Quantitative Biology 2020, 8(2): 143–154
https://doi.org/10.1007/s40484-020-0200-y

RESEARCH ARTICLE

Confidence intervals for Markov chain transition probabilities based on next generation sequencing reads data

Lin Wan[1], Xin Kang[2], Jie Ren[3], Fengzhu Sun[3,*]

[1] NCMIS, LSC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[2] School of Mathematical Sciences, Fudan University, Shanghai 200433, China
[3] Quantitative and Computational Biology Program, University of Southern California, Los Angeles, CA 90089, USA
* Correspondence: fsun@usc.edu

Received November 6, 2019; Revised January 8, 2020; Accepted February 3, 2020

Background: **Markov chains (MC) have been widely used to model molecular sequences. The estimations of MC transition matrix and confidence intervals of the transition probabilities from long sequence data have been intensively studied in the past decades. In next generation sequencing (NGS), a large amount of short reads are generated. These short reads can overlap and some regions of the genome may not be sequenced resulting in a new type of data. Based on NGS data, the transition probabilities of MC can be estimated by moment estimators. However, the classical asymptotic distribution theory for MC transition probability estimators based on long sequences is no longer valid.**

Methods: **In this study, we present the asymptotic distributions of several statistics related to MC based on NGS data. We show that, after scaling by the effective coverage *d* defined in a previous study by the authors, these statistics based on NGS data approximate to the same distributions as the corresponding statistics for long sequences.**

Results: **We apply the asymptotic properties of these statistics for finding the theoretical confidence regions for MC transition probabilities based on NGS short reads data. We validate our theoretical confidence intervals using both simulated data and real data sets, and compare the results with those by the parametric bootstrap method.**

Conclusions**: We find that the asymptotic distributions of these statistics and the theoretical confidence intervals of transition probabilities based on NGS data given in this study are highly accurate, providing a powerful tool for NGS data analysis.**

Keywords: Markov chains; next generation sequencing; transition probabilities; confidence intervals

Author summary: Markov chains (MC) have been widely used to model molecular sequences. We present the asymptotic distributions of several statistics related to MC based on next generation sequencing (NGS) short reads data. We show that, after scaling by an effective coverage *d* proposed by the authors, these statistics based on NGS data approximate to the same distributions as the corresponding statistics for long sequences. The asymptotic properties of these statistics can be applied to find the theoretical confidence regions for MC transition probabilities based on NGS short reads data with high accuracy.

INTRODUCTION

Markov chains (MC) have been widely used to model molecular sequences [1,2]. They have been used to study the dependencies between the bases [3], the enrichment and depletion of certain word patterns [4], prediction of occurrences of long word patterns from short patterns [5,6], and the detection of signals in introns [7]. Narlikar *et al*. [8] studied the effect of the order of MCs on several biological problems including phylogenetic analysis, assignment of sequence fragments to different genomes in meta-genomic studies, motif discovery, and functional

classification of promoters. Usually the transition probabilities of the Markov chain are estimated from data, and hence the uncertainty of the estimated transition probabilities has to be taken into account. This is done for example in [2] for word counts in long sequences.

Often DNA samples do not consist of a few long sequences, but of many short sequences, for example based on next generation sequencing (NGS) technology. For short sequences the standard asymptotic results do not apply, see for example [9]. In [9] it was shown that the word count standardization includes an additional scaling factor, the effective coverage d. While [9] considered word count statistics under a Markov chain with estimated transition probabilities, in this study we are interested in finding confidence regions for the transition probabilities. We derive a normal and a chi-square approximation for several statistics related to MC based on many short sequences (also called reads) from a long underlying sequence that is assumed to be a realization of a Markov chain; and again the effective coverage d appears.

Consider an r-th order Markov chain $\mathbf{A} = A_1A_2...A_n$ of length n. The maximum likelihood estimate of transition probability from a word $\mathbf{w}$ of length r to a letter b is given by

$$\hat{P}_{\mathbf{w},b} = f_{\mathbf{w}b}/f_{\mathbf{w}},$$

where $f_{\mathbf{w}}$ is the number of occurrences of word $\mathbf{w}$ within the sequence. Many investigators have studied the limit distribution of $\hat{P}_{\mathbf{w},b}$ and other related statistics when the sequence length n tends to infinity [10,11]. In the classical papers [10,11], Billingsley studied many important problems related to MC. In particular, he showed the following important results.

Suppose the long sequence follows a simple first order Markov chain, with finite state space $\mathcal{A}$ of size s. We call the elements of $\mathcal{A}$ letters, as we will be thinking of a Markov chain on the set of nucleotides or amino acids. Denote the transition probability from letter i to letter j by p_{ij} , and let

$$\xi_{ij} = (f_{ij} - f_i p_{ij})/f_i^{\frac{1}{2}}. \quad (1)$$

Theorem 3.1 in [11] shows that if the Markov chain is stationary and ergodic, then as the sequence length $n \to \infty$ the distribution of the s^2-dimensional random vector $\xi = (\xi_{i,j})$ converges to the normal distribution with mean 0 and covariance matrix $(\lambda_{i,j;k,l})$, where

$$\lambda_{i,j;k,l} = \delta_{i,k}(\delta_{j,l}p_{ij} - p_{ij}p_{kl}), \quad (2)$$

where $\delta_{u,v}$ is the Kronecker delta that is 1 when $u = v$, and 0 otherwise. Let $\zeta = (\zeta_1, \ldots, \zeta_s)$ be the random vector with components

$$\zeta_i = (f_i - np_i)/n^{\frac{1}{2}}, \quad (3)$$

Theorem 3.3 in [11] also shows that the distribution of the random vector ζ converges to the normal distribution with covariance matrix $\alpha_{ij} + O\left(\frac{1}{n}\right)$, where

$$\alpha_{ij} = \delta_{ij}p_i - p_ip_j + p_i\sum_{m=1}^{\infty}(p_{ij}^{(m)} - p_j) + p_j\sum_{m=1}^{\infty}(p_{ji}^{(m)} - p_i).$$

In [11] it is also shown that the statistic

$$S = \sum_{ij}\frac{(f_{ij} - f_ip_{ij})^2}{f_ip_{ij}}. \quad (4)$$

is asymptotically chi-square in distribution, and the number of degrees of freedom is $s(s - 1)$ (assuming the transition probabilities p_{ij} are all positive), where s is the size of the alphabet set $\mathcal{A}$.

These results were derived under the assumption of one sample consisting of a long realisation of the Markov chain. With the development of next generation sequencing technologies, instead of one long sequence, typically short fragments from the long genome sequence are sampled, while assembly of these genome fragments as in [12,13] can be challenging.

