A Supplemental materials

A.1 Verification of Theorem 8 using simulations

Similar to Table 1, we repeated the experiment for the same settings except with two different scale factors. The results are shown in this section.

	L = 10 K			L = 100 K			L = 1 M		
p =	0.001	0.1	0.2	0.001	0.1	0.2	0.001	0.1	0.2
k = 21	95.4	95.3	94.7	95.0	95.2	95.06	95.0	95.0	94.6
k = 51	95.4	94.8	N/A	94.8	94.6	N/A	94.9	95.1	94.4
k = 100	94.7	N/A	N/A	94.6	N/A	N/A	95.4	93.7	N/A

Table S1: The percentage of experiments that resulted in the true mutation rate falling within the 95% confidence interval given in Theorem 8 when using various mutation rates across multiple k-mer sizes and L values. A scale factor of 0.2 was used. The results show an average over 10,000 simulations for each setting. N/A entries indicate that the parameters are not particularly meaningful and will not produce interpretable results, either because $E[N_{\text{mut}}] \approx L$ in these cases (almost all k-mers are mutated), or because the scale factor is too small to differentiate between the two FracMinHash sketches.

	L = 10 K			L = 100 K			L = 1 M		
p =	0.001	0.1	0.2	0.001	0.1	0.2	0.001	0.1	0.2
k = 21	96.3	95.0	96.0	95.1	95.0	95.3	95.0	95.2	94.9
k = 51	94.9	94.5	N/A	94.7	95.3	N/A	94.7	95.0	N/A
k = 100	95.2	N/A	N/A	95.2	N/A	N/A	94.5	N/A	N/A

Table S2: The percentage of experiments that resulted in the true mutation rate falling within the 95% confidence interval given in Theorem 8 when using various mutation rates across multiple k-mer sizes and L values. A scale factor of 0.05 was used. The results show an average over 10,000 simulations for each setting. N/A entries indicate that the parameters are not particularly meaningful and will not produce interpretable results, either because $E[N_{\text{mut}}] \approx L$ in these cases (almost all k-mers are mutated), or because the scale factor is too small to differentiate between the two FracMinHash sketches.

A.2 Expected number of non-mutated k-mers in different scenarios

To explain the N/A entries in Figures 1, S1 and S2, we show the expected number of non-mutated k-mers after undergoing the simple mutation process in Table S3. Depending on the scale factor, only a fraction of these non-mutated k-mers show up in the FracMinHash sketch. Therefore, if the number of non-mutated k-mers is too small, it would be meaningless to run the experiment.

A.3 Theorems and proofs

Theorem 1. *For $0 < s < 1$, if A and B are two non-empty sets such that $A \setminus B$ and $A \cap B$ are non-empty, the following holds:*

$$\mathrm{E}\left[\hat{C}_{frac}(A,B)\mathbb{1}_{|\mathbf{FRAC}_s(A)|>0}\right] = \frac{|A \cap B|}{|A|}\left(1-(1-s)^{|A|}\right).$$

	$L = 10$ K			$L = 100$ K			$L = 1$ M		
$p =$	0.001	0.1	0.2	0.001	0.1	0.2	0.001	0.1	0.2
$k = 21$	9792.1	1094.2	92.2	97920.9	10941.9	922.3	979208.7	109419.0	9223.3
$k = 51$	9502.5	46.4	0.11	95025.4	463.8	1.1	950254.4	4638.4	11.4
$k = 100$	9047.9	0.26	2.04E-6	90479.2	2.7	2.04E-5	904792.1	26.6	2.04E-4

Table S3: The expected number of non-mutated k-mers after undergoing the simple mutation process, shown across multiple k-mer sizes, L values, and mutation rates.

Proof. Using the notation introduced previously, observe that

$$\hat{C}_{\text{frac}}(A,B)\mathbb{1}_{|\mathbf{FRAC}_s(A)|>0} = \frac{X_{A\cap B}}{X_{A\cap B}+X_{A\setminus B}}\mathbb{1}_{X_{A\cap B}+X_{A\setminus B}>0},$$

and that the random variables $X_{A\cap B}$ and $X_{A\setminus B}$ are independent (which follows directly from the fact that $A\cap B$ and $A\setminus B$ are non-empty, distinct sets). We will use the following fact from standard calculus:

$$\int_0^1 xt^{x+y-1}\,dt = \frac{x}{x+y}\mathbb{1}_{x,y>0}. \tag{7}$$

Then using the moment generating function of the binomial distribution, we have

$$\mathrm{E}\left[t^{X_{A\cap B}}\right] = (1-s+st)^{|A\cap B|} \tag{8}$$

$$\mathrm{E}\left[t^{X_{A\setminus B}}\right] = (1-s+st)^{|A\setminus B|} \tag{9}$$

