Alignment-Free Sequence Comparison Based on Next Generation Sequencing Reads: Supplementary Materials

Kai Song[1], Jie Ren[1], Zhiyuan Zhai[2], Xuemei Liu[3], Minghua Deng[1*], and Fengzhu Sun[4,5*]

[1] School of Mathematics, Peking University, Beijing, PR China
[2] School of Mathematics, Shandong University, PR China
[3] School of Physics, South China University of Technology, Guangzhou, PR China
[4] TNLIST/Department of Automation, Tsinghua University, Beijing, PR China
[5] Molecular and Computational Biology Program, University of Southern California, Los Angeles, California, USA.
*Corresponding authors (dengmh@pku.edu.cn, fsun@usc.edu)

Proofs of the Propositions and Theorems

In this section, we give proofs for Propositions 1-2 and Theorems 1-5. First, we present a method for the calculation of $\mathbf{E}_{\beta,\eta}(\mathbf{u},\mathbf{v})$.

1). For the calculation of $\mathbf{E}_{\beta,0}(\mathbf{u},\mathbf{v})=\mathbf{E}(X_{\mathbf{u}}[1,\beta]X_{\mathbf{v}}[1,\beta])$, we first define $I_t(\mathbf{u})=I(A_tA_{t+1}\cdots A_{t+k-1}=\mathbf{u})$. Then we have

$$
\begin{aligned}
&\mathbf{E}\Big(X_{\mathbf{u}}[1,\beta]X_{\mathbf{v}}[1,\beta]\Big)=\mathbf{E}\Big(\sum_{i=1}^{\beta-k+1}(I_i(\mathbf{u})+I_i(\bar{\mathbf{u}}))\sum_{j=1}^{\beta-k+1}(I_j(\mathbf{v})+I_j(\bar{\mathbf{v}}))\Big)\\
&=\sum_{i=1}^{\beta-k+1}\sum_{j=1}^{\beta-k+1}\mathbf{E}\Big(I_i(\mathbf{u})I_j(\mathbf{v})+I_i(\mathbf{u})I_j(\bar{\mathbf{v}})+I_i(\bar{\mathbf{u}})I_j(\mathbf{v})+I_i(\bar{\mathbf{u}})I_j(\bar{\mathbf{v}})\Big)\\
&=(\beta-k+1)\Big(\beta_{\mathbf{u},\mathbf{v}}(0)P(\mathbf{u}_0\mathbf{v})+\beta_{\mathbf{u},\bar{\mathbf{v}}}(0)P(\mathbf{u}_0\bar{\mathbf{v}})+\beta_{\bar{\mathbf{u}},\mathbf{v}}(0)P(\bar{\mathbf{u}}_0\mathbf{v})+\beta_{\bar{\mathbf{u}},\bar{\mathbf{v}}}(0)P(\bar{\mathbf{u}}_0\bar{\mathbf{v}})\Big)\\
&+\sum_{j=1}^{k-1}(\beta-k-j+1)\Big(\beta_{\mathbf{u},\mathbf{v}}(j)P(\mathbf{u}_j\mathbf{v})+\beta_{\mathbf{u},\bar{\mathbf{v}}}(j)P(\mathbf{u}_j\bar{\mathbf{v}})+\beta_{\bar{\mathbf{u}},\mathbf{v}}(j)P(\bar{\mathbf{u}}_j\mathbf{v})+\beta_{\bar{\mathbf{u}},\bar{\mathbf{v}}}(j)P(\bar{\mathbf{u}}_j\bar{\mathbf{v}})\Big)\\
&+\sum_{j=1}^{k-1}(\beta-k-j+1)\Big(\beta_{\mathbf{v},\mathbf{u}}(j)P(\mathbf{v}_j\mathbf{u})+\beta_{\mathbf{v},\bar{\mathbf{u}}}(j)P(\mathbf{v}_j\bar{\mathbf{u}})+\beta_{\bar{\mathbf{v}},\mathbf{u}}(j)P(\bar{\mathbf{v}}_j\mathbf{u})+\beta_{\bar{\mathbf{v}},\bar{\mathbf{u}}}(j)P(\bar{\mathbf{v}}_j\bar{\mathbf{u}})\Big)\\
&+\sum_{j=0}^{\beta-2k}(\beta-2k-j+1)\Big(\mathbf{E}I_1(\mathbf{u})I_{k+1+j}(\mathbf{v})+\mathbf{E}I_1(\mathbf{u})I_{k+1+j}(\bar{\mathbf{v}})+\mathbf{E}I_1(\bar{\mathbf{u}})I_{k+1+j}(\mathbf{v})+\mathbf{E}I_1(\bar{\mathbf{u}})I_{k+1+j}(\bar{\mathbf{v}})\Big)\\
&+\sum_{j=0}^{\beta-2k}(\beta-2k-j+1)\Big(\mathbf{E}I_1(\mathbf{v})I_{k+1+j}(\mathbf{u})+\mathbf{E}I_1(\mathbf{v})I_{k+1+j}(\bar{\mathbf{u}})+\mathbf{E}I_1(\bar{\mathbf{v}})I_{k+1+j}(\mathbf{u})+\mathbf{E}I_1(\bar{\mathbf{v}})I_{k+1+j}(\bar{\mathbf{u}})\Big),
\end{aligned}
$$

where $\mathbf{u}_j\mathbf{v}=\mathbf{u}_1\cdots\mathbf{u}_j\mathbf{v}_1\cdots\mathbf{v}_k$.

2). For $0 \le \eta \le \beta - 1$, we have

$$\begin{aligned}
&\mathbf{E}_{\beta,\eta}(\mathbf{u},\mathbf{v}) = \mathbf{E}_{\beta-\eta,0}(\mathbf{u},\mathbf{v}) \\
&+ \sum_{j=1}^{k-1} \Big(\eta - (\eta - j + 1) \vee 1 + 1\Big)\Big(\beta_{\mathbf{u},\mathbf{v}}(l)P(\mathbf{u}_j\mathbf{v}) + \beta_{\mathbf{u},\bar{\mathbf{v}}}(l)P(\mathbf{u}_j\bar{\mathbf{v}}) + \beta_{\bar{\mathbf{u}},\mathbf{v}}(l)P(\bar{\mathbf{u}}_j\mathbf{v}) + \beta_{\bar{\mathbf{u}},\bar{\mathbf{v}}}(l)P(\bar{\mathbf{u}}_j\bar{\mathbf{v}})\Big) \\
&+ \sum_{j=1}^{k-1} \Big((\beta - \eta - k - j + 1) \wedge (\beta - k + 1) - (\beta - k + 2 - j) \vee (\eta + 1) + 1\Big)\Big(\beta_{\mathbf{u},\mathbf{v}}(l)P(\mathbf{u}_j\mathbf{v}) \\
&+ \beta_{\mathbf{u},\bar{\mathbf{v}}}(l)P(\mathbf{u}_j\bar{\mathbf{v}}) + \beta_{\bar{\mathbf{u}},\mathbf{v}}(l)P(\bar{\mathbf{u}}_j\mathbf{v}) + \beta_{\bar{\mathbf{u}},\bar{\mathbf{v}}}(l)P(\bar{\mathbf{u}}_j\bar{\mathbf{v}})\Big) \\
&+ \sum_{j=0}^{\beta+\eta-2k} h_j\Big(\mathbf{E}I_1(\mathbf{u})I_{k+1+j}(\mathbf{v}) + \mathbf{E}I_1(\mathbf{u})I_{k+1+j}(\bar{\mathbf{v}}) + \mathbf{E}I_1(\bar{\mathbf{u}})I_{k+1+j}(\mathbf{v}) + \mathbf{E}I_1(\bar{\mathbf{u}})I_{k+1+j}(\bar{\mathbf{v}})\Big) \\
&+ \sum_{j=0}^{\beta-2k} (R(j) - L(j) + 1)\Big(\mathbf{E}I_1(\mathbf{u})I_{k+1+j}(\mathbf{v}) + \mathbf{E}I_1(\mathbf{u})I_{k+1+j}(\bar{\mathbf{v}}) + \mathbf{E}I_1(\bar{\mathbf{u}})I_{k+1+j}(\mathbf{v}) + \mathbf{E}I_1(\bar{\mathbf{u}})I_{k+1+j}(\bar{\mathbf{v}})\Big),
\end{aligned}$$

where the summations are 0 when the upper limit is smaller than the lower limit,

$$h_j = \begin{cases} j + k, & 0 \le j \le \eta - k; \\ \eta, & \eta - k < j \le \beta - 2k + 1; \\ \beta + \eta - 2k - j + 1, & \beta - 2k + 1 < j \le \beta + \eta - 2k, \end{cases}$$

and we use the abbreviation $L(j) = (\beta - 2k - j + 2) \vee (\eta + 1)$ and $R(j) = (\beta - k + 1) \wedge (\eta + \beta - 2k - j + 1)$. The $\beta_{\mathbf{u},\mathbf{v}}(l)$ is the overlap bit which equals 1 if $\mathbf{u}_{l+i} = \mathbf{v}_i, i = 1, 2, \cdots, k - l$, and $\beta_{\mathbf{u},\mathbf{v}}(l) = 0$ otherwise.