In this study, we investigate the asymptotic distributions of similar statistics described above as in [10,11], but under the new model proposed in [9,14] for the NGS short reads data (see Fig. 1 and Section of "Methods" for the details of the model). We demonstrate that, after scaling by the effective coverage d proposed in [9], these statistics converge to the same corresponding distributions as those statistics for long sequences. We further apply the asymptotic properties of these statistics for finding the theoretical confidence regions for MC

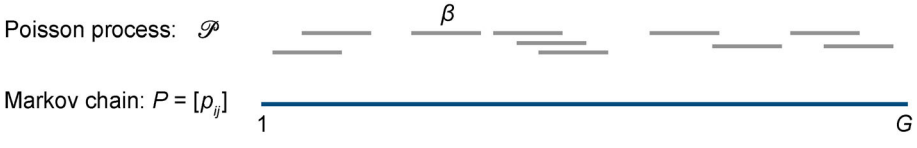

Figure 1. The model of generating NGS short reads data. Two random processes are involved in the generation of the NGS short reads data: i) the underlying genome sequence with length G in bp is generated by a stationary and ergodic Markov chain with transition probability matrix $P = [p_{ij}]$; ii) short reads of length β are randomly sampled from the genome by the Poisson process $\mathcal{P}$ with rate $c(x) = c(x, G)$ at position x. The Poisson process $\mathcal{P}$ is independent of the Markov chain.

transition probabilities based on NGS short reads data. We validate these asymptotic distributions and our theoretical confidence intervals using simulated as well as real data sets. In addition, we propose a parametric bootstrap method, and compare the bootstrap confidence intervals with our theoretical confidence intervals. We find that, the asymptotic properties of these statistics and the theoretical confidence intervals given in this study are highly accurate for the NGS short reads data, providing a powerful tool for NGS data analysis.

RESULTS

Our theoretical results are mainly presented in Section of "Methods": we first introduce the probabilistic model for NGS data based on a MC sequence and random sampling of short reads; we then derive the statistics of interest based on NGS reads and the effective coverage d; we then develop two propositions on the asymptotic distributions of these statistics, and finally we propose the theoretical confidence intervals as well as the parametric bootstrap confidence intervals for MC transition probabilities based on NGS short reads data. Although we cannot rigorously prove the major propositions developed in this study, we validate statistical properties of these NGS related statistics and theoretical confidence intervals of transition probabilities using simulated NGS data sets as well as a real viral genome sequence data in this section.

Validating the theoretical results using simulated NGS data sets

We validate Proposition 1, Proposition 2 and the theoretical confidence intervals for MC transition probabilities of Eq. (16) in Section of "Methods" using simulated NGS short reads data sets.

For each simulated NGS data set, we first simulate the underlying genome sequence with MC using the same set of parameters of transition probabilities $P = [p_{ij}]$ as in [9], which are listed in Supplementary Table S1. The initial position of the simulated genome is sampled based on the stationary distribution of the MC. We set the genome length $G = 100,000$ for all simulated data sets. We then randomly sample M reads of length β bps from the simulated underlying genome. The reads are sampled based on the Poisson process as described in Section of "Methods". The Poisson process can be homogeneous or inhomogeneous.

We simulate three NGS data sets with reads sampled by homogeneous Poisson processes. The parameters of the three homogeneous data sets are (1) $M = 500, \beta = 100$ (denoted as H1 thereafter); (2) $M = 1000, \beta = 200$ (denoted as H2 thereafter); and (3) $M = 1000, \beta = 300$ (denoted as H3 thereafter).

For each of these homogeneous data sets, the rate c of Poisson processes is calculated by

$$c = \frac{M}{G-\beta} \approx \frac{M}{G}.$$

Noting Remark 1, the sequencing coverage $C = \beta c = \beta M/G$. We follow Eq. (13) and calculate the corresponding effective coverage d as

$$d = 1 + \beta c = 1 + \frac{\beta M}{G}.$$

For real NGS data, the reads are generally generated inhomogeneously from the genome. To test our results for inhomogeneous cases, we also simulate three NGS data sets with reads sampled by inhomogeneous Poisson processes. To achieve this, we divide the long underlying genome sequence into 100 consecutive non-overlapped bins $b_1,..., b_t,..., b_{100}$ with the same size; the sampling rates along the positions within the same bin b_t are equal; and the rates of the different bins are proportional to 100 random variables drawn from the gamma distribution $\Gamma(1, 20)$ [15]. The simulated three inhomogeneous data sets of different read numbers and read lengths are (1) $M = 500, \beta = 100$ (denoted as IH1 thereafter); (2) $M = 1000, \beta = 200$ (denoted as IH2 thereafter); and (3) $M = 1000, \beta = 300$ (denoted as IH3 thereafter).

Validating the approximate normal distribution of $\xi_{ij}^{(R)}$ in Proposition 1

To validate the approximate normal distribution of $\xi_{ij}^{(R)}$ by Proposition 1, we first standardize the $\xi_{ij}^{(R)}$ as

$$\widetilde{\xi_{ij}^{(R)}} = \frac{\xi_{ij}^{(R)}}{\sqrt{dp_{ij}(1-p_{ij})}} \to N(0, 1),$$

where the estimated values of d and $p_{i,j}$ are used in the standardization of $\xi_{ij}^{(R)}$. For each of the six data sets (H1, H2, H3, IH1, IH2, and IH3), we repeat the simulating processes by the same parameters 2,000 times to generate 2,000 samples; then calculate the $\widetilde{\xi_{ij}^{(R)}}$ value for each sample; and check whether 2,000 values of $\widetilde{\xi_{ij}^{(R)}}$ have the standard normal distribution by the Kolmogorov-Smirnov (KS) test.

For the homogeneous data sets, for all ij, $\widetilde{\xi_{ij}^{(R)}}$'s have approximately mean of 0 and variance of 1, with p-values being uniformly distributed in the region of $(0, 1)$ (Fig. 2), indicating that all $\widetilde{\xi_{ij}^{(R)}}$'s follow the standard normal distribution. We obtain similar results for the inhomoge-