We also know by continuity that

$$\mathrm{E}\left[X_{A\cap B}\,t^{X_{A\cap B}-1}\right] = \frac{d}{dt}(1-s+st)^{|A\cap B|} \tag{10}$$

$$= |A\cap B|s(1-s+st)^{|A\cap B|-1}. \tag{11}$$

Using these observations, we can then finally calculate that

$$\mathrm{E}\left[\frac{X_{A\cap B}}{X_{A\cap B}+X_{A\setminus B}}\mathbb{1}_{X_{A\cap B}+X_{A\setminus B}>0},\right] = \mathrm{E}\left[\int_0^1 X_{A\cap B}\,t^{X_{A\cap B}+X_{A\setminus B}-1}\,dt\right] \tag{12}$$

$$= \int_0^1 \mathrm{E}\left[X_{A\cap B}\,t^{X_{A\cap B}+X_{A\setminus B}-1}\,dt\right] \tag{13}$$

$$= \int_0^1 \mathrm{E}\left[X_{A\cap B}\,t^{X_{A\cap B}-1}\right]\mathrm{E}\left[t^{X_{A\setminus B}}\right]\,dt \tag{14}$$

$$= |A\cap B|s\int_0^1 (1-s+st)^{|A\cap B|+|A\setminus B|-1}\,dt \tag{15}$$

$$= \left.\frac{|A\cap B|s(1-s+st)^{|A|}}{|A|s}\right|_{t=0}^{t=1} \tag{16}$$

$$= \frac{|A\cap B|}{|A|}\left(1-(1-s)^{|A|}\right), \tag{17}$$

where Fubini's theorem is used in Equation (13) and independence in Equation (14). ■

Theorem 3. *For $n = |A \cap B|$ and $m = |A \setminus B|$ where both m and n are non-zero, a first order Taylor series approximation gives*

$$\operatorname{Var}\left[\hat{C}_{frac}(A,B)\right] \approx \frac{mn(1-s)}{s(m+n)^3}.$$

Proof. Let $g(x,y) = \frac{x}{x+y}$, $\mu_x = ns$, $\mu_y = ms$ and use subscripts to denote partial derivatives:

$$\begin{aligned} g_x(x,y) &= \frac{y}{(x+y)^2} \\ g_y(x,y) &= \frac{-x}{(x+y)^2} \end{aligned}$$

We then have the first order Taylor series:

$$\begin{aligned} \operatorname{Var}\left(g\left(X_{A\cap B}, X_{A\setminus B}\right)\right) &= g_x^2(\mu_x,\mu_y)\operatorname{Var}(X_{A\cap B}) \\ &\quad + 2g_x(\mu_x,\mu_y)g_y(\mu_x,\mu_y)\mathrm{E}[X_{A\cap B}-\mu_x]\mathrm{E}[X_{A\setminus B}-\mu_y] \\ &\quad + g_y^2(\mu_x,\mu_y)\operatorname{Var}(X_{A\setminus B}) \qquad (18) \\ &= \frac{m^2}{s^2(m+n)^4}ns(1-s) + \frac{n^2}{s^2(m+n)^4}ms(1-s) \\ &= \frac{mn(1-s)}{(m+n)^3 s}, \end{aligned}$$

with the middle term of eq. (18) factoring due to independence.

■

Theorem 4. *For $g(x,y) = \frac{x}{x+y}$, $n = |A \cap B|$ and $m = |A \setminus B|$ where both m and n are non-zero,*

$$\sqrt{n+m}\left(g(X_{A\cap B}, X_{A\setminus B}) - g(n,m)\right) \xrightarrow[n,m\to\infty]{\mathscr{D}} \mathscr{N}\left(0, \frac{mn(1-s)}{(m+n)^3 s}\right).$$

Proof. The covariance matrix is calculated as

$$\Sigma = \begin{bmatrix} ns(1-s) & 0 \\ 0 & ms(1-s) \end{bmatrix}.$$

Using the same notation as in Theorem 3, let

$$\phi = \begin{bmatrix} g_x(\mu_x,\mu_y) \\ g_y(\mu_x,\mu_y) \end{bmatrix} = \begin{bmatrix} \frac{m}{s(n+m)^2} \\ \frac{-n}{s(n+m)^2} \end{bmatrix}.$$

The delta method then uses the first order Taylor series from Theorem 3 to obtain that $\sqrt{n+m}\left(g(X_{A\cap B}, X_{A\setminus B}) - g(n,m)\right)$ converges in distribution to a centered normal with variance

$$\phi' \Sigma \phi = \frac{mn(1-s)}{(m+n)^3 s}.$$

■

Theorem 5. *For $0 < s < 1$, if A and B are respectively distinct sets of k-mers of a sequence S and a sequence S' derived from S under the simple mutation model with mutation probability p such that $A \cap B$ is non-empty, then the expectation of $C_{frac}(A, B)$ in the product space $\mathcal{P}, \mathcal{S}$ is given by*

$$E_{\mathcal{P},\mathcal{S}}[C_{frac}(A,B)] = (1-p)^k, \tag{5}$$

where $\mathcal{P} = (\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ and $\mathcal{S} = (\Omega_2, \mathcal{F}_2, \mathbf{P}_2)$ are the probability spaces corresponding to the mutation and FracMinHash sketching random processes, respectively.

Proof.

$$\begin{aligned}
E_{\mathcal{P},\mathcal{S}}[C_{\text{frac}}(A,B)] &= \int_{\mathcal{P},\mathcal{S}} C_{\text{frac}}(A,B)\ d_{\mu_1 \times \mu_2} = \int_P \int_S C_{\text{frac}}(A,B)\ d\mu_2\ d\mu_1 \\
&= \mathrm{E}_P\left[\mathrm{E}_S\left[C_{\text{frac}}(A,B)\right]\right] = \mathrm{E}_P\left[1 - \frac{N_{\text{mut}}}{L}\right] \\
&= 1 - \frac{Lq}{L} = 1 - (1-(1-p)^k) \\
&= (1-p)^k.
\end{aligned}$$

Here, we used Fubini's theorem in the second step. We also used the expectation of N_{mut} from (Blanca et al 2022), where $q = 1 - (1-p)^k$. ■

