Information Theory of DNA Shotgun Sequencing

Abolfazl S. Motahari, Guy Bresler, *Student Member, IEEE*, and David N. C. Tse, *Fellow, IEEE*

***Abstract*—DNA sequencing is the basic workhorse of modern day biology and medicine. Shotgun sequencing is the dominant technique used: many randomly located short fragments called reads are extracted from the DNA sequence, and these reads are assembled to reconstruct the original sequence. A basic question is: given a sequencing technology and the statistics of the DNA sequence, what is the minimum number of reads required for reliable reconstruction? This number provides a fundamental limit to the performance of *any* assembly algorithm. For a simple statistical model of the DNA sequence and the read process, we show that the answer admits a critical phenomenon in the asymptotic limit of long DNA sequences: if the read length is below a threshold, reconstruction is impossible no matter how many reads are observed, and if the read length is above the threshold, having enough reads to cover the DNA sequence is sufficient to reconstruct. The threshold is computed in terms of the Renyi entropy rate of the DNA sequence. We also study the impact of noise in the read process on the performance.**

***Index Terms*—DNA sequencing, de novo assembly, information theory.**

I. Introduction

A. Background and Motivation

DNA sequencing is the basic workhorse of modern day biology and medicine. Since the sequencing of the Human Reference Genome ten years ago, there has been an explosive advance in sequencing technology, resulting in several orders of magnitude increase in throughput and decrease in cost. This advance allows the generation of a massive amount of data, enabling the exploration of a diverse set of questions in biology and medicine that were beyond reach even several years ago. These questions include discovering genetic variations across different humans (such as single-nucleotide polymorphisms), identifying genes affected by mutation in cancer tissue genomes, sequencing an individual's genome for diagnosis (personal genomics), and understanding DNA regulation in different body tissues.

Manuscript received May 21, 2012; revised May 24, 2013; accepted June 04, 2013. Date of publication June 20, 2013; date of current version September 11, 2013. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and in part by the Center for Science of Information, an NSF Science and Technology Center, under Grant CCF-0939370. This work was done at the University of California, Berkeley. This paper was presented in part at the 2012 IEEE International Symposium of Information Theory.

A. S. Motahari and D. N. C. Tse are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94704 USA (e-mail: motahari@eecs.berkeley.edu; dtse@eecs.berkeley.edu).

G. Bresler is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: gbresler@mit.edu).

Communicated by O. Milenkovic, Associate Editor for Coding Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2013.2270273

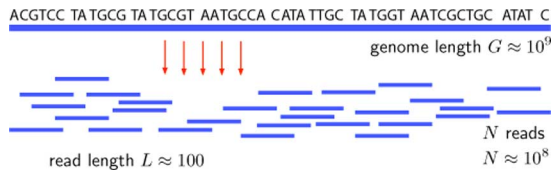

Fig. 1. Schematic for shotgun sequencing.

Shotgun sequencing is the dominant method currently used to sequence long strands of DNA, including entire genomes. The basic shotgun DNA sequencing setup is shown in Fig. 1. Starting with a DNA molecule, the goal is to obtain the sequence of nucleotides (A, C, G, or T) comprising it. (For humans, the DNA sequence has about 3×10^9 nucleotides, or base pairs.) The sequencing machine extracts a large number of reads from the DNA; each read is a randomly located fragment of the DNA sequence, of lengths of the order of 100–1000 base pairs, depending on the sequencing technology. The number of reads can be of the order of 10s of millions to billions. The *DNA assembly problem* is to reconstruct the DNA sequence from the many reads.

When the human genome was sequenced in 2001, there was only one sequencing technology, the Sanger platform [24]. Since 2005, there has been a proliferation of "next generation" platforms, including Roche/454, Life Technologies SOLiD, Illumina Hi-Seq 2000 and Pacific Biosciences RS. Compared to the Sanger platform, these technologies can provide massively parallel sequencing, producing far more reads per instrument run and at a lower cost, although the reads are shorter in lengths. Each of these technologies generates reads of different lengths and with different noise profiles. For example, the 454 machines have read lengths of about 400 base pairs, while the SOLiD machines have read lengths of about 100 base pairs. At the same time, there has been a proliferation of a large number of assembly algorithms, many tailored to specific sequencing technologies. (Recent survey articles [17], [19], [21] discuss no less than 20 such algorithms, and the Wikipedia entry on this topic listed 42 [31].)

The design of these algorithms is based primarily on *computational* considerations. The goal is to design efficient algorithms that can scale well with the large amount of sequencing data. Current algorithms are often tailored to particular machines and are designed based on heuristics and domain knowledge regarding the specific DNA being sequenced. This makes it difficult to compare different algorithms, not to mention the difficulty of defining what is meant by an "optimal" assembly algorithm for a given sequencing problem. One reason for the heuristic approach taken toward the problem is that various formulations of the assembly problem are known to be NP-hard (see for example [12]).

An alternative to the computational view is the *information-theoretic* view. In this view, the genome sequence is regarded as a random string to be estimated based on the read data. The basic question is: what is the minimum number of reads needed to reconstruct the DNA sequence with a given reliability? This minimum number can be used as a benchmark to compare different algorithms, and an optimal algorithm is one that achieves this minimum number. It can also provide an algorithm-independent basis for comparing different sequencing technologies and for designing new technologies.

This information-theoretic view falls in the realm of DNA sequencing theory [30]. A well-known lower bound on the number of reads needed can be obtained by a *coverage analysis*, an approach pioneered by Lander and Waterman [13]. This lower bound is the number of reads $N_{\rm cov}$ such that with a desired probability, say $1-\epsilon$, the randomly located reads cover the entire genome sequence. The number $N_{\rm cov}$ can be easily approximated:

$$N_{\rm cov}(\epsilon, G, L) \approx \frac{G}{L} \ln\left(\frac{G}{L\epsilon}\right),$$

where G and L are DNA and read length, respectively. While this is clearly a lower bound on the minimum number of reads needed, it is in general not tight: only requiring the reads to cover the entire genome sequence does not guarantee that consecutive reads can actually be stitched back together to recover the entire sequence. The ability to do that depends on other factors such as the repeat statistics of the DNA sequence and also the noise profile in the read process. Thus, characterizing the minimum number of reads required for reconstruction is, in general, an open question.

B. Main Contributions

In this paper, we make progress on this basic problem. We first focus on a very simple model:

1) The DNA sequence is modeled as an i.i.d. random process of length G with each symbol taking values according to a probability distribution $\mathbf{p}$ on the alphabet $\{A, C, G, T\}$.
2) Each read is of length L symbols and begins at a uniformly distributed location on the DNA sequence and the locations are independent from one read to another.
3) The read process is noiseless.

Fix an $\epsilon \in (0, 1/2)$ and let $N_{\rm min}(\epsilon, G, L)$ be the minimum number of reads required to reconstruct the DNA with probability at least $1-\epsilon$. We would like to know how $\frac{N_{\rm min}(\epsilon,G,L)}{N_{\rm cov}(\epsilon,G,L)}$ behaves in the asymptotic regime when G and L grow to infinity. It turns out that in this regime, the ratio depends on G and L through a normalized parameter:

$$\bar{L} := \frac{L}{\ln G}.$$

We define

$$c_{\rm min}(\bar{L}) = \lim_{G\to\infty, L=\bar{L}\ln G} \frac{N_{\rm min}(\epsilon, G, L)}{N_{\rm cov}(\epsilon, G, L)}.$$

Let $H_2(\mathbf{p})$ be the Renyi entropy of order 2, defined to be

$$H_2(\mathbf{p}) := -\ln \sum_i p_i^2. \tag{1}$$

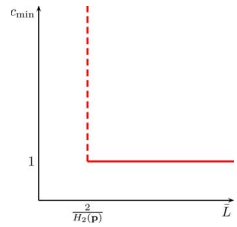

Fig. 2. Critical phenomenon.

Our main result, Theorem 1, yields a *critical phenomenon*: when $\bar{L}$ is below the threshold $2/H_2(\mathbf{p})$, reconstruction is impossible, i.e., $c_{\rm min}(\bar{L}) = \infty$, but when $\bar{L}$ is above that threshold, the obvious necessary condition of coverage is also sufficient for reconstruction, i.e., $c_{\rm min}(\bar{L}) = 1$. A simple greedy algorithm is able to reconstruct using this minimum coverage depth. The significance of the threshold is that when $\bar{L} < 2/H_2(\mathbf{p})$, with high probability, there are many repeats of length L in the DNA sequence, while when $\bar{L} > 2/H_2(\mathbf{p})$, with high probability, there are no repeats of length L. Thus, another way to interpret the result is that $\bar{L} < 2/H_2(\mathbf{p})$ is a *repeat-limited* regime, while $\bar{L} > 2/H_2(\mathbf{p})$ is a *coverage-limited* regime. The result is summarized in Fig. 2.

A standard measure of data requirements in DNA sequencing projects is the *coverage depth* NL/G, which is the average number of reads covering each base pair. Thus, $N_{\rm cov}(\epsilon, G, L) \times L/G$ is the coverage depth required to cover the DNA sequence with probability $1-\epsilon$ (as predicted by Lander–Waterman), and $N_{\rm min}(\epsilon, G, L) \times L/G$ is the minimum coverage depth required to reconstruct the DNA sequence with probability $1-\epsilon$. The quantity $c_{\rm min}(\bar{L})$ can, therefore, be interpreted as the (asymptotic) normalized minimum coverage depth required to reconstruct the DNA sequence.

In a related work, Arratia *et al.* [2] showed that $\bar{L} > 2/H_2(\mathbf{p})$ is a necessary and sufficient condition for reconstruction of the i.i.d. DNA sequence if *all* length L subsequences of the DNA sequence are given as reads. This arises in a technology called *sequencing by hybridization*. Obviously, for the same read length L, having all length L subsequences provides more information than any number of reads from shotgun sequencing, where the reads are randomly sampled. Hence, it follows that $\bar{L} > 2/H_2(\mathbf{p})$ is also a necessary condition for shotgun sequencing. What our result says is that this condition together with coverage is sufficient for reconstruction asymptotically.

The basic model of i.i.d. DNA sequence and noiseless reads is very simplistic. We provide two extensions to our basic result: 1) Markov DNA statistics and 2) noisy reads. In the first case, we show that the same result as the i.i.d. case holds except that the Renyi entropy $H_2(\mathbf{p})$ is replaced by the Renyi entropy rate of the Markov process. In the second case, we analyze the performance of a modification of the greedy algorithm to deal with noisy reads and show that the effect of noise is to increase the threshold on the read length below which reconstruction is impossible.

Even with these extensions, our models still miss several important aspects of real DNA and read data. Perhaps the most important aspect is the presence of long repeats in the DNA sequences of many organisms, ranging from bacteria to humans. These long repeats are poorly captured by i.i.d. or even Markov

models due to their short-range correlation. Another aspect is the nonuniformity of the sampling of reads from the DNA sequences. At the end of this paper, we will discuss how our results can be used as a foundation to tackle these and other issues.

C. Related Work

Li [14] has also posed the question of minimum number of reads for the i.i.d. equiprobable DNA sequence model. He showed that if $\bar{L} > 4/\ln 2$, then the number of reads needed is $O\big((G/L)\ln G\big)$, i.e., a constant multiple of the number needed for coverage. Specializing our result to the equiprobable case, we see that reconstruction is possible with probability $1-\epsilon$ if and only if $\bar{L} > 1/\ln 2$ and the number of reads is $(G/L)\ln(G/L\epsilon)$. Not only is our characterization necessary and sufficient, we have a much weaker condition on the read length L, and we get the correct prelog constant on the number of reads needed. As will be seen later, many different algorithms have the same scaling behavior in the number of reads they need, but it is the prelog constant that distinguishes them.

A common formulation of DNA assembly is the shortest common superstring (SCS) problem. The SCS problem is the problem of finding the shortest string containing a set of strings, where in the DNA assembly context, the given strings are the reads and the superstring is the estimate of the original DNA sequence. While the general SCS problem with arbitrary instances is NP-hard [12], the greedy algorithm is known to achieve a constant factor approximation in the worst case (see, e.g., [8] and [26]). More related to our work, the greedy algorithm has been shown to be optimal for the SCS problem under certain probabilistic settings [7], [15]. Thus, the reader may have the impression that our results overlap with these previous works. However, there are significant differences.

First, at a basic problem formulation level, the SCS problem and the DNA sequence reconstruction problem are not equivalent: there is no guarantee that the SCS containing the given reads is the original DNA sequence. Indeed, it has already been observed in the assembly literature (see, e.g., [16]) that the SCS of the reads may be a significant compression of the original DNA sequence, especially when the latter has a lot of repeats, since finding the SCS tends to merge these repeats. For example, in the case of very short reads, the resulting SCS is definitely not the original DNA sequence. In contrast, we formulate the problem directly in terms of reconstructing the original sequence, and a lower bound on the required read length emerges as part of the result.

Second, even if we assume that the SCS containing the reads is the original DNA sequence, one cannot recover our result from either [7] or [15], for different reasons. The main result (Theorem 1) in [15] says that if one models the DNA sequence as an arbitrary sequence perturbed by mutating each symbol independently with probability p and the reads are arbitrarily located, the average length of the sequence output by the greedy algorithm is no more than a factor of $1+3\delta$ of the length of the SCS, provided that $p > 2\ln(GL)/(\delta L)$, i.e., $p > 2/(\delta\bar{L})$. However, since $p \le 1$, the condition on p in their theorem implies that $\delta \ge \frac{2}{\bar{L}}$. Thus, for a fixed $\bar{L}$, they actually only showed that the greedy algorithm is approximately optimal to within a factor of $1+6/\bar{L}$, and optimal only under the further condition that $\bar{L}\to\infty$. In contrast, our result shows that the greedy algorithm is optimal for *any* $\bar{L} > 2/H_2(\mathbf{p})$, albeit under a weaker model for the DNA sequence (i.i.d. or Markov) and read locations (uniform random).

Regarding [7], the probabilistic model they used does not capture the essence of the DNA sequencing problem. In their model, the given reads are all independently distributed and not from a single "mother" sequence, as in our model. In contrast, in our model, even though the original DNA sequence is assumed to be i.i.d., the reads will be highly *correlated*, since many of the reads will be physically overlapping. In fact, it follows from [7] that, given N reads and the read length L scaling like $\ln N$, the length of the SCS scales like $N\ln N$. On the other hand, in our model, the length of the reconstructed sequence would be proportional to N. Hence, the length of the SCS is much longer for the model studied in [7], a consequence of the reads being independent and therefore much harder to merge. So the two problems are completely different, although coincidentally the greedy algorithm is optimal for both problems.

