ISIT'2013 1569720421

Reference-Based DNA Shotgun Sequencing: Information Theoretic Limits

Soheil Mohajer, Abolfazl Motahari, and David Tse
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
{mohajer, motahari, dtse}@eecs.berkeley.edu

***Abstract*—The reference-based DNA shotgun assembly problem is studied from an information-theoretic point of view. The entire sequence has to be assembled based on a reference sequence which is a *noisy* version of the desired one, and a set of short reads sampled from the desired sequence. Two necessary conditions on the underlying parameters for reconstruction are obtained. A reference-based assembly algorithm is proposed, and it is shown that under these conditions the algorithm can reconstruct the sequence with high probability.**

I. Introduction

Deoxyribonucleic acid (DNA) is a molecule containing the genetic instructions used in the development and functioning of all known living organisms. It is a rather long (roughly 3 billion base pairs for humans) sequence of nucleotides (A,C,G, and T). DNA sequencing, referring to the process of determining the precise order of nucleotides within a DNA molecule, is the basic workhorse of modern day biology and medicine. Here the main challenge is that the sequence is observed only through its randomly located short fragments called *reads*. *De novo* assembly, the process of putting the reads in the right order and recover the original sequence is a hard and time-consuming task.

As the cost of DNA sequencing falls and new full genome sequencing technologies emerge, more genome sequences continue to be generated. This has revealed a high level of similarity between individual genome, namely, any two people share more than 99% of their DNAs [1]. This allows the use of a reference genome, assembled by scientists as a representative example of a species' set of genes, as a guide on which new genomes are built. *Reference-based* shotgun sequencing, referring to the reconstruction of DNA based on both a reference and set of reads, allows the assembly to be performed much more quickly and cheaply compared to de novo assembly.

The variations across genomes of individuals include single nucleotide polymorphisms (SNPs), block substitutions, heterozygous and homozygous insertion/deletion, as well as numerous segmental duplications and copy number variation regions [1]. It is known that SNPs, referring to a single mutation (change of a single base) in the DNA sequence, occur on the average about once every 100 to 300 bases [2], SNPs account for 90% of all human sequence variants [3], and are therefore, the major source of heterogeneity. As a consequence, finding SNPs is an important task in reference-based sequencing. We refer to [4] and references therein for a general review on reference-based assembly.

The problem of SNP calling is also motivated by medical genetics and personalized medicine, where research has enabled a more detailed understanding of the impact of genetics in disease. It is revealed that SNPs account for some of diseases, and therefore detecting them in human genome allows for examining genetic variation and risk for many common diseases. For example, a single base mutation in the *Apolipoprotein E* gene is associated with a higher risk for Alzheimer's disease [5]. It is believed that the SNP profile of a variety of diseases will be eventually characterized, and then it will only be a matter of time before physicians can screen individuals for susceptibility to a disease just by analyzing their DNA samples for specific SNP patterns [6].

In this work we study the reference-based assembly and SNP calling problems from an information-theoretic perspective, to obtain a tight characterization of the fundamental limits of read length and data rate for a simple probabilistic model. Our results parallel those in [7] for de novo assembly.

II. problem Statement

Let $\mathbf{t}$ be a DNA sequence [1] of length G, consist of elements from the alphabet set $\mathcal{X} = \{A, C, G, T\}$. This sequence is called *target* sequence, assumed to be i.i.d. with marginal distributed uniform on $\mathcal{X}$. A reference sequence $\mathbf{r}$ is generated by passing $\mathbf{t}$ through a discrete memoryless quaternary symmetric channel (QSC) with crossover probability ϵ, that is

$$Q(i|j) = \begin{cases} 1-\epsilon & i = j \\ \epsilon/3 & i \neq j, \end{cases} \qquad i, j \in \{A, C, G, T\}. \tag{1}$$

A (noiseless) *read* sampled at position s is the L-segment $\mathcal{S} = \mathbf{t}_L(s) \triangleq \big(\mathbf{t}(s), \mathbf{t}(s+1), \ldots, \mathbf{t}(s+L-1)\big)$ of string $\mathbf{t}$ started at position s. We denote by $\mathbf{S}$ the set of N reads $\{\mathcal{S}_1, \ldots, \mathcal{S}_N\}$ which are sampled from $\mathbf{t}$ with starting positions drawn independently and uniformly from $\{1, 2, \ldots, G\}$.

The aim in reference-based sequencing is to reconstruct the target sequence, given the reference and set of reads:

$$\hat{\mathbf{t}} = \Phi\left(\mathbf{r}, \mathbf{S}\right), \tag{2}$$

where $\hat{\mathbf{t}}$ is the reconstructed sequence. This is equivalent to characterizing the set of SNP positions, and their actual values, since the reference and the target share the same values on other positions, i.e. SNP calling.

A SNP caller may commit two types of error, namely, false alarm and mis-detection. More precisely, the false alarm error and misdetection error can be respectively defined as

$$\Delta_{\mathsf{f}} = |\{i \in \{1, \ldots, G\} : \hat{\mathbf{t}}(i) \neq \mathbf{t}(r), \mathbf{t}(i) = \mathbf{r}(i)\}| \tag{3}$$

$$\Delta_{\mathsf{m}} = |\{i \in \{1, \ldots, G\} : \hat{\mathbf{t}}(i) \neq \mathbf{t}(r), \mathbf{t}(i) \neq \mathbf{r}(i)\}|. \tag{4}$$

[1] More precisely, we model $\mathbf{t}$ as a cyclic sequence to avoid the minor effects caused by the end points.