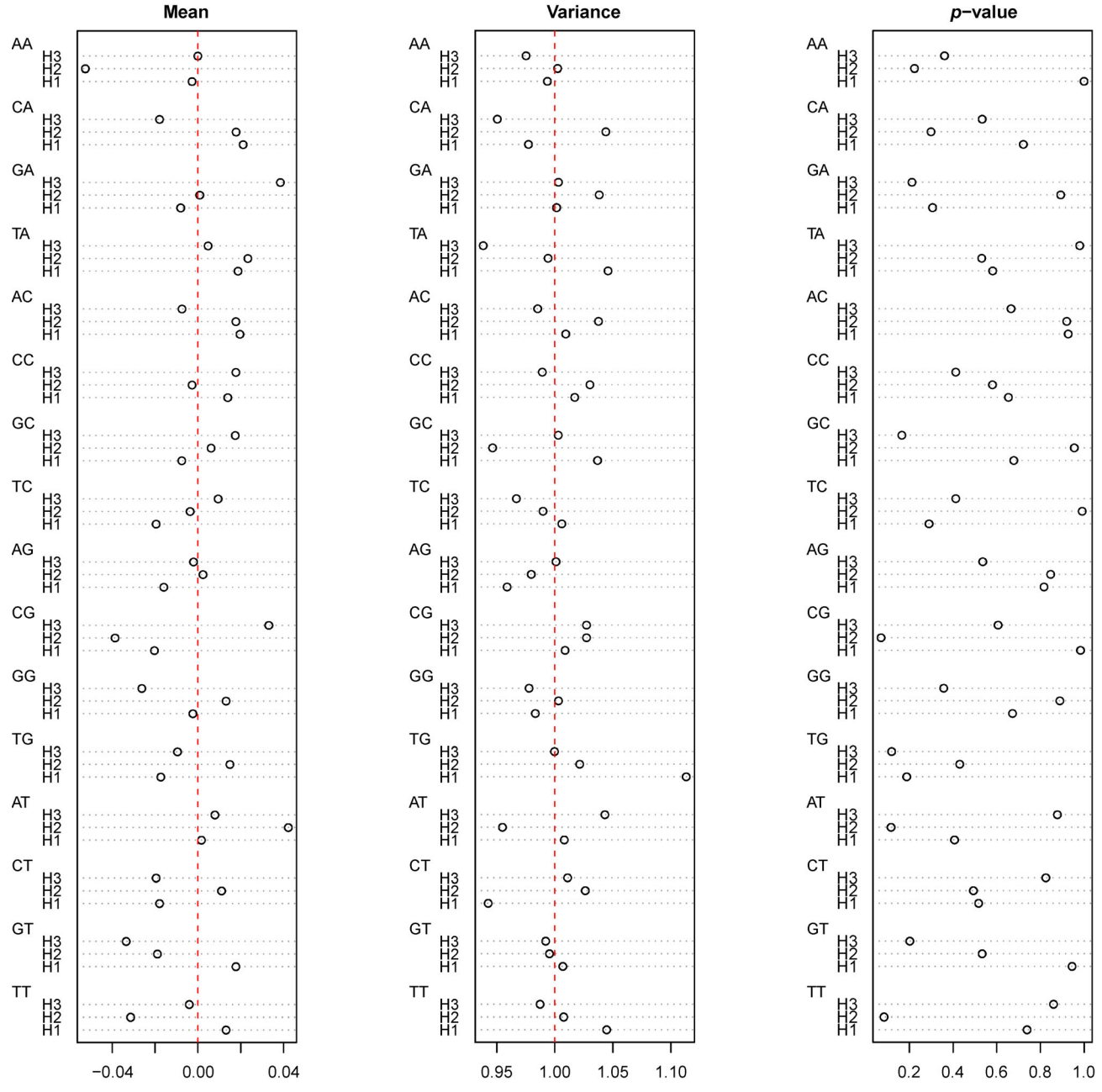

Figure 2. The mean, variance and *p*-value of statistic $\xi_{ij}^{(R)}$ for the simulated homogeneous data sets of H1, H2, and H3. The *p*-value is calculated by Kolmogorov-Smirnov test to whether $\widetilde{\xi_{ij}^{(R)}}$ approximates a standard normal distribution.

neous data sets (Supplementary Fig. S1). These results show strong validations of Proposition 1.

Validating the approximate distributions of $\zeta_i^{(R)}$ and $S^{(R)}$ in Proposition 2

To validate the approximate distributions of $\zeta_i^{(R)}$ and $S^{(R)}$ in Proposition 2, we first standardize them as follows:

$$\widetilde{\zeta_i^{(R)}} = \frac{\zeta_i^{(R)}}{\sqrt{d\alpha_{ii}}} \to N(0,\ 1),$$

and

$$\widetilde{S^{(R)}} = \frac{S^{(R)}}{d} \to \chi^2\ \ (s(s-1)).$$

We then check on the simulated data sets, whether the

new statistic $\widetilde{\zeta_i^{(R)}}$ has the standard normal distribution for all $i \in A$, and the new statistic $\widetilde{S^{(R)}}$ has the chi-square distribution with $df = s(s-1) = 4 \times 3 = 12$, respectively.

Our simulation results show that: 1) for all $i \in A$, $\widetilde{\zeta_i^{(R)}}$ approximates to the standard normal distribution quite well for both homogeneous data sets (Supplementary Fig. S2) and inhomogeneous data sets (Supplementary Fig. S3); 2) the $\widetilde{S^{(R)}}$ approximates a chi-square distribution with $df = 12$ (Mean = 12 and Variance = 24) on both homogeneous data sets and inhomogeneous data sets as well: the calculated values of (Mean, Variance and p-value) are (12.02, 25.15, 0.402) for H1, (12.00, 24.75, 0.690) for H2, (12.14, 24.82, 0.556) for H3, (11.81, 22.92, 0.058) for IH1, (12.15, 23.65, 0.064) for IH2, and (11.88, 24.18, 0.453) for IH3.

Validating the theoretical confidence intervals of MC transition probabilities using NGS data

One of the major applications of Proposition 1 is to construct the theoretical confidence intervals of MC transition probabilities using NGS data. We validate the derived theoretical level $1-\alpha$ confidence interval Eq. (16) by simulation study.

For each of the simulated homogeneous (H1, H2 and H3) and inhomogeneous (IH1, IH2 and IH3) data sets, we

1. Simulate one sample, estimate $\hat{p}_{ij}$s;
2. Calculate theoretical level $1-\alpha$ ($\alpha \in \{0.01, 0.05, 0.1\}$) confidence interval for p_{ij} by Eq. (16);
3. Repeat Steps 1 and 2 for 1,000 times, and calculate the fraction of times that the theoretical level $1-\alpha$ confidence intervals cover the true p_{ij}.

The simulation results in Fig. 3 and Supplementary Fig. S4 show that, the fractions of times that the theoretical confidence intervals cover the true p_{ij} are very close to $1-\alpha$ (α = 0.01, 0.05, 0.1) for both homogeneous and inhomogeneous data sets.

Calculating theoretical confidence intervals of p_{ij} based the estimated $\hat{d}$

In real study, the effective coverage d is generally not available for the NGS data sets. To deal with data sets with unknown d, we adopt the approach proposed in [9], and estimate the effective coverage d as follows:

$$\hat{d} = \text{median}\{(Z_{\mathbf{w}}^{(R)})^2,\ \mathrm{w} \in \mathcal{A}^k\}/0.456 \tag{5}$$

where $Z_{\mathbf{w}}^{(R)}$ is given in Eq. (14).

We find that, by plug-in the estimated effective coverage $\hat{d}$ to the theoretical confidence intervals for p_{ij} in Eq. (16), we can still achieve similar accuracies in calculating the theoretical confidence intervals (Supplementary Figs. S5, S6).

Validating the theoretical confidence interval for p_{ij} by parametric bootstrap

For each of the homogeneous NGS data sets (H1, H2, H3), we calculate the confidence intervals by parametric bootstrap method. We repeat 100 bootstrap calculation for each data set. We find that the fraction of times that the bootstrap confidence intervals covering the true transition probabilities approximate the level $1-\alpha$ well (Supplementary Figs. S7–S9). Meanwhile, the theoretical confidence intervals by Eq. (16) achieve similar results as the parametric bootstrap confidence intervals both in 1) the fraction of time cover the true transition probability (Supplementary Figs. S7–S9), and 2) lengths of confidence intervals (Supplementary Fig. S10).