D. Notation and Outline

A brief remark on notation is in order. Sets (and probabilistic events) are denoted by calligraphic type, e.g., $\mathcal{A}$, $\mathcal{B}$, $\mathcal{E}$, vectors by boldface, e.g., $\mathbf{s}$, $\mathbf{x}$, $\mathbf{y}$, and random variables by capital letters such as S, X, Y. Random vectors are denoted by capital boldface, such as $\mathbf{S}$, $\mathbf{X}$, $\mathbf{Y}$. The exceptions to these rules, for the sake of consistency with the literature, are the (nonrandom) parameters G, N, and L. The natural logarithm is denoted by $\ln(\cdot)$. Unless otherwise stated, all logarithms are natural and all entropies are in units of nats.

The rest of this paper is organized as follows. Section II-A gives the precise formulation of the problem. Section II-B explains why reconstruction is impossible for read length below the stated threshold. For read length above the threshold, an optimal algorithm is presented in Section II-C, where a heuristic argument is given to explain why it performs optimally. Sections III and IV describe extensions of our basic result to incorporate read noise and a more complex model for DNA statistics, respectively. Section V discusses future work. Appendixes contain the formal proofs of all the results in the paper.

II. I.I.D. DNA Model

This section states the main result of this paper, addressing the optimal assembly of i.i.d. DNA sequences. We first formulate the problem and state the result. Next, we compare the performance of the optimal algorithm with that of other existing algorithms. Finally, we discuss the computational complexity of the algorithm.

A. Formulation and Result

The DNA sequence $\mathbf{S} = S_1S_2\ldots S_G$ is modeled as an i.i.d. random process of length G with each symbol taking values according to a probability distribution $\mathbf{p} = (p_1, p_2, p_3, p_4)$ on the alphabet $\{A, C, G, T\}$. For notational convenience, we instead denote the letters by numerals, i.e., $S_i \in \{1, 2, 3, 4\}$. To avoid boundary effects, we assume that the DNA sequence is circular,

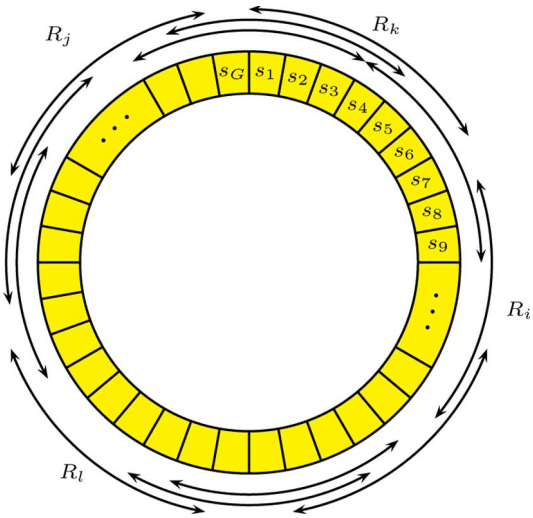

Fig. 3. Circular DNA sequence which is sampled randomly.

with $S_i = S_j$ if $i = j \bmod G$; this simplifies the exposition, and all results apply with appropriate minor modification to the noncircular case as well.

The objective of DNA sequencing is to reconstruct the whole sequence $\mathbf{s}$ based on N *reads* drawn randomly from the sequence (see Fig. 3). A read is a substring of length L from the DNA sequence. The set of reads is denoted by $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N\}$. The starting position of read i is denoted by T_i, so $R_i = \mathbf{S}[T_i, T_i + L - 1]$. The set of starting positions of the reads is denoted $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$. We assume that the starting position of each read is uniformly distributed on $\{1, \ldots, G\}$ and the positions are independent from one read to another.

An *assembly algorithm* takes a set of N reads $\mathcal{R} = \{\mathbf{R}_1, \ldots, \mathbf{R}_N\}$ and returns an estimated sequence $\hat{\mathbf{S}} = \hat{\mathbf{S}}(\mathcal{R})$. We require *perfect reconstruction*, which presumes that the algorithm makes an error if $\hat{\mathbf{S}} \neq \mathbf{S}$.[1] We let $\mathbb{P}$ denote the probability model for the (random) DNA sequence $\mathbf{S}$ and the sample positions $\mathcal{T}$, and $\mathcal{E} := \{\hat{\mathbf{S}} \neq \mathbf{S}\}$ the error event. A question of central interest is: what are the conditions on the read length L and the number of reads N such that the reconstruction error probability is less than a given target ϵ? Unfortunately, this is in general a difficult question to answer. We instead ask an easier *asymptotic* question: what is the ratio of the minimum number of reads $N_{\min}$ to the number of reads needed to cover the sequence N_{cov} as $L, G \to \infty$ with $\bar{L} = L/\ln G$ being a constant, and which algorithm achieves the optimal performance asymptotically? More specifically, we are interested in $c_{\min}(\bar{L})$, which is defined as

$$c_{\min}(\bar{L}) = \lim_{G\to\infty, L=\bar{L}\ln G} \frac{N_{\min}(\epsilon, G, L)}{N_{\text{cov}}(\epsilon, G, L)}. \tag{2}$$

The main result for this model is the following.

[1]The notion of perfect reconstruction can be thought of as a mathematical idealization of the notion of "finishing" a sequencing project as defined by the National Human Genome Research Institute [18], where finishing a chromosome requires at least 95% of the chromosome to be represented by a contiguous sequence.

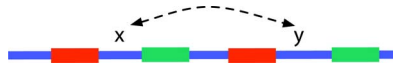

Fig. 4. Two pairs of interleaved repeats of length $L-1$ create ambiguity: from the reads, it is impossible to know whether the sequences $\mathbf{x}$ and $\mathbf{y}$ are as shown, or swapped.

Theorem 1: Fix an $\epsilon < 1/2$. The minimum normalized coverage depth $c_{\min}(\bar{L})$ is given by

$$c_{\min}(\bar{L}) = \begin{cases} \infty & \text{if } \bar{L} < 2/H_2(\mathbf{p}), \\ 1 & \text{if } \bar{L} > 2/H_2(\mathbf{p}), \end{cases} \tag{3}$$

where $H_2(\mathbf{p})$ is the Renyi entropy of order 2 defined in (1).

Section II-B proves the first part of the theorem that reconstruction is impossible for $\bar{L} < 2/H_2(\mathbf{p})$. Section II-C shows how a simple greedy algorithm can achieve optimality for $\bar{L} > 2/H_2(\mathbf{p})$.

B. $\bar{L} < \frac{2}{H_2(\mathbf{p})}$: Repeat-Limited Regime

The random nature of the DNA sequence gives rise to a variety of patterns. The key observation in [27] is that there are two patterns whose appearance in the DNA sequence precludes reconstruction from an arbitrary set of reads of length L. In other words, reconstruction is not possible even if the L-spectrum, the multiset of all substrings of length L appearing in the DNA sequence, is given. The first part of Theorem 1 is proved by showing that if $\bar{L} < 2/H_2(\mathbf{p})$, then one of the two patterns exists in the DNA sequence and reconstruction is impossible.

The first pattern is the three-wise repeat of a substring of length $L-1$. The second pattern is interleaved repeats of length $L-1$, depicted in Fig. 4. Arratia *et al.* [2] carried out a thorough analysis of randomly occurring repeats for the same i.i.d. DNA model as ours, and showed that the interleaved repeats pattern is the typical event causing reconstruction to be impossible. The following lemma is a consequence of [2, Th. 7] (see also [6]).

Lemma 2 (see [2]): Fix $\bar{L} < 2/H_2(\mathbf{p})$. An i.i.d. random DNA sequence contains interleaved repeats of length $L = \bar{L}\ln G$ with probability $1 - o(1)$.

We give a heuristic argument for the lemma, following [2], based on the expected number of repeats. Denoting by $\mathbf{S}_i^L$ the length-L subsequence starting at position i, we have

$$\mathbb{E}(\#\text{ of length } L \text{ repeats}) = \sum_{1\le i<j\le G} \mathbb{P}(\mathbf{S}_i^L = \mathbf{S}_j^L). \tag{4}$$

Now, the probability that two specific *physically disjoint* length-ℓ subsequences are identical is

$$\Big(\sum_i p_i^2\Big)^\ell = e^{-\ell H_2(\mathbf{p})},$$

where $H_2(\mathbf{p}) = -\ln\big(\sum_i p_i^2\big)$ is the Rényi entropy of order 2. Ignoring the GL terms in (4) in which $\mathbf{S}_i^L$ and $\mathbf{S}_j^L$ overlap gives the lower bound

$$\begin{aligned}\mathbb{E}(\#\text{ of repeats}) &> \left(\frac{G^2}{2} - GL\right) e^{-LH_2(\mathbf{p})} \\ &\approx \frac{G^2}{2}\cdot e^{-LH_2(\mathbf{p})}.\end{aligned} \tag{5}$$

Taking $G, L \to \infty$ with $L = \bar{L} \ln G$, this quantity approaches zero if $\bar{L} > 2/H_2(\mathbf{p})$, infinity if $\bar{L} < 2/H_2(\mathbf{p})$.

The heuristic step is to use the expected value as a surrogate for the actual number of repeats, allowing us to deduce that if $\bar{L} < 2/H_2(\mathbf{p})$, there are many repeats. Now, as shown in [2], the positions of the repeats are distributed uniformly on the DNA, so it is likely that some pair among the many pairs of repeats is interleaved. This is exactly the statement of Lemma 2.

Let us see why $\bar{L} = 2/H_2(\mathbf{p})$ is actually the correct threshold for the existence of interleaved repeats. First, as a consequence of Lemma 11 in Appendix A, we may safely neglect the terms in (4) due to the (significantly fewer) physically overlapping subsequences, implying that the right-hand side of (5) is a good estimate for the expected number of repeats. Next, as noted immediately after (5), if $\bar{L} > 2/H_2(\mathbf{p})$, then the right-hand side—and hence the expected number of repeats—vanishes asymptotically. But if there are no repeats, there can be no interleaved repeats.

We now prove the first part of Theorem 1, which states that if $\bar{L} < 2/H_2(\mathbf{p})$, then $c_{\min}(\bar{L}) = \infty$, i.e., reliable reconstruction is impossible.

Proof of Theorem 1, Part 1: Having observed a sequence of reads $\mathbf{R}_1, \ldots, \mathbf{R}_N$, the optimal guess for the DNA sequence is given by the maximum *a posteriori* (MAP) rule

$$\hat{\mathbf{S}} = \operatorname*{argmax}_{\mathbf{s}} \mathbb{P}(\mathbf{s}|\mathbf{R}_1, \ldots, \mathbf{R}_N). \tag{6}$$

To show the result, it thus suffices to argue that the MAP rule (6) is in error, $\hat{\mathbf{S}} \neq \mathbf{S}$, with probability at least $1/2$.

The probability of observing reads $\mathbf{r}_1, \ldots, \mathbf{r}_N$ given a DNA sequence $\mathbf{s}$ is

$$\begin{aligned}\mathbb{P}(\mathbf{r}_1, \ldots, \mathbf{r}_N|\mathbf{s}) &= \prod_{i=1}^{N} \mathbb{P}(\mathbf{r}_i|\mathbf{s}) \\ &= \prod_{i=1}^{N} \frac{\#\text{ of occurrences of } \mathbf{r}_i \text{ in } \mathbf{s}}{G}.\end{aligned}$$

Now suppose the DNA sequence $\mathbf{s}$ has interleaved repeats of length $L-1$ as in Fig. 4. If $\mathbf{s}'$ denotes the sequence obtained from $\mathbf{s}$ by swapping $\mathbf{x}$ and $\mathbf{y}$, then the number of occurrences of each read $\mathbf{r}_i$ in $\mathbf{s}$ and $\mathbf{s}'$ is the same, and hence

$$\mathbb{P}(\mathbf{r}_1, \ldots, \mathbf{r}_N|\mathbf{s}) = \mathbb{P}(\mathbf{r}_1, \ldots, \mathbf{r}_N|\mathbf{s}').$$

Moreover, $\mathbb{P}(\mathbf{s}) = \mathbb{P}(\mathbf{s}')$, so

$$\mathbb{P}(\mathbf{s}|\mathbf{r}_1, \ldots, \mathbf{r}_N) = \mathbb{P}(\mathbf{s}'|\mathbf{r}_1, \ldots, \mathbf{r}_N).$$

It follows that the MAP rule has probability of reconstruction error of at least $1/2$ conditional on the DNA sequence having interleaved repeats of length $L-1$, regardless of the number of reads. By Lemma 2, this latter event has probability approaching 1 as $G \to \infty$, for $\bar{L} < 2/H_2(\mathbf{p})$. Since $\epsilon < 1/2$, this implies that for sufficiently large G, $N_{\min}(\epsilon, G, L) = \infty$, proving the result.

Remark: Note that for any *fixed* read length L, the probability of the interleaved repeat event will approach 1 as the DNA length $G \to \infty$. This means that if we had defined the minimum normalized coverage depth for a fixed read length L, then for *any* value of L, the minimum normalized coverage depth would have been ∞. It follows that in order to get a meaningful result, one must scale L with G, and Lemma 2 suggests that letting L and G grow, while fixing $\bar{L}$ is the correct scaling.

C. $\bar{L} > \frac{2}{H_2(\mathbf{P})}$: Coverage-Limited Regime

In this section, we show that if $\bar{L} > 2/H_2(\mathbf{p})$, then as stated in Theorem 1, $c_{\min}(\bar{L}) = 1$. For this range of $\bar{L}$, reads are sufficiently long that repeats no longer pose a problem. The bottleneck, it turns out, is covering the sequence. We first review the coverage analysis of Lander and Waterman [13] and then show how a simple greedy algorithm can reconstruct reliably when the sequence is covered by the reads.

In order to reconstruct the DNA sequence, it is necessary for the reads to cover the DNA sequence (see Fig. 5): Clearly one must observe each of the nucleotides, but worse than the missing nucleotides, gaps in coverage create ambiguity in the order of the contiguous pieces. Thus, $N_{\text{cov}}(\epsilon, G, L)$, the minimum number of reads needed in order to cover the entire DNA sequence with probability $1-\epsilon$, is a lower bound to $N_{\min}(\epsilon, G, L)$, the minimum number of reads needed to reconstruct with probability $1-\epsilon$. The classical 1988 paper of Lander and Waterman [13] studied the coverage problem in the context of DNA sequencing, and from their results, one can deduce the following asymptotics for $N_{\text{cov}}(\epsilon, G, L)$.