The problem can be characterized by parameters (G, N, L, ϵ). We are interested in the right scaling law of N and L for which there exist a reconstruction method such that $(\Delta_{\mathsf{f}} + \Delta_{\mathsf{m}})/G\epsilon \to 0$ holds with probability approaching 1 (over the randomness of the target sequence, SNPs, and read sampling) as $G \to \infty$. We focus on an asymptotic regime where $G, L \to \infty$, with $\bar{L} = L/\log G$ fixed. As shown in [7], this is the natural limit for studying the assembly problem. The parameter ϵ represents the target error rate; since it should be small, a natural scaling turns out to be $\epsilon = G^{-\alpha}$ for some $\alpha > 0$. Our goal is to characterize the values of $\lambda = N/G$ and $\bar{L} = L/\log G$ in order for reliable reconstruction.

In the rest of this work we use the following notation: $(a, b]$ denotes an interval including all integer numbers more than a and not exceeding b, and $\mathbf{t}_a(s)$ denotes $\mathbf{t}(s, s+1, \ldots, s+a-1)$. Finally, $a \doteq b$ indicates that a and b are asymptotically the same up to an exponential order.

III. Lower Bounds on the Feasible Region

In this section we derive some bounds on the scaling law of parameters required for a reliable SNP calling.

A. Coverage Bound

As stated above the reconstruction problem is equivalent to SNP calling. Assume that the positions of all SNPs are revealed by a genie, we just need to find out the correct value for each of such positions. This information can only be provided by the reads covering the SNPs. Therefore, each SNP position has to be covered by at least one read, or equivalently, at least one arrival is required in interval $I_x \triangleq (x - L, x]$, for each SNP position x. Using the Poisson model for the read arrivals [7], the expected number of non-covered reads can be written as $G\epsilon e^{-\lambda L}$, which is growing with G if $\lambda \bar{L} < (1-\alpha)\ln 2$. Therefore, we have the following theorem, whose proof is along the same lines as that of Lemma 2 in [7]. We skip the proof due to lack of space.

Theorem 1. *A reliable SNP calling is possible only if* $\lambda \bar{L} > (1-\alpha)\ln 2$.

B. A Lower bound on L

In this part we prove a lower bound for the required read length. We argue that if $\bar{L} < 1/2$ then, any SNP calling algorithm fails with high probability.

Theorem 2. *A reliable reconstruction is possible only if* $\bar{L} = \frac{L}{\log G} > \frac{1}{2}$.

The rest of this section is dedicated to the proof of Theorem 2. The proof is based on a counting argument. Let $\bar{L} = \frac{1-\gamma}{2}$, for some $\gamma > 0$, which implies $G4^{-L} = G^{\gamma}$.

We index all quaternary L-tuples by integer number in $i = 1, \ldots, M$, where $M = 4^L$ is the total number of quaternary sequences of length L, and denote them by $\{\mathcal{T}_1, \ldots, \mathcal{T}_M\}$. The *spectrum* of a sequence represents the number of occurrence of each L-subsequence. More precisely, for a sequence $\mathbf{x}$, we denote by its spectrum by $\mathsf{spec}(\mathbf{x}) = (n_1(\mathbf{x}), n_2(\mathbf{x}), \ldots, n_M(\mathbf{x}))$, where $n_i(\mathbf{x})$ is the number of times that the L-tuple $\mathcal{T}_i$ appears in $\mathbf{x}$. It is clear there is a total of G segments of length L in the entire sequence, and hence, $\sum_{i=1}^{M} n_i(\mathbf{x}) = G$. The following lemma, which is proved in Appendix, is useful for the rest of this argument.

Lemma 1. *Let* $\mathbf{r}$ *be a sequence of length* G *with elements drawn independently and uniformly from* $\{A, C, G, T\}$. *If* $\bar{L} = (1-\gamma)/2$, *then with probability at least* $1 - o(1)$ *we have*

$$\min_i n_i(\mathbf{r}) \geq G^{\gamma}/2L, \quad \textit{and} \quad \max_i n_i(\mathbf{r}) \leq 3G^{\gamma}/2.$$

With a slight abuse of notation, we may also denote the result of read sampling by $\mathsf{spec}(\mathbf{S}) = (N_1, N_2, \ldots, N_M)$, where N_i is the number of times $\mathcal{T}_i$ is observed in sampling. For a sampling with N trials, we have $\sum_{i=1}^{M} N_i = N$.