Theorem 6. *For $0 < s < 1$, if A and B are respectively distinct sets of k-mers of a sequence S and a sequence S' derived from S under the simple mutation model with mutation probability p such that $A \cap B$ is non-empty, then the variance of $C_{frac}(A, B)$ in the product space $\mathcal{P}, \mathcal{S}$ is given by*

$$\operatorname*{Var}_{\mathcal{P},\mathcal{S}}[C_{frac}(A,B)] = \frac{(1-s)}{sL^3\left(1-(1-s)^L\right)^2}(L\mathrm{E}_{\mathcal{P}}[N] - \mathrm{E}_{\mathcal{P}}[N^2]) + \frac{1}{L^2}\operatorname*{Var}_{\mathcal{P}}(N_{mut}) \tag{6}$$

where $\mathcal{P} = (\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ and $\mathcal{S} = (\Omega_2, \mathcal{F}_2, \mathbf{P}_2)$ are the probability spaces corresponding to the mutation and FracMinHash sketching random processes, respectively.

Proof. First, we calculate the second moment of $C_{\text{frac}}(A, B)$ in the product space as follows:

$$\begin{aligned}
\mathrm{E}_{\mathcal{P},\mathcal{S}}[C_{\text{frac}}(A,B)^2] &= \int_{\mathcal{P},\mathcal{S}} C_{\text{frac}}(A,B)^2\ d_{\mu_1 \times \mu_2} = \int_{\mathcal{P}} \int_S C_{\text{frac}}(A,B)^2\ d\mu_2\ d\mu_1 \\
&= \int_{\mathcal{P}} \left[\frac{mn(1-s)}{s(m+n)^3\left(1-(1-s)^L\right)^2} + \left(\frac{L - N_{mut}}{L}\right)^2\right] d\mu_1 \\
&= \mathrm{E}_{\mathcal{P}}\left[\frac{N(L-N)(1-s)}{sL^3\left(1-(1-s)^L\right)^2} + \frac{1}{L^2}(L^2 - 2LN + N^2)\right] \\
&= \frac{(1-s)}{sL^3\left(1-(1-s)^L\right)^2}(L\mathrm{E}_{\mathcal{P}}[N] - \mathrm{E}_{\mathcal{P}}[N^2]) \\
&\quad + \frac{1}{L^2}(L^2 - 2L\mathrm{E}_{\mathcal{P}}[N] + \mathrm{E}_{\mathcal{P}}[N^2])
\end{aligned}$$

Therefore, we calculate the variance in the product space as follows.

$$\begin{aligned}
\underset{\mathcal{P},\mathcal{S}}{\mathrm{Var}}\,(C_{\text{frac}}(A,B)) &= \mathrm{E}_{\mathcal{P},\mathcal{S}}[C_{\text{frac}}(A,B)^2] - E_{\mathcal{P},\mathcal{S}}[C_{\text{frac}}(A,B)]^2 \\
&= \frac{(1-s)}{sL^3\left(1-(1-s)^L\right)^2}(L\mathrm{E}_{\mathcal{P}}[N] - \mathrm{E}_{\mathcal{P}}[N^2]) \\
&\quad + \frac{1}{L^2}(L^2 - 2L\mathrm{E}_{\mathcal{P}}[N] + \mathrm{E}_{\mathcal{P}}[N^2]) \\
&\quad - \frac{1}{L^2}(L - \mathrm{E}_{\mathcal{P}}[N])^2 \\
&= \frac{(1-s)}{sL^3\left(1-(1-s)^L\right)^2}(L\mathrm{E}_{\mathcal{P}}[N] - \mathrm{E}_{\mathcal{P}}[N^2]) \\
&\quad + \frac{1}{L^2}(L^2 - 2L\mathrm{E}_{\mathcal{P}}[N] + \mathrm{E}_{\mathcal{P}}[N^2]) \\
&\quad - \frac{1}{L^2}(L^2 - 2L\mathrm{E}_{\mathcal{P}}[N] + \mathrm{E}_{\mathcal{P}}[N]^2) \\
&= \frac{(1-s)}{sL^3\left(1-(1-s)^L\right)^2}(L\mathrm{E}_{\mathcal{P}}[N] - \mathrm{E}_{\mathcal{P}}[N^2]) + \frac{1}{L^2}\underset{\mathcal{P}}{\mathrm{Var}}(N_{mut})
\end{aligned}$$

■

Theorem 7. *Let $0 < s < 1$, let A and B be two distinct sets of k-mers, respectively of a sequence S and a sequence S' derived from S under the simple mutation model with mutation probability p, such that $A \cap B$ is non-empty.*

Also, let $0 < \alpha < 1$, and C_{low} and C_{high} be defined as follows.

$$C_{low} = (1-p)^k - z_\alpha \sqrt{\frac{(1-s)}{sL^3\left(1-(1-s)^2\right)}(L\mathrm{E}_P[N_{mut}] - \mathrm{E}_P[{N_{mut}}^2]) + \frac{1}{L^2}\underset{P}{\mathrm{Var}}(N_{mut})}$$

$$C_{high} = (1-p)^k + z_\alpha \sqrt{\frac{(1-s)}{sL^3\left(1-(1-s)^2\right)}(L\mathrm{E}_P[N_{mut}] - \mathrm{E}_P[{N_{mut}}^2]) + \frac{1}{L^2}\underset{P}{\mathrm{Var}}(N_{mut})}.$$

Then, the following holds as $L \to \infty$ and when p and k are independent of L:

$$\Pr[C_{low} \le C_{frac}(A,B) \le C_{high}] = 1 - \alpha.$$

Proof. As discussed in the Methods section, $C_{\text{frac}}(A,B)$ is asymptotically normal when the required conditions are met. Therefore, the hypothesis test for a random variable following the Gaussian distribution holds for $C_{\text{frac}}(A,B)$. Using the expectation and the variance proved in Theorems 5 and 6, we have the results stated in the theorem.