Lemma 3 (see [13]): For any $\epsilon \in (0,1)$:

$$\lim_{L,G\to\infty, L/\ln G = \bar{L}} \frac{N_{\text{cov}}(\epsilon, G, L)}{G/\bar{L}} = 1.$$

Note that the lemma is consistent with the estimate $N_{\text{cov}} \approx (G/L)\ln(G/L\epsilon)$ (see Section I).

A standard coupon collector-style argument proves Lemma 3 in [13]. An intuitive justification of the lemma, which will be useful in the sequel, is as follows. To a very good approximation, the starting positions of the reads are given according to a Poisson process with rate $\lambda = N/G$, which means that each offset has an exponential (λ) distribution. It follows that the probability that there is a gap between two successive reads is approximately $e^{-\lambda L}$ and the expected number of gaps is approximately

$$Ne^{-\lambda L}.$$

Asymptotically, this quantity is bounded away from zero if $N < G/\bar{L}$ and approaches zero otherwise, in agreement with Lemma 3.

We now show that in the parameter regime under consideration, $\bar{L} > 2/H_2(\mathbf{p})$, a simple greedy algorithm (perhaps surprisingly) attains the coverage lower bound. The greedy algorithm merges the reads repeatedly into *contigs*,[2] and the merging is done greedily according to an overlap score defined on pairs of strings. For a given score, the algorithm is described as follows.

Greedy Algorithm: Input: the set of reads $\mathcal{R}$.

1) Initialize the set of contigs as the given reads.
2) Find two contigs with largest overlap score, breaking ties arbitrarily, and merge them into one contig.

[2]Here, a contig means a contiguous fragment formed by overlapping sequenced reads.

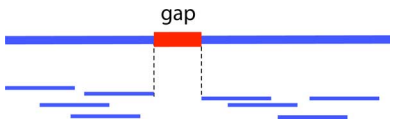

Fig. 5. Reads must cover the sequence.

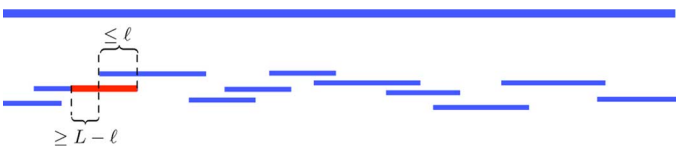

Fig. 6. Greedy algorithm merges reads into contigs according to the amount of overlap. At stage ℓ, the algorithm has already merged all reads with overlap greater than ℓ. The red segment denotes a read at the boundary of two contigs; the neighboring read must be offset by at least $L-\ell$.

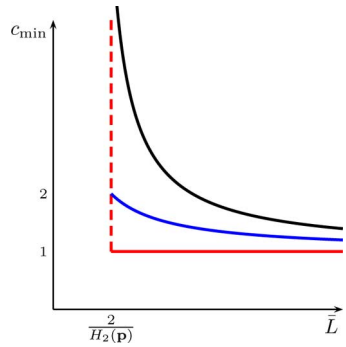

Fig. 7. Minimum normalized coverage depth obtained by the sequential algorithm is in the middle, given by $c_{\text{seq}} = \frac{\bar{L}H_2(\mathbf{p})}{\bar{L}H_2(\mathbf{p})-1}$; the minimum normalized coverage depth obtained by the K-mers-based algorithm is at top, given by $c_{K-\text{mer}} = \frac{\bar{L}H_2(\mathbf{p})}{\bar{L}H_2(\mathbf{p})-2}$.

3) Repeat Step 2 until only one contig remains.

For the i.i.d. DNA model and noiseless reads, we use the overlap score $W(\mathbf{s_1}, \mathbf{s_2})$ defined as the length of the longest suffix of $\mathbf{s_1}$ identical to a prefix of $\mathbf{s_2}$.

Showing optimality of the greedy algorithm entails showing that if the reads cover the DNA sequence and there are no repeats of length L, then the greedy algorithm can reconstruct the DNA sequence. In the remainder of this section, we heuristically explain the result, and we give a detailed proof in Appendix A.

Since the greedy algorithm merges reads according to overlap score, we may think of the algorithm as working in stages, starting with an overlap score of L down to an overlap score of 0. At stage ℓ, the merging is between contigs with overlap score ℓ. The key is to find the typical stage at which the *first* error in merging occurs. Assuming no errors have occurred in stages $L, L-1, \ldots, \ell+1$, consider the situation in stage ℓ, as depicted in Fig. 6. The algorithm has already merged the reads into a number of contigs. The boundary between two neighboring contigs is where the overlap between the neighboring reads is less than or equal to ℓ; if it were larger than ℓ, the two contigs would have been merged already. Hence, the expected number of contigs at stage ℓ is the expected number of pairs of successive reads with spacing greater than $L-\ell$. Again invoking the Poisson approximation, this is roughly equal to

$$Ne^{-\lambda(L-\ell)},$$

where $\lambda = N/G$.

Two contigs will be merged in error in stage ℓ if the length ℓ suffix of one contig equals the length ℓ prefix of another contig from a different location. Assuming these substrings are physically disjoint, the probability of this event is

$$e^{-\ell H_2(\mathbf{p})}.$$

Hence, the expected number of pairs of contigs for which this confusion event happens is approximately

$$\left[Ne^{-\lambda(L-\ell)}\right]^2 \cdot e^{-\ell H_2(\mathbf{p})}. \tag{7}$$

This number is largest either when $\ell = L$ or $\ell = 0$. This suggests that, typically, errors occur in stage L or stage 0 of the algorithm. Errors occur at stage L if there are repeats of length L substrings in the DNA sequence. Errors occur at stage 0 if there are still leftover unmerged contigs. The no-repeat condition ensures that the probability of the former event is small and the coverage condition ensures that the probability of the latter event is small. Hence, the two necessary conditions are also sufficient for reconstruction.

D. Performance of Existing Algorithms

The greedy algorithm was used by several of the most widely used genome assemblers for Sanger data, such as phrap, TIGR Assembler [25], and CAP3 [9]. More recent software aimed at assembling short-read sequencing data uses different algorithms. We will evaluate the normalized coverage depth of some of these algorithms on our basic statistical model and compare them to the information-theoretic limit. The goal is not to compare between different algorithms; that would have been unfair since they are mainly designed for more complex scenarios including noisy reads and repeats in the DNA sequence. Rather, the aim is to illustrate our information-theoretic framework and make some contact with the existing assembly algorithm literature.

1) Sequential Algorithm: By merging reads with the largest overlap first, the greedy algorithm discussed above effectively grows the contigs *in parallel*. An alternative greedy strategy, used by software like SSAKE [29], VCAKE [11], and SHARCGS [5], grows one contig *sequentially*. An unassembled read is chosen to start a contig, which is then repeatedly extended (say toward the right) by identifying reads that have the largest overlap with the contig until no more extension is possible. The algorithm succeeds if the final contig is the original DNA sequence.

The following proposition gives the normalized coverage depth of this algorithm.

Proposition 4: The minimum normalized coverage depth for the sequential algorithm is $c_{\text{seq}}(\bar{L}) = \frac{\bar{L}H_2(\mathbf{p})}{\bar{L}H_2(\mathbf{p})-1}$ if $\bar{L} > 2/H_2(\mathbf{p})$.

The result is plotted in Fig. 7. The performance is strictly worse than that of the greedy algorithm. We give only a heuristic argument for Proposition 4.

Motivated by the discussion in the previous section, we seek the typical overlap ℓ at which the first error occurs in merging

a read; unlike the greedy algorithm, where this overlap corresponds to a specific stage of the algorithm, for the sequential algorithm, this error can occur anytime between the first and last merging.

Let us compute the expected number of pairs of reads which can be merged in error at overlap ℓ. To begin, a read has the potential to be merged to an incorrect successor at overlap ℓ if it has overlap less than or equal to ℓ with its true successor, since otherwise the sequential algorithm discovers the read's true successor. By the Poisson approximation, there are roughly $Ne^{-\lambda(L-\ell)}$ reads with physical overlap less than or equal to ℓ with their successors. In particular, if $\ell < L - \lambda^{-1}\ln N$, there will be no such reads, and so we may assume that ℓ lies between $\max\{0, L - \lambda^{-1}\ln N\}$ and L.

Note furthermore that in order for an error to occur, the second read must not yet have been merged when the algorithm encounters the first read, and thus, the second read must be positioned later in the sequence. This adds a factor one-half. Combining this reasoning with the preceding paragraph, we see that there are approximately

$$\tfrac{1}{2}N^2 e^{-\lambda(L-\ell)}$$

pairs of reads which may potentially be merged incorrectly at overlap ℓ.

For such a pair, an erroneous merging actually occurs if the length-ℓ suffix of the first read equals the length-ℓ prefix of the second. Assuming (as in the greedy algorithm calculation) that these substrings are physically disjoint, the probability of this event is $e^{-\ell H_2(\mathbf{p})}$.

The expected number of pairs of reads that are merged in error at overlap ℓ, for $L - \lambda^{-1}\ln N \le \ell \le L$, is thus approximately

$$\tfrac{1}{2}N^2 e^{-\lambda(L-\ell)} e^{-\ell H_2(\mathbf{p})}. \tag{8}$$

This number is largest when $\ell = L$ or $\ell = \max\{0, L - \lambda^{-1}\ln N\}$. Plugging $\ell = L$ into (8), we see that the expression approaches zero whenever $\bar{L} > 2/H_2(\mathbf{p})$. Plugging into the latter value and performing some arithmetic manipulations, we conclude that the expression in (8) approaches zero whenever $\frac{N}{N_{\text{cov}}} > \frac{\bar{L}H_2(\mathbf{p})}{\bar{L}H_2(\mathbf{p})-1}$ and $\bar{L} > 2/H_2(\mathbf{p})$, as in Proposition 4.

2) K-mer-Based Algorithms: Many recent assembly algorithms operate on K-mers instead of directly on the reads themselves. K-mers are length K subsequences of the reads; from each read, one can generate $L-K+1$ K-mers. One of the early works which pioneer this approach is the sort-and-extend technique in ARACHNE [23]. By lexicographically sorting the set of all the K-mers generated from the collection of reads, identical K-mers from physically overlapping reads will be adjacent to each other. This enables the overlap relation between the reads (so called overlap graph) to be computed in $O(N\ln N)$ time (time to sort the set of K-mers) as opposed to the $O(N^2)$ time needed if pairwise comparisons between the reads were done.

Another related approach is the de Bruijn graph approach [10], [20]. In this approach, the $(K+1)$-mers are represented as vertices of a de Bruijn graph and there is an edge between two vertices if they represent adjacent $(K+1)$-mers in some read (here adjacency means their positions are offset by one). A naive de Bruijn algorithm discards the reads after construction of the de Bruijn graph (all actual de Bruijn-based algorithms use the read information to some extent, including [10] and [20]). The DNA sequence reconstruction problem is then formulated as finding an Eulerian cycle traversing all the edges of the de Bruijn graph.

The performance of these algorithms on the basic statistical model can be analyzed by observing that two conditions must be satisfied for them to work. First, K should be chosen such that with high probability, K-mers from physically disjoint parts of the DNA sequence should be distinct, i.e., there are no repeated length-K subsequences in the DNA sequence. In the sort-and-extend technique, this will ensure that two identical adjacent K-mers in the sorted list belong to two physically overlapping reads rather than two physically disjoint reads. Similarly, in the de Bruijn graph approach, this will ensure that the Eulerian cycle will be connecting $(K+1)$-mers that are physically overlapping. This minimum K can be calculated as we did to justify Lemma 2:

$$\frac{K}{\ln G} > \frac{2}{H_2(\mathbf{p})}. \tag{9}$$

The second condition for success is that all successive reads should have a physical overlap of at least K base pairs. This is needed so that the reads can be assembled via the K-mers. According to the Poisson approximation, the expected number of successive reads with spacing greater than $L-K$ base pairs is roughly $Ne^{-\lambda(L-K)}$. To ensure that with high probability, all successive reads overlap by at least K base pairs, this expected number should be small, i.e.,

$$N > \frac{G\ln N}{L-K} \approx \frac{G\ln G}{L-K}. \tag{10}$$

Substituting (9) into (10) and using the definition $\bar{L} = L/\ln G$, we obtain

$$\frac{N}{N_{\text{cov}}} > \frac{\bar{L}H_2(\mathbf{p})}{\bar{L}H_2(\mathbf{p}) - 2}.$$

The minimum normalized coverage depth of this algorithm is plotted in Fig. 7. Note that the performance of the K-mer-based algorithms is strictly less than the performance achieved by the greedy algorithm. The reason is that for $\bar{L} > 2/H_2(\mathbf{p})$, while the greedy algorithm only requires the reads to cover the DNA sequence, the K-mer-based algorithms need more, that successive reads have (normalized) overlap at least $2/H_2(\mathbf{p})$.

E. Complexity of the Greedy Algorithm

A naive implementation of the greedy algorithm would require an all-to-all pairwise comparison between all the reads. This would require $O(N^2)$ comparisons, resulting in unacceptably high computational cost for N in the order of tens of millions. However, drawing inspiration from the sort-and-extend technique discussed in the previous section, a more clever implementation would yield a complexity of $O(LN\ln N)$. Since $L \ll N$, this represents a huge saving. We compute the complexity as follows. Recall that in stage ℓ of the greedy algorithm, successive reads with overlap ℓ are considered. Instead of doing many pairwise comparisons to obtain such reads, one can simply extract all the ℓ-mers from the reads and perform a sort-and-extend to find all the reads with overlap ℓ. Since we have to apply

sort-and-extend in each stage of the algorithm, the total complexity is $O(LN \ln N)$.

An idea similar to this and resulting in the same complexity was described by Turner [26] (in the context of the SCS problem), with the sorting effectively replaced with a suffix tree data structure. Ukkonen [28] used a more sophisticated data structure, which essentially computes overlaps between strings in parallel, to reduce the complexity to $O(NL)$.

III. Markov DNA Model

In this section, we extend the results for the basic i.i.d. DNA sequence model to a Markov sequence model.

A. Formulation and Result

The problem formulation is identical to the one in Section II-A except that we assume the DNA sequence is correlated and model it by a Markov source with transition matrix $Q = [q_{ij}]_{i,j\in\{1,2,3,4\}}$, where $q_{ij} = \mathbb{P}(S_k = i|S_{k-1} = j)$.