Having a set of reads $\mathbf{S}$ and a reference $\mathbf{r}$, the maximum likelihood (ML) reconstruction reveals sequence $\hat{\mathbf{t}}$, where

$$\hat{\mathbf{t}}_{\mathsf{ML}} = \arg\max_{\mathbf{x}} \mathbb{P}[\mathbf{r}, \mathbf{S}|\mathbf{x}] = \arg\max_{\mathbf{x}} \mathbb{P}[\mathbf{r}|\mathbf{x}]\mathbb{P}[\mathbf{S}|\mathbf{x}] \quad (5)$$

Note that the second equality above holds since given the original sequence, $\mathbf{r}$ and $\mathbf{S}$ are independent. This can be further simplified by evaluating the terms above as

$$\mathbb{P}[\mathbf{r}|\mathbf{x}] = \epsilon^{d_H(\mathbf{r},\mathbf{x})}(1-\epsilon)^{G - d_H(\mathbf{r},\mathbf{x})} \quad (6)$$

where $d_H(\mathbf{r}, \mathbf{x})$ is the number of locations at which $\mathbf{r}$ is differ from $\mathbf{x}$. Moreover,

$$\begin{aligned}
\mathbb{P}[\mathsf{spec}(\mathbf{S})|\mathbf{x}] &= \binom{N}{N_1, \ldots, N_M} \prod_{i=1}^{M} \left(\frac{n_i(\mathbf{x})}{G}\right)^{N_i} \\
&= \binom{N}{N_1, \ldots, N_M} \exp\left[\sum_{i=1}^{M} N_i \ln \frac{n_i(\mathbf{x})}{G}\right] \\
&= \binom{N}{N_1, \ldots, N_M} \exp\left[-ND(\mathcal{P}_{\mathbf{S}} \parallel \mathcal{P}_{\mathbf{x}}) - NH(\mathcal{P}_{\mathbf{S}})\right],
\end{aligned}$$

where $D(P \parallel Q)$ is the Kullback-Leibler divergence between distributions P and Q, and $H(P)$ is the entropy of distribution P (both in natural base). Moreover, $\mathcal{P}_{\mathbf{x}} = \frac{1}{G}\mathsf{spec}(\mathbf{x})$ and $\mathcal{P}_{\mathbf{S}} = \frac{1}{N}\mathsf{spec}(\mathbf{S})$ are the empirical distributions of L-tuples in sequence $\mathbf{x}$ and sampled reads $\mathbf{S}$, respectively. Hence, by dropping terms independent from the choice of $\mathbf{x}$, we can write

$$\hat{\mathbf{t}}_{\mathsf{ML}} = \arg\max_{\mathbf{x}} \left(\frac{\epsilon}{1-\epsilon}\right)^{d_H(\mathbf{r},\mathbf{x})} \exp[-ND(\mathcal{P}_{\mathbf{S}} \parallel \mathcal{P}_{\mathbf{x}})]. \quad (7)$$

Let us examine the true target sequence versus the reference sequence as two potential solutions for (7).

$$\frac{\mathbb{P}[\mathbf{r}, \mathcal{P}_{\mathbf{S}}|\mathbf{t}]}{\mathbb{P}[\mathbf{r}, \mathcal{P}_{\mathbf{S}}|\mathbf{r}]} = \left(\frac{\epsilon}{1-\epsilon}\right)^{d_H(\mathbf{r},\mathbf{t})} \exp\Big\{N\Big[D(\mathcal{P}_{\mathbf{S}} \,\|\, \mathcal{P}_{\mathbf{r}}) - D(\mathcal{P}_{\mathbf{S}} \,\|\, \mathcal{P}_{\mathbf{t}})\Big]\Big\}. \quad (8)$$

Proposition 1. *If* $\bar{L} = \frac{1-\gamma}{2}$ *for some* $\gamma > 0$, *then with probability* $1 - o(1)$ *we have* $\mathbb{P}[\mathbf{r}, \mathcal{P}_{\mathbf{S}}|\mathbf{t}] < \mathbb{P}[\mathbf{r}, \mathcal{P}_{\mathbf{S}}|\mathbf{r}]$.

In the following we focus on upper bounding the exponent term in (8). Consider a fixed tuple $\mathcal{T}_i$ which appears $n_i(\mathbf{r})$ times within the reference sequence. Some of these $n_i(\mathbf{r})$ tuples may be converted to a different tuple due to substitutions at SNP positions. On the other hand some other tuples $\mathcal{T}_j$ may be converted to $\mathcal{T}_i$ because of SNPs, and increase the number of appearance of $\mathcal{T}_i$ in the target sequence. We denote these numbers by δ_i^- and δ_i^+, respectively. Hence, we have $n_i(\mathbf{t}) = n_i(\mathbf{r}) + \delta_i^+ - \delta_i^-$. With $N_\epsilon \simeq G\epsilon$ be the number of SNPs occurred in the system, we have $\sum_i \delta_i^+ = \sum_i \delta_i^- \leq LN_\epsilon$, since each substitution may destroy at most L existing L-tuples, and generate at most L new L-tuples.

On the other hand, we can write $\frac{N_i}{N} = \frac{n_i(\mathbf{t})+\rho_i}{G}$, and

$$D(\mathcal{P}_{\mathbf{S}} \parallel \mathcal{P}_{\mathbf{r}}) - D(\mathcal{P}_{\mathbf{S}} \parallel \mathcal{P}_{\mathbf{t}}) = \sum_{i=1}^{M} \frac{N_i}{N} \ln \frac{n_i(\mathbf{t})/G}{n_i(\mathbf{r})/G} \tag{9}$$

$$= \sum_{i}^{M} \frac{n_i(\mathbf{t}) + \rho_i}{G} \ln \frac{n_i(\mathbf{t})}{n_i(\mathbf{r})}$$

$$= \sum_{i}^{M} \frac{n_i(\mathbf{r}) + \delta_i^+ - \delta_i^- + \rho_i}{G} \ln \frac{n_i(\mathbf{r}) + \delta_i^+ - \delta_i^-}{n_i(\mathbf{r})}$$