■

Theorem 8. *Let A and B be two distinct sets of k-mers, respectively of a sequence S and a sequence S' derived from S under the simple mutation model with mutation rate p, such that $A \cap B$ is non-empty. Let $\mathrm{E}_{p_{fixed}}[X]$ and $\mathrm{Var}_{p_{fixed}}[X]$ denote the expectation and variance of a given random variable X under the randomness from the mutation process with fixed mutation rate p_{fixed}, respectively. Then, for fixed α, s, k and an observed $C_{frac}(A,B)$, there exists an L large enough such that there exist unique solutions $p = p_{low}$ and $p = p_{high}$ to the following equations, respectively,*

$$C_{frac}(A,B) = (1-p_{low})^k + z_\alpha \sqrt{\frac{(1-s)}{sL^3\left(1-(1-s)^L\right)^2}(L\mathrm{E}_{p_{low}}[N_{mut}] - \mathrm{E}_{p_{low}}[{N_{mut}}^2]) + \frac{1}{L^2}\underset{p_{low}}{\mathrm{Var}}(N_{mut})},$$

$$C_{frac}(A,B) = (1-p_{high})^k - z_\alpha \sqrt{\frac{(1-s)}{sL^3\left(1-(1-s)^L\right)^2}\left(L\mathrm{E}_{p_{high}}[N_{mut}] - \mathrm{E}_{p_{high}}[{N_{mut}}^2]\right) + \frac{1}{L^2}\underset{p_{high}}{\mathrm{Var}}(N_{mut})},$$

such that the following holds:

$$\lim_{L\to\infty} \Pr[p_{low} \le p \le p_{high}] = 1-\alpha.$$

Proof. Given the results in Theorem 7, we only need to prove that p_{low} and p_{high} are well defined. It suffices to show that

$$(1-p_{\mathrm{low}})^k + z_\alpha \sqrt{\frac{(1-s)}{sL^3\left(1-(1-s)^L\right)^2}\left(L\mathrm{E}_{p_{\mathrm{low}}}[N_{\mathrm{mut}}] - \mathrm{E}_{p_{\mathrm{low}}}[{N_{\mathrm{mut}}}^2]\right) + \frac{1}{L^2}\underset{p_{\mathrm{low}}}{\mathrm{Var}}(N_{\mathrm{mut}})}$$

and

$$(1-p_{\mathrm{high}})^k - z_\alpha \sqrt{\frac{(1-s)}{sL^3\left(1-(1-s)^2\right)}\left(L\mathrm{E}_{p_{\mathrm{high}}}[N_{\mathrm{mut}}] - \mathrm{E}_{p_{\mathrm{high}}}[{N_{\mathrm{mut}}}^2]\right) + \frac{1}{L^2}\underset{p_{\mathrm{high}}}{\mathrm{Var}}(N_{\mathrm{mut}})}$$

are strictly monotonic in p_{low} and p_{high}, respectively under the Stated conditions.

Let us first investigate the function of p_{low}. For simplicity, we will write p instead of p_{low}, z instead of z_α and N instead of N_{mut}. We observe the following:

$$\begin{aligned}
&\frac{\partial}{\partial p}\left[(1-p)^k + z_\alpha \sqrt{\frac{(1-s)}{sL^3\left(1-(1-s)^L\right)^2}\left(L\mathrm{E}_p[N] - \mathrm{E}_p[N^2]\right) + \frac{1}{L^2}\underset{p}{\mathrm{Var}}(N)}\right]\\
&= -k(1-p)^{-1+k} - ((\frac{1}{L^2}(- kL(- 2k + (1-(1-p)^k)(- 1 + 2k + \frac{2}{p}))(1-p)^{-1+k} +\\
&L(k(- 1 + 2k + \frac{2}{p})(1-p)^{-1+k} - \frac{2(1-(1-p)^k)}{p^2})(1-p)^k - 2(-1+k)k^2(1-p)^{-1+2k} -\\
&\frac{4(1-p)^k(- 1 + (1-p)^k + (1 + (-1+k)(1-p)^k)p)}{p^3} -\\
&\frac{2k(1-p)^{-1+k}(- 1 + (1-p)^k + (1 + (-1+k)(1-p)^k)p)}{p^2} +\\
&\frac{2(1-p)^k(1 - k(1-p)^{-1+k} + (-1+k)(1-p)^k - (-1+k)k(1-p)^{-1+k}p)}{p^2}) +\\
&\frac{1}{L^3\left(1-(1-s)^L\right)^2 s}(kL^2(1-p)^{-1+k} + kL(- 2k + (1-(1-p)^k)(- 1 + 2k + \frac{2}{p}))(1-p)^{-1+k} -\\
&2\mathrm{k}L^2(1-(1-p)^k)(1-p)^{-1+k} - L(k(- 1 + 2k + \frac{2}{p})(1-p)^{-1+k} - \frac{2(1-(1-p)^k)}{p^2})(1-p)^k +\\
&2(-1+k)k^2(1-p)^{-1+2k} + \frac{4(1-p)^k(- 1 + (1-p)^k + (1 + (-1+k)(1-p)^k)p)}{p^3} +\\
&\frac{2k(1-p)^{-1+k}(- 1 + (1-p)^k + (1 + (-1+k)(1-p)^k)p)}{p^2} -\\
&\frac{2(1-p)^k(1 - k(1-p)^{-1+k} + (-1+k)(1-p)^k - (-1+k)k(1-p)^{-1+k}p)}{p^2})(1-s))z)/2\sqrt{f},
\end{aligned}$$