Application to real viral genome sequence data

We apply the theoretical results developed in this study on a real viral genome sequence *Bacillus phage SP-beta*. The length G of the genomic sequence, which was downloaded from NCBI, is 134, 416 in bp. The frequencies of A, G, C and T are 45,580 (33.91%), 21,078 (15.68%), 25,477 (18.95%) and 42,281 (31.46%), respectively. We model the *Bacillus* genome sequence with a 2nd-order MC after estimating MC order by method developed in [9]. We estimate the transition probability based on the genome sequence as

$$\tilde{p}_{ij,k} = \frac{f_{ijk}}{f_{ij}},$$

which are shown in Supplementary Table S2. The stationary distribution π of the 2nd-order MC is calculated based on the estimated transition probabilities (Supplementary Table S2). Based on Theorem 3.1 in Billingsley [11], we derive the theoretical confidence for long genome sequence as follows

$$\mathrm{CI}_{1-\alpha}^{\mathrm{Long}} = \left(\tilde{p}_{ij,k} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_{ij,k}(1-\tilde{p}_{ij})}{f_{ij}}},\ \tilde{p}_{ij,k} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_{ij,k}(1-\tilde{p}_{ij,k})}{f_{ij}}} \right) \tag{6}$$

See Table 1 for an example of the confidence intervals for $p_{AA,A}$ by Eq. (6).

We then simulate NGS data sets by generating short reads from the *Bacillus phage SP-beta* genome sequence

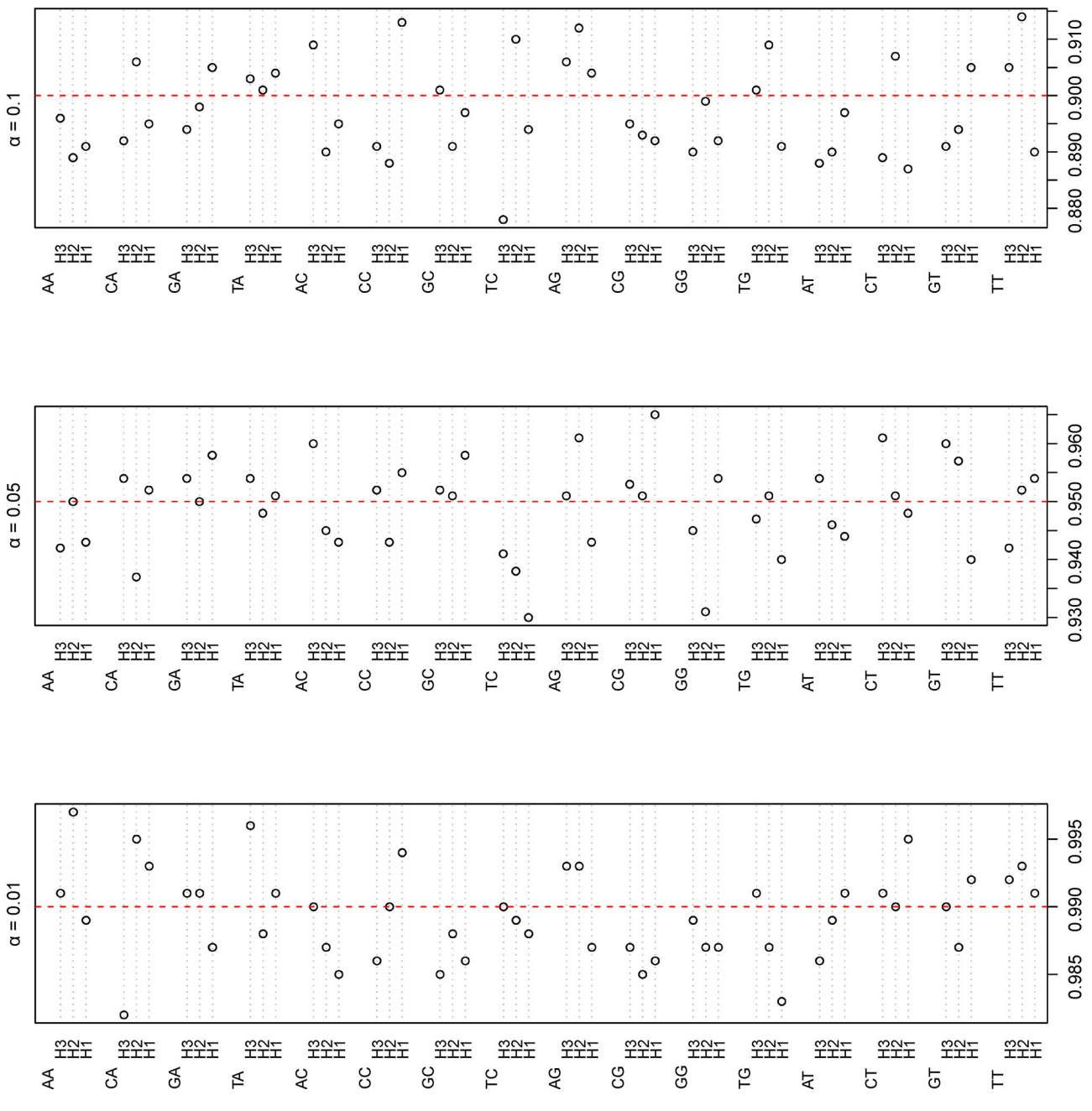

Figure 3. The fraction of times that the theoretical confident intervals of p_{ij} cover the true transition probability p_{ij} on homogeneous data sets H1, H2, and H3.

based on both homogeneous and inhomogeneous Poisson processes. The read length is β = 200 in bp and we generate NGS data sets using coverage $C = \beta c \in \{0.5, 1, 2, 5, 10\}$. We then estimate the MC transition matrix using NGS reads as

$$\hat{p}_{ij,k} = \frac{f_{ijk}^{(R)}}{f_{ij}^{(R)}},$$

and extend the theoretical confidence intervals of Eq. (16) for NGS to the 2nd-order MC transition probabilities $p_{ij,k}$ as follows

$$\mathrm{CI}_{1-\alpha}^{\mathrm{NGS}} = \left(\hat{p}_{ij,k} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{d\hat{p}_{ij,k}\left(1-\hat{p}_{ij,k}\right)}{f_{ij}^{(R)}}},\ \hat{p}_{ij,k} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{d\hat{p}_{ij,k}\left(1-\hat{p}_{ij,k}\right)}{f_{ij}^{(R)}}} \right). \quad (7)$$

We calculate the theoretical confidence intervals by Eq. (7) as well as the bootstrap confidence intervals by Eq. (19) for transition probability $p_{AA,A}$ on homogeneous NGS data sets (Table 1). Although we do not have true transition probabilities to validate these confidence intervals, we find that the theoretical confidence intervals by Eq. (7) as well as the parametric bootstrap by Eq. (19) are very close to the confidence intervals estimated from the long sequence by Eq. (6) (Table 1).