$$\overset{(a)}{\leq} \sum_{i}^{M} \frac{n_i(\mathbf{r}) + \delta_i^+ - \delta_i^- + \rho_i}{G} \left(\frac{\delta_i^+ - \delta_i^-}{n_i(\mathbf{r})} \right)$$

$$= \frac{1}{G} \left[\sum_{i}^{M} (\delta_i^+ - \delta_i^-) + \sum_{i=1}^{M} \frac{(\delta_i^+ - \delta_i^-)^2}{n_i(\mathbf{r})} + \sum_{i=1}^{M} \frac{\rho_i(\delta_i^+ - \delta_i^-)}{n_i(\mathbf{r})} \right]$$

where in (a) we used the fact that $\ln(1+x) \leq x$. Note that the first term in (9) is zero. We will bound the other terms in the following.

We first present the following lemmas, that are useful to bound the second term in (9). These are proved in Appendix.

Lemma 2. *Let* $\mathbf{r}$ *be a random i.i.d sequence and* $\mathbf{t}$ *is generated by passing* $\mathbf{r}$ *through a memoryless* $\mathsf{QSC}(G^{-\alpha})$*. Then, there exists a constant* $\kappa = \kappa(\alpha, \gamma)$ *(not growing with G) such that*

$$\max_i \delta_i^- \leq \kappa L \quad \textit{and} \quad \max_i \delta_i^+ \leq \kappa L \tag{10}$$

holds with probability $1 - o(1)$.

Lemma 3. *Under the same assumptions as Lemma 2, with probability* $1 - o(1)$ *there exist a constant* κ *such that*

$$\sum_{i=1}^{M} (\delta_i^-)^2 \leq L\kappa N_\epsilon, \quad \textit{and} \quad \sum_{i=1}^{M} (\delta_i^+)^2 \leq L\kappa N_\epsilon.$$

From Lemmas 1 and 3 and we can show that

$$\sum_{i=1}^{M} \frac{(\delta_i^+ - \delta_i^-)^2}{n_i(\mathbf{r})} \leq \sum_{i=1}^{M} \frac{(\delta_i^+)^2 + (\delta_i^-)^2}{n_i(\mathbf{r})}$$

$$\leq \frac{1}{\min_i n_i(\mathbf{r})} \left(\sum_{i=1}^{M} (\delta_i^+)^2 + \sum_{i=1}^{M} (\delta_i^-)^2 \right)$$

$$\leq \frac{2L\kappa N_\epsilon}{G^\gamma/2L} = 4L^2 G^{-\gamma} \kappa N_\epsilon, \tag{11}$$

where the last inequality holds with high probability.

Bounding the last term in (9) is more involved, because it depends on the sampling randomness. The following lemma, proved in Appendix, plays a key role in bounding this term.

Lemma 4. *With probability* $1 - o(1)$*, we have*

$$\max_i |\rho_i| \leq 4LG^{-\gamma/2}\sqrt{\lambda \ln 2} \cdot n_i(\mathbf{t}). \tag{12}$$

Next, we using Lemma 4 and write

$$\sum_{i=1}^{M} \frac{\rho_i(\delta_i^+ - \delta_i^+)}{n_i(\mathbf{r})} \leq \sum_{i=1}^{M} \frac{|\rho_i||\delta_i^+ - \delta_i^+|}{n_i(\mathbf{r})}$$

$$\leq \sum_{i=1}^{M} \frac{4LG^{-\gamma/2}\sqrt{\lambda \ln 2} n_i(\mathbf{t})}{n_i(\mathbf{r})} |\delta_i^+ - \delta_i^+|$$

$$\leq 12L^2 G^{-\gamma/2}\sqrt{\lambda \ln 2} \sum_{i=1}^{M} |\delta_i^+ - \delta_i^+|$$

$$\leq 12L^2 G^{-\gamma/2}\sqrt{\lambda \ln 2} \cdot 2LN_\epsilon \tag{13}$$

where we used Lemma 1 in the third inequality, and the fact that $|a-b| \leq |a| + |b|$ in the last one. Finally, by replacing (11) and (13) in (9), we get

$$N\big[D(\mathcal{P}_{\mathbf{S}} \parallel \mathcal{P}_{\mathbf{r}}) - D(\mathcal{P}_{\mathbf{S}} \parallel \mathcal{P}_{\mathbf{t}})\big] \tag{14}$$
$$\leq \lambda \left[4L^2 G^{-\gamma} \kappa N_\epsilon + 24L^3 G^{-\gamma/2}\sqrt{\lambda \ln 2} N_\epsilon \right]$$