3). For $\eta \ge \beta$, we have

$$\mathbf{E}_{\beta,\eta}(\mathbf{u},\mathbf{v}) = \sum_{j=\eta-\beta}^{\eta+\beta-2k} f_\eta(j)\Big(\mathbf{E}I_1(\mathbf{u})I_{k+1+j}(\mathbf{v}) + \mathbf{E}I_1(\mathbf{u})I_{k+1+j}(\bar{\mathbf{v}}) + \mathbf{E}I_1(\bar{\mathbf{u}})I_{k+1+j}(\mathbf{v}) + \mathbf{E}I_1(\bar{\mathbf{u}})I_{k+1+j}(\bar{\mathbf{v}})\Big),$$

where $f_\eta(j) = (\beta + j + 1) \wedge (\eta + \beta - k + 1) - j - k - (\eta + 1 - j - k) \vee 1 + 1$.

Proof of Proposition 1:

For the calculation of the covariance of $X_{\mathbf{u}}$ and $X_{\mathbf{v}}$, we have

$$\begin{aligned}
\mathrm{Cov}(X_{\mathbf{u}}, X_{\mathbf{v}}) &= \Big(M + M(M-1)\sum_{i=1}^{n-\beta+1} \lambda_i^2\Big)\Big(\mathbf{E}_{\beta,0}(\mathbf{u},\mathbf{v}) - (\beta - k + 1)^2(P(\mathbf{u}) + P(\bar{\mathbf{u}}))(P(\mathbf{v}) + P(\bar{\mathbf{v}}))\Big) \\
&+ M(M-1)\sum_{i=1}^{n-\beta+1} \lambda_i \sum_{\eta=1}^{n-i-\beta+1} \lambda_{i+\eta}\Big(\mathbf{E}_{\beta,\eta}(\mathbf{u},\mathbf{v}) + \mathbf{E}_{\beta,\eta}(\mathbf{v},\mathbf{u}) - 2(\beta - k + 1)^2(P(\mathbf{u}) + P(\bar{\mathbf{u}}))(P(\mathbf{v}) + P(\bar{\mathbf{v}}))\Big).
\end{aligned}$$

So if $\lim_{n\to\infty}(n - \beta - \eta + 1)\sum_{i=1}^{n-\beta-\eta+1} \lambda_i\lambda_{i+\eta} = r_\eta$ and M depends on n such that $\lim_{n\to\infty} M/n = \theta$, then

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\mathrm{Cov}(X_{\mathbf{u}}, X_{\mathbf{v}})}{M} \\
&= (1 + \theta r_0)\Big(\mathbf{E}_{\beta,0}(\mathbf{u},\mathbf{v}) - (\beta - k + 1)^2(P(\mathbf{u}) + P(\bar{\mathbf{u}}))(P(\mathbf{v}) + P(\bar{\mathbf{v}}))\Big) \\
&+ \theta\sum_{\eta=1}^{\infty} r_\eta\Big(\mathbf{E}_{\beta,\eta}(\mathbf{u},\mathbf{v}) + \mathbf{E}_{\beta,\eta}(\mathbf{v},\mathbf{u}) - 2(\beta - k + 1)^2(P(\mathbf{u}) + P(\bar{\mathbf{u}}))(P(\mathbf{v}) + P(\bar{\mathbf{v}}))\Big).
\end{aligned} \tag{1}$$

Proposition 2 can be proved easily by the Central Limit Theorem.

Proof of Theorem 1:

It is obvious that a word pattern $\mathbf{w}$ occurs in a read or it's complement if and only $\mathbf{w}$ or $\bar{\mathbf{w}}$ occurs in the read from the forward strand. So we have $\mathbf{E}X_{\mathbf{w}} = M(\beta - k + 1)(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}))$.

$$\mathbf{E}D_2 = \sum \mathbf{E}X_{\mathbf{w}}\mathbf{E}Y_{\mathbf{w}} = M^2(\beta - k + 1)^2 \sum (P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}))^2,$$

Also we have $\mathbf{E}\tilde{X}_{\mathbf{w}} = M(\beta - k + 1)(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}) - p_{\mathbf{w}} - p_{\bar{\mathbf{w}}})$. Thus,

$$\mathbf{E}D_2^* = \sum \frac{\mathbf{E}\tilde{X}_{\mathbf{w}}\mathbf{E}\tilde{Y}_{\bar{\mathbf{w}}}}{M(\beta - k + 1)(p_{\mathbf{w}} + p_{\bar{\mathbf{w}}})} = M(\beta - k + 1) \sum \frac{(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}) - p_{\mathbf{w}} - p_{\bar{\mathbf{w}}})^2}{p_{\mathbf{w}} + p_{\bar{\mathbf{w}}}}.$$

For the proof of the approximate mean of D_2^S, we first use the Taylor expansion,

$$\frac{(x + a)(y + a)}{\sqrt{(x + a)^2 + (y + a)^2}} = \frac{|a|}{\sqrt{2}} + \frac{\text{sign}(a)}{2\sqrt{2}}(x + y) + O(x^2 + y^2).$$

For each $\mathbf{w}$, let $a = (\beta - k + 1)(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}) - p_{\mathbf{w}} - p_{\bar{\mathbf{w}}})$, $x = X_{\mathbf{w}}/M - (\beta - k + 1)(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}))$, and $y = Y_{\mathbf{w}}/M - (\beta - k + 1)(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}))$. Then with the Taylor expansion,

$$\begin{aligned}
&\frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{M\sqrt{\tilde{X}_{\mathbf{w}}^2 + \tilde{Y}_{\mathbf{w}}^2}} = \frac{(\beta - k + 1)}{\sqrt{2}}|(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}) - p_{\mathbf{w}} - p_{\bar{\mathbf{w}}})| \\
&+ \frac{(\beta - k + 1)\text{sign}(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}) - p_{\mathbf{w}} - p_{\bar{\mathbf{w}}})}{2\sqrt{2}}\left(\frac{X_{\mathbf{w}} + Y_{\mathbf{w}}}{M} - 2(\beta - k + 1)(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}))\right) \\
&+ O\Big(\big(X_{\mathbf{w}}/M - (\beta - k + 1)(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}))\big)^2 + \big(Y_{\mathbf{w}}/M - (\beta - k + 1)(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}))\big)^2\Big).
\end{aligned}$$

Taking expectation, we have

$$\lim_{n\to\infty} \frac{\mathbf{E}D_2^S}{M} = \frac{\beta - k + 1}{\sqrt{2}} \sum \big|\big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big) - \big(p_{\mathbf{w}} + p_{\bar{\mathbf{w}}}\big)\big|.$$

Further, we refine the Taylor expansion to

$$\frac{(x + a)(y + a)}{\sqrt{(x + a)^2 + (y + a)^2}} = \frac{|a|}{\sqrt{2}} + \frac{\text{sign}(a)}{2\sqrt{2}}(x + y) - \frac{3}{8\sqrt{2}|a|}\big(x - y\big)^2.$$

Thus, we have

$$\lim_{n\to\infty} M\left(\frac{\mathbf{E}D_2^S}{M} - \frac{\beta - k + 1}{\sqrt{2}} \sum \big|\big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big) - \big(p_{\mathbf{w}} + p_{\bar{\mathbf{w}}}\big)\big|\right) = -\frac{3\sqrt{2}}{8(\beta - k + 1)} \sum \frac{\sigma_\rho^2(\mathbf{w})}{\big|\big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big) - \big(p_{\mathbf{w}} + p_{\bar{\mathbf{w}}}\big)\big|}.$$

Proof of Theorem 2:

From the definition of D_2, we have

$$\begin{aligned}
\frac{D_2}{M^2} &= \sum \frac{X_{\mathbf{w}}}{M}\frac{Y_{\mathbf{w}}}{M} = \sum \left(\frac{X_{\mathbf{w}}}{M} - (\beta - k + 1)\big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big) + (\beta - k + 1)\big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big)\right) \\
&\quad \left(Y_{\mathbf{w}} - (\beta - k + 1)\big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big) + (\beta - k + 1)\big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big)\right) \\
&= \sum \left(\frac{X_{\mathbf{w}}}{M} - (\beta - k + 1)\big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big)\right)\left(Y_{\mathbf{w}} - (\beta - k + 1)\big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big)\right) \\
&\quad + (\beta - k + 1) \sum \big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big)\left(\frac{X_{\mathbf{w}}}{M} - (\beta - k + 1)\big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big)\right) \\
&\quad + (\beta - k + 1) \sum \big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big)\left(\frac{Y_{\mathbf{w}}}{M} - (\beta - k + 1)\big(P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}})\big)\right) \\
&\quad + (\beta - k + 1)^2 \sum (P_\rho(\mathbf{w}) + P_\rho(\bar{\mathbf{w}}))^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sqrt{M}\Big(\frac{D_2}{M^2}-(\beta-k+1)^2\sum\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)^2\Big)\\
&=\frac{1}{\sqrt{M}}\sum\sqrt{M}\Big(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\Big)\sqrt{M}\Big(\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\Big)\\
&+(\beta-k+1)\sum\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\Big(\sqrt{M}\Big(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\Big)\\
&+\sqrt{M}\Big(\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\Big)\Big).
\end{aligned}\tag{2}$$

For $0<\rho\le 1$, it has been shown in Proposition 2 that in distribution,

$$\lim_{n\to\infty}\sqrt{M}\Big(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\Big)=N(0,\sigma_\rho^2(\mathbf{w})),$$

$$\lim_{n\to\infty}\sqrt{M}\Big(\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\Big)=N(0,\sigma_\rho^2(\mathbf{w})),$$

where $\sigma_\rho^2(\mathbf{w})=\lim_{n\to\infty}\frac{\mathrm{Var}(X_{\mathbf{w}})}{M}$. Therefore, the first term in (2) tends to 0 when $M\to\infty$. Thus,

$$\Big(\sqrt{M}\Big(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\Big)+\sqrt{M}\Big(\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\Big)\Big)\to N(0,2\sigma_\rho^2(\mathbf{w})).$$

Let $\sigma_\rho(\mathbf{w},\mathbf{w}')=\lim_{n\to\infty}\frac{\mathrm{Cov}(X_{\mathbf{w}},X_{\mathbf{w}'})}{M}$ which can be calculated as in (4) in the main text. Since $\{X_{\mathbf{w}},\mathbf{w}\in\mathcal{A}^k\}$ and $\{Y_{\mathbf{w}},\mathbf{w}\in\mathcal{A}^k\}$ are independent, the second term in equation (2) is asymptotically normal with mean 0 and variance $2(\Sigma_\rho)^2$. The first term in equation (2) tends to 0. Theorem 2 is proved.