The theoretical confidence intervals by Eq. (7) for inhomogeneous NGS data sets are listed in Supplementary Table S3.

DISCUSSION

In this study, we propose statistics for estimations of MC transition probabilities and their corresponding confidence intervals based on next generation sequence reads. Although we did not rigorously prove the two propositions (Propositions 1 and 2) on the asymptotic distributions of $\xi_{ij}^{(R)}$, $\zeta_i^{(R)}$ and $S^{(R)}$, our simulation studies show that the propositions and the derived formulas for the confidence intervals of the transition probabilities perform decently well on simulated NGS datasets with wide regions of sequencing coverage.

In addition, we propose a parametric bootstrap method for the confidence intervals of MC transition probabilities based on NGS reads data, and show that both theoretical confidence intervals derived from the two propositions and the bootstrap methods achieve similar results with high accuracies, further supporting the validity of Propositions 1 and 2. However, it is worth noting that the parametric bootstrap based approach is computationally expensive in sampling data sets many times; while the theoretical confidence intervals bypass the need of the time consuming sampling step, thereby making it highly applicable to large-scale NGS datasets.

The developed theory and methods implicitly assume error-free for the NGS short reads data. Since the sequencing error rates of bulk sequencing are lower than 1% for Illumina sequencing machines, the error-free assumption on NGS data will not be a big issue. The rapid advances in single-cell sequencing (SCS) [16,17] provide unprecedented insights of cell heterogeneity. The statistical theory and methods developed in this study will be promising for the quantitative characterizations of SCS data. However, due to limited molecular materials within

Table 1 The estimated transition probability of $p_{AA,A}$, the theoretical confidence intervals as well as the parametric bootstrap confidence intervals on simulated homogeneous NGS data sets of *Bacillus phage SP-beta* genome.

			99%		95%		90%	
	$C = \beta c$	$p_{AA,A}$	Lower	Upper	Lower	Upper	Lower	Upper
$\mathrm{CI}_{1-\alpha}^{\mathrm{Long}}$ by Eq. (6)		0.3739	0.3644	0.3834	0.3667	0.3812	0.3678	0.3800
$\mathrm{CI}_{1-\alpha}^{\mathrm{NGS}}$ by Eq. (7)	0.5	0.3714	0.3551	0.3884	0.3590	0.3844	0.3611	0.3823
	1.0	0.3771	0.3640	0.3911	0.3672	0.3879	0.3689	0.3862
	2.0	0.3728	0.3610	0.3845	0.3638	0.3817	0.3653	0.3803
	5.0	0.3707	0.3375	0.4044	0.3454	0.3964	0.3496	0.3923
	10.0	0.3734	0.3633	0.3835	0.3657	0.3811	0.3670	0.3798
$\mathrm{CI}_{1-\alpha}^{B}$ by Eq. (19)	0.5	0.3781	0.3570	0.3903	0.3618	0.3869	0.3635	0.3852
	1.0	0.3736	0.3591	0.3885	0.3632	0.3847	0.3649	0.3831
	2.0	0.3723	0.3619	0.3858	0.3641	0.3825	0.3660	0.3814
	5.0	0.3758	0.3629	0.3835	0.3654	0.3814	0.3670	0.3807
	10.0	0.3753	0.3631	0.3834	0.3661	0.3811	0.3677	0.3801

a single cell as well as high dropout events, the SCS data are error-prone. For SCS data analysis, the sequencing errors are non-negligible and will blur the estimators, making their asymptotic distributions biased. We will study the asymptotic distribution theory of statistics based on the error-prone NGS data in our future work.

METHODS

Probabilistic modeling of a MC sequence and random sampling of the reads using NGS

In NGS, a large number of reads are randomly sampled from the genome. Hence two random processes are involved in the generation of the short reads data: i) the generation of the underlying genome sequence with genome length G base pairs (bps), and ii) random sampling of the M reads with length β from the simulated genome in i). See Fig. 1 for an illustration.

In this study, we assume the underlying genome sequence is generated by a Markov chain. We treat mainly the case of a first-order Markov chain, as the results can be easily adapted to higher order Markov chains. We thus use a first-order stationary and ergodic MC to model the underlying genome sequence with each letter taking values in a finite alphabet set $\mathcal{A}$ of size s. Since our study is based on genomic sequences, $\mathcal{A} = \{A, C, G, T\}$ and $s = 4$. We denote the transition matrix of MC as $P = [p_{ij}]_{4\times 4}$.

We model the random sampling of the reads using a Poisson process as in the Lander-Waterman model [18]. The Poisson process $\mathcal{P}$ is a process on $[0, \infty)$ and is assumed to be independent of the Markov chain. As in [14,18–21], we assume that the genome is a sequence of contiguous bases and that the distribution of reads along the genome follows a potentially *inhomogeneous* Poisson process with rate $c(x) = c(x, G)$ at position x. We suppress the argument G unless we take limits. If $c(x) = c$ for all x, we refer to the sampling of the reads as *homogeneous*. Allowing for inhomogeneity in the sampling Poisson process reflects observations about NGS data in [22] that the sampling rates of reads at different positions are not homogeneous.

We assume that all sampled reads have the same length of β bps. This length β does not depend on the length of the underlying sequence, nor on the parameters of the Markov chain or the Poisson process. For NGS data, β is given by the technology which is applied to obtain the reads.

A total of M reads are independently sampled from the genome of length G bps. The M reads are our data; they are a realisation of the above Poisson process and the Markov chain sample $\mathbf{A}$, as follows. Suppose that an event in $\mathcal{P}$ occurs at t. Let s be the closest integer $\leqslant t$. Then the Poisson event results in a read $R = (A_s, A_{s+1}, ..., A_{s+\beta})$ from the underlying Markov chain $\mathbf{A} = (A_1,..., A_G)$.

Statistics related to Markov chains based on NGS data

We extend the statistics ξ_{ij}, ζ_i and S for long sequence in [11] to $\xi_{ij}^{(R)}$, $\zeta_i^{(R)}$ and $S^{(R)}$ for the NGS short sequence/reads data accordingly as follows.

Let $f_{ij}^{(R)}$ be the number of occurrences of dinucleotide (ij) within the M reads and $f_i^{(R)} = \sum_j f_{ij}^{(R)}$. Define

$$\xi_{ij}^{(R)} = \frac{f_{ij}^{(R)} - f_i^{(R)} p_{ij}}{\sqrt{f_i^{(R)}}}. \tag{8}$$

Note that the moment estimation of p_{ij} from NGS data is $\hat{p}_{ij}^{(R)} = \frac{f_{ij}^{(R)}}{f_i^{(R)}}$, thus $\xi_{ij}^{(R)}$ can also be written as

$$\xi_{ij}^{(R)} = \sqrt{f_i^{(R)}}\left(\hat{p}_{ij}^{(R)} - p_{ij}\right).$$

Similarly, we define

$$\zeta_i^{(R)} = \frac{f_i^{(R)} - n^{(R)} p_i}{\sqrt{n^{(R)}}}, \tag{9}$$

and

$$S^{(R)} = \sum_{ij} \frac{(f_{ij}^{(R)} - f_i^{(R)} p_{ij})^2}{f_i^{(R)} p_{ij}}, \tag{10}$$

where

$$n^{(R)} = M\beta,$$

with M being the number of reads.