Replacing (14) in (8), we have

$$\ln \frac{\mathbb{P}[\mathbf{r}, \mathcal{P}_{\mathbf{S}}|\mathbf{t}]}{\mathbb{P}[\mathbf{r}, \mathcal{P}_{\mathbf{S}}|\mathbf{r}]} \leq N_\epsilon \ln \frac{\epsilon}{1-\epsilon} + N_\epsilon \lambda \left[\frac{4\kappa L^2}{G^\gamma} + \frac{24\sqrt{\lambda \ln 2} L^3}{G^{\gamma/2}} \right]$$
$$\doteq N_\epsilon \left[-\ln G + G^{-\gamma/2} \right], \tag{15}$$

where in the last expression we only kept the dominating terms and ignored the rest, and used the fact that $\epsilon = G^{-\alpha}$. Note that the first term is negative and growing logarithmically with G, while the positive term vanishes inverse polynomially. Therefore, for large enough G, the negative term overcomes the positive one, and the LLR becomes negative, which proves the claim in Proposition 1. This implies that the even an optimum algorithm cannot reveal the true sequence. A similar inequality ($\mathbb{P}[\mathbf{r}, \mathcal{P}_{\mathbf{S}}|\mathbf{t}'] < \mathbb{P}[\mathbf{r}, \mathcal{P}_{\mathbf{S}}|\mathbf{r}]$) holds for any $\mathbf{t}'$ with $d_H(\mathbf{t}', \mathbf{t})/G\epsilon \to 0$, whose proof is beyond the available space of this paper.

IV. Algorithm and Analysis

In this section we propose an algorithm to detect SNPs, and analyze the performance of the algorithm.

A. Algorithm

We fix constant k with $k > \lfloor \frac{1}{\alpha} \rfloor + 1$. For all $i \in \{1, \ldots, N\}$, we find all the locations on the reference sequence which are within Hamming distance k to the read $\mathcal{R}_i$. If there exists no such location or there are more than one such, we discard the read. Otherwise, we modify all the mismatching bases of the uniquely aligned location of the reference to the read $\mathcal{R}_i$.

B. Analysis

A read is called *identifier* if it is uniquely mapped to the reference sequence, that is there is a unique L-segment of the reference within distance k of the read. Let $\mathcal{E}$ be the event that the algorithm cannot assemble the target sequence. We will show that if $\mathbb{P}(\mathcal{E}) \to 0$ as G grows, provided that $\bar{L} > \frac{1}{2}$ and $\lambda \bar{L} > (1-\alpha)\ln 2$.

Let $\mathcal{E}_{\mathsf{f}}$ and $\mathcal{E}_{\mathsf{m}}$ be the events that the algorithm falsely calls at least one non-SNP locations and misses at least on SNPs, respectively. Clearly,

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}(\mathcal{E}_{\mathsf{f}} \cup \mathcal{E}_{\mathsf{m}}) \leq \mathbb{P}(\mathcal{E}_{\mathsf{f}}) + \mathbb{P}(\mathcal{E}_{\mathsf{m}}).$$

We first provide an upper bound on $\mathbb{P}(\mathcal{E}_{\mathsf{f}})$. The following lemma, proved in Appendix, provides the condition under which no false SNP is called by the algorithm.

Lemma 5. *The proposed algorithm has no false positive calls if the number of SNPs in any interval of length L on the target sequence is less than* k.

Using Lemma 5 and applying the union bound, we obtain

$$\mathbb{P}(\mathcal{E}_{\mathsf{f}}) \leq \sum\nolimits_{i=1}^{G} \mathbb{P}(\mathcal{D}_i) = G\mathbb{P}(\mathcal{D}_1), \tag{16}$$

where $\mathcal{D}_i$ is the event that more than $(k-1)$ SNPs occur in the interval $[i, i+L)$. One can show

$$\mathbb{P}(\mathcal{D}_i) = \sum_{j=k}^{L} \binom{L}{j} \epsilon^j (1-\epsilon)^{L-j} \leq L\binom{L}{k}\epsilon^j(1-\epsilon)^{L-j} \leq L^{k+1}\epsilon^k. \quad (17)$$

This implies $\mathbb{P}(\mathcal{E}_{\mathsf{f}}) \leq GL^{k+1}\epsilon^k \leq G^{1-\alpha k}L^{k+1}$. Since $k > \lfloor \frac{1}{\alpha} \rfloor + 1$, we conclude that $\mathbb{P}(\mathcal{E}_{\mathsf{f}}) \to 0$ as $G \to \infty$.

Next, we need to upper bound $\mathbb{P}(\mathcal{E}_{\mathsf{m}})$. Let x be a SNP position, and define $\mathcal{C}_x$ as the event that x is not called by the algorithm. It can be shown that $\mathcal{C}_x$'s have the same probability. Applying the union bound, we obtain

$$\mathbb{P}(\mathcal{E}_{\mathsf{m}}) \leq \sum\nolimits_{i:\mathbf{t}(i)\neq\mathbf{r}(i)} \mathbb{P}(\mathcal{C}_i) = N_\epsilon \mathbb{P}(\mathcal{C}_x), \quad (18)$$

Note that $\mathcal{C}_x$ occurs only if all the reads of x, the reads sampled from $I_x \triangleq (x-L, x]$, are discarded by the algorithm. Let us define $\mathcal{B}_x$ the event that interval $(x-L, x+L)$ contains more than $k-1$ SNPs. In this way, we can write

$$\begin{aligned}\mathbb{P}(\mathcal{C}_x) &= \mathbb{P}(\mathcal{C}_x|\mathcal{B}_x)\mathbb{P}(\mathcal{B}_x) + \mathbb{P}(\mathcal{C}_x|\mathcal{B}_x^c)\mathbb{P}(\mathcal{B}_x^c) \\ &\leq \mathbb{P}(\mathcal{B}_x) + \mathbb{P}(\mathcal{C}_x|\mathcal{B}_x^c) \overset{(b)}{\leq} (2L)^{k+1}\epsilon^k + \mathbb{P}(\mathcal{C}_x|\mathcal{B}_x^c), \end{aligned} \quad (19)$$

where (b) follows from similar calculations as in (17). It remains to compute $\mathbb{P}(\mathcal{C}_x|\mathcal{B}_x^c)$.