Proof of Theorem 3:

a). For the case of $\rho=1$, we have

$$\lim_{n\to\infty}\sqrt{M}\Big(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)\Big)=N(0,\sigma_1^2(\mathbf{w})),\tag{3}$$

$$\lim_{n\to\infty}\sqrt{M}\Big(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)\Big)=N(0,\sigma_1^2(\mathbf{w})).\tag{4}$$

From the definition of D_2^*, we have

$$\begin{aligned}
\lim_{n\to\infty}D_2^*&=\lim_{n\to\infty}\sum\frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{M(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)}\\
&=\lim_{n\to\infty}\sum\frac{\big(X_{\mathbf{w}}-M(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)\big)\big(Y_{\mathbf{w}}-M(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)\big)}{M(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)}\\
&=\lim_{n\to\infty}\sum\sqrt{M}\Big(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)\Big)\sqrt{M}\Big(\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)\Big)/(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)\\
&=\sum\frac{Z_{\mathbf{w}}^{(1)}Z_{\mathbf{w}}^{(2)}}{(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)},
\end{aligned}$$

where $\{Z_{\mathbf{w}}^{(1)},\mathbf{w}\in\mathcal{A}_k\}$ and $\{Z_{\mathbf{w}}^{(2)},\mathbf{w}\in\mathcal{A}_k\}$ are independent and have the same mean 0 normal distribution.

b). To prove the second part, we note that

$$\begin{aligned}\frac{D_2^*}{M} &= \sum \frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{M^2(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)}\\ &= \sum \frac{1}{(\beta-k+1)\big(p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}\big)}\bigg(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\bigg)\bigg(\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\bigg)\\ &+ \sum \frac{(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}})}{p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}}\bigg(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)+\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\bigg)\\ &+ (\beta-k+1)\sum \frac{(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}})^2}{p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}}.\end{aligned}$$

It follows from the normal approximation for individual word counts that in distribution,

$$\lim_{n\to\infty}\sqrt{M}\sum \frac{1}{p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}}\bigg(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\bigg)\bigg(\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\bigg)\to 0.$$

Therefore, in distribution,

$$\begin{aligned}&\lim_{n\to\infty}\sqrt{M}\bigg(\frac{D_2^*}{M}-(\beta-k+1)\sum\frac{(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}})^2}{p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}}\bigg)\\ &=\lim_{n\to\infty}\sum\frac{(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}})}{p_{\mathbf{w}}+p_{\bar{\mathbf{w}}}}\sqrt{M}\bigg(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\\ &+\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)\big(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})\big)\bigg).\end{aligned}$$

So the expression has a normal distribution with mean 0 and variance $2(\Sigma_\rho^*)^2$, where Σ_ρ^* is given in the main text of Theorem 3. The theorem is proved.

Proof of Theorem 4:

a). The proof of this part is similar to the proof of the first part of Theorem 3 and is omitted here.

b). For this part, using Taylor expansion, it is straightforward to show that for any $a\neq 0$ and (x,y) in the neighborhood of $(0,0)$,

$$\frac{(x+a)(y+a)}{\sqrt{(x+a)^2+(y+a)^2}}=\frac{|a|}{\sqrt{2}}+\frac{\operatorname{sign}(a)}{2\sqrt{2}}(x+y)+O(x^2+y^2).$$

For each word $\mathbf{w}$, let $a=(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}}), x=\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))$, and $y=\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))$. Then, with the Taylor expansion,

$$\begin{aligned}\frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{M\sqrt{\tilde{X}_{\mathbf{w}}^2+\tilde{Y}_{\mathbf{w}}^2}} &= \frac{(\beta-k+1)|P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}}|}{\sqrt{2}}\\ &+\frac{\operatorname{sign}(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}})}{2\sqrt{2}}\bigg(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))\\ &+\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))\bigg)\\ &+O\bigg(\bigg(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))\bigg)^2+\bigg(\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))\bigg)^2\bigg).\end{aligned}$$

Taking expectations we obtain that

$$\begin{aligned}\mathbf{E}_\rho\bigg(\frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{M\sqrt{\tilde{X}_{\mathbf{w}}^2+\tilde{Y}_{\mathbf{w}}^2}}\bigg) &= \frac{(\beta-k+1)|P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}}|}{\sqrt{2}}\\ &+O\bigg(\mathbf{E}_\rho\bigg(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))\bigg)^2+\mathbf{E}_\rho\bigg(\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))\bigg)^2\bigg).\end{aligned}$$

As $E_\rho(\frac{X_{\mathbf{w}}}{M} - (\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))^2 = \frac{1}{M}\mathrm{Var}_\rho(\frac{X_{\mathbf{w}}}{\sqrt{M}}) = O(M^{-1})$, we obtain that the asymptotic mean of $\frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{M\sqrt{\tilde{X}^2_{\mathbf{w}}+\tilde{Y}^2_{\mathbf{w}}}}$ equals $(\beta-k+1)|P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}}|/\sqrt{2}$. Moreover, we have

$$
\begin{aligned}
&\sqrt{M}\Big(\frac{D_2^S}{M} - (\beta-k+1)\sum\frac{|P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}}|}{\sqrt{2}}\Big)\\
&= \sqrt{M}\sum\Big(\frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{M\sqrt{\tilde{X}^2_{\mathbf{w}}+\tilde{Y}^2_{\mathbf{w}}}} - (\beta-k+1)\frac{|P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}}|}{\sqrt{2}}\Big)\\
&= \sum\frac{\mathrm{sign}(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}})}{2\sqrt{2}}\sqrt{M}\Big(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))+\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))\Big)\\
&+\frac{1}{\sqrt{M}}\sum O\Big(M\big(\frac{X_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))\big)^2+M\big(\frac{Y_{\mathbf{w}}}{M}-(\beta-k+1)(P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}}))\big)^2\Big).
\end{aligned}
$$

Similar to the proof of Theorem 3, we see that $\sqrt{M}\big(\frac{D_2^S}{M} - (\beta-k+1)\sum\frac{|P_\rho(\mathbf{w})+P_\rho(\bar{\mathbf{w}})-p_{\mathbf{w}}-p_{\bar{\mathbf{w}}}|}{\sqrt{2}}\big)$ is asymptotically normal with mean 0 and variance $2(\Sigma_\rho^S)^2$.

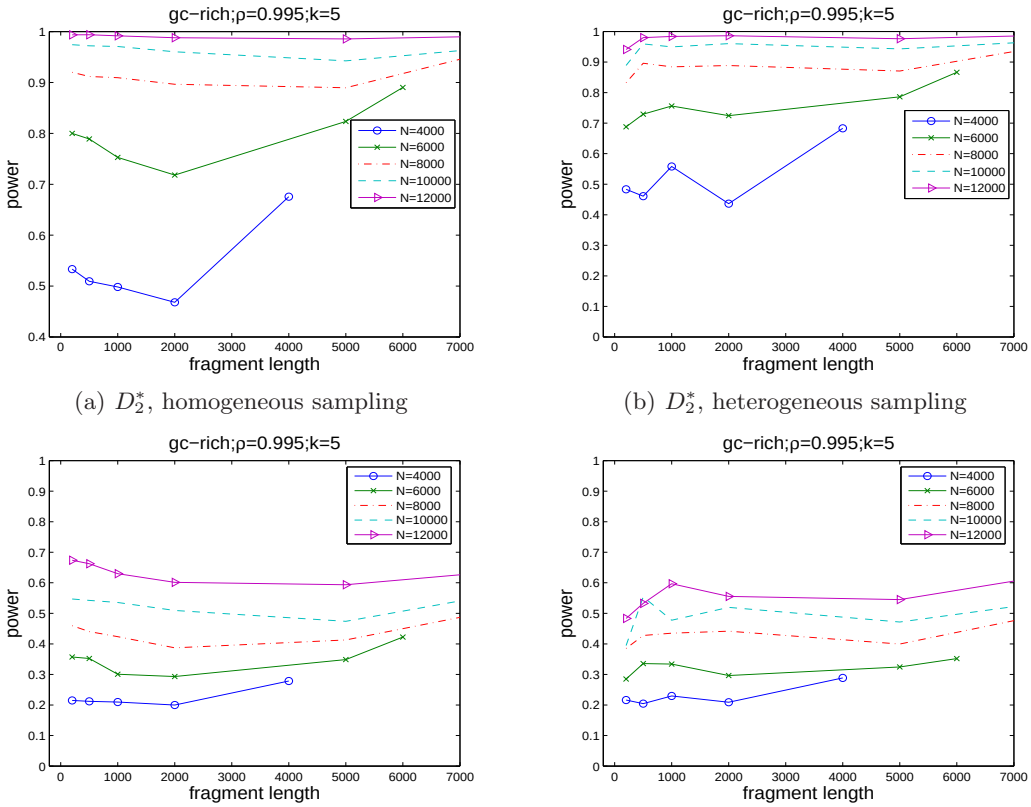
Proof of Theorem 5:

The proof of the three equation are roughly the same, and thus we only give the proof for the first equation. We consider one-sided test. For fixed type I error α, based on Theorem 2 (a), we find z_α such that $P\{Z_1 \geq z_\alpha\} = \alpha$. We reject the null hypothesis if $Z_1 \geq z_\alpha$ which is approximately equivalent to $D_2 > M^2(\beta-k+1)^2A(1) + z_\alpha\sqrt{M^3}$. Thus, the power for D_2 is approximately

$$
\begin{aligned}
1-\beta &= P_\rho\big(D_2 > M^2(\beta-k+1)^2A(1)+z_\alpha\sqrt{M^3}\big)\\
&= P_\rho\Big(\frac{\sqrt{M}\big(\frac{D_2}{M^2}-(\beta-k+1)^2A(\rho)\big)}{\sqrt{2}\Sigma_\rho} > \frac{\sqrt{M}\big(\frac{z_\alpha}{\sqrt{M}}-(\beta-k+1)^2(A(\rho)-A(1))\big)}{\sqrt{2}\Sigma_\rho}\Big)\\
&\approx 1-\Phi(C(\rho)),
\end{aligned}
$$

where $A(\rho) = \sum_{\mathbf{w}\in\mathcal{A}^k}(P_\rho(\mathbf{w})+P_\rho(\mathbf{w}))^2$. The last approximation holds because of Theorem 2 (b).