To see the centering in the above equations, we first calculate the expectation of $f_i^{(R)}$ and of $f_{ij}^{(R)}$. For this purpose we introduce some notations. For a read $R_k = (r_{k,1},..., r_{k,\beta})$, we denote its starting position on genome G by $S(R_k)$, rounded to the lower integer. The left-hand end of the read is a Poisson process which creates the read starts at t such that $s \leqslant t < s + 1$ for a nonnegative integer s, then we set $S(R_k) = s$. Let f_i be the number of occurrences of letter i in the genome, we have that

$$f_i = \sum_{s=1}^{G} 1(A_s = i), \quad \mathbb{E}f_i = Gp_i,$$

where 1 denotes the indicator function which is 1 when the argument is true, and 0 otherwise.

To calculate the counts of letter i in the M reads, the $f_i^{(R)}$ can be written in the form of $S(R_k)$ as follows

$$f_i^{(R)} = \sum_{s=1}^{G-\beta} \sum_{k=1}^{M} 1(S(R_k)=s) \sum_{\ell=1}^{\beta} 1(r_{k,\ell}=i)$$

$$= \sum_{s=1}^{G-\beta} \sum_{k=1}^{M} 1(S(R_k)=s) \sum_{\ell=1}^{\beta} 1(A_{s+\ell-1}=i).$$

By the independence of the Poisson process and the Markov chain, we have

$$\begin{aligned}\mathbb{E}f_i^{(R)} &= p_i\beta \sum_{s=1}^{G-\beta} \mathbb{E}\left(\sum_{k=1}^{M} 1(S(R_k)=s)\right)\\ &= p_i\beta \sum_{s=1}^{G-\beta} \mathbb{E}(\mathcal{P}([s,s+1)))\\ &= p_i\beta \int_0^{G-\beta} c(x)dx\\ &= (G-\beta)\overline{c}\beta p_i = M\beta p_i = n^{(R)}p_i,\end{aligned}$$

where $\overline{c}$ is the averaged coverage, which is calculated as follows

$$\overline{c} = \frac{1}{G-\beta}\int_0^{G-\beta} c(x)dx = M/(G-\beta).$$

This explains the centering in Eq. (9).

For the centering in Eq. (8) we first calculate

$$f_{ij}^{(R)} = \sum_{s=1}^{G-\beta} \sum_{k=1}^{M} 1(S(R_k)=s) \sum_{\ell=1}^{\beta-1} 1(r_{k,\ell}=i,\, r_{k,\ell+1}=j)$$

$$= \sum_{s=1}^{G-\beta} \sum_{k=1}^{M} 1(S(R_k)=s) \sum_{\ell=1}^{\beta-1} 1(A_{s+\ell-1}=i,\, A_{s+\ell}=j).$$

Note that

$$\mathbb{E}(1(A_{s+\ell-1}=i,\, A_{s+\ell}=j)|A_1,\ldots,A_{s+\ell-1}) = p_{ij}1(A_{s+\ell-1}=i),$$

we then have

$$\mathbb{E}f_{ij}^{(R)} = p_{ij} \sum_{s=1}^{G-\beta} \sum_{k=1}^{M} 1(S(R_k)=s) \sum_{\ell=1}^{\beta} 1(A_{s+\ell-1}=i)$$

$$= p_{ij}f_i^{(R)}.$$

This conditional expectation calculation makes the centering in Eq. (8) reasonable. Note that we do not center by $\mathbb{E}f_{ij}^{(R)} = M\beta p_i p_{ij}$. The centering in Eq. (10) follows the centering in Eq. (8).

The effective coverage d of NGS data

In order to state our results, we need to introduce some more notations. For a position $s \in \{1, 2,..., G\}$, let $L(s)$ denote the (random) number of reads which cover position s. For n = 1, 2,...,

$$C(n) = \{s \in \{1,2,...,G\} : L(s) = n\}$$

denote the set of positions that is covered by exactly n reads. As in [9], we also call $C(n)$ the region of coverage n. Note that a region does not need to be contiguous; we group all the parts of the genome with coverage n into the n-th region.

We define

$$d_n = \frac{1}{G}\mathbb{E}|C(n)|, \tag{11}$$

which is the expected percentage of the genome that has coverage n. We then define the **effective coverage** d as follows

$$d = d(G) = \frac{\sum_{n=1}^{\infty} n^2 d_n}{\sum_{n=1}^{\infty} n d_n}. \tag{12}$$

To calculate d, we note that

$$\mathbb{E}|C(n)| = \sum_{s=1}^{G} P(L(s)=n) = \sum_{s=1}^{G} P(\mathcal{P}(s-\beta,s]=n).$$

can be with an explicit expression, since $\mathcal{P}(s-\beta,s]$ has a Poisson distribution with parameter $\int_{s-\beta}^{s} c(x)dx$. In the homogeneous case, $\mathcal{P}(s-\beta,s]$ has Poisson distribution with parameter βc. Then d_n = P($Poi(\beta c)$ = n), where $Poi(\beta c)$ denotes a Poisson random variable with parameter βc. Thus in the homogeneous case, we have

$$\sum_{n=1}^{\infty} n d_n = \mathbb{E}Poi(\beta c) = \beta c,$$

and

$$\sum_{n=1}^{\infty} n^2 d_n = \mathbb{E}Poi^2(\beta c) = \beta c + (\beta c)^2.$$

Thus, in the homogeneous case, d has a simple expression as

$$d = 1 + \beta c. \tag{13}$$

Remark 1. The results in [9] also include d, but the notation is slightly different. The rate $c(x)$ in this paper translates into the rate $c(x)/\beta$ in [9] so that βc in this paper is c in [9]. We argue that in the setup, it is more intuitive to let the rate of the Poisson process not be scaled by the length of the reads, and hence we change in notation.

Asymptotic distributions of statistics related to Markov chains based on NGS data

The previous study [9] showed that under the same NGS short reads generative model as in the Section of "Methods" which models the underlying long genome sequence as a MC and uses the LanderWaterman model for random sampling of the reads, the traditional standardization of word count as

$$Z_{\mathbf{w}}^{(R)} = \frac{f_{\mathbf{w}}^{(R)} - \mathbb{E}_{\mathbf{w}}^{(R)}}{\hat{\sigma}_{\mathbf{w}}^{(R)}}, \tag{14}$$

where $(\hat{\sigma}_{\mathbf{w}}^{(R)})^2 = \mathbb{E}_{\mathbf{w}}^{(R)}\left(1 - \frac{f_{\mathbf{w}}^{(R)}}{f_{-\mathbf{w}-}^{(R)}}\right)\left(1 - \frac{f_{\mathbf{w}-}^{(R)}}{f_{-\mathbf{w}-}^{(R)}}\right)$, for NGS short reads does not converge to the standard normal distribution as that for a long sequence. Thus [9], proposed the concept of effective coverage d for NGS short reads data given in Eq. (12), and proved that $\frac{Z_{\mathbf{w}}^{R}}{\sqrt{d}}$ approximates to the standard normal distribution.