where

$$\begin{aligned}
f = {} & \frac{L\big(-2k+\big(1-(1-p)^k\big)\big(-1+2k+\frac{2}{p}\big)\big)(1-p)^k+(-1+k)k(1-p)^{2k}}{L^2}+ \\
& \frac{2(1-p)^k\big(-1+(1-p)^k+\big(1+(-1+k)(1-p)^k\big)p\big)}{L^2p^2}+\frac{1}{L^3\big(1-(1-s)^L\big)^2 s}\big(L^2\big(1-(1-p)^k\big) \\
& -L^2\big(1-(1-p)^k\big)^2-L\big(-2k+\big(1-(1-p)^k\big)\big(-1+2k+\frac{2}{p}\big)\big)(1-p)^k- \\
& (-1+k)k(1-p)^{2k}-\frac{2(1-p)^k\big(-1+(1-p)^k+\big(1+(-1+k)(1-p)^k\big)p\big)}{p^2}\big)(1-s).
\end{aligned}$$

After a tedious, but straightforward (due to the polynomial and rational terms) series expansion of this derivative about $L=\infty$, we obtain that the derivative is

$$-k(1-p)^{k-1}+O(L^{-1/2})$$

Therefore, as L approaches ∞, the derivative is always negative, which gives us that the function $(1-p_{\text{low}})^k+z_\alpha\sqrt{\frac{(1-s)}{sL^3(1-(1-s)^2)}(L\mathrm{E}_{p_{\text{low}}}[N]-\mathrm{E}_{p_{\text{low}}}[N^2])+\frac{1}{L^2}\operatorname{Var}_{p_{\text{low}}}(N_{mut})}$ is monotonically decreasing in p_{low} in the asymptotic case.

The proof that $(1-p_{\text{high}})^k-z_\alpha\sqrt{\frac{(1-s)}{sL^3(1-(1-s)^2)}(L\mathrm{E}_{p_{\text{high}}}[N]-\mathrm{E}_{p_{\text{high}}}[N^2])+\frac{1}{L^2}\operatorname{Var}_{p_{\text{high}}}(N_{mut})}$ is monotonically decreasing in p_{high} proceeds in an entirely analogous manner.

■

A.4 Dynamic Programming algorithm to compute the PMF of N_{mut}

Here, we will continue to use the notations of the simple mutation model for this algorithm, namely the parameters L, k and p. Let a string $\mathscr{S}$ of length l undergo the simple mutation process. For ease of understanding, we will represent the mutations introduced to $\mathscr{S}$ using a binary string $\mathscr{B}$ of length l, where $\mathscr{B}[i]=1$ if position i in $\mathscr{S}$ was mutated, and 0 otherwise. Therefore, each 1 in this binary string comes from a point mutation, occurring with a probability of p, and each 0 with a probability of $1-p$. Note that there are $l-k+1$ k-mers in $\mathscr{S}$. If we could account for all such binary string $\mathscr{B}$'s that result in a total of x mutated k-mers, we can accumulate the probabilities associated with each of these strings and compute $\Pr[N_{\text{mut}}=x]$ by letting $l=L+k-1$ (which is the length of S and S'). We do this efficiently by defining the following indicator variable:

$$I[i]=\begin{cases}1 & \text{if k-span } K_i \text{ in } \mathscr{S} \text{ is a mutated } k\text{-mer}\\ 0 & \text{otherwise}\end{cases}$$

for $i=1$ up to $l-k+1$, and making use of the following subproblems:

$$\mathbf{P}(l,x,z)=\Pr\left[\left(\sum_{i=1}^{l-k+1}I[i]\right)=x,\forall j \text{ s.t. } l-z+1\le j\le l,\ \mathscr{B}[j]=0\right] \tag{19}$$

where $0\le z<k$, $l\ge k$, $0\le x\le l-k+1$. Put another way, $\mathbf{P}(l,x,z)$ is the probability of having x mutated k-mers in a string of length l with z trailing zeros. Here, $l\ge k$ is required to make sure there is at least one k-mer. Equation (19) covers the cases where a string can have at most $k-1$ trailing zeros. For the rest of the cases, we define the following subproblem:

$$\mathbf{P}(l,x,k) = \Pr\left[\left(\sum_{i=1}^{l-k+1} I[i]\right) = x, \forall j \text{ s.t. } l-\mu+1 \leq j \leq l,\ \mathscr{B}[j]=0, \mu \geq k\right] \tag{20}$$

where $l \geq k$, $0 \leq x \leq l-k+1$. Put another way, $\mathbf{P}(l,x,k)$ is the probability of having x mutated k-mers in a string of length l with k or more trailing zeros.

The base cases of these subproblems are when the string has a length of k, and there can only be one k-mer. This k-mer will be non-mutated when the corresponding binary string has k zeros, giving us a probability of $\mathbf{P}(l=k,0,k) = (1-p)^k$. On the other hand, if we have $z < k$ trailing zeros, all we need is a 1 preceding these zeros for the k-mer to be mutated, giving us $\mathbf{P}(l=k,1,z) = p(1-p)^z$ for $0 \leq z < k$. It is straightforward to verify that summing these probabilities indeed gives us 1.