Supplementary Figures



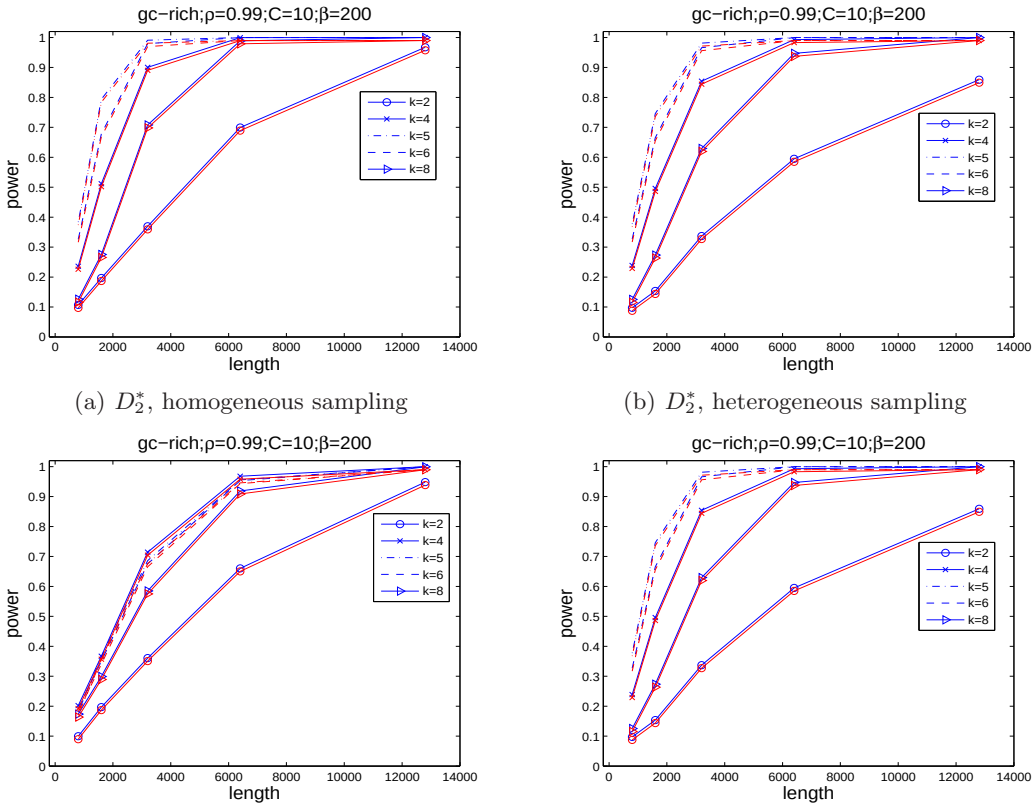
(a) D_2^*, homogeneous sampling

(b) D_2^*, heterogeneous sampling

(c) D_2^S, homogeneous sampling

(d) D_2^S, heterogeneous sampling

Fig. 1. (Supplementary Figure S1). The power of D_2^* (a,b), and D_2^S (c,d) under the homogeneous (left column) and the heterogeneous (right column) NGS read sampling as a function of sequence length and read length β. Here GC-rich distribution, $\rho = 0.99$, k = 5, and $\beta = 200$.



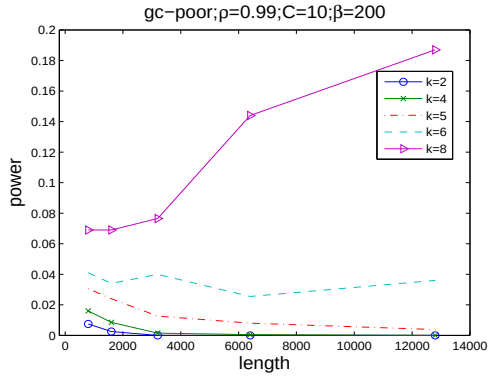
(a) D_2^*, homogeneous sampling

(b) D_2^*, heterogeneous sampling

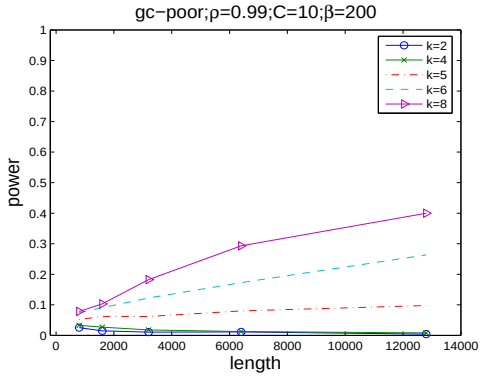
(c) D_2^S, homogeneous sampling

(d) D_2^S, heterogeneous sampling

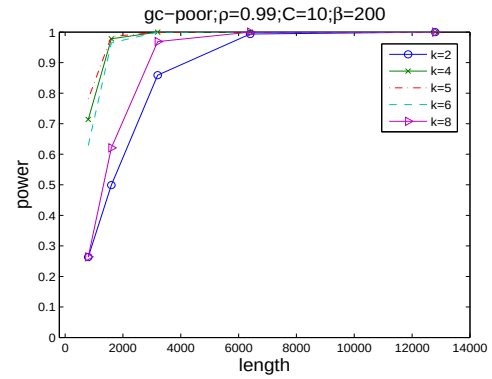
Fig. 2. (Supplementary Figure S2). The power of D_2^* (a,b), and D_2^S (c,d) under the homogeneous (left column) and the heterogeneous (right column) NGS read sampling as a function of sequence length and with/without sequencing errors. Here GC-rich distribution, $\rho = 0.99$, k = 5, and $\beta = 200$.