In this study, we show through simulations that the effective coverage d is also critical in the asymptotic distribution theory of MC transition probability estimators based on NGS short reads. We have the following two propositions for the asymptotic distributions of $\xi_{ij}^{(R)}$, $\zeta_i^{(R)}$ and $S^{(R)}$, all of which are related to MC for NGS short reads data.

Proposition 1. Assuming the underlying genomic sequence follows a stationary and ergodic MC with transition probability matrix $P = [p_{ij}]_{4\times 4}$. We also assume that a set of reads by NGS are randomly sampled according to the Poisson process, possibly inhomogeneous, with rate function $c(x)$. Then, as the genome length G tends to infinity, $\left(\frac{1}{\sqrt{d}}\xi_{ij}^{(R)}\right)_{i,j\in\mathcal{A}}$ converge in distribution to a mean zero normal vector with covariance Σ with entries $\sum_{i,j;k,l} = \lambda_{i,j;k,l}$ where d is the scaling factor given in Eq. (12) and $\lambda_{i,j;k,l}$ is given in Eq. (2).

Proposition 2. Under the same assumptions of Proposition 1, as $G \to \infty$,

1. The vector $\left(\frac{1}{\sqrt{d}}\zeta_i^{(R)}\right)_{i\in\mathcal{A}}$ converges in distribution to a mean zero normal vector with covariance matrix $(\alpha_{ij})_{i,j\in A}$.

2. The statistic $S^{(R)}/d$ has an approximate χ^2 -distribution with $s(s-1)$ degrees of freedom.

Remark 2. In this study, we only show the validity of the two propositions by simulation studies. Currently, we cannot provide rigorous mathematical proofs for the two propositions, due to the complex intercorrelations of words within reads. We leave them as our conjectures for future mathematical justifications.

Theoretical confidence intervals for MC transition probabilities based on NGS reads data

We show in Proposition 1 that, ξ_{ij}^{R} has an asymptotic normal distribution

$$\xi_{ij}^{R} = \frac{f_{ij}^{R} - f_i^R p_{ij}}{\sqrt{f_i^{(R)}}} \sim N(0,\ dp_{ij}(1-p_{ij})).$$

We apply this property to obtain the theoretical confidence interval for p_{ij} using NGS data. By dividing the numerator and denominator by $f_i^{(R)}$, we have

$$\xi_{ij}^{R} = \frac{\hat{p}_{ij} - p_{ij}}{\sqrt{1/f_i^{(R)}}} \sim N(0,\ dp_{ij}(1-p_{ij})),$$

where, $\hat{p}_{ij} = f_{ij}^R/f_i^R$, and hence,

$$\frac{\hat{p}_{ij} - p_{ij}}{\sqrt{dp_{ij}(1-p_{ij})/f_i^{(R)}}} \sim N(0,1).$$

For $0 < \alpha < 1$, let $z_{1-\alpha}$ be the number such that the area under the standard normal density function to the left of $z_{1-\alpha}$ is $1-\alpha$. Thus, we have

$$P\left(-z_{1-\frac{\alpha}{2}} \leqslant \frac{\hat{p}_{ij} - p_{ij}}{\sqrt{dp_{ij}(1-p_{ij})/f_i^{(R)}}} \leqslant z_{1-\frac{\alpha}{2}}\right) = 1-\alpha. \tag{15}$$

Manipulation of the inequalities gives

$$P\left(\frac{1}{1+\gamma}\left(\hat{p}_{ij} + \frac{1}{2}\gamma - \sqrt{\gamma\hat{p}_{ij}(1-\hat{p}_{ij}) + \gamma^2}\right)\right.$$

$$\left. \leqslant p_{ij} \leqslant \frac{1}{1+\gamma}\left(\hat{p}_{ij} + \frac{1}{2}\gamma + \sqrt{\gamma\hat{p}_{ij}(1-\hat{p}_{ij}) + \gamma^2}\right)\right) = 1-\alpha,$$

wher $\gamma = \frac{z_{1-\frac{\alpha}{2}}^2 d}{f_i^{(R)}}$. Therefore, we obtain the level $1-\alpha$ confidence interval for p_{ij} as

$$\mathrm{CI}_{1-\alpha} = \frac{1}{1+\gamma}\left(\hat{p}_{ij} + \frac{1}{2}\gamma - \sqrt{\gamma\hat{p}_{ij}(1-\hat{p}_{ij}) + \gamma^2},\right.$$

$$\left.\hat{p}_{ij} + \frac{1}{2}\gamma + \sqrt{\gamma\hat{p}_{ij}(1-\hat{p}_{ij}) + \gamma^2}\right).$$

In practice, $f_i^{(R)}$ can be very large, making $f_i^{(R)} \gg z_{1-\frac{\alpha}{2}}^2 d$ and $\gamma \approx 0$. We thus discard terms of resulting in a simplified confidence interval for p_{ij} as

$$\mathrm{CI}_{1-\alpha}^{\mathrm{NGS}}=\left(\hat{p}_{ij}-z_{1-\frac{\alpha}{2}}\sqrt{\frac{d\hat{p}_{ij}\left(1-\hat{p}_{ij}\right)}{f_i^{(R)}}},\right.$$

$$\left.\hat{p}_{ij}+z_{1-\frac{\alpha}{2}}\sqrt{\frac{d\hat{p}_{ij}\left(1-\hat{p}_{ij}\right)}{f_i^{(R)}}}\right). \quad (16)$$

Parametric bootstrap for obtaining the confidence intervals for transition probabilities and the stationary distribution of MC

Let X be a stationary and ergodic Markov chain with transition matrix $P=[p_{ij}]_{s\times s}$. Assume that a set of reads are randomly sampled from X according to the Poisson process, denoted by $\{R_1, R_2, ..., R_M\}$. We can estimate the transition matrix P and stationary distribution π by $\hat{P}=[\hat{p}_{ij}]$ and $\hat{\pi}=[\hat{\pi}_i]$, where,

$$\hat{p}_{ij}=\frac{f_{ij}^{(R)}}{f_i^{(R)}}, \quad (17)$$

$$\hat{\pi}_i=\frac{\sum_{j\in S}f_{ij}^{(R)}}{\sum_{m=1}^{M}(|R_m|-1)}, \quad (18)$$

where $|R_m|$ is the length of read R_m, m = 1, 2,..., M . Moreover,

$$\hat{\pi}\hat{P}=\hat{\pi},$$

$$\sum_{i\in S}\hat{\pi}_i=1.$$

When $G\to\infty$,

$$\lim_{G\to\infty}\hat{P}=P,\ \lim_{G\to\infty}\hat{\pi}_i=\pi_i.$$

We provide a parametric bootstrap method for finding the confidence intervals for p_{ij} and π_i as follows:

B1 Simulate the NGS data set. Simulate a genome sequence of length G by the stationary and ergodic Markov chain P, and the sequence starts from the stationary distribution π generate NGS data set by sampling M reads $\{R_1,R_2, ..., R_M\}$ from the simulated genome by the Poisson process. Note that, for a real data set, we skip Step B1 and go directly to B2.