Not that a read $\mathcal{S}_i$ sampled from $x - t \in I_x$ is discarded if and only if there exists a position y in the reference sequence such that $d_H(\mathcal{S}_i, \mathbf{r}_L(y)) < k$ or $d_H(\mathcal{S}_i, \mathbf{r}_L(x-t)) \geq k$. However, the latter cannot happen under condition $\mathcal{B}^c$. So, discarding read $\mathcal{S}_i$ sampled from $x-t$ because of position y occurs only due to a similarity between the L-interval I_{x-t} and I_y. Therefore, the next read $\mathcal{S}_j$ sampled from $x-t+t'$ is likely to be discarded for location $y+t'$ for small values of t', since a large fraction of the corresponding L-segments are known to be similar. This implies that events of discarding reads covering x are not independent.

As illustrated in Fig. 1, we can analyze the reads covering x by considering those aligned to the same regions on the reference sequence. The entire interval $(x-L, x]$ can be partitioned into short segments. Each segment either contains no arrival, or all its sampled reads are mapped the same region on the reference. Let $\ell_1, \ell_2, \ldots, \ell_n$ be the length of short segments with at least one arrival, for some $n \geq 0$. Note that a pile of read who are mapped to the same location generate a segment of length at most $2L$, with Hamming distance at most $2k$ from the destination interval. Therefore probability of discarding all reads from an interval of length ℓ starting at $x-t$ because of position y can be founded as

$$\begin{aligned}\pi_\ell &\leq \sum_{y\neq x-t} \mathbb{P}[d_H(\mathbf{t}_{L+\ell}(x-t), \mathbf{r}_{L+\ell}(y)) \leq 2k] \\ &\leq G4^{-(L+\ell)} \sum_{d=0}^{2k} \binom{L+\ell-1}{d} 3^d \leq G4^{-(L+\ell)}(6L)^{2k+1}.\end{aligned}$$

Therefore, the probability that all reads covering x are discarded can be bounded as

$$\begin{aligned}\mathbb{P}[\mathcal{C}_x|\mathcal{B}_x^c, \{\ell_1, \ldots, \ell_n\}] &\leq \left[\exp\left[-\lambda(L - \sum\nolimits_{i=1}^n \ell_i)\right] \prod_i \pi_{\ell_i}\right] \\ &\leq \left((6L)^{2k+1}G4^{-(L)}\right)^n \exp\left[-\lambda\left(L - \sum\nolimits_{i=1}^n \ell_i\right)\right] 4^{-\sum_{i=1}^n \ell_i} \\ &\overset{(c)}{\leq} e^{-\lambda(L-\sum_i \ell_i)} 4^{-\sum_i \ell_i} \leq \max_{0\leq u\leq L} e^{-\lambda(L-u)}4^{-u} \end{aligned} \quad (20)$$

where in (c) we used the fact that $(6L)^k G4^{-L} < 1$ for large enough G. Therefore,

$$\begin{aligned}\mathbb{P}[\mathcal{C}_x|\mathcal{B}_x^c] &= \sum_{n=0}^{L} \sum_{\ell_1,\ldots,\ell_n} \mathbb{P}[\mathcal{C}_x|\mathcal{B}_x^c, \{\ell_1, \ldots, \ell_n\}] \cdot \mathbb{P}[\ell_1, \ldots, \ell_n|\mathcal{B}_x^c] \\ &\leq \max_{0\leq u\leq L} e^{-\lambda(L-u)}4^{-u}, \end{aligned} \quad (21)$$

where the last equality holds since the probability terms add up to 1. By replacing (21) in (19), and combining the result with (18), we can conclude

$$\mathbb{P}(\mathcal{E}_{\mathsf{m}}) \leq 2G(2L\epsilon)^{k+1} + G\epsilon\left(\max_{0\leq u\leq L} e^{-\lambda(L-u)}4^{-u}\right). \quad (22)$$

It is clear that the first term is vanishing as $G \to \infty$ since $k > \lfloor \frac{1}{\alpha} \rfloor + 1$. The second term has a linear exponent in u, and therefore it takes the maximum at one of the extreme points. For $u = 0$, we have $G\epsilon e^{-\lambda L}$ which is vanishing as G grows, since $\lambda \bar{L} > (1-\alpha)\ln 2$. On the hand, for $u = L$, we have $G\epsilon 4^{-L}$, which also goes to zero because $\bar{L} > 1/2$. This completes the analysis of the algorithm.

V. Discussions

We studied the reference-based sequencing and SNP-calling based on noiseless reads, and derive matching bounds for having reliable sequencing. The reads may be corrupted by noise in a more realistic model, and SNPs might occur as insertion/deletion. Using smarter and more algorithm for dealing with reading noise is inevitable. We are currently working on extensions of this work towards more general models.