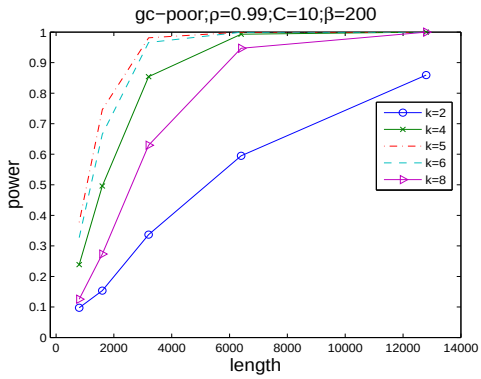
(a) D_2, homogeneous sampling



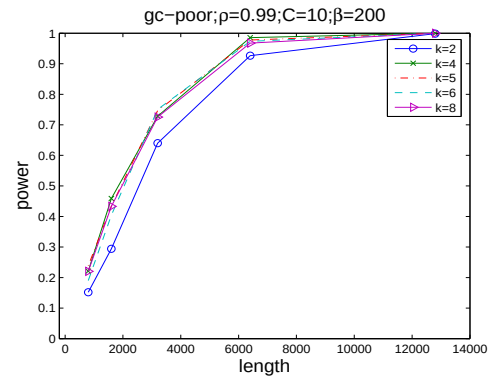
(b) D_2, heterogeneous sampling



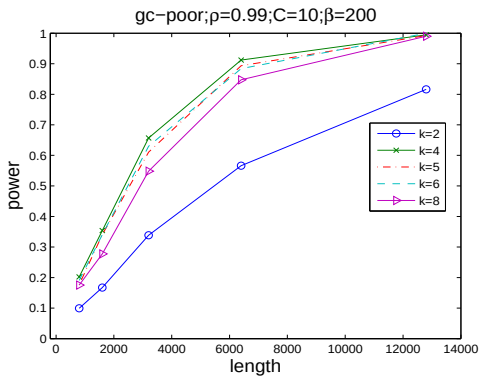
(c) D_2^*, homogeneous sampling



(d) D_2^*, heterogeneous sampling

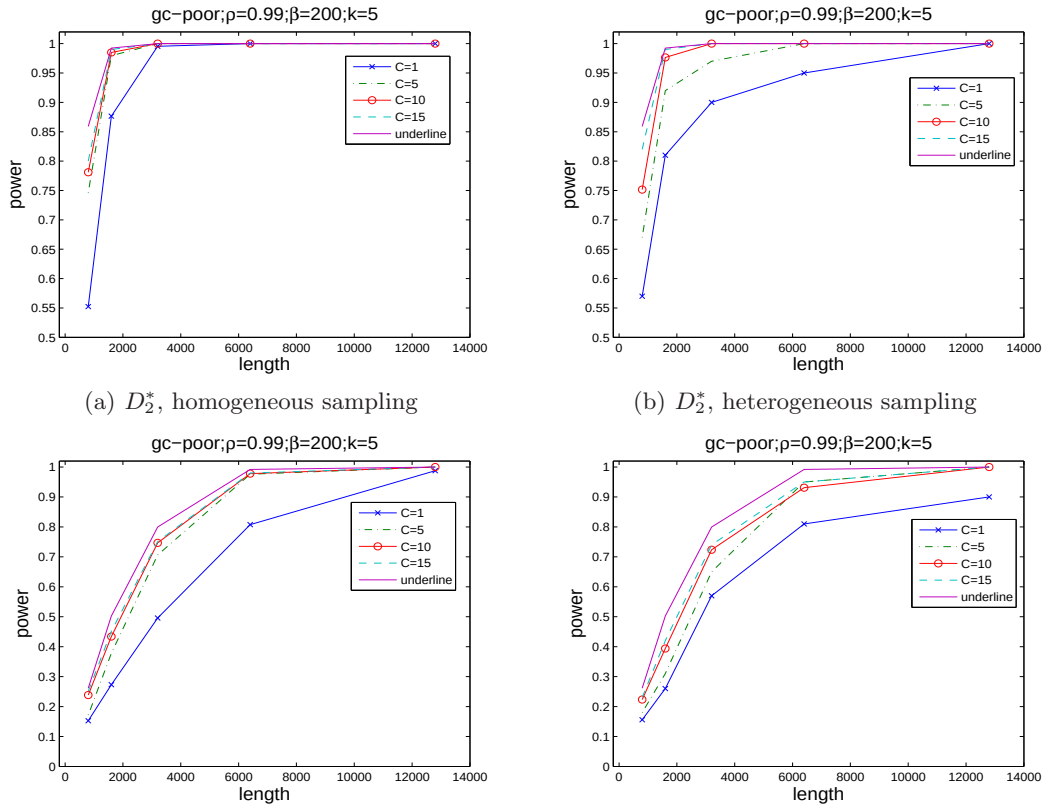


(e) D_2^S, homogeneous sampling



(f) D_2^S, heterogeneous sampling

Fig. 3. (Supplementary Figure S3). The power of D_2(a,b), D_2^*(c,d), and D_2^S(e,f) under the homogeneous (left column) and the heterogeneous (right column) NGS read sampling as a function of sequence length and word size k. Here GC-poor distribution: $\rho = 0.99$, coverage $= 10$ and $\beta = 200$.



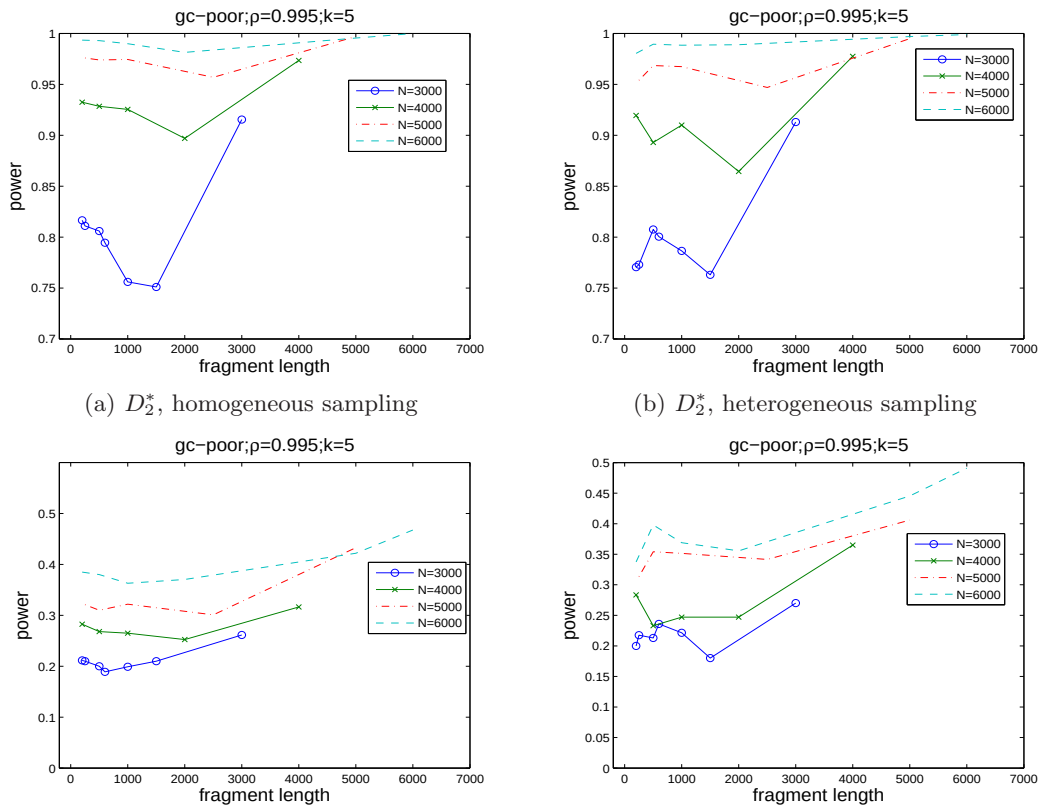
(a) D_2^*, homogeneous sampling

(b) D_2^*, heterogeneous sampling

(c) D_2^S, homogeneous sampling

(d) D_2^S, heterogeneous sampling

Fig. 4. (Supplementary Figure S4). The power of D_2^* (a,b), and D_2^S (c,d) under the homogeneous (left column) and the heterogeneous (right column) NGS read sampling as a function of sequence length and coverage. For comparison, the power of the statistics when the whole genome sequences are known is also shown (underline). Here GC-poor distribution, $\rho = 0.99$, k = 5, and $\beta = 200$.



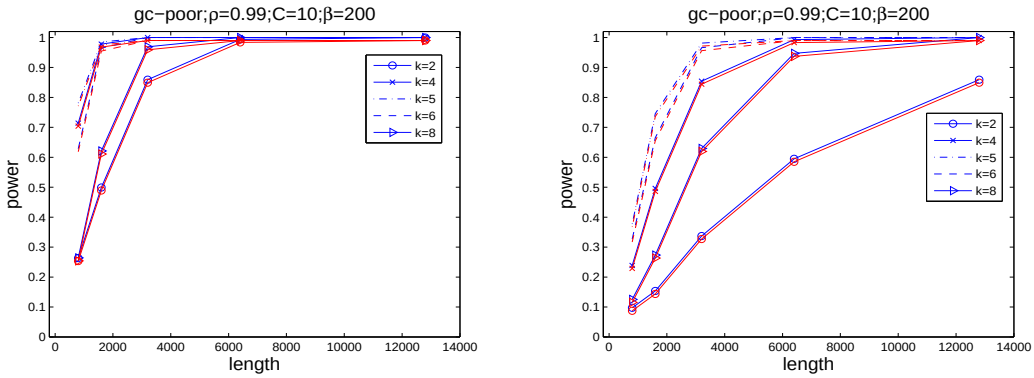
(a) D_2^*, homogeneous sampling

(b) D_2^*, heterogeneous sampling

(c) D_2^S, homogeneous sampling

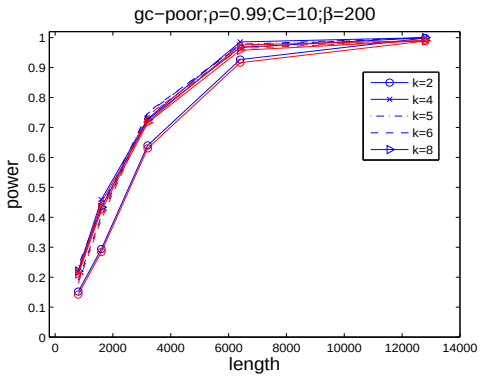
(d) D_2^S, heterogeneous sampling

Fig. 5. (Supplementary Figure S5). The power of D_2^* (a,b), and D_2^S (c,d) under the homogeneous (left column) and the heterogeneous (right column) NGS read sampling as a function of sequence length and read length β. Here GC-poor distribution, $\rho = 0.99$, k = 5, and $\beta = 200$.