We next turn to using the smaller subproblems to solve the larger ones. The core idea is that if we append a 1 at the end of a binary string, then the number of mutated k-mers will increase by one, and there are no trailing zeros in the resulting string. On the other hand, if we append a 0 at the end of the string, then the number of mutated k-mers will stay the same if the total number of trailing zeros is k or more. Appending a 0 at the end will increase the number of mutated k-mers by one if the total number of trailing zeros is less than k. In both of these latter scenarios, the number of trailing zeros will increase by one. These observations lead to the following recurrence relation:

$$\mathbf{P}(l,x,z) = \begin{cases} \Big(\sum_{z'=0}^{k-1} \mathbf{P}(l-1,x-1,z') + \mathbf{P}(l-1,x-1,k)\Big) \times p & \text{if } z=0 \\ \mathbf{P}(l-1,x-1,z-1) \times (1-p) & \text{if } 1 \leq z < k \\ \Big(\mathbf{P}(l-1,x,k-1) + \mathbf{P}(l-1,x,k)\Big) \times (1-p) & \text{if } z=k. \end{cases} \tag{21}$$

For our parameters L, k and p, we would need to solve the subproblems for $l = L+k-1$. Finally, we would compute $\Pr[N_{\text{mut}} = x]$, $x = 0$ up to L as follows:

$$\Pr[N_{\text{mut}} = x] = \mathbf{P}(L+k-1,x,k) + \sum_{z'=0}^{k-1} \mathbf{P}(L+k-1,x,z'). \tag{22}$$

These base cases and recurrence relations give us Algorithm 1 to compute the PMF of N_{mut}. The loop at Step 5 of the algorithm iterates L times. The inner loop at Step 6 iterates at most L times. It is straightforward to count that Steps 7 – 11 take $O(k)$ number of operations. These observations give us a running time of $O(L^2 k)$. Note that k is usually in the magnitude of 20 to 50. Considering $k << L$, we have an $O(L^2)$ algorithm to compute the PMF of N_{mut}.

A.5 Theoretical guarantees to accurately estimate containment index

In this section, we present theoretical evidence that $C_{\text{frac}}(A,B)$ is able to estimate the true containment index $C(A,B)$ with high accuracy. Let the elements in $A \cup B$ be e_i for $i = 1$ to N. We define an indicator variable Y_i associated with an element e_i as follows:

$$Y_i = \begin{cases} 1 & \text{if } \ e_i \in \mathbf{FRAC}_s(A) \cap \mathbf{FRAC}_s(B) \\ 0 & \text{otherwise} \end{cases}.$$

Let Y be the number of elements in $\mathbf{FRAC}_s(A) \cap \mathbf{FRAC}_s(B)$. Naturally, $Y = \sum_{i=1}^{N} Y_i$. The probability of Y_i being 1 is $\frac{|A \cap B| s}{|A \cup B|}$. Therefore, we have:

Algorithm 1 : PMF – Nmut

```
*Input:*
  $L$, total number of $k$-mers
  $k$, length of a $k$-mer
  $p$, mutation rate
*Initialization:*
  $\mathbf{P}(l, x, z) = 0$ for $l = k$ up to $L + k - 1$, $x = 0$ up to $L$, $z = 0$ up to $k$
*Steps:*
  $\mathbf{P}(k, 0, k) = (1 - p)^k$
  for $z = 0, \dots, k - 1$ do
      $p(k, 1, z) = \mathbf{P}(1 - p)^z$
  end for

  for $l = k + 1, \dots, L + k - 1$ do
      for $x = 0, \dots, l - k + 1$ do
          $\mathbf{P}(l, x, 0) = \Big( \sum_{z'=0}^{k-1} \mathbf{P}(l - 1, x - 1, z') + \mathbf{P}(l - 1, x - 1, k) \Big) \times p$
          for $z = 1, \dots, k - 1$ do
              $\mathbf{P}(l, x, z) = \mathbf{P}(l - 1, x - 1, z - 1) \times (1 - p)$
          end for
          $\mathbf{P}(l, x, k) = \Big( \mathbf{P}(l - 1, x, k - 1) + \mathbf{P}(l - 1, x, k) \Big) \times (1 - p)$
      end for
  end for

  for $x = 0, \dots, L$ do
      PMF$[x] = \mathbf{P}(L + k - 1, x, k) + \sum_{z'=0}^{k-1} \mathbf{P}(L + k - 1, x, z')$
  end for
*Output:*
  PMF, where PMF$[x]$ = $\Pr[N_{\text{mut}} = x]$
```

$$E[Y] = \sum_{i=1}^{N} \Pr[Y_i = 1] = |A \cap B| s.$$

Let us make a simplifying assumption that the exact cardinality of the set A is known. Let us define Y' as $Y' = \frac{Y}{|A|s}$. Therefore, $E[Y'] = |A \cap B| / |A| = C(A, B)$. If we use Y' as the estimator to measure $C(A, B)$, then we have

$$\Pr\Big[\Big|\frac{Y' - C(A,B)}{C(A,B)}\Big| \geq \delta\Big] = \Pr\Big[\Big|\frac{Y - |A \cap B|s}{|A \cap B|s}\Big| \geq \delta\Big] \leq 2e^{-\delta^2 |A \cap B| s/3},$$

where we used Chernoff bound for a sum of Bernoulli random variables in the last step. The results are trivial, stating that when the two sets have more in common, or when we work with a larger scale factor, the estimate Y' performs better. This is expected, and conforms to the concept of using a scale factor. $C_{\text{frac}}(A, B)$ estimates $C(A, B)$ slightly differently than Y', and further investigations are required to narrow down the theoretical guarantees of $C_{\text{frac}}(A, B)$ estimating $C(A, B)$.