Appendix

Proof of Lemma 1. Fix an $m \in \{1, \ldots, L\}$ and partition the entire sequence into $K = G/L$ short segments of length L by cutting it at positions $kL + m$ for $k = 1, \ldots, K$. We denote these strings by $\mathbf{t}(m;k)$ for $k = 1, \ldots, K$, and by $\tilde{n}_i^{(m)}(\mathbf{r})$ the number of occurrences of L-tuple $\mathcal{T}_i$ among these K strings. It is clear that $n_i(\mathbf{r}) = \sum_{m=1}^{L} \tilde{n}_i^{(m)}(\mathbf{r})$. We also define the indicators $U_i(m;k)$ as $U_i(m;k) = \mathbb{1}\{\mathbf{t}(m;k) = \mathcal{T}_i\}$. This splitting is useful to bound the number of occurrences, since random variables $U_i(m;j)$ and $U_i(m;k)$ are independent. Note that $\mathbb{P}[U_i(m;j) = 1] = 4^{-L}$, and $\tilde{n}_i^{(m)}(\mathbf{r}) = \sum_{k=1}^{M} U_i(k,m)$. Therefore, using Chernoff bound [8, Corollary 4.6] we have

$$\mathbb{P}\left[|\tilde{n}_i^{(m)}(\mathbf{r}) - K4^{-L}| \geq \frac{1}{2}K4^{-L}\right] \leq 2\exp\left[-G4^{-L}/12L\right].$$

This together with $n_i(\mathbf{r}) = \sum_{m=1}^{L} \tilde{n}_i^{(m)}(\mathbf{r})$ implies

$$\begin{aligned}\mathbb{P}\big[\min_i n_i(\mathbf{r}) \leq G4^{-L}/2L\big] &= \mathbb{P}\left[\bigcup\nolimits_{i=1}^{M} n_i(\mathbf{r}) \leq G4^{-L}/2L\right] \\ &\leq \sum\nolimits_{i=1}^{M} \mathbb{P}[n_i(\mathbf{r}) \leq G4^{-L}/2L] \\ &\leq \sum\nolimits_{i=1}^{M} \mathbb{P}[\tilde{n}_i^{(1)}(\mathbf{r}) \leq G4^{-L}/12L] \leq 2 \cdot 4^L \exp[-G^\gamma/12L].\end{aligned}$$

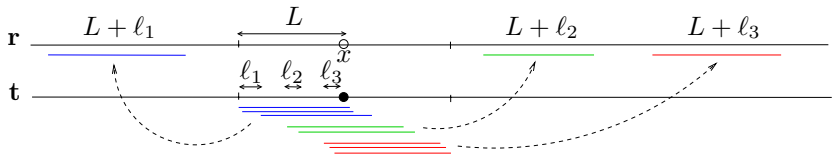

Fig. 1. A SNP position misses all its marking reads: the L-interval $(x-L, x]$ can be split into short segments such that each segment either has no arrival read, or all reads from the segment are mapped to a common region on the reference.

On the other hand

$$\mathbb{P}\big[\max_i n_i(\mathbf{r}) \geq 3G4^{-L}/2\big] \leq \sum_{i=1}^{M} \mathbb{P}[n_i(\mathbf{r}) \geq 3G4^{-L}/2]$$
$$\leq \sum_{i=1}^{M} \mathbb{P}\left[\bigcup_{m=1}^{L} \tilde{n}_i^{(m)}(\mathbf{r}) \geq 3G4^{-L}/2L\right]$$
$$\leq \sum_{i=1}^{M}\sum_{m=1}^{L} \mathbb{P}[\tilde{n}_i^{(m)}(\mathbf{r}) \geq 3G4^{-L}/2L] \leq 2L4^L \exp[-G^\gamma/12L].$$

Therefore, both bounds on the minimum and maximum hold with probability $1-\Theta(e^{G^{-\gamma}})$ which approaches 1 as $G \to \infty$, provided that $\gamma > 0$. □

Proof of Lemma 2. Consider a tuple $\mathcal{T}_i$ which appears in $n_i(\mathbf{r})$ positions in the reference, and therefore the total length of $\mathbf{r}$ covered by $\mathcal{T}_i$ is at most $Ln_i(\mathbf{r})$. Each SNP can reduce $n_i(\mathbf{r})$ by at most L units, because it does not modify more than L tuples in general. Therefore, the probability that more than Lu of the existing $n_i(\mathbf{r})$ occurrences of $\mathcal{T}_i$ disappear by SNPs can be upper bounded by

$$\mathbb{P}[\delta_i^- \geq Lu] \leq \mathbb{P}[\text{at least } u \text{ SNPs occur in } Ln_i(\mathbf{r}) \text{ bases}]$$
$$\leq \sum_{k=u}^{Ln_i(\mathbf{r})} \binom{Ln_i(\mathbf{r})}{k} \epsilon^k (1-\epsilon)^{Ln_i(\mathbf{r})-k}$$
$$\leq Ln_i(\mathbf{r}) \cdot (Ln_i(\mathbf{r})\epsilon)^u, \qquad (23)$$

where the last inequality holds since the first term dominates all other terms in the summation. Using union bound, we get

$$\mathbb{P}[\max_i \delta_i^- \geq Lu] \leq \sum_{i=1}^{M} \mathbb{P}[\delta_i^- \geq Lu]$$
$$\leq 4^L \max_i Ln_i(\mathbf{r}) \cdot (Ln_i(\mathbf{r})\epsilon)^u. \qquad (24)$$