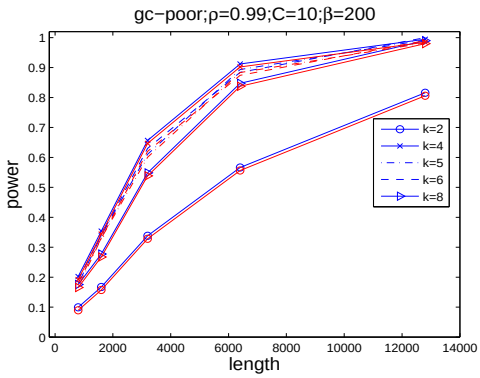


(a) D_2^*, homogeneous sampling

(b) D_2^*, heterogeneous sampling

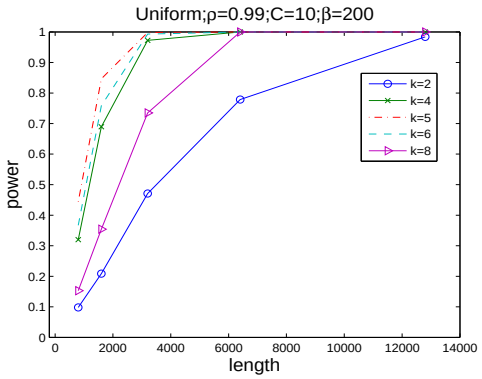


(c) D_2^S, homogeneous sampling

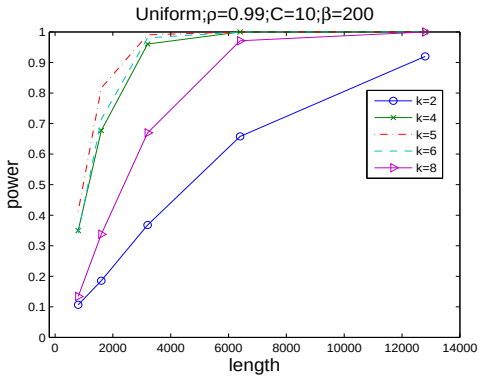


(d) D_2^S, heterogeneous sampling

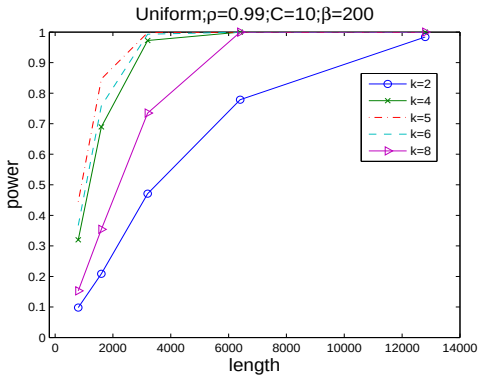
Fig. 6. (Supplementary Figure S6). The power of D_2^* (a,b), and D_2^S (c,d) under the homogeneous (left column) and the heterogeneous (right column) NGS read sampling as a function of sequence length and with/without sequencing errors. Here GC-poor distribution, $\rho = 0.99$, k = 5, and $\beta = 200$.



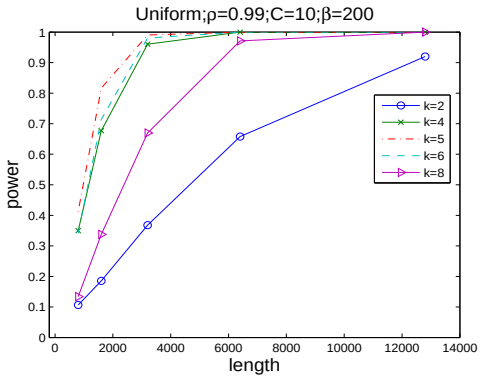
(a) D_2, homogeneous sampling



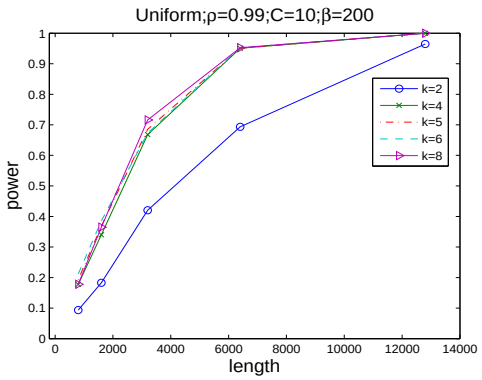
(b) D_2, heterogeneous sampling



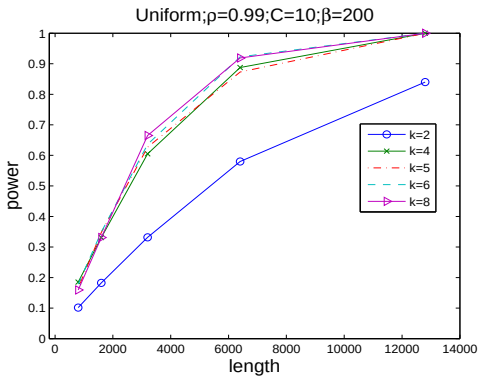
(c) D_2^*, homogeneous sampling



(d) D_2^*, heterogeneous sampling

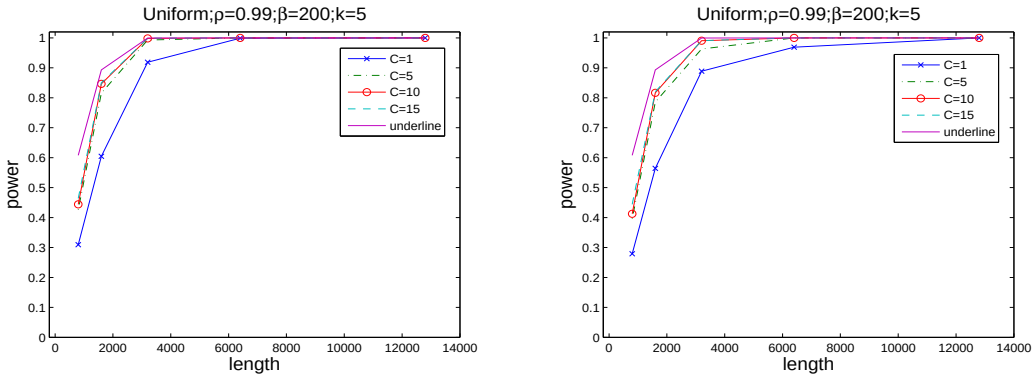


(e) D_2^S, homogeneous sampling



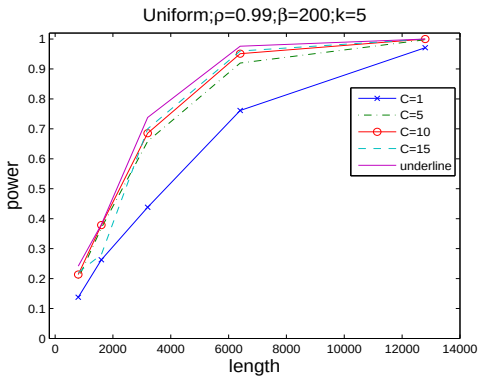
(f) D_2^S, heterogeneous sampling

Fig. 7. (Supplementary Figure S7). The power of D_2(a,b), D_2^*(c,d), and D_2^S(e,f) under the homogeneous (left column) and the heterogeneous (right column) NGS read sampling as a function of sequence length and word size k. Here uniform distribution: $\rho = 0.99$, coverage $= 10$ and $\beta = 200$.

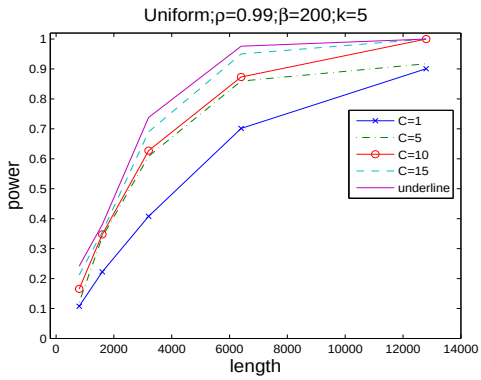


(a) D_2^*, homogeneous sampling

(b) D_2^*, heterogeneous sampling

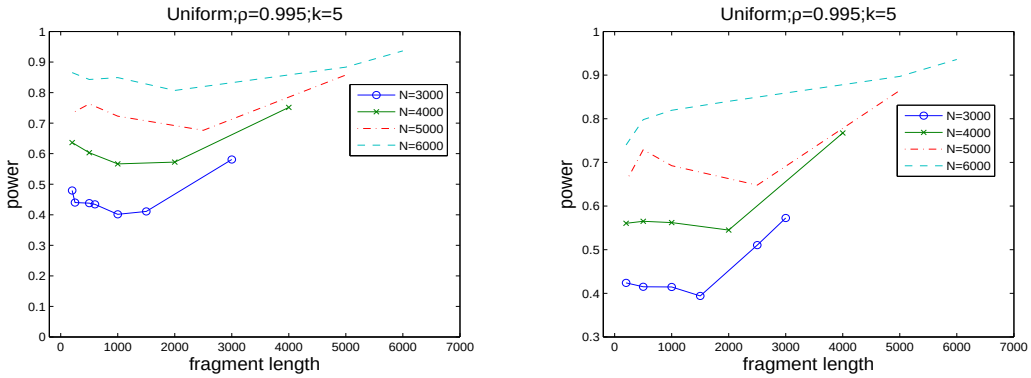


(c) D_2^S, homogeneous sampling



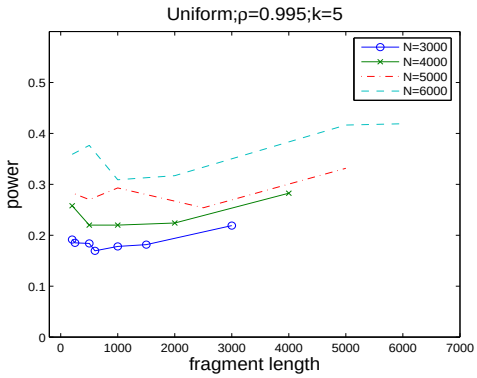
(d) D_2^S, heterogeneous sampling

Fig. 8. (Supplementary Figure S8). The power of D_2^* (a,b), and D_2^S (c,d) under the homogeneous (left column) and the heterogeneous (right column) NGS read sampling as a function of sequence length and coverage. For comparison, the power of the statistics when the whole genome sequences are known is also shown (underline). Here uniform distribution, $\rho = 0.99$, k = 5, and $\beta = 200$.