Remark 5: We assume that the DNA is a Markov process of order 1, but the result can be generalized to Markov processes of order m as long as m is constant and does not grow with G.

In the basic i.i.d. model, we observed that the minimum normalized coverage depth depends on the DNA statistics through the Rényi entropy of order 2. We prove that a similar dependence holds for Markov models. In [22], it is shown that the Rényi entropy *rate* of order 2 for a stationary ergodic Markov source with transition matrix Q is given by

$$H_2(Q) := \ln\left(\frac{1}{\rho_{\max}(\bar{Q})}\right),$$

where $\rho_{\max}(\bar{Q}) \triangleq \max\{|\rho| : \rho \text{ eigenvalue of } \bar{Q}\}$, and $\bar{Q} = [q_{ij}^2]_{i,j\in\{1,2,3,4\}}$. In terms of this quantity, we state the following theorem.

Theorem 6: The minimum normalized coverage depth of a stationary ergodic Markov DNA sequence is given by

$$c_{min}(\bar{L}) = \begin{cases} \infty & \text{if } \bar{L} < 2/H_2(Q), \\ 1 & \text{if } \bar{L} > 2/H_2(Q). \end{cases} \tag{11}$$

B. Sketch of Proof

Similar to the i.i.d. case, it suffices to show the following statements:

1) If $\bar{L} < \frac{2}{H_2(Q)}$, $N_{\min}(\epsilon, G, L) = \infty$ for sufficiently large G.
2) If $\bar{L} > \frac{2}{H_2(Q)}$, then $c_{\min} = 1$.

The following lemma is the analog of Lemma 2 for the Markov case and is used in a similar way to prove statement 1.

Lemma 7: If $\bar{L} < 2/H_2(Q)$, then a Markov DNA sequence contains interleaved repeats with probability $1 - o(1)$.

To justify Lemma 7, we use a similar heuristic argument as for the i.i.d. model, but with a new value for the probability that two physically disjoint sequences $\mathbf{S}_i^L$ and $\mathbf{S}_j^L$ are equal:

$$\mathbb{P}(\mathbf{S}_i^L = \mathbf{S}_j^L) \approx e^{-L\ln\left(\rho_{\max}(\bar{Q})\right)}.$$

The lemma follows from the fact that there are roughly G^2 such pairs in the DNA sequence. A formal proof of the lemma is provided in Appendix B.

Statement 2 is again a consequence of the optimality of the greedy algorithm, as shown in the following lemma.

Lemma 8: Suppose $\bar{L} > \frac{2}{H_2(Q)}$. Then, the greedy algorithm with exactly the same overlap score as used for the i.i.d. model can achieve minimum normalized coverage depth $c_{\min} = 1$.

Lemma 8 is proved in Appendix B. The key technical step of the proof entails showing that the effect of physically overlapping reads does not affect the asymptotic performance of the algorithm, just as in the i.i.d. case.

IV. Noisy Reads

In our basic model, we assumed that the read process is noiseless. In this section, we assess the impact of noise on the greedy algorithm.

A. Formulation and Result

The problem formulation here differs from Section II-A in two ways. First, we assume that the read process is noisy and consider a simple probabilistic model for the noise. A nucleotide s is read to be r with probability $Q(r|s)$. The nucleotides within a read are perturbed independently, i.e., if $\mathbf{r}$ is a read from the physical underlying subsequence $\mathbf{s}$, then

$$\mathbb{P}(\mathbf{r}|\mathbf{s}) = \prod_{i=1}^{L} Q(r_i|s_i).$$

Additionally, the noise is assumed to be independent for different reads.

Second, we require a weaker notion of reconstruction. Instead of *perfect reconstruction*, we aim for *perfect layout*. By perfect layout, we mean that all the reads are mapped correctly to their true locations. Note that perfect layout does not imply perfect reconstruction, since the consensus sequence may not be identical to the DNA sequence. On the other hand, since coverage implies that most positions on the DNA are covered by $O(\ln G)$ many reads, the consensus sequence will be correct in *most* positions if we achieve perfect layout.

Remark: In the jargon of information theory, we are modeling the noise in the read process as a discrete memoryless channel with transition probability $Q(\cdot|\cdot)$. Noise processes in actual sequencing technologies can be more complex than this model. For example, the amount of noise can increase as the read process proceeds, or there may be insertions and deletions in addition to substitutions. Nevertheless, understanding the effect of noise on the assembly problem in this model provides considerable insight into the problem.

We now evaluate the performance of the greedy algorithm for the noisy read problem.

To tailor the greedy algorithm to the noisy reads, the only requirement is to define the overlap score between two distinct reads. Given two reads $\mathbf{r}_i$ and $\mathbf{r}_j$, we would like to know whether they are physically overlapping with length ℓ. Let $\mathbf{X}$ and $\mathbf{Y}$ of length ℓ be the suffix of $\mathbf{r}_i$ and prefix of $\mathbf{r}_j$, respectively. We have the following hypotheses for $\mathbf{X}$ and $\mathbf{Y}$:

1) H_0: $\mathbf{X}$ and $\mathbf{Y}$ are noisy reads from the same physical source subsequence;
2) H_1: $\mathbf{X}$ and $\mathbf{Y}$ are noisy reads from two disjoint source subsequences.

The decision rule that is optimal in trading off the two types of error is the MAP rule, obtained by a standard large deviations calculation (see, for example, [4, Chs. 11.7 and 11.9].) In log-likelihood form, the MAP rule for this hypothesis testing problem is

$$\text{Decide } H_0 \text{ if } \ln \frac{P(\mathbf{x},\mathbf{y})}{P(\mathbf{x})P(\mathbf{y})} = \sum_{j=1}^{\ell} \ln \frac{P_{X,Y}(x_j,y_j)}{P_X(x_j)P_Y(y_j)} \geq \ell\theta, \tag{12}$$

where $P_{X,Y}(x,y)$, $P_X(x)$, and $P_Y(y)$ are the marginals of the joint distribution $P_S(s)Q(x|s)Q(y|s)$, and θ is a parameter reflecting the prior distribution of H_0 and H_1.

We can now define the overlap score, whereby two reads $\mathbf{R}_i$ and $\mathbf{R}_j$ have overlap at least ℓ if the MAP rule on the length ℓ suffix of $\mathbf{R}_i$ and the length ℓ prefix of read $\mathbf{R}_j$ decides H_0. The performance of the greedy algorithm using this score is given in the following theorem.

Theorem 9: The modified greedy algorithm can achieve normalized coverage depth $c(\bar{L}) = 1$ if $\bar{L} > 2/I^*$, where

$$I^* = \max_{\theta} \min(2D(P_\mu \| P_{X,Y}), D(P_\mu \| P_X \cdot P_Y)),$$

and the distribution P_μ is given by

$$P_\mu(x,y) := \frac{[P_{X,Y}(x,y)]^\mu [P_X(x)P_Y(y)]^{1-\mu}}{\sum_{a,b}[P_{X,Y}(a,b)]^\mu [P_X(a)P_Y(b)]^{1-\mu}}$$

with μ the solution to the equation

$$D(P_\mu \| P_X \cdot P_Y) - D(P_\mu \| P_{X,Y}) = \theta.$$

The statement of Theorem 9 uses the Kullback–Leibler divergence $D(P\|Q)$ of the distribution P relative to Q, defined as

$$D(P\|Q) = \sum_a P(a) \ln \frac{P(a)}{Q(a)}. \tag{13}$$

The details of the proof of the theorem are in Appendix C. To illustrate the main ideas, we sketch the proof for the special case of uniform source and symmetric noise.

B. Sketch of Proof for Uniform Source and Symmetric Noise

In this section, we provide an argument to justify Theorem 9 in the case of uniform source and symmetric noise. Concretely, $\mathbf{p} = (1/4, 1/4, 1/4, 1/4)$ and the noise is symmetric with transition probabilities:

$$Q(i|j) = \begin{cases} 1-\epsilon & \text{if } i = j \\ \epsilon/3 & \text{if } i \neq j. \end{cases} \tag{14}$$

The parameter ϵ is often called the error rate of the read process. It ranges from 1% to 10% depending on the sequencing technology.

Corollary 10: The greedy algorithm with the modified definition of overlap score between reads can achieve normalized coverage depth $c(\bar{L}) = 1$ if $\bar{L} > 2/I^*(\epsilon)$, where

$$I^*(\epsilon) = D(\alpha^* \| \tfrac{3}{4})$$

and α^* satisfies

$$D(\alpha^* \| \tfrac{3}{4}) = 2D(\alpha^* \| 2\epsilon - \tfrac{4}{3}\epsilon^2).$$

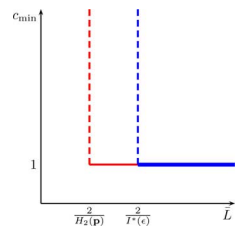

Fig. 8. Performance of the modified greedy algorithm with noisy reads.

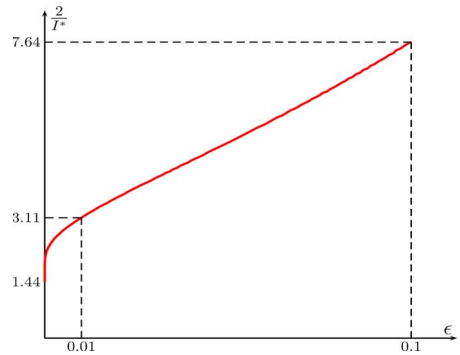

Fig. 9. Plot of $\frac{2}{I^*(\epsilon)}$ as a function of ϵ for the uniform source and symmetric noise model.

Here, $D(\alpha\|\beta)$ is the divergence between a $\mathrm{Bern}(\alpha)$ and a $\mathrm{Bern}(\beta)$ random variable.

Proof: The proof follows by applying Theorem 9. For uniform source and symmetric noise, the optimum I^* is attained when $2D(P_\mu\|P_{X,Y}) = D(P_\mu\|P_X \cdot P_Y)$. The result is written in terms of α^* which is a function of the optimal value θ^*. ■

The performance of this algorithm is shown in Fig. 8. The only difference between the two curves, one for noiseless reads and one with noisy reads, is the larger threshold $2/I^*(\epsilon)$ at which the minimum normalized coverage depth becomes one. A plot of this threshold as a function of ϵ is shown in Fig. 9. It can be seen that when $\epsilon = 0$, $2/I^*(\epsilon) = 2/H_2(\mathbf{p}) = 1/\ln 2$, and increases continuously as ϵ increases.

We justify the corollary by the following argument. In the noiseless case, two reads overlap by at least ℓ if the length ℓ prefix of one read is identical to the length ℓ suffix of the other read. The overlap score is the largest such ℓ. When there is noise, this criterion is not appropriate. Instead, a natural modification of this definition is that two reads overlap by at least ℓ if the Hamming distance between the prefix and the suffix strings is less than a fraction α of the length ℓ. The overlap score between the two reads is the largest such ℓ. The parameter α controls how stringent the overlap criterion is. By optimizing over the value of α, we can obtain the following result.

We picture the greedy algorithm as working in stages, starting with an overlap score of L down to an overlap score of 0. Since the spacing between reads is independent of the DNA sequence and noise process, the number of reads at stage ℓ given no errors have occurred in previous stages is again roughly

$$Ne^{-\lambda(L-\ell)}.$$

To pass this stage without making an error, the greedy algorithm should correctly merge those reads having spacing of length ℓ

to their successors. Similar to the noiseless case, the greedy algorithm makes an error if the overlap score between two non-consecutive reads is ℓ at stage ℓ; in other words

1. the Hamming distance between the length ℓ suffix of the present read and the length ℓ prefix of some read which is not the successor is less than $\alpha\ell$ by random chance.
   A standard large deviations calculation shows that the probability of this event is approximately
   $$e^{-\ell D(\alpha\|\frac{3}{4})},$$
   which is the probability that two independent strings of length ℓ have Hamming distance less than $\alpha\ell$. Hence, the expected number of pairs of contigs for which this confusion event happens is approximately
   $$\left[Ne^{-\lambda(L-\ell)}\right]^2 e^{-\ell D(\alpha\|\frac{3}{4})}. \tag{15}$$
   Unlike the noiseless case, however, there is another important event affecting the performance of the algorithm. The *missed detection* event is defined as
2. the Hamming distance between the length ℓ suffix of the present read and the length ℓ prefix of the successor read is larger than $\alpha\ell$ due to an excessive amount of noise.

Again, a standard large deviations calculation shows that the probability of this event for a given read is approximately

$$e^{-\ell D(\alpha\|\eta)},$$

where $\eta = 2\epsilon - \frac{4}{3}\epsilon^2$ is the probability that the ith symbol in the length ℓ suffix of the present read does not match the ith symbol in the length ℓ prefix of the successor read (here we are assuming that $\alpha > \eta$). Thus, the expected number of contigs missing their successor contig at stage ℓ is approximately

$$Ne^{-\lambda(L-\ell)}e^{-\ell D(\alpha\|\eta)}. \tag{16}$$

Both (15) and (16) are largest when either $\ell = L$ or $\ell = 0$. Similarly to the noiseless case, errors do not occur at stage 0 if the DNA sequence is covered by the reads. The coverage condition guarantees no gap exists in the assembled sequence. From (15) and (16), we see that no errors occur at stage L if

$$\bar{L} = \frac{L}{\ln G} > \frac{2}{\min(2D(\alpha\|\eta), D(\alpha\|\frac{3}{4}))}.$$

Selecting α to minimize the right-hand side results in the two quantities within the minimum being equal, thereby justifying Corollary 10.

The algorithm described in this section is only one particular scheme to handle noisy reads. A natural question is whether there is any better scheme. We answer this question in the affirmative in a subsequent work [32]. In particular, we showed that for any noisy read channel, there is a threshold on the noise level such that below this threshold, noiseless performance can be asymptotically achieved. In the example of uniform source and symmetric noise, this threshold is 19%. This is in contrast to the performance of the greedy algorithm, which degrades as soon as the noise level is non-zero.

V. Discussions and Future Work

This paper seeks to understand the basic data requirements for shotgun sequencing, and we obtain results for a few simple models. The models for the DNA sequence and read process in this paper serve as a starting point from which to pursue extensions to more realistic models. We discuss a few of the many possible extensions.