B2 Given the set of reads of the NGS data set (either a simulated one from Step B1 or a real one), estimate parameters of transition matrix $\hat{P}$ and stationary distribution $\hat{\pi}$ by Eqs. (17) and (18).

B3 Generate bootstrap data set. Simulate a new Markov genome sequence of length G based on the estimated parameters $\hat{P}$ and $\hat{\pi}$ in B2; and generate a bootstrap data set with M reads $\{R_1^*,R_2^*, ..., R_M^*\}$ by the Poisson process.

B4 Estimate parameters of transition matrix $\hat{P}^*$ and stationary distribution $\hat{\pi}^*$ from the simulated bootstrap data set in B3 by Eqs. (17) and (18).

B5 Repeat Steps B3–B4 B times to get B estimators $\hat{p}_{ij}^*$ and $\hat{\pi}_i^*$; For each dinucleotide, order $\hat{p}_{ij}^*$ from the smallest to largest values as $\hat{p}_{ij}^*(1)$, $\hat{p}_{ij}^*(2)$, ..., $\hat{p}_{ij}^*(B)$, and the level $1-\alpha$ confidence interval for p_{ij} is given by

$$\mathrm{CI}_{1-\alpha}^{B}=\left(\hat{p}_{ij}^*\left(\left\lfloor B\times\frac{\alpha}{2}\right\rfloor\right),\hat{p}_{ij}^*\left(\left\lceil B\times\left(1-\frac{\alpha}{2}\right)\right\rceil\right)\right); \quad (19)$$

For each i, order $\hat{\pi}_i^*$ from the smallest to largest values as $\hat{\pi}_i^*(1)$, $\hat{\pi}_i^*(2)$, ..., $\hat{\pi}_i^*(B)$, and the level $1-\alpha$ confidence interval for π_i is given by

$$\mathrm{CI}_{1-\alpha}^{\mathrm{b}}=\left(\hat{\pi}_i^*\left(\left\lfloor B\times\frac{\alpha}{2}\right\rfloor\right),\hat{\pi}_i^*\left(\left\lceil B\times\left(1-\frac{\alpha}{2}\right)\right\rceil\right)\right). \quad (20)$$

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://10.1007/s40484-020-0200-y.

ACKNOWLEDGEMENTS

We thank Dr. Gesine Reinert of Oxford University for discussions and help related to the problems studied in this paper. LW is supported by NSFC grants (Nos.11571349 and 91630314), the National Key R&D Program of China under Grant 2018YFB0704304, NCMIS of CAS, LSC of CAS, and the Youth Innovation Promotion Association of CAS. JR and FS were supported by US National Science Foundation (NSF) (DMS-1518001) and National Institutes of Health (NIH) (R01GM120624, 1R01GM131407).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Lin Wan, Xin Kang, Jie Ren and Fengzhu Sun declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Almagor, H. (1983) A Markov analysis of DNA sequences. J. Theor. Biol., 104, 633–645
2. Reinert, G., Schbath, S. and Waterman, M. S. (2005) Statistics on words with applications to biological sequences. In: Applied Combinatorics on Words, Lothaire, M. (ed.), ch. 6, pp. 268–352 New York: Cambridge University Press
3. Blaisdell, B. E. (1985) Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. J. Mol. Evol., 21, 278–288
4. Pevzner, P. A., Borodovsky, M.Y., and MironovA. A. (1989) Linguistics of nucleotide sequences. I: The significance of deviations from mean statistical characteristics and prediction of

the frequencies of occurrence of words. J. Biomol. Struct. Dyn., 6, 1013–1026
5. Hong, J. (1990) Prediction of oligonucleotide frequencies based upon dinucleotide frequencies obtained from the nearest neighbor analysis. Nucleic Acids Res., 18, 1625–1628
6. Arnold, J., Cuticchia, A. J., Newsome, D. A., Jennings, III, W. W. and Ivarie, R. (1988) Mono-through hexanucleotide composition of the sense strand of yeast DNA: a Markov chain analysis. Nucleic Acids Res., 16, 7145–7158
7. Avery, P. J. (1987) The analysis of intron data and their use in the detection of short signals. J. Mol. Evol., 26, 335–340
8. Narlikar, L., Mehta, N., Galande, S. and Arjunwadkar, M. (2013) One size does not fit all: on how Markov model order dictates performance of genomic sequence analyses. Nucleic Acids Res., 41, 1416–1424
9. Ren, J., Song, K., Deng, M., Reinert, G., Cannon, C. H. and Sun, F. (2016) Inference of Markovian properties of molecular sequences from NGS data and applications to comparative genomics. Bioinformatics, 32, 993–1000
10. Billingsley, P. (1961) Statistical Inference for Markov Processes, vol. 2. Chicago: University of Chicago Press Chicago
11. Billingsley, P. (1961) Statistical methods in Markov chains. Ann. Math. Stat., 32, 12–40
12. Pevzner, P. A., Tang, H. and Waterman, M. S. (2001) An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. USA, 98, 9748–9753
13. Zerbino, D. R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using *de Bruijn* graphs. Genome Res., 18, 821–829
14. Zhai, Z., Reinert, G., Song, K., Waterman, M. S., Luan, Y. and Sun, F. (2012) Normal and compound Poisson approximations for pattern occurrences in NGS reads. J. Comput. Biol., 19, 839–854
15. Song, K., Ren, J., Zhai, Z., Liu, X., Deng, M. and Sun, F. (2013) Alignment-free sequence comparison based on next-generation sequencing reads. J. Comput. Biol., 20, 64–79
16. Sun, F., Arnheim, N. and Waterman, M. S. (1995) Whole genome amplification of single cells: mathematical analysis of PEP and tagged PCR. Nucleic Acids Res., 23, 3034–3040
17. Daley, T.and Smith, A. D. (2014) Modeling genome coverage in single-cell sequencing. Bioinformatics, 30, 22, 3159–3165
18. Lander, E. S. and Waterman, M. S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics, 2, 231–239
19. Zhang, Z. D., Rozowsky, J., Snyder, M., Chang, J. and Gerstein, M. (2008) Modeling ChIP sequencing *in silico* with applications. PLOS Comput. Biol., 4, e1000158
20. Daley, T. and Smith, A. D. (2013) Predicting the molecular complexity of sequencing libraries. Nat. Methods, 10, 325–327
21. Simpson, J. T. (2014) Exploring genome characteristics and sequence quality without a reference. Bioinformatics, 30, 1228–1235
22. Schwartz, S., Oren, R. and Ast, G. (2011) Detection and removal of biases in the analysis of next-generation sequencing reads. PLoS One, 6, e16685