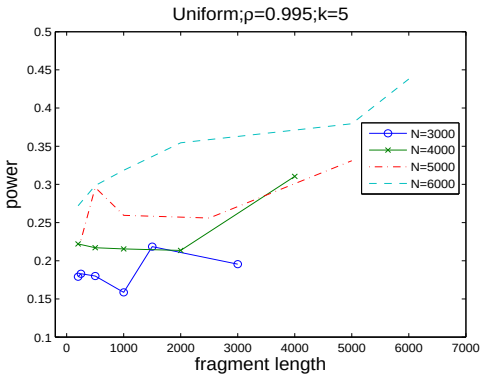


(a) D_2^*, homogeneous sampling

(b) D_2^*, heterogeneous sampling

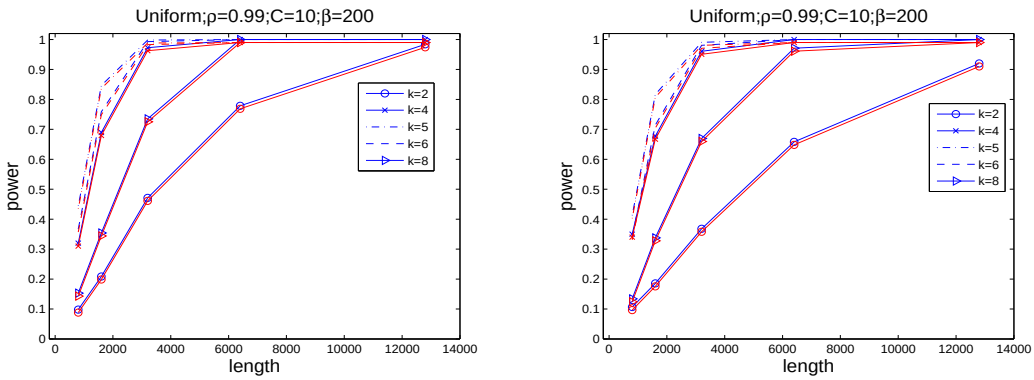


(c) D_2^S, homogeneous sampling



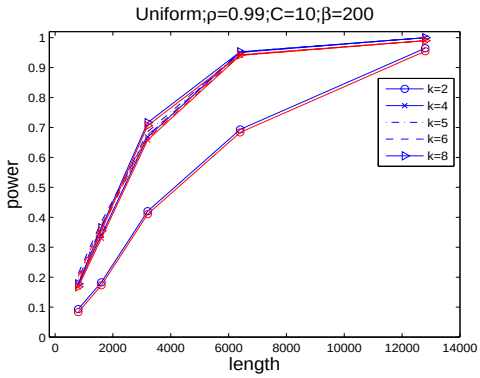
(d) D_2^S, heterogeneous sampling

Fig. 9. (Supplementary Figure S9). The power of D_2^* (a,b), and D_2^S (c,d) under the homogeneous (left column) and the heterogeneous (right column) NGS read sampling as a function of sequence length and read length β. Here uniform distribution, $\rho = 0.99$, k = 5, and $\beta = 200$.

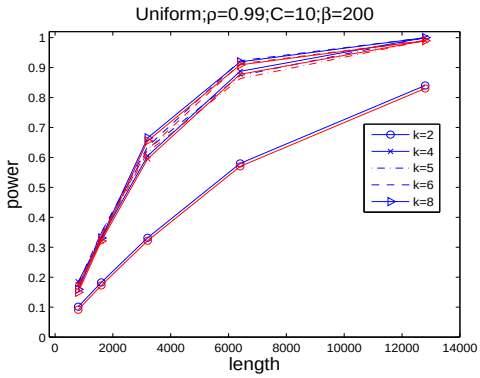


(a) D_2^*, homogeneous sampling

(b) D_2^*, heterogeneous sampling



(c) D_2^S, homogeneous sampling



(d) D_2^S, heterogeneous sampling

Fig. 10. (Supplementary Figure S10). The power of D_2^* (a,b), and D_2^S (c,d) under the homogeneous (left column) and the heterogeneous (right column) NGS read sampling as a function of sequence length and with/without sequencing errors. Here uniform distribution, $\rho = 0.99$, k = 5, and $\beta = 200$.

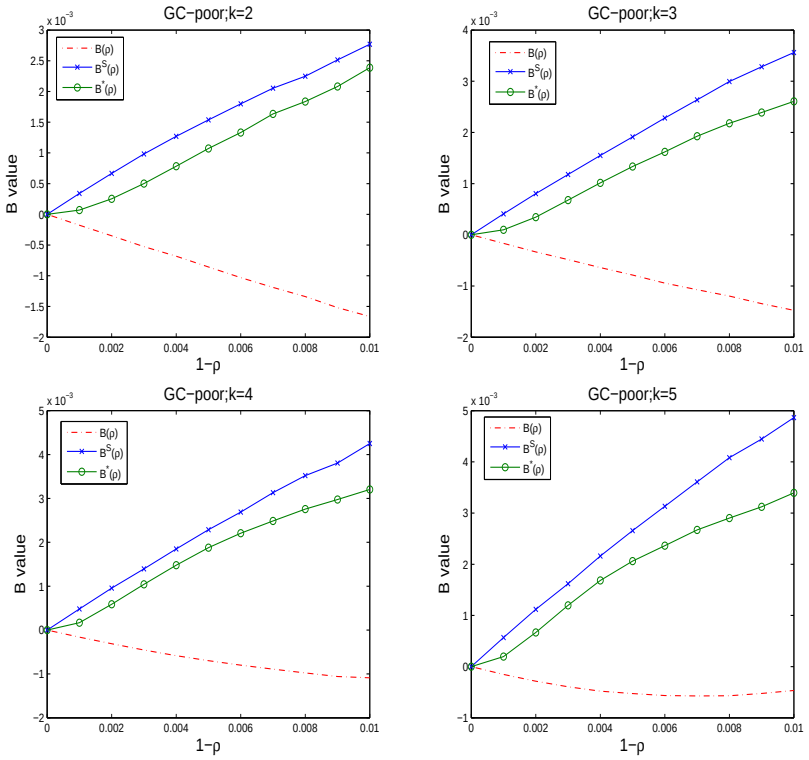


Fig. 11. (Supplementary Figure S11). The values of $B(\rho), B^*(\rho)$ and $B^S(\rho)$ as a function of $1-\rho$ from 0 to 0.01 for the GC-poor background distribution and $\beta = 200$ under homogeneous read sampling.

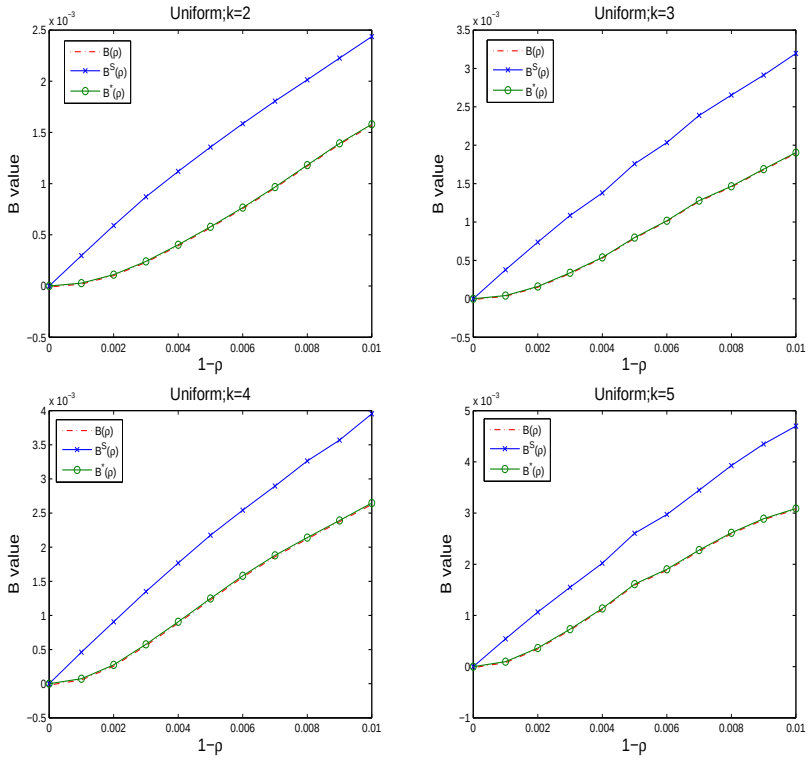
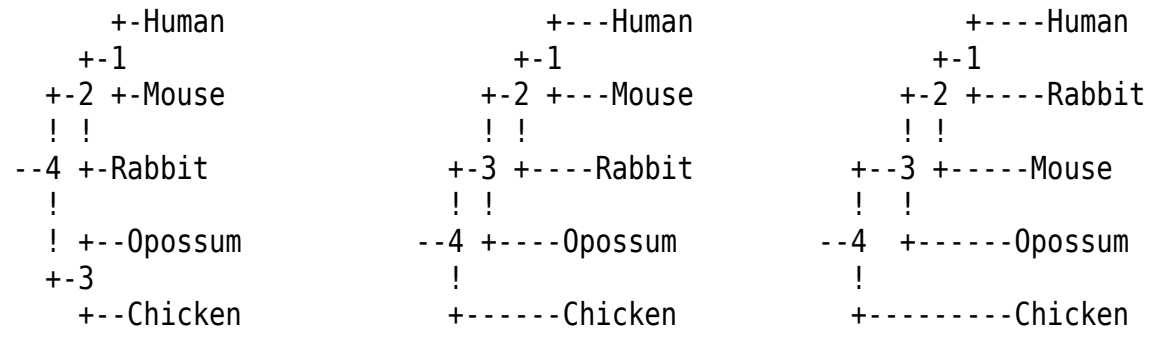
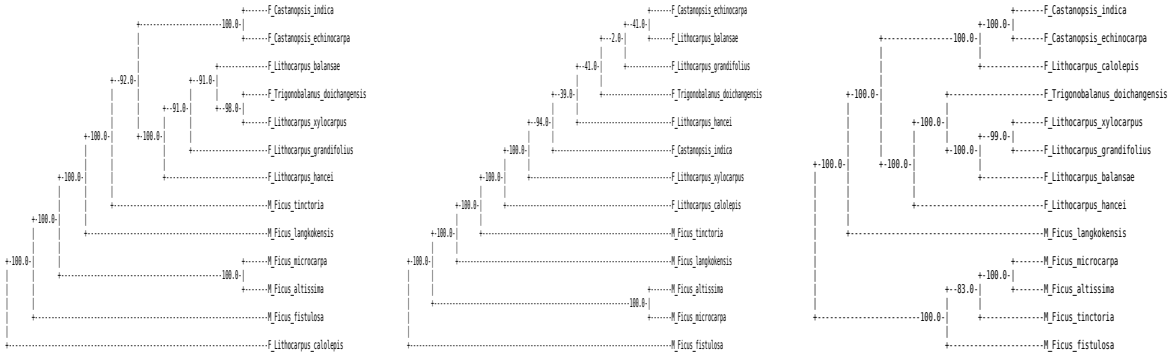