1) Long repeats: Long repeats occur in many genomes, from bacteria to human. The repetitive nature of real genomes is understood to be a major bottleneck for sequence assembly. Thus, a caveat of this paper is that the DNA sequence models we have considered, both i.i.d. and Markov, exhibit only short-range correlations, and therefore fail to capture the long-range correlation present in complex genomes. Motivated by this issue, a follow-up work [3] extends the approach of this paper to *arbitrary* repeat statistics, in particular the statistics of actual genomes, overcoming the difficulties posed by the lack of a good probability model for DNA. The read model considered in [3] is the same uniform noiseless model we consider.

We briefly summarize the results and approach of [3]. First, Ukkonen's condition that there is no interleaved or triple repeats of length at least $L-1$ is generalized to give a lower bound on the read length and the coverage depth required for reconstruction in terms of repeat statistics of the genome. Next, they design a de Brujin graph-based assembly algorithm that can achieve very close to the lower bound for repeat statistics of a wide range of sequenced genomes. The approach results in a pipeline, which takes as input a genome sequence and desired success probability $1-\epsilon$, computes a few simple repeat statistics, and from these statistics computes a feasibility plot that indicates for which L and N reconstruction is possible.

2) Double-strandedness: The DNA sequence is double stranded and consists of a sequence $\mathbf{s}$ and its reverse complement $\tilde{\mathbf{s}}$. Reads are obtained from either of the two strands, and a natural concern is whether this affects the results. It turns out that for the i.i.d. sequence model considered in this paper (as well as the Markov model), the asymptotic minimum normalized coverage depth remains the same. The optimal greedy algorithm is modified slightly by including the reverse complements of the reads as well as the originals, and stopping when there are two reconstructed sequences $\mathbf{s}$ and $\tilde{\mathbf{s}}$. The heuristic argument follows by observing that the probability of error at stage ℓ given in (7) is changed only by a factor 2, which does not change the asymptotic result. The rigorous proof involves showing that the contribution from overlapping reads is negligible, where the notion of reads overlapping accounts for both the sequence and its reverse complement.

3) Read process: There are a number of important properties of the read process which can be incorporated into more accurate models. Beyond the substitution noise considered in this paper, some sequencing technologies (such as PacBio) produce insertions and deletions. Often bases come with quality scores, and these scores can be used to mitigate the effect of noise. Other interesting aspects include correlation in the noise from one base to another (e.g., typically producing several errors in a row), nonuniformity of the error rate within a read, and correlation of the noise process with the read content. Aside from

noise, a serious practical difficulty arises due to the positions of reads produced by some sequencing platforms being biased by the sequence, e.g., by the GC content. Noise and sampling bias in the reads make assembly more difficult, but another important direction is to incorporate mate-pairs into the read model. Mate-pairs (or paired-end reads) consisting of two reads with an approximately known separation help to resolve long repeats using short reads.

4) Partial reconstruction: In practice, the necessary conditions for perfect reconstruction are not always satisfied, but it is still desirable to produce the best possible assembly. While the notion of perfect reconstruction is relatively simple, defining what "best" means is more delicate for partial reconstructions; one must allow for multiple contigs in the output as well as errors (misjoins). Thus, an optimal algorithm is one which trades off optimally between the number of contigs and number of errors.

Appendix A
Proof of Theorem 1, Part 2

We first state and prove the following lemma. This result can be found in [2], but for ease of generalization to the Markov case later, we include the proof.

Lemma 11: For any distinct substrings $\mathbf{X}$ and $\mathbf{Y}$ of length ℓ of the i.i.d. DNA sequence:

1) If the strings have no physical overlap, the probability that they are identical is $e^{-\ell H_2(\mathbf{p})}$.
2) If the strings have nonzero physical overlap, the probability that they are identical is bounded above by $e^{-\ell H_2(\mathbf{p})/2}$.

Proof: We give notation for the probability that any k distinct bases in the DNA sequence are identical:

$$\pi_k \triangleq \sum_{i=1}^{4} p_i^k.$$

The proof of part 1 is immediate: Consider $\mathbf{X} = S_{i+1} \ldots S_{i+\ell}$ and $\mathbf{Y} = S_{j+1} \ldots S_{j+\ell}$ having no physical overlap. In this case, the events $\{S_{i+m} = S_{j+m}\}$ for $m \in \{1, \ldots, \ell\}$ are independents and equiprobable. Therefore, the probability that $\mathbf{X} = \mathbf{Y}$ is given by

$$\pi_2^{\ell} = e^{-\ell H_2(\mathbf{p})}.$$

We now prove part 2. For the case of overlapping strings $\mathbf{X}$ and $\mathbf{Y}$, suppose that a substring of length $k < \ell$ from the DNA sequence is shared between the two strings. Without loss of generality, we assume that $\mathbf{X}$ and $\mathbf{Y}$ are, respectively, the prefix and suffix of $\mathbf{S}_1^{2\ell-k}$. Let q and r be the quotient and remainder of ℓ divided by $\ell - k$, i.e., $\ell = q(\ell - k) + r$, where $0 \leq r < \ell - k$. It can be shown that $\mathbf{X} = \mathbf{Y}$ if and only if $\mathbf{S}_1^{2\ell-k}$ is a string of the form $\mathbf{UVUV} \ldots \mathbf{UVU}$, where $\mathbf{U}$ and $\mathbf{V}$ have length r and $\ell - k - r$. Since the number of copies of $\mathbf{U}$ and $\mathbf{V}$ are, respectively, $q + 2$ and $q + 1$, the probability of observing a structure of the form $\mathbf{UVUV} \ldots \mathbf{UVU}$ is given by

$$\begin{aligned}(\pi_{q+2})^r \times (\pi_{q+1})^{\ell-k-r} &\overset{(a)}{\leq} (\pi_2)^{\frac{r(q+2)}{2}} \times (\pi_2)^{\frac{(\ell-k-r)(q+1)}{2}} \\ &= (\pi_2)^{\ell - \frac{k}{2}},\end{aligned}$$

where (a) follows from the fact that $(\pi_q)^{\frac{1}{q}} \leq (\pi_2)^{\frac{1}{2}}$ for all $q \geq 2$. Since $k < \ell$, we have $(\pi_2)^{\ell - \frac{k}{2}} \leq (\pi_2)^{\frac{\ell}{2}}$. Therefore, the probability that $\mathbf{X} = \mathbf{Y}$ for two overlapping strings is bounded above by

$$(\pi_2)^{\frac{\ell}{2}} = e^{-\ell H_2(\mathbf{p})/2}.$$

This completes the proof. ■

Proof of Theorem 1, Part 2: The greedy algorithm finds a contig corresponding to a substring of the DNA sequence if each read $\mathbf{R}_i$ is correctly merged to its successor read $\mathbf{R}_i^s$ with the correct amount of *physical overlap* between them, which is $V_i = L - (T_i^s - T_i)_{\mathrm{mod}\, G}$.[3] If, in addition, the whole sequence is covered by the reads, then the output of the algorithm is exactly the DNA sequence $\mathbf{S}$.

Let $\mathcal{E}_1$ be the event that some read is merged incorrectly; this includes merging to the read's valid successor but at the wrong relative position, as well as merging to an impostor. Let $\mathcal{E}_2$ be the event that the DNA sequence is not covered by the reads. The union of these events, $\mathcal{E}_1 \cup \mathcal{E}_2$, contains the error event $\mathcal{E}$. We first focus on event $\mathcal{E}_1$.

Since the greedy algorithm merges reads with highest overlap score, we may think of the algorithm as working in stages starting with an overlap score of L down to an overlap score of 0. Thus, $\mathcal{E}_1$ naturally decomposes as $\mathcal{E}_1 = \cup_\ell \mathcal{A}_\ell$, where $\mathcal{A}_\ell$ is the event that the first error in merging occurs at stage ℓ.

Recall that the greedy algorithm uses overlap score given by $W_{ij} := W(\mathbf{R}_i, \mathbf{R}_j)$ defined as the length of the longest suffix of $\mathbf{R}_i$ identical to a prefix of $\mathbf{R}_j$. Additionally, for a read $\mathbf{R}_j$ that is the successor to $\mathbf{R}_k$, let $U_j = V_k$ denote the physical overlap between them. Thus, U_j is the physical overlap between $\mathbf{R}_j$ and its predecessor.

Now, we claim that

$$\mathcal{A}_\ell \subseteq \mathcal{B}_\ell \cup \mathcal{C}_\ell, \tag{17}$$

where

$$\mathcal{B}_\ell \triangleq \{\mathbf{R}_j \neq \mathbf{R}_i^s, U_j \leq \ell, V_i \leq \ell, W_{ij} = \ell \text{ for some } i \neq j\} \tag{18}$$

$$\mathcal{C}_\ell \triangleq \{\mathbf{R}_j = \mathbf{R}_i^s, U_j = V_i < \ell, W_{ij} = \ell \text{ for some } i \neq j\}. \tag{19}$$

If the event $\mathcal{A}_\ell$ occurs, then either there are two reads $\mathbf{R}_i$ and $\mathbf{R}_j = \mathbf{R}_i^s$ such that $\mathbf{R}_i$ is merged to its successor $\mathbf{R}_j$, but at an overlap larger than their physical overlap, or there are two reads $\mathbf{R}_i$ and $\mathbf{R}_j$ such that $\mathbf{R}_i$ is merged to $\mathbf{R}_j$, an impostor. The first case implies the event $\mathcal{C}_\ell$. In the second case, in addition to $W_{ij} = \ell$, it must be true that the physical overlaps $V_i, U_j \leq \ell$, since otherwise at least one of these two reads would have been merged at an earlier stage. (By definition of $\mathcal{A}_\ell$, there were no errors before stage ℓ.) Hence, in this second case, the event $\mathcal{B}_\ell$ occurs.

Now we will bound $\mathbb{P}(\mathcal{B}_\ell)$ and $\mathbb{P}(\mathcal{C}_\ell)$. First, let us consider the event $\mathcal{B}_\ell$. This is the event that two reads that are not neighbors with each other were merged by mistake. Intuitively, event $\mathcal{B}_\ell$ says that the pairs of reads that can potentially cause such confusion at stage ℓ are limited to those with short physical overlap with their own neighboring reads, since the ones with

[3]Note that the physical overlap can take negative values.

large physical overlaps have already been successfully merged to their correct neighbor by the algorithm in the early stages. In Fig. 6, these are the reads at the ends of the contigs that are formed by stage ℓ.

For any two distinct reads $\mathbf{R}_i$ and $\mathbf{R}_j$, we define the event

$$\mathcal{B}_\ell^{ij} \triangleq \{\mathbf{R}_j \neq \mathbf{R}_i^s, U_j \leq \ell, V_i \leq \ell, W_{ij} = \ell\}.$$

From the definition of $\mathcal{B}_\ell$ in (18), we have $\mathcal{B}_\ell \subseteq \cup_{ij}\mathcal{B}_\ell^{ij}$. Applying the union bound and considering the fact that $\mathcal{B}_\ell^{ij}$'s are equiprobable yields

$$\mathbb{P}(\mathcal{B}_\ell) \leq N^2\mathbb{P}(\mathcal{B}_\ell^{12}).$$

Let $\mathcal{D}$ be the event that the two reads $\mathbf{R}_1$ and $\mathbf{R}_2$ have no physical overlap. Using the law of total probability, we obtain

$$\mathbb{P}(\mathcal{B}_\ell^{12}) = \mathbb{P}(\mathcal{B}_\ell^{12}|\mathcal{D})\mathbb{P}(\mathcal{D}) + \mathbb{P}(\mathcal{B}_\ell^{12}|\mathcal{D}^c)\mathbb{P}(\mathcal{D}^c).$$

Since $\mathcal{D}^c$ happens only if $T_2 \in [T_1 - L + 1, T_1 + L - 1]$, $\mathbb{P}(\mathcal{D}^c) \leq \frac{2L}{G}$. Hence,

$$\mathbb{P}(\mathcal{B}_\ell^{12}) \leq \mathbb{P}(\mathcal{B}_\ell^{12}|\mathcal{D}) + \mathbb{P}(\mathcal{B}_\ell^{12}|\mathcal{D}^c)\frac{2L}{G}. \tag{20}$$

We proceed with bounding $\mathbb{P}(\mathcal{B}_\ell^{12}|\mathcal{D})$ as follows:

$$\begin{aligned}\mathbb{P}(\mathcal{B}_\ell^{12}|\mathcal{D}) &= \mathbb{P}(U_2 \leq \ell, V_1 \leq \ell, W_{12} = \ell|\mathcal{D}) \\ &\overset{(a)}{=} \mathbb{P}(U_2 \leq \ell, V_1 \leq \ell|\mathcal{D})\mathbb{P}(W_{12} = \ell|\mathcal{D}) \\ &\overset{(b)}{=} \mathbb{P}(U_2 \leq \ell, V_1 \leq \ell|\mathcal{D})e^{-\ell H_2(\mathbf{p})},\end{aligned}$$

where (a) follows from the fact that given $\mathcal{D}$, the events $\{U_2 \leq \ell, V_1 \leq \ell\}$ and $\{W_{12} = \ell\}$ are independent, and (b) follows from Lemma 11 part 1.