A.6 Estimating number of distinct k-mers from FracMinHash

In this section, we detail a simple method to estimate the total number of distinct k-mers in a given set from its FracMinHash. This can be useful for applications such as the de-biasing in eq. (3) when the set under consideration is small. We have already observed that for $X_A := |\mathbf{FRAC}_s(A)|$ the size of the sketch, X_A is distributed as a binomial random variable: $X_A \sim \text{Binom}(|A|, s)$. Hence $E[X_A/s] = |A|$, so a point estimate of the number of distinct k-mers $|A|$ can be had by dividing the sketch size by the scale factor. As the underlying distribution is binomial, we can easily obtain a Chernoff bound for the probability of deviating from this expected value by some relative error δ:

$$P\left(\left|\frac{X_a/s - |A|}{|A|}\right| < \delta\right) > 1 - 2e^{-\delta^2|A|s/3}.$$

So, for example, if using a scale factor of $s = 1/1000$, if one wants to be at least 95% sure that the estimate X_A/s is off by less than $\delta = 5\%$, this would require that $|A| \geq 4.4 \times 10^6$.

A.7 Jaccard calculated using FracMinHash sketches has bias

The theoretical analyses of $C_{\text{frac}}(A, B)$ presented in this work reveal the bias in containment index when computed from two FracMinHash sketches. Similarly, a bias in the Jaccard index computed from two FracMinHash sketches can also be proved. Please note that a similar confidence interval can *not* be derived for the Jaccard index as we were able to for the containment index. This is primarily because we found that the Jaccard index cannot be proved to be asymptotically Normal. Nonetheless, the following analysis proves that Jaccard version has a bias associated with it as well. Let us define

$$\hat{J}_{\text{frac}} := \frac{|\mathbf{FRAC}_s(A) \cap \mathbf{FRAC}_s(B)|}{|\mathbf{FRAC}_s(A) \cup \mathbf{FRAC}_s(B)|} \tag{23}$$

and investigate how well $\hat{J}_{\text{frac}}$ approximates the Jaccard index

$$J := \frac{|A \cap B|}{|A \cup B|}. \tag{24}$$

Using the same notations introduced previously, we have the following theorem.

Theorem 9. *For $0 < s < 1$, if A and B are two non-empty sets such that $A \setminus B$ and $A \cap B$ are non-empty and $B \not\subset A$ as well as $A \not\subset B$, the following holds:*

$$\mathrm{E}\left[\hat{J}_{frac}\mathbb{1}_{|\mathbf{FRAC}_s(A)\cup\mathbf{FRAC}_s(B)|>0}\right] = \frac{|A \cap B|}{|A \cup B|}\left(1 - (1-s)^{|A\cup B|}\right).$$

Proof. We observe that

$$\hat{J}_{\text{frac}}\mathbb{1}_{|\mathbf{FRAC}_s(A)\cup\mathbf{FRAC}_s(B)|>0} = \frac{X_{A\cap B}}{X_{A\cap B} + X_{A\ominus B}}\mathbb{1}_{X_{A\cap B}+X_{A\ominus B}>0},$$

and that the random variables $X_{A\cap B}$ and $X_{A\ominus B}$ are independent assuming the conditions of the theorem. Here, $A \ominus B = (A \setminus B) \cup (B \setminus A)$. From standard calculus, we have:

$$\int_0^1 x t^{x+y-1}\,dt = \frac{x}{x+y}\mathbb{1}_{x+y>0}. \tag{25}$$

Then using the moment generating function of the binomial distribution, we have

$$\begin{aligned}
\mathrm{E}\left[t^{X_{A\cap B}}\right] &= (1-s+st)^{|A\cap B|} \qquad (26)\\
\mathrm{E}\left[t^{X_{A\ominus B}}\right] &= (1-s+st)^{|A\ominus B|}. \qquad (27)
\end{aligned}$$

We also know by continuity that

$$\begin{aligned}
\mathrm{E}\left[X_{A\cap B}\,t^{X_{A\cap B}-1}\right] &= \frac{d}{dt}(1-s+st)^{|A\cap B|} \qquad (28)\\
&= |A\cap B|s(1-s+st)^{|A\cap B|-1}. \qquad (29)
\end{aligned}$$

Using these observations, we can then finally calculate that

$$\begin{aligned}
\mathrm{E}\left[\frac{X_{A\cap B}}{X_{A\cap B}+X_{A\ominus B}}\mathbb{1}_{X_{A\cap B}+X_{A\ominus B}>0},\right] &= \mathrm{E}\left[\int_0^1 X_{A\cap B}\,t^{X_{A\cap B}+X_{A\ominus B}-1}\,dt\right] \qquad (30)\\
&= \int_0^1 \mathrm{E}\left[X_{A\cap B}\,t^{X_{A\cap B}+X_{A\ominus B}-1}\,dt\right] \qquad (31)\\
&= \int_0^1 \mathrm{E}\left[X_{A\cap B}\,t^{X_{A\cap B}-1}\right]\mathrm{E}\left[t^{X_{A\ominus B}}\right]\,dt \qquad (32)\\
&= |A\cap B|\int_0^1 (1-s+st)^{|A\cap B|+|A\ominus B|-1}\,dt \qquad (33)\\
&= \left.\frac{|A\cap B|(1-s+st)^{|A\cup B|}}{|A\cup B|}\right|_{t=0}^{t=1} \qquad (34)\\
&= \frac{|A\cap B|}{|A\cup B|}\left(1-(1-s)^{|A\cup B|}\right), \qquad (35)
\end{aligned}$$

where Fubini's theorem is used in eq. (31) and independence in eq. (32).

