shotgun sequencing," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6273–6289, Oct. 2013.
[18] G. Bresler, M. Bresler, and D. Tse, "Optimal assembly for high throughput shotgun sequencing," *BMC Bioinformat.*, vol. 14, no. 5, Apr. 2013.
[19] E. Ukkonen, "Approximate string-matching with q-grams and maximal matches," *Theor. Comput. Sci.*, vol. 92, no. 1, pp. 191–211, 1992.
[20] P. A. Pevzner, "DNA physical mapping and alternating Eulerian cycles in colored graphs," *Algorithmica*, vol. 13, no. 1, pp. 77–105, 1995.
[21] A. Motahari, K. Ramchandran, D. Tse, and N. Ma, "Optimal DNA shotgun sequencing: Noisy reads are as good as noiseless reads," in *Proc. IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, 2013, pp. 1640–1644.
[22] K.-K. Lam, A. Khalak, and D. Tse, "Near-optimal assembly for shotgun sequencing with noisy reads," *BMC Bioinformat.*, vol. 15, no. 9, p. S4, 2014.
[23] I. Shomorony, S. H. Kim, T. A. Courtade, and D. Tse, "Information-optimal genome assembly via sparse read-overlap graphs," *Bioinformatics*, vol. 37, no. 17, pp. i494–i502, 2016.
[24] I. Shomorony, T. Courtade, and D. Tse, "Do read errors matter for genome assembly?" in *Proc. ISIT*, Hong Kong, 2015, pp. 919–923.
[25] J. Hui, I. Shomorony, K. Ramchandran, and T. A. Courtade, "Overlap-based genome assembly from variable-length reads," in *Proc. ISIT*, Barcelona, Spain, 2016, pp. 1018–1022.

Ilan Shomorony received the B.S. degree in mathematics and ECE from the Worcester Polytechnic Institute, Worcester, MA, USA, in 2009, and the Ph.D. degree in ECE from Cornell University, Ithaca, NY, USA, in 2014. He is currently a Post-Doctoral Fellow with the NSF Center for Science of Information, University of California, Berkeley, CA, USA. His current research interests are in the area of computational biology, particularly in the error correction and assembly of genomic data. He was a recipient of the Olin Fellowship from the School of Electrical and Computer Engineering at Cornell in 2009, the Qualcomm Innovation Fellowship in 2013, and the Simons Research Fellowship in 2014.

Thomas A. Courtade (S'06–M'13) received the B.Sc. (*summa cum laude*) degree in electrical engineering from Michigan Technological University in 2007, and the M.S. and Ph.D. degrees from University of California, Los Angeles, in 2008 and 2012, respectively. In 2014, he was a Post-Doctoral Fellow supported by the NSF Center for Science of Information. He is an Assistant Professor with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. He was a recipient of the Distinguished Ph.D. Dissertation Award, the Excellence in Teaching Award from the UCLA Department of Electrical Engineering, and the Jack Keil Wolf Student Paper Award for the 2012 International Symposium on Information Theory.

David Tse (F'09) received the B.A.Sc. degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 1989, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1991 and 1994, respectively. From 1994 to 1995, he was a Post-Doctoral Technical Staff Member with AT&T Bell Laboratories. From 1995 to 2014, he was on the faculty of the University of California at Berkeley. He is currently a Professor with Stanford University, Stanford, CA, USA. He has co-authored, with Pramod Viswanath, the book entitled *Fundamentals of Wireless Communication* (Cambridge, U.K.: Cambridge Univ. Press, 2005), which has been used in over 60 institutions around the world. He was a recipient of the 1967 NSERC Graduate Fellowship from the Government of Canada in 1989, the NSF CAREER Award in 1998, the Best Paper Awards at the Infocom 1998 and Infocom 2001 conferences, the Erlang Prize in 2000 from the INFORMS Applied Probability Society, the IEEE Communications and Information Theory Society Joint Paper Awards in 2001 and 2013, the Information Theory Society Paper Award in 2003, the 2009 Frederick Emmons Terman Award from the American Society for Engineering Education, the Gilbreth Lectureship from the National Academy of Engineering in 2012, the Signal Processing Society Best Paper Award in 2012, and the Stephen O. Rice Paper Award in 2013.