Note that the event $\{\mathbf{R}_2 \neq \mathbf{R}_1^s, U_2 \leq \ell, V_1 \leq \ell\}$ implies that no reads start in the intervals $[T_1, T_1 + L - \ell - 1]$ and $[T_2 - L + \ell + 1, T_2]$. We claim that given $\mathcal{D}$, the two intervals are disjoint: otherwise (since the intervals are empty) $\mathbf{R}_2$ is the successor of $\mathbf{R}_1$, contradicting $\mathbf{R}_2 \neq \mathbf{R}_1^s$. Thus, the probability that there is no read starting in the intervals is given by

$$\left(1 - \frac{2(L-\ell)}{G}\right)^{N-2}.$$

Using the inequality $1 - a \leq e^{-a}$, we obtain

$$\mathbb{P}(\mathcal{B}_\ell^{12}|\mathcal{D}) \leq e^{-2\lambda(L-\ell)(1-2/N)}e^{-\ell H_2(\mathbf{p})}. \tag{21}$$

We now turn to $\mathbb{P}(\mathcal{B}_\ell^{12}|\mathcal{D}^c)$. We first observe that $\mathcal{B}_\ell^{12}$ implies that the length of the physical overlap between $\mathbf{R}_1$ and $\mathbf{R}_2$ is strictly less than ℓ. Conditional on $\mathcal{D}^c$, this overlap must be strictly between zero and ℓ, an event we denote by $\mathcal{D}_1$. Thus

$$\begin{aligned}\mathbb{P}(\mathcal{B}_\ell^{12}|\mathcal{D}^c) &\leq \mathbb{P}(\mathcal{B}_\ell^{12}, \mathcal{D}_1, \mathcal{D}^c)/\mathbb{P}(\mathcal{D}_1, \mathcal{D}^c) \\ &\leq \mathbb{P}(U_2 \leq \ell, V_1 \leq \ell, W_{12} = \ell|\mathcal{D}_1) \\ &\leq \mathbb{P}(V_1 \leq \ell, W_{12} = \ell|\mathcal{D}_1) \\ &\overset{(a)}{\leq} \mathbb{P}(V_1 \leq \ell|\mathcal{D}_1)\mathbb{P}(W_{12} = \ell|\mathcal{D}_1) \\ &\overset{(b)}{\leq} \mathbb{P}(V_1 \leq \ell|\mathcal{D}_1)e^{-\ell H_2(\mathbf{p})/2}\end{aligned}$$

where (a) follows from the fact that given $\mathcal{D}_1$, the events $\{V_1 \leq \ell\}$ and $\{W_{12} = \ell\}$ are independent, and (b) follows from Lemma 11 part 2. Since $\{V_1 \leq \ell\}$ corresponds to the event that there is no read starting in the interval $[T_1, T_1 + L - \ell - 1]$, we obtain

$$\mathbb{P}(V_1 \leq \ell|\mathcal{D}_1) = \left(1 - \frac{L-\ell}{G}\right)^{N-2}.$$

Applying the inequality $1 - a \leq e^{-a}$, we obtain

$$\mathbb{P}(\mathcal{B}_\ell^{12}|\mathcal{D}^c) \leq e^{-\lambda(L-\ell)(1-2/N)}e^{-\ell H_2(\mathbf{p})/2}.$$

Putting all terms together, we have

$$\mathbb{P}(\mathcal{B}_\ell) \leq q_\ell^2 + 2\lambda L q_\ell \tag{22}$$

where

$$q_\ell = \lambda G e^{-\lambda(L-\ell)(1-2/N)}e^{-\ell H_2(\mathbf{p})/2}. \tag{23}$$

The first term reflects the contribution from the reads with no physical overlap and the second term from the reads with physical overlap. Even though there are lots more of the former than the latter, the probability of confusion when the reads are physically overlapping can be much larger. Hence, both terms have to be considered.

Let us define

$$\mathcal{C}_\ell^i \triangleq \{V_i < \ell, W(\mathbf{R}_i, \mathbf{R}_i^s) = \ell\}.$$

From the definition of $\mathcal{C}_\ell$ in (19), we have $\mathcal{C}_\ell \subseteq \cup_i \mathcal{C}_\ell^i$. Applying the union bound and using the fact that the $\mathcal{C}_\ell^i$'s are equiprobable yields $\mathcal{C}_\ell \leq N\mathbb{P}(\mathcal{C}_\ell^1)$; hence

$$\mathbb{P}(\mathcal{C}_\ell) \leq N\mathbb{P}(W(\mathbf{R}_1, \mathbf{R}_1^s) = \ell|V_1 < \ell)\mathbb{P}(V_1 < \ell).$$

Applying Lemma 11 part 2, we obtain

$$\mathbb{P}(\mathcal{C}_\ell) \leq N e^{-\ell H_2(\mathbf{p})/2}\left(1 - \frac{L-\ell}{G}\right)^{N-1}.$$

Using the inequality $1 - a \leq e^{-a}$, we obtain

$$\mathbb{P}(\mathcal{C}_\ell) \leq \lambda G e^{-\lambda(L-\ell)(1-1/N)}e^{-\ell H_2(\mathbf{p})/2} \leq q_\ell. \tag{24}$$

Using the bounds (22) and (24), we get

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1) = \mathbb{P}(\cup_\ell \mathcal{A}_\ell) &\leq \sum_{\ell=0}^{L}\mathbb{P}(\mathcal{A}_\ell) = \sum_{\ell=0}^{L}\mathbb{P}(\mathcal{B}_\ell) + \mathbb{P}(\mathcal{C}_\ell) \\ &\leq \sum_{\ell=0}^{L} q_\ell^2 + (2\lambda L + 1)q_\ell,\end{aligned}$$

where q_ℓ is defined in (23). Since q_ℓ is monotonic in ℓ, we can further bound $\mathbb{P}(\mathcal{E}_1)$ by

$$\mathbb{P}(\mathcal{E}_1) \leq (L+1)\max\{q_0^2 + (2\lambda L + 1)q_0, q_L^2 + (2\lambda L + 1)q_L^2\}. \tag{25}$$

Since $\bar{L} > \frac{2}{H_2(\mathbf{p})}$, q_L vanishes exponentially in L and the second term on the right-hand side of (25) has no contribution asymptotically. Now, choose

$$N = \frac{G}{L}\ln(GL^3).$$

A direct computation shows that for this choice of N, $q_0^2 + (2\lambda L+1)q_0 = O(\frac{1}{L^2})$. Hence, the bound (25) implies that $\mathbb{P}(\mathcal{E}_1) \to 0$. Moreover, the probability of no coverage $\mathbb{P}(\mathcal{E}_2)$ also goes to zero with this choice of N. Hence, the probability of error in reconstruction $\mathbb{P}(\mathcal{E})$ also goes to zero. This implies that the minimum number of reads required to meet the desired reconstruction error probability of at most ϵ satisfies

$$N_{\min}(\epsilon, G, L) \leq \frac{G}{L}\ln(GL^3)$$

for sufficiently large G and L with $L/\ln G = \bar{L}$. Writing $\ln(GL^3) = \ln G + 3\ln L$ and noting that $\ln L/\ln G \to 0$ in our scaling, in the limit, we have

$$\limsup_{L,G\to\infty, L/\ln G=\bar{L}} = \frac{N_{\min}(\epsilon, G, L)}{G/\bar{L}} \leq 1.$$

Combining this with Lemma 3, we get

$$\limsup_{L,G\to\infty, L/\ln G=\bar{L}} = \frac{N_{\min}(\epsilon, G, L)}{N_{\text{cov}}(\epsilon, G, L)} \leq 1.$$

But since $N_{\min}(\epsilon, G, L) \geq N_{\text{cov}}(\epsilon, G, L)$, it follows that

$$\lim_{L,G\to\infty, L/\ln G=\bar{L}} = \frac{N_{\min}(\epsilon, G, L)}{N_{\text{cov}}(\epsilon, G, L)} = 1,$$

completing the proof.

Appendix B
Proof of Theorem 6

The stationary distribution of the source is denoted by $\mathbf{p} = (p_1, p_2, p_3, p_4)^t$. Since $\bar{Q}$ has positive entries, the Perron–Frobenius theorem implies that its largest eigenvalue $\rho_{\max}(\bar{Q})$ is real and positive and the corresponding eigenvector $\boldsymbol{\pi}$ has positive components. We have

$$\begin{aligned}
&\sum_{i_1 i_2 \ldots i_\ell} q_{i_2 i_1}^2 q_{i_3 i_2}^2 \cdots q_{i_\ell i_{\ell-1}}^2 \\
&\leq \max_{i_1 \in \{1,2,3,4\}} \left\{\frac{1}{\pi_{i_1}}\right\} \sum_{i_1 i_2 \ldots i_\ell} \pi_{i_1} q_{i_2 i_1}^2 q_{i_3 i_2}^2 \cdots q_{i_\ell i_{\ell-1}}^2 \quad (26)\\
&= \max_{i\in\{1,2,3,4\}} \left\{\frac{1}{\pi_i}\right\} \|\bar{Q}^{\ell-1}\boldsymbol{\pi}\|_1 \\
&= \max_{i\in\{1,2,3,4\}} \left\{\frac{1}{\pi_i}\right\} \left(\rho_{\max}(\bar{Q})\right)^{\ell-1} \|\boldsymbol{\pi}\|_1 \\
&= \gamma \left(\rho_{\max}(\bar{Q})\right)^{\ell} \quad (27)
\end{aligned}$$

where $\gamma = \max_{i\in\{1,2,3,4\}} \left\{\frac{\|\boldsymbol{\pi}\|_1}{\pi_i \rho_{\max}(\bar{Q})}\right\}$. This completes the proof.

1) Proof of Lemma 7: In [2], Arratia *et al.* showed that interleaved repeats are the dominant term preventing reconstruction. They also used Poisson approximation to derive bounds on the event that $\mathbf{S}$ is recoverable from its L-spectrum. We take a similar approach to obtain an upper bound under the Markov model. First, we state the following theorem regarding Poisson approximation of the sum of indicator random variables; cf., [1].

Theorem 12 (Chen–Stein Poisson Approximation): Let $W = \sum_{\alpha\in I} \chi_\alpha$, where χ_αs are indicator random variables for some index set I. For each α, $B_\alpha \subseteq I$ denotes the set of indices where χ_α is independent from the σ-algebra generated by all χ_β with $\beta \in I - B_\alpha$. Let

$$b_1 = \sum_{\alpha\in I}\sum_{\beta\in B_\alpha} \mathbb{E}[\chi_\alpha]\mathbb{E}[\chi_\beta], \quad (28)$$

$$b_2 = \sum_{\alpha\in I}\sum_{\beta\in B_\alpha, \beta\neq\alpha} \mathbb{E}[\chi_\alpha\chi_\beta]. \quad (29)$$

Then

$$d_{\text{TV}}(W, W') \leq \frac{1-e^{-\theta}}{\theta}(b_1+b_2), \quad (30)$$

where $\theta = \mathbb{E}[W]$ and $d_{\text{TV}}(W, W')$ is the total variation distance[4] between W and Poisson random variable W' with the same mean.

2) Proof of Lemma 7: Let $\mathcal{U}$ denote the event that there is no two pairs of interleaved repeats in the DNA sequence. Given the presence of k repeats in $\mathbf{S}$, the probability of $\mathcal{U}$ can be found by using the Catalan numbers [2]. This probability is $2^k/(k+1)!$. If Z denotes the random variable indicating the number of repeats in the DNA sequence, we obtain,

$$\mathbb{P}(\mathcal{U}) = \sum_k \frac{2^k}{(k+1)!}\mathbb{P}(Z=k).$$

To approximate $\mathbb{P}(\mathcal{U})$, we partition the sequence as

$$\mathbf{S} = S_1\mathbf{X}_1 S_{L+2}\mathbf{X}_2 S_{2(L+1)+1}\mathbf{X}_3 \ldots S_{(K-1)(L+1)+1}\mathbf{X}_K$$

where $\mathbf{X}_i = \mathbf{S}[(i-1)(L+1)+2, i(L+1)]$ and $K = \frac{G}{L+1}$. Each $\mathbf{X}_i$ has length L and will be denoted by $\mathbf{X}_i = X_{i1}\cdots X_{iL}$. We write $\mathbf{X}_i \sim \mathbf{X}_j$ with $i \neq j$ to mean $X_{i1} \neq X_{j1}$ and $X_{ik} = X_{jk}$ for $2 \leq k \leq L$. In other words, $\mathbf{X}_i \sim \mathbf{X}_j$ means that there is a repeat of length at least $L-1$ starting from locations $(i-1)(L+1)+3$ and $(j-1)(L+1)+3$ in the DNA sequence and the repeat cannot be extended from left. The requirement $X_{i1} \neq X_{j1}$ is due to the fact that allowing left extension ruins accuracy of Poisson approximation as repeats appear in clumps.

Let $I = \{(i,j) | 1 \leq i < j \leq K\}$. Let χ_α with $\alpha \in I$ denote the indicator random variable for a repeat at $\alpha = (i,j)$, i.e., $\chi_\alpha = \mathbf{1}(\mathbf{X}_i \sim \mathbf{X}_j)$. Let $W = \sum_{\alpha\in I} X_\alpha$. Clearly,

$$\mathbb{P}(\mathcal{U}) \leq \sum_k \frac{2^k}{(k+1)!}\mathbb{P}(W=k).$$

Letting $\mathbf{Y} = S_1 S_{L+2}\ldots S_{(K-1)(L+1)+1}$, we obtain

$$\mathbb{P}(\mathcal{U}) \leq \sum_{\mathbf{Y}}\sum_k \frac{2^k}{(k+1)!}\mathbb{P}(W=k|\mathbf{Y})\mathbb{P}(\mathbf{Y}).$$

[4]The total variation distance between two distributions W and W' is defined by $d_{\text{TV}}(W,W') = \sup_{A\in\mathcal{F}} |\mathbb{P}_W(A) - \mathbb{P}_{W'}(A)|$, where $\mathcal{F}$ is the σ-algebra defined for W and W'

For any $\mathbf{Y}$, let ϵ be the total variation distance between W and its corresponding Poisson distribution W' with mean $\theta_{\mathbf{Y}} = \mathbb{E}[W|\mathbf{Y}]$. Then, we obtain

$$\begin{aligned}\mathbb{P}(\mathcal{U}) &\le \sum_{\mathbf{Y}} \left(\epsilon + e^{-\theta_{\mathbf{Y}}} \sum_{k=0}^{\infty} \frac{(2\theta_{\mathbf{Y}})^k}{k!(k+1)!}\right) \mathbb{P}(\mathbf{Y}) \\ &= \epsilon + \sum_{\mathbf{Y}} e^{-\theta_{\mathbf{Y}}} \sum_{k=0}^{\infty} \frac{(2\theta_{\mathbf{Y}})^k}{k!(k+1)!} \mathbb{P}(\mathbf{Y}) \\ &\le \epsilon + \sum_{\mathbf{Y}} e^{-\theta_{\mathbf{Y}}} \sum_{k=0}^{\infty} \left(\frac{(\sqrt{2\theta_{\mathbf{Y}}})^k}{k!}\right)^2 \mathbb{P}(\mathbf{Y}) \\ &\le \epsilon + \sum_{\mathbf{Y}} e^{-\theta_{\mathbf{Y}}} \left(\sum_{k=0}^{\infty} \frac{(\sqrt{2\theta_{\mathbf{Y}}})^k}{k!}\right)^2 \mathbb{P}(\mathbf{Y}) \\ &\le \epsilon + \sum_{\mathbf{Y}} e^{-\theta_{\mathbf{Y}} + 2\sqrt{2\theta_{\mathbf{Y}}}} \mathbb{P}(\mathbf{Y}).\end{aligned}$$