Theorem 9 gives us the following result:

$$J_{\text{frac}} = \frac{\hat{J}_{\text{frac}}}{1-(1-s)^{|A\cup B|}} \tag{36}$$

is an unbiased estimator of the Jaccard index J. Using the same terminologies introduces in the Methods section, the expectation of this unbiased estimator (considering only the sletching random process) is given by

$$\mathrm{E}[J_{\text{frac}}] = \frac{L-N_{\text{mut}}}{L+N_{\text{mut}}} \tag{37}$$

A.8 Point estimate of mutation rate p can be calculated using the Jaccard index

Much like the analysis shown in Theorem 5, we can also obtain a point estimate of the mutation rate p using the Jaccard index estimator J_{frac}.

Theorem 10. *For $0 < s < 1$, if A and B are respectively distinct sets of k-mers of a sequence S and a sequence S' derived from S under the simple mutation model with mutation probability p such that $A \cap B$ is non-empty, then the expectation of J_{frac} in the product space $\mathcal{P}, \mathcal{S}$ is approximately given by*

$$E_{\mathcal{P},\mathcal{S}}[J_{frac}] = \frac{(1-p)^k}{2-(1-p)^k}, \tag{38}$$

where $\mathcal{P} = (\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ and $\mathcal{S} = (\Omega_2, \mathcal{F}_2, \mathbf{P}_2)$ are the probability spaces corresponding to the mutation and FracMinHash sketching random processes, respectively.

Proof.

$$\begin{aligned}
E_{\mathcal{P},\mathcal{S}}[J_{\text{frac}}] &= \int_{\mathcal{P},\mathcal{S}} J_{\text{frac}}\ d_{\mu_1 \times \mu_2} = \int_P \int_S J_{\text{frac}}\ d\mu_2\ d\mu_1 \\
&= \mathrm{E}_P\left[\mathrm{E}_S\left[J_{\text{frac}}\right]\right] = \mathrm{E}_P\left[\frac{L - N_{\text{mut}}}{L + N_{\text{mut}}}\right] \\
&\approx \frac{L - \mathrm{E}_P[N_{\text{mut}}]}{L + \mathrm{E}_P[N_{\text{mut}}]} \\
&= \frac{L - Lq}{L + Lq} \\
&= \frac{(1-p)^k}{2-(1-p)^k}
\end{aligned}$$

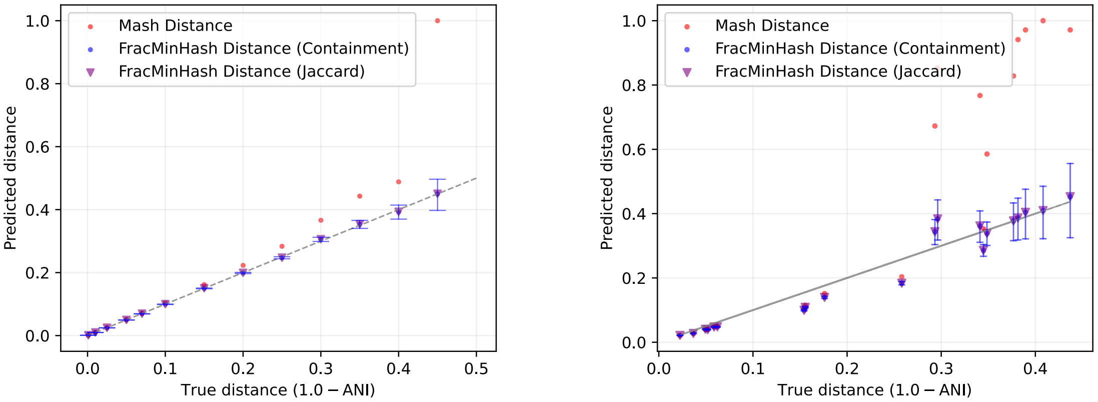
Here, we used Fubini's theorem in the second step. Using First-Order Taylor approximation, we approximated $\mathrm{E}[X/Y]$, the expectation of ratio of two random variables, using $\mathrm{E}[X]/\mathrm{E}[Y]$, the ratio of expectations of the same random variables. We also used the expectation of N_{mut} from (Blanca et al 2022), where $q = 1 - (1-p)^k$. ■

Theorem 10 allows us to obtain the following point estimate of mutation rate p using an observed Jaccard index J_{frac} obtained through FracMinHash sketches.

$$p = 1 - \left(\frac{2J_{\text{frac}}}{1 + J_{\text{frac}}}\right)^{1/k} \tag{39}$$

Unfortunately, the random variable $\frac{L-N_{\text{mut}}}{L+N_{\text{mut}}}$ cannot be expressed as Normally distributed – which is the core reason why we could not obtain a confidence interval similar to Theorem 8 using the Jaccard index.

Nevertheless, the point estimate shown above can still be useful. To demonstrate the usefulness of the point estimate obtained in Equation (39), we ran the same set of experiments presented in Figure 3a and Figure 3b. The results are shown in Figure S1 – which show that the point estimate obtained using the Jaccard index is pretty close to the point estimate obtained using the containment index, both in the cases of simulated and real data.



(a) Estimates of evolutionary distances between original and mutated *Staphylococcus* genome

(b) Estimates of evolutionary distances between pairs of real bacterial genomes

Fig. S1: Mash distances and FracMinHash estimates of evolutionary distance (given in terms of one minus the average nucleotide identity: ANI) when (a) introducing point mutations to a *Staphylococcus* genome at a known rate, and (b) between pairs of real bacterial genomes.