Fig. 12. (Supplementary Figure S12). The values of $B(\rho), B^*(\rho)$ and $B^S(\rho)$ as a function of $1-\rho$ from 0 to 0.01 for the uniform background distribution and $\beta = 200$ under homogeneous read sampling.



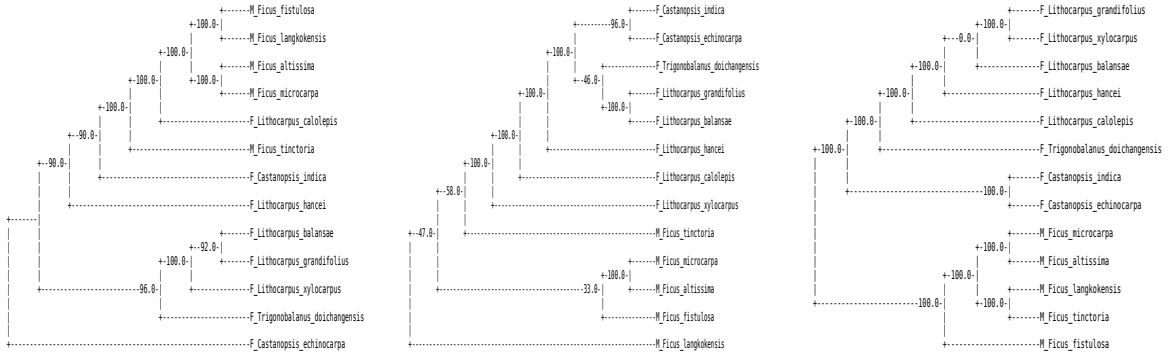
(a) d_2 (b) d_2^* (c) d_2^S

Fig. 13. (Supplementary Figure S13). The clusterings of the five mammalian species with dissimilarity measures, from left to right d_2(a), d_2^*(b), and d_2^S(c) with $k = 7$ using only the reads with $X_{\mathbf{w}}/EX_{\mathbf{w}} \leq 2$.



(a) d_2 (b) d_2^* (c) d_2^S

Fig. 14. (Supplementary Figure S14). The clusterings of the 13 tree species with dissimilarity measures, from left to right d_2(a), d_2^*(b), and d_2^S(c) with $k = 7$ using all the reads. The number on each internal branch is the fraction of times the branch occurs in 100 random sampling using $d = 5\%$ of the reads.



(a) d_2 (b) d_2^* (c) d_2^S

Fig. 15. (Supplementary Figure S15). The clusterings of the 13 tree species with dissimilarity measures, from left to right d_2(a), d_2^*(b), and d_2^S(c) with $k = 11$ using all the reads. The number on each internal branch is the fraction of times the branch occurs in 100 random sampling using $d = 5\%$ of the reads.

Supplementary Tables

	D_2		D_2^*		D_2^S	
$n \times 10^{-4}$	$\frac{\mathbf{E}ND_2\times 10^3}{\sqrt{M}(\beta-k+1)^2}$	$\sigma(ND_2)\times 10^{-5}$	$\frac{\mathbf{E}ND_2^*\times 10}{\sqrt{M}(\beta-k+1)}$	$\sigma(ND_2^*)\times 10^{-2}$	$\frac{\mathbf{E}ND_2^S\times 10^2}{\sqrt{M}(\beta-k+1)}$	$\sigma(ND_2^S)\times 10^{-2}$
0.32	6.50	8.65	3.17	3.25	5.86	6.02
0.64	6.50	8.92	3.15	2.72	6.23	6.63
1.28	6.50	8.62	3.16	2.54	6.89	7.14
2.56	6.50	8.54	3.15	2.54	7.67	7.32
10	6.50	8.17	3.15	2.34	9.51	7.67
20	6.50	8.32	3.15	2.22	9.61	7.92
Theory	6.50	8.45	3.15	2.09	13.66	8.28

Table 1. (Supplementary Table S1). Comparison of simulated means and variances of $ND_2, ND_2^*,$ and ND_2^S for different sequence length n with the corresponding theoretical limits, with GC-poor background and motif=AGCCA, $C = 1$, $\rho = 0.99$ and for the expectation, word length $k = 5$, and for the covariance, word length $k = 2$. The number of simulations is 10,000.

	D_2		D_2^*		D_2^S	
$n \times 10^{-4}$	Theory	Simulated	Theory	Simulated	Theory	Simulated
0.1	1.1	3.3	69.5	67.3	53.4	17.3
0.2	1.4	2.5	90.0	92.4	89.0	27.9
0.32	1.1	1.6	98.0	99.5	98.8	49.6
0.4	0.8	1.5	99.4	99.9	99.5	59.0
0.5	1.1	1.4	100	100	99.7	72.2
0.64	1.1	1.3	100	100	99.9	90.8
1.28	0.6	0.7	100	100	99.9	98.4

Table 2. (Supplementary Table S2). Comparison of the theoretical and the simulated power for D_2, D_2^* and D_2^S for different sequence length n with GC-poor background and motif=AGCCA, $C = 1$, $\rho = 0.99$ and word length $k = 5$. The type I error $\alpha = 0.05$. The number of simulations is 10,000.

	D_2		D_2^*		D_2^S	
$n \times 10^{-4}$	$\frac{\mathbf{E}ND_2\times 10^3}{\sqrt{M}(\beta-k+1)^2}$	$\sigma(ND_2)\times 10^{-3}$	$\frac{\mathbf{E}ND_2^*\times 10}{\sqrt{M}(\beta-k+1)}$	$\sigma(ND_2^*)\times 10^{-2}$	$\frac{\mathbf{E}ND_2^S\times 10^2}{\sqrt{M}(\beta-k+1)}$	$\sigma(ND_2^S)$
0.32	4.20	9.53	1.51	5.44	5.04	15.40
0.64	4.20	7.01	1.51	6.15	5.48	11.33
1.28	4.20	5.59	1.50	6.83	6.01	9.04
2.56	4.20	4.68	1.50	6.95	6.75	7.57
10	4.20	4.35	1.50	7.62	8.46	7.32
20	4.20	4.22	1.50	7.93	8.61	7.11
Theory	4.20	4.11	1.50	8.31	12.80	6.10

Table 3. (Supplementary Table S3). Comparison of simulated mean and variance of $ND_2, ND_2^*,$ and ND_2^S for different sequence length n with the corresponding theoretical limits, with uniform background and motif=AGCCA, $C = 1$, $\rho = 0.99$ and for the expectation, word length $k = 5$, and for the covariance, word length $k = 2$. The number of simulations is 10,000.

	D_2		D_2^*		D_2^S	
$n \times 10^{-4}$	Theory	Simulated	Theory	Simulated	Theory	Simulated
0.1	35.1	35.5	35.1	35.5	54.2	12.1
0.2	71.7	68.2	71.7	68.2	90.9	27.2
0.32	90.1	91.9	90.1	91.9	98.3	41.1
0.4	94.4	95.9	94.4	95.9	99.7	48.5
0.5	97.7	99.4	97.7	99.4	99.9	63.0
0.64	99.2	99.9	99.2	99.9	99.9	74.8
1.28	100	100	100	100	100	96.6

Table 4. (Supplementary Table S4). Comparison of the theoretical and the simulated power for D_2, D_2^* and D_2^S for different sequence length n with uniform background and motif=AGCCA, $C = 1$, $\rho = 0.99$ and word length $k = 5$. The type I error $\alpha = 0.05$. The number of simulations is 10,000.

Group	Species Name	SRA accession No.
Fagaceae:	1 Lithocarpus balansae	SRR035946
	2 Lithocarpus grandifolius	SRR037157
	3 Lithocarpus hancei	SRR037158
	4 Lithocarpus xylocarpus	SRR037437
	5 Lithocarpus calolepis	SRR037484
	6 Trigonobalanus doichangensis	SRR037802
	7 Castanopsis echinocarpa	SRR037535
	8 Castanopsis indica	SRR037537
Moraceae:	1 Ficus altissima	SRR037748
	2 Ficus langkokensis	SRR037751
	3 Ficus microcarpa	SRR037888
	4 Ficus tinctoria	SRR038268
	5 Ficus fistulosa	SRR035147

Table 5. (Supplementary Table S5). The names of the 13 tree species and their NCBI short read archive (SRA) accession numbers.