We assume $\theta_{\mathbf{Y}} \ge 8$ for all $\mathbf{Y}$ and let $\theta = \min_{\mathbf{Y}} \theta_{\mathbf{Y}}$. For this region, the exponential factor within the summation is monotonically decreasing and

$$\mathbb{P}(\mathcal{U}) \le \epsilon + e^{-\theta + 2\sqrt{2\theta}}. \tag{31}$$

To calculate the bound, we need to obtain an upper bound for ϵ and a lower bound for θ. We start with the lower bound on θ. From Markov property and for a given $\alpha = (i, j)$,

$$\begin{aligned}\mathbb{E}[\chi_\alpha|\mathbf{Y}] &= \sum_{i_1 i_2 \ldots i_L} \mathbb{P}(X_{i1} \ne X_{j1}|\mathbf{Y}) q_{i_2 i_1}^2 q_{i_3 i_2}^2 \ldots q_{i_L i_{L-1}}^2 \\ &\ge \min\left\{\frac{\mathbb{P}(X_{i1} \ne X_{j1}|\mathbf{Y})}{\pi_{i_1}}\right\} \sum_{i_1 i_2 \ldots i_L} \pi_{i_1} q_{i_2 i_1}^2 \ldots q_{i_L i_{L-1}}^2 \\ &= \zeta \left(\rho_{\max}(\bar{Q})\right)^L\end{aligned}$$

where $\zeta = \min\left\{\frac{\mathbb{P}(X_{i1} \ne X_{j1}|\mathbf{Y})}{\pi_{i_1} \rho_{\max}(\bar{Q})}\right\}$. Therefore,

$$\theta_{\mathbf{Y}} = \sum_{\alpha \in I} \mathbb{E}[\chi_\alpha|Y] \ge \binom{K}{2} \zeta \left(\rho_{\max}(\bar{Q})\right)^L = \theta. \tag{32}$$

To bound ϵ, we make use of the Chen–Stein method. Let $B_\alpha = \{(i', j') \in I | i' = i \text{ or } j' = j\}$. Note that B_α has cardinality $2K - 3$. Since given $\mathbf{Y}$, χ_α is independent of the sigma-algebra generated by all χ_β, $\beta \in I - B_\alpha$, we can use Theorem 12 to obtain

$$d_{\mathrm{TV}}(W, W'|\mathbf{Y}) \le \frac{b_1 + b_2}{\theta_{\mathbf{Y}}}, \tag{33}$$

where b_1 and b_2 are defined in (28) and (29), respectively. Since $\mathbb{E}[X_\alpha X_\beta] = \mathbb{E}[X_\alpha]\mathbb{E}[X_\beta]$ for all $\alpha \ne \beta \in B_\alpha$, we can conclude that $b_2 \le b_1$. Therefore,

$$d_{\mathrm{TV}}(W, W'|\mathbf{Y}) \le \frac{2b_1}{\theta_{\mathbf{Y}}}.$$

Since $\theta \le \theta_{\mathbf{Y}}$,

$$d_{\mathrm{TV}}(W, W'|\mathbf{Y}) \le \frac{2b_1}{\theta}.$$

In order to compute b_1, we need an upper bound on $\mathbb{E}[\chi_\alpha|\mathbf{Y}]$. By using (27), we obtain

$$\begin{aligned}\mathbb{E}[\chi_\alpha|\mathbf{Y}] &= \sum_{i_1 i_2 \ldots i_L} \mathbb{P}(X_{i1} \ne X_{j1}|\mathbf{Y}) q_{i_2 i_1}^2 q_{i_3 i_2}^2 \ldots q_{i_L i_{L-1}}^2 \\ &\le \sum_{i_1 i_2 \ldots i_L} q_{i_2 i_1}^2 q_{i_3 i_2}^2 \ldots q_{i_L i_{L-1}}^2 \\ &\le \gamma \left(\rho_{\max}(\bar{Q})\right)^L.\end{aligned}$$

Hence,

$$\begin{aligned}b_1 &= \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} \mathbb{E}[\chi_\alpha|\mathbf{Y}]\mathbb{E}[\chi_\beta|\mathbf{Y}], \\ &\le (2K-3)\binom{K}{2}\gamma^2 \left(\rho_{\max}(\bar{Q})\right)^{2L} \\ &= \frac{\gamma^2\theta^2(2K-3)}{\zeta^2\binom{K}{2}} \\ &\le \frac{4\gamma^2\theta^2}{\zeta^2 K}.\end{aligned}$$

Using the bound for b_1, we have the following bound for the total variation distance:

$$d_{\mathrm{TV}}(W, W'|\mathbf{Y}) \le \frac{8\gamma^2\theta}{\zeta^2 K}.$$

Form the above inequality, we can choose $\epsilon = \frac{8\gamma^2\theta}{\zeta^2 K}$. Substituting into (31) yields

$$\mathbb{P}(\mathcal{U}) \le \frac{8\gamma^2\theta}{\zeta^2 K} + e^{-\theta + 2\sqrt{2\theta}}. \tag{34}$$

From the definition of θ in (32), we have

$$\theta = \frac{\zeta(K-1)(L+1)^2}{2K} G^{2 - \bar{L}\ln\left(\frac{1}{\rho_{\max}(\bar{Q})}\right)}.$$

Therefore, if $2 > \bar{L}\ln\left(\frac{1}{\rho_{\max}(\bar{Q})}\right)$, then θ and $\frac{\theta}{K}$ go, respectively, to infinity and zero exponentially fast. Since the right-hand side of (34) approaches zero, we can conclude that with probability $1 - o(1)$, there exists a two pairs of interleaved repeats in the sequence. This completes the proof.

3) Proof of Lemma 8: The proof follows closely from that of the i.i.d. model. In fact, we only need to replace Lemma 11 with the following lemma.

Lemma 13: For any distinct substrings $\mathbf{X}$ and $\mathbf{Y}$ of length ℓ of the Markov DNA sequence:

1) If the strings have no physical overlap, the probability that they are identical is bounded above by $\gamma e^{\ell \ln\left(\rho_{\max}(\bar{Q})\right)}$.
2) If the strings have physical overlap, the probability that they are identical is bounded above by $\sqrt{\gamma} e^{\ell \ln\left(\rho_{\max}(\bar{Q})\right)/2}$.

Proof: For the first part, the Markov property gives

$$\begin{aligned}\mathbb{P}(\mathbf{X} = \mathbf{Y}) &= \sum_{i_1 i_2 \ldots i_\ell} \mathbb{P}(X_1 = Y_1 = i_1) q_{i_2 i_1}^2 q_{i_3 i_2}^2 \ldots q_{i_\ell i_{\ell-1}}^2 \\ &\le \sum_{i_1 i_2 \ldots i_\ell} q_{i_2 i_1}^2 q_{i_3 i_2}^2 \ldots q_{i_\ell i_{\ell-1}}^2 \\ &\le \gamma \left(\rho_{\max}(\bar{Q})\right)^\ell,\end{aligned}$$

where the last line follows from (27).

We now prove the second part. Without loss of generality, we assume that $\mathbf{X} = \mathbf{S}[1,\ell]$ and $\mathbf{Y} = \mathbf{S}[\ell-k+1, 2\ell-k]$ for some $k \in \{1,\dots,\ell-1\}$. Let q and r be the quotient and remainder of dividing $2\ell-k$ by $\ell-k$. From decomposition of $\mathbf{S}[1,2\ell-k]$ as $\mathbf{U}_1\mathbf{U}_2\dots\mathbf{U}_q\mathbf{V}$, where $|\mathbf{U}_i| = \ell-k$ for all $i \in \{1,\dots,q\}$ and $|\mathbf{V}| = r$, one can deduce that $\mathbf{X} = \mathbf{Y}$ if and only if $\mathbf{U}_i = S_1S_2\dots S_{\ell-k}$ for all $i \in \{1,\dots,q\}$ and $\mathbf{V} = S_1S_2\dots S_r$. Hence, we have

$$\begin{aligned}
&\mathbb{P}(\mathbf{X}=\mathbf{Y}) = \mathbb{P}(\mathbf{S}[1,2\ell-k] = \mathbf{U}\mathbf{U}\dots\mathbf{U}\mathbf{V}) \\
&= \sum_{i_1 i_2 \dots i_{\ell-k}} p_{i_1} \left(q_{i_2 i_1} q_{i_3 i_2} \dots q_{i_1 i_{\ell-k}}\right)^q \left(q_{i_2 i_1} q_{i_3 i_2} \dots q_{i_r i_{r-1}}\right) \\
&\overset{(a)}{\le} \sqrt{\sum_{i_1} p_{i_1}^2} \times \\
&\quad \sqrt{\sum_{i_1 i_2 \dots i_{\ell-k}} \left(q_{i_2 i_1}^2 q_{i_3 i_2}^2 \dots q_{i_1 i_{\ell-k}}^2\right)^q \left(q_{i_2 i_1}^2 q_{i_3 i_2}^2 \dots q_{i_r i_{r-1}}^2\right)} \\
&\overset{(b)}{\le} \sqrt{\sum_{i_1 i_2 \dots i_{\ell-k}} \left(q_{i_2 i_1}^2 q_{i_3 i_2}^2 \dots q_{i_1 i_{\ell-k}}^2\right)^q \left(q_{i_2 i_1}^2 q_{i_3 i_2}^2 \dots q_{i_r i_{r-1}}^2\right)} \\
&\overset{(c)}{\le} \sqrt{\sum_{i_1 i_2 \dots i_{2\ell-k}} q_{i_2 i_1}^2 q_{i_3 i_2}^2 \dots q_{i_{2\ell-k} i_{2\ell-k-1}}^2} \\
&\overset{(d)}{\le} \sqrt{\gamma \left(\rho_{\max}(\bar{Q})\right)^{2\ell-k}} \\
&= \sqrt{\gamma} \left(\rho_{\max}(\bar{Q})\right)^{\ell-\frac{k}{2}} \\
&\overset{(e)}{\le} \sqrt{\gamma} \left(\rho_{\max}(\bar{Q})\right)^{\frac{\ell}{2}},
\end{aligned}$$

where (a) follows from the Cauchy–Schwarz inequality and (b) follows from the fact that $\sum_i P_i^2 \le 1$. In (c), some extra terms are added to the inequality. (d) comes from (27), and finally, (e) comes from the fact that $k < \ell$ and $\rho_{\max}(\bar{Q}) \le 1$. ■

Appendix C
Proof of Theorem 9

As explained in Section IV-A, the criterion for overlap scoring is based on the MAP rule for deciding between two hypotheses: H_0 and H_1. The null hypothesis H_0 indicates that two reads are from the same physical source subsequence. Formally, we say that two reads $\mathbf{R}_i$ and $\mathbf{R}_j$ have overlap score $W_{ij} = w$ if w is the longest suffix of $\mathbf{R}_i$ and prefix of $\mathbf{R}_j$ passing the criterion (12).

Let $f(\ell) = (1+\ell)^{|\mathcal{X}|}$, where $|\mathcal{X}|$ is the cardinality of the channel's output symbols. The following theorem is a standard result in the hypothesis testing problem; cf., [4, Ch. 11.7].

Theorem 14: Let $\mathbf{X}$ and $\mathbf{Y}$ be two random sequences of length ℓ. For the given hypotheses H_0 and H_1 and their corresponding MAP rule (12),

$$\mathbb{P}(H_0|H_1) \le f(\ell) e^{-\ell D(P_\mu || P_X \cdot P_Y)}$$

and

$$\mathbb{P}(H_1|H_0) \le f(\ell) e^{-\ell D(P_\mu || P_{X,Y})},$$

where

$$P_\mu(x,y) := \frac{[P_{X,Y}(x,y)]^\mu [P_X(x)P_Y(y)]^{1-\mu}}{\sum_{a,b}[P_{X,Y}(a,b)]^\mu [P_X(a)P_Y(b)]^{1-\mu}}$$

and μ is the solution of

$$D(P_\mu \| P_X \cdot P_Y) - D(P_\mu \| P_{X,Y}) = \theta.$$

Paralleling the proof of the noiseless case, we first prove the following lemma concerning erroneous merging due to imposter reads.

Lemma 15 (False Alarm): For any distinct ℓ-mers $\mathbf{X}$ and $\mathbf{Y}$ from the set of reads:

1) If the two ℓ-mers have no physical overlap, the probability that H_0 is accepted is
$$f(\ell) e^{-\ell D(P_\mu \| P_X \cdot P_Y)}. \tag{35}$$
2) If the two ℓ-mers have physical overlap, the probability that H_0 is accepted is
$$\gamma f(\ell) e^{-\ell D(P_\mu \| P_X \cdot P_Y)/2}, \tag{36}$$
where γ is a constant.

Proof: The proof of the first statement is an immediate consequence of Theorem 14.

We now turn to the second statement. We only consider the case $\ell = 2k$, and note that the more general statement can be deduced easily by following similar steps. Let $\chi_j = \ln \frac{P(x_j,y_j)}{P(x_j)P(y_j)}$. Since χ_js are not independent, we cannot directly use Theorem 14 to compute $\mathbb{P}\left(\sum_{j=1}^{\ell} \chi_j \ge \ell\theta\right)$. However, we claim that χ_js can be partitioned into two disjoint sets J_1 and J_2 of the same size, where the χ_js within each set are independent. Assuming the claim,

$$\begin{aligned}
\mathbb{P}\left(\sum_{j=1}^{\ell} \chi_j \ge \ell\theta\right) &= \mathbb{P}\left(\sum_{j\in J_1} \chi_j + \sum_{j\in J_1} \chi_j \ge \ell\theta\right) \\
&\overset{(a)}{\le} \mathbb{P}\left(\sum_{j\in J_1} \chi_j \ge \frac{\ell}{2}\theta\right) + \mathbb{P}\left(\sum_{j\in J_2} \chi_j \ge \frac{\ell}{2}\theta\right) \\
&\le 2\mathbb{P}\left(\sum_{j\in J_1} \chi_j \ge \frac{\ell}{2}\theta\right),
\end{aligned}$$

where (a) follows from the union bound. Since $|J_1| = \frac{\ell}{2}$, one can use Theorem 14 to show (36).