Recall from Lemma 1 that $\max_i n_i(\mathbf{r}) \leq 2G4^{-L}$ with probability $1-o(1)$. Under this assumption, we have

$$\mathbb{P}[\max_i \delta_i^- \geq Lu] \leq 4^L (2G4^{-L})^{u+1}\epsilon^u$$
$$\doteq G^{2\bar{L}+\gamma-u(\alpha-\gamma)} < G^{-\gamma} \qquad (25)$$

where the last inequality holds for any $u \geq 2(\bar{L}+\gamma)/(\alpha-\gamma)$. Bounding δ_i^+ is based on bounding the number $\mathcal{T}_i$'s generated by SNPs, and can be done using a similar argument, which we skip here fort the sake of brevity. □

Proof of Lemma 3. We know from Lemma 2 that with probability $1-o(1)$ all δ_i^-s are upper bounded by κL. Under this assumption, we are interested in finding $\max \sum_{i=1}^{M} (\delta_i^-)^2$ subject to $0 \leq \delta_i^- \leq \kappa L$, $\forall i$, and $\sum_{i=1}^{M} \delta_i^- \leq LN_\epsilon$. It is easy to show that the maximum value is achieved by setting δ_i^-'s at the extreme values (κL or 0). This yields to the maximum value of

$$\sum_{i=1}^{M} (\delta_i^-)^2 \leq \left(\frac{LN_\epsilon}{L\kappa}\right)(L\kappa)^2 = \kappa L^2 N_\epsilon. \qquad (26)$$

Bounding $\sum_{i=1}^{M} (\delta_i^+)^2$ goes through a similar argument. □

Proof of Lemma 4. Recall that $\mathcal{P}_\mathbf{S} = \frac{1}{N}(N_1, N_2, \ldots, N_M)$ is the empirical distribution generated by sampling N i.i.d. samples from distribution $\mathbf{t}$, and $\rho_i = N_i/\lambda - n_i(\mathbf{r})$. For a fixed i, define $X_j = \mathbb{1}\{\}\mathcal{S}_j = \mathcal{T}_i$, for $j = 1, \ldots, N$. It is clear that $N_i = \sum_{i=1}^{N} X_j$, and X_j's are independent of each other with $\mathbb{P}[X_j = 1] = n_i(\mathbf{t})/G$. Therefore, we can use Chernoff bound and write

$$\mathbb{P}[\rho_i \geq \mu n_i(\mathbf{r})] = \mathbb{P}\left[N_i \geq (1+\mu)N\frac{n_i(\mathbf{t})}{G}\right]$$
$$\leq \exp\left[-\mu^2 N n_i(\mathbf{t})/3G\right], \qquad (27)$$

where the probability is over the randomness of sampling. By union bound, we have

$$\mathbb{P}\left[\exists i : \rho_i \geq \mu n_i(\mathbf{t})\right] \leq \sum_{i=1}^{M} \mathbb{P}\left[\rho_i \geq \mu n_i(\mathbf{t})\right]$$
$$\leq 4^M \exp\left[-\mu^2/3\lambda \cdot \min_i n_i(\mathbf{t})\right]$$
$$\leq \exp\left[-\mu^2 G^\gamma/2\lambda L + 2L\ln 2\right], \qquad (28)$$

where the last inequality holds with probability $1-o(1)$ due to Lemma 1. It is clear that this probability is vanishing when G grows for $\mu = 4LG^{-\gamma/2}\sqrt{\lambda \ln 2}$. Therefore with probability $1-2^{-6L}$ we have

$$\max_i \rho_i \leq 4LG^{-\gamma/2}\sqrt{\lambda \ln 2} \cdot n_i(\mathbf{t}). \qquad (29)$$

We can bound the probability that $-\rho_i < -\mu n_i(\mathbf{t})$ in a similar way, which together with (29) conclude the desired bound. □

Proof of Lemma 5. For any false positive call, we need an identifier read. However, the Hamming distance of any read to its true location is at most $k-1$. This implies that a read can either be mapped to its sampling location or will be discarded by the algorithm. Hence, the algorithm does not mis-align any identifier, and therefore no non-SNP will be called. □

REFERENCES

[1] S. Levy *et al.*, "The diploid genome sequence of an individual human," *PLoS biology*, vol. 5, no. 10, p. e254, 2007.

[2] L. Jorde and S. Wooding, "Genetic variation, classification and 'race'," *Nature genetics*, vol. 36, pp. S28–S33, 2004.

[3] T. Philippe, K. Roman, F. Laura, H. Martin, and F. Christoph, "Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers," *BMC Bioinformatics*, vol. 12, 2011.

[4] S. Pabinger *et al.*, "A survey of tools for variant analysis of next-generation genome sequencing data," *Briefings in bioinformatics*, 2013.

[5] A. Wolf, R. Caselli, E. Reiman, and J. Valla, "Apoe and neuroenergetics: an emerging paradigm in alzheimer's disease," *Neurobiology of Aging*, 2012.

[6] P. Elmer. (2013, jan) Snp and genotyping overview. Available at http://shop.perkinelmer.com/content/snps/genotyping.asp.

[7] A. Motahari, G. Bresler, and D. Tse, "Information theory of dna sequencing," *Arxiv preprint arXiv:1203.6233*, 2012.

[8] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.