It remains to prove the claim. To this end, let k be the amount of physical overlap between $\mathbf{X}$ and $\mathbf{Y}$. Without loss of generality, we assume that $S_1S_2\cdots S_{2\ell+k}$ is the shared DNA sequence. Let q and r be the quotient and remainder of ℓ divided by $2(\ell-k)$, i.e., $\ell = 2q(\ell-k)+r$ where $0 \le r < 2(\ell-k)$. Since ℓ is even, r is even. Let J_1 be the set of indices j where either $(j \bmod 2(\ell-k)) \in \{0,1,\dots,\ell-k-1\}$ for $j \in \{1,\dots,2q(\ell-k)\}$ or $j \in \{2q(\ell-k)+1,\dots,2q(\ell-k)+\frac{\ell}{2}\}$. We claim that the random variables χ_js with $j \in J_1$ are independent. We observe that χ_j depends only on s_j and $s_{j+(\ell-k)}$. Consider two indices $j_1 < j_2 \in J_1$. The pairs $(s_{j_1}, s_{j_1+(\ell-k)})$ and $(s_{j_2}, s_{j_2+(\ell-k)})$ are disjoint iff $j_1 + (\ell-k) \ne j_2$. By the

construction of J_1, one can show that $j_1+(\ell-k)\neq j_2$ for any $j_1<j_2\in J_1$. Hence, χ_js with $j\in J_1$ are independent. A similar argument shows χ_js with $j\in J_2=\{1,\ldots,2\ell-k\}-J_1$ are independent. This completes the proof. ■

Due to noise, two physically overlapping reads may not pass the criterion. To deal with this event, we state the following lemma.

Lemma 16 (Missed Detection): Let $\mathbf{X}$ and $\mathbf{Y}$ be two distinct ℓ-mers from the same physical location. The probability that H_1 is accepted is bounded by

$$f(\ell)e^{-\ell D(P_\mu\|P_{X,Y})}.$$

Proof: This is an immediate consequence of Theorem 14. ■

Proof of Theorems 9: Similar to the proof of achievability result in the noiseless case, we decompose the error event $\mathcal{E}$ into $\mathcal{E}_1\cup\mathcal{E}_2$, where $\mathcal{E}_1$ is the event that some read is merged incorrectly and $\mathcal{E}_2$ is the event that the DNA sequence is not covered by the reads. The probability of the second event, similar to the noiseless case, goes to zero exponentially fast if $R>\bar{L}$. We only need to compute $\mathbb{P}(\mathcal{E}_1)$. Again, $\mathcal{E}_1$ can be decomposed as $\mathcal{E}_1=\cup_\ell\mathcal{A}_\ell$, where $\mathcal{A}_\ell$ is the event that the first error in merging occurs at stage ℓ. Moreover,

$$\mathcal{A}_\ell\subseteq\mathcal{B}_\ell\cup\mathcal{C}_\ell, \tag{37}$$

where

$$\mathcal{B}_\ell\triangleq\{\mathbf{R}_j\neq\mathbf{R}_i^s, U_j\leq\ell, V_i\leq\ell, W_{ij}=\ell \text{ for some } i\neq j.\} \tag{38}$$

$$\mathcal{C}_\ell\triangleq\{\mathbf{R}_j=\mathbf{R}_i^s, U_j=V_i\neq\ell, W_{ij}=\ell \text{ for some } i\neq j.\} \tag{39}$$

Note that here the definition of $\mathcal{C}_\ell$ is different from that of (19) as for the noiseless reads the overlap score is never less than the physical overlap. However, in the noisy reads, there is a chance for observing this event due to misdetection.

The analysis of $\mathcal{B}_\ell$ follows closely from that of the noiseless case. In fact, using Lemma 15 which is a counterpart of Lemma 11 and following similar steps in calculation of $\mathbb{P}(\mathcal{B}_\ell)$ in the noiseless case, one can obtain

$$\mathbb{P}(\mathcal{B}_\ell)\leq f(\ell)(q_\ell^2+2\gamma\lambda L q_\ell), \tag{40}$$

where

$$q_\ell=\lambda G e^{-\lambda(L-\ell)(1-2/N)}e^{-\ell D(P_\mu\|P_X\cdot P_Y)/2}. \tag{41}$$

To compute $\mathbb{P}(\mathcal{C}_\ell)$, we note that $\mathcal{C}_\ell\subseteq\cup_i\mathcal{C}_\ell^i$, where

$$\mathcal{C}_\ell^i\triangleq\{V_i=\ell, W(\mathbf{R}_i,\mathbf{R}_i^s)\neq\ell\}.$$

Applying the union bound and considering the fact that $\mathcal{C}_\ell^i$'s are equiprobable yields

$$\mathcal{C}_\ell\leq N\mathbb{P}(\mathcal{C}_\ell^1).$$

Hence,

$$\mathbb{P}(\mathcal{C}_\ell)\leq N\mathbb{P}(V_i<\ell)\times (\mathbb{P}(W(\mathbf{R}_i,\mathbf{R}_i^s)>\ell|V_i=\ell)+\mathbb{P}(W(\mathbf{R}_i,\mathbf{R}_i^s)<\ell|V_i=\ell)).$$

Using Lemma 15 part 2 and Lemma 16 yields

$$\begin{aligned}\mathbb{P}(\mathcal{C}_\ell)&\leq\lambda G f(\ell)\left(\gamma e^{-\ell D(P_\mu\|P_X\cdot P_Y)/2}+e^{-\ell D(P_\mu\|P_{X,Y})}\right)\times e^{-\lambda(L-\ell)(1-1/N)}\\&\leq f(\ell)\left(\gamma q_\ell+q'_\ell\right),\end{aligned}$$

where

$$q'_\ell=\lambda G e^{-\ell D(P_\mu\|P_{X,Y})}e^{-\lambda(L-\ell)(1-1/N)}.$$

Combining all the terms, we obtain

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1)&\leq\sum_{\ell=0}^{L}\mathbb{P}(\mathcal{B}_\ell)+\mathbb{P}(\mathcal{C}_\ell)\\&\leq\sum_{\ell=0}^{L}f(\ell)(q_\ell^2+\gamma(2\lambda L+1)q_\ell+q'_\ell).\end{aligned}$$

To show that that $\mathbb{P}(\mathcal{E}_1)\to 0$, it is sufficient to argue that q_0, q'_0, q_L, and q'_L go to zero exponentially in L. Considering first q_0 and q'_0, they vanish exponentially in L if $N>G\ln G/L$ which implies $c_{\min}(\bar{L})=1$. The terms q_L and q'_L vanish exponentially in L if

$$\bar{L}>\frac{2}{\min(2D(P_\mu\|P_{X,Y}),D(P_\mu\|P_X\cdot P_Y))}.$$

Since $\mathbb{P}(\mathcal{E}_1)=o(1)$ and $\mathbb{P}(\mathcal{E}_2)=o(1)$ for any choice of θ, one can optimize over θ to obtain the result given in the theorem. This completes the proof.

ACKNOWLEDGMENT

The authors would like to thank Prof. Y. Song for discussions in the early stage of this project, and R. Pedarsani for discussions about the complexity of the greedy algorithm.

REFERENCES

[1] R. Arratia, L. Goldstein, and L. Gordon, "Poisson approximation and the Chen-Stein method," *Statist. Sci.*, vol. 5, no. 4, pp. 403–434, 1990.
[2] R. Arratia, D. Martin, G. Reinert, and M. S. Waterman, "Poisson process approximation for sequence repeats, and sequencing by hybridization," *J. Comput. Biol.*, vol. 3, pp. 425–463, 1996.
[3] G. Bresler, M. Bresler, and D. N. C. Tse, Optimal assembly for high throughput shotgun sequencing 2013, arXiv preprint arXiv:1301.0068.
[4] T. M. Covar and J. A. Thomas, *Elements of Information Theory*. Oxford, U.K.: Oxford Univ. Press, 2006.
[5] C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "SHARCGS: A fast and highly accurate short-read assembly algorithm for de novo genomic sequencing," *Genome Res.*, vol. 17, pp. 1697–1706, 2007.
[6] M. Dyer, A. Frieze, and S. Suen, "The probability of unique solutions of sequencing by hybridization," *J. Comput. Biol.*, vol. 1, no. 2, pp. 105–110, 1994.
[7] A. Frieze and W. Szpankowski, "Greedy algorithms for the shortest common superstring that are asymptotically optimal," *Algorithmica*, vol. 21, no. 1, pp. 21–36, 1998.
[8] J. K. Gallant, "String compression algorithms," Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 1982.

[9] X. Huang and A. Madan, "CAP3: A DNA sequence assembly program," *Genome Res.*, vol. 9, no. 9, pp. 868–877, 1999.
[10] M. S. Waterman and R. M. Idury, "A new algorithm for DNA sequence assembly," *J. Comput. Biol.*, vol. 2, pp. 291–306, 1995.
[11] W. R. Jeck, J. A. Reinhardt, D. A. Baltrus, M. T. Hickenbotham, V. Magrini, E. R. Mardis, J. L. Dangl, and C. D. Jones, "Extending assembly of short DNA sequences to handle error," *Bioinformatics*, vol. 23, pp. 2942–2944, 2007.
[12] H. Kaplan and N. Shafrir, "The greedy algorithm for shortest superstrings," *Inf. Process. Lett.*, vol. 93, no. 1, pp. 13–17, 2005.
[13] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: A mathematical analysis," *Genomics*, vol. 2, no. 3, pp. 231–239, 1988.
[14] M. Li, "Towards a DNA sequencing theory (learning a string)," *Found. Comput. Sci.*, vol. 1, pp. 125–134, Oct. 1990.
[15] B. Ma, "Why greed works for shortest common superstring problem," *Theor. Comput. Sci.*, vol. 410, no. 51, pp. 5374–5381, 2009.
[16] P. Medvedev and M. Brudno, "Maximum likelihood genome assembly," *J. Comput. Biol.*, vol. 16, no. 8, pp. 1101–1116, 2009.
[17] J. Miller, S. Koren, and G. Sutton, "Assembly algorithms for next-generation sequencing data," *Genomics*, vol. 95, pp. 315–327, 2010.
[18] Human Genome Sequence Quality Standards NIH National Human Genome Research Institute, Dec. 2012 [Online]. Available: http://www.genome.gov/10000923
[19] K. Paszkiewicz and D. J. Studholme, "De novo assembly of short sequence reads," *Brief. Bioinformat.*, vol. 11, no. 5, pp. 457–472, 2010.
[20] P. A. Pevzner, H. Tang, and M. S. Waterman, "An Eulerian path approach to DNA fragment assembly," *Proc. Nat. Acad. Sci. USA*, vol. 98, pp. 9748–9753, 2001.
[21] M. Pop, "Genome assembly reborn: Recent computational challenges," *Brief. Bioinformat.*, vol. 10, no. 4, pp. 354–366, 2009.
[22] Z. Rached, F. Alajaji, and L. L. Campbell, "Renyi's divergence and entropy rates for finite alphabet Markov sources," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1553–1561, May 2001.
[23] S. Batzoglou *et al.*, "Arachne: A whole-genome shotgun assembler," *Genome Res.*, vol. 12, pp. 177–189, 2002.
[24] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc. Nat. Acad. Sci. USA*, vol. 74, no. 12, pp. 5463–5467, 1977.
[25] G. G. Sutton, O. White, M. D. Adams, and A. Kerlavage, "TIGR Assembler: A new tool for assembling large shotgun sequencing projects," *Genome Sci. Technol.*, vol. 1, pp. 9–19, 1995.
[26] J. S. Turner, "Approximation algorithms for the shortest common superstring problem," *Inf. Comput.*, vol. 83, no. 1, pp. 1–20, 1989.
[27] E. Ukkonen, "Approximate string matching with q-grams and maximal matches," *Theor. Comput. Sci.*, vol. 92, no. 1, pp. 191–211, 1992.
[28] E. Ukkonen, "A linear-time algorithm for finding approximate shortest common superstrings," *Algorithmica*, vol. 5, pp. 313–323, 1990.
[29] R. L. Warren, G. G. Sutton, S. J. Jones, and R. A. Holt, "Assembling millions of short DNA sequences using SSAKE," *Bioinformatics*, vol. 23, pp. 500–501, 2007.
[30] *DNA Sequencing Theory—Wikipedia, The Free Encyclopedia,* Wikipedia, 2012.
[31] *Sequence Assembly—Wikipedia, The Free Encyclopedia,* Wikipedia, 2012.
[32] A. S. Motahari, K. Ramchandran, D. N. C. Tse, and N. Ma, "Optimal DNA shotgun sequencing: Noisy reads are as good as noiseless reads," arXiv preprint arXiv:1304.2798 (2013).

Abolfazl S. Motahari received the B.Sc. degree from the Iran University of Science and Technology (IUST), Tehran, in 1999, the M.Sc. degree from Sharif University of Technology, Tehran, in 2001, and the Ph.D. degree from University of Waterloo, Waterloo, Canada, in 2009, all in electrical engineering. From August 2000 to August 2001, he was a Research Scientist with the Advanced Communication Science Research Laboratory, Iran Telecommunication Research Center (ITRC), Tehran. From October 2009 to September 2010, he was a Postdoctoral Fellow with the University of Waterloo, Waterloo. Currently, he is a Postdoctoral Fellow with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley. His research interests include multiuser information theory and Bioinformatics. He received several awards including Natural Science and Engineering Research Council of Canada (NSERC) Post-Doctoral Fellowship.

Guy Bresler (S'07) received the B.S. degree in electrical and computer engineering and the M.S. degree in mathematics from the University of Illinois at Urbana-Champaign, both in 2006. He received the Ph.D. degree in 2012 from the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. Dr. Bresler is currently a postdoctoral associate at the Laboratory for Information and Decision Systems in the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology. Guy is the recipient of an NSF Graduate Research Fellowship, a Vodafone Graduate Fellowship, the Barry M. Goldwater Scholarship, a Vodafone Undergraduate Scholarship, the E. C. Jordan Award from the ECE department at UIUC, and the Roberto Padovani Scholarship from Qualcomm.

David N. C. Tse (M'96–SM'07–F'09) received the B.A.Sc. degree in systems design engineering from University of Waterloo in 1989, and the M.S. and Ph.D. degrees in electrical engineering from Massachusetts Institute of Technology in 1991 and 1994 respectively. From 1994 to 1995, he was a postdoctoral member of technical staff at A.T. & T. Bell Laboratories. Since 1995, he has been at the Department of Electrical Engineering and Computer Sciences in the University of California at Berkeley, where he is currently a Professor. He received a 1967 NSERC graduate fellowship from the government of Canada in 1989, a 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, a 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. He was an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY from 2001 to 2003, the Technical Program co-chair in 2004 and the General co-chair in 2015 of the International Symposium on Information Theory. He is a coauthor, with Pramod Viswanath, of the text "Fundamentals of Wireless Communication", which has been used in over 60 institutions around the world.