NUI projection recapture model

Walfred MA
Kwok Lab, UCSF
wangfei.ma@ucsf.edu
kwoklab.ucsf.edu

To estimate the diversities among populations and to predict future sequencing effectiveness, we project our current data on bigger sample sizes base on a computational resampling model we inferred. We first expand our dataset by down-sampling to obtain sufficient points to generate distribution curves, as explained in the following subsection. Then, based on the resampling model and mathematical reasoning on the current dataset, we propose an recapturing model to simulate sequencing process, which is also explained in another subsection. Projection is performed by fitting the last 10 points of our expanded dataset on the model. The package can be found at url: https://github.com/WalfredMA/NUI_Projection.

Down-sampling:

We down sample the current SVs based on the mean of all possible subsamples at certain lower sample size.

$Let\ S\ denotes\ the\ set\ of\ all\ sequenced\ samples.(|S| = s)$

$X_i\ denotes\ the\ SV\ set\ the\ sample\#i\ has\ and\ \overline{X_i}\ denotes\ the\ SV\ set\ the sample\#i\ does\ not\ have.$

$V_T^f\ denotes\ the\ number\ of\ SVs\ at\ frequnecy = f\ in\ sample\ set\ T.$

$V_t^f\ denotes\ mean\ number\ of\ SVs\ at\ \ frequency = f\ when\ sample\ size = t.\ Thus,\ Mean(V_{T|T\subset S}^f) = V_t^f.$

$For\ any$ not intersect $sample\ set\ pair\ \ M, N(|M| = m, |N| = n,\ \ M \cup N = T, m + n = |T| = t).$

$By\ the\ defination, V_T^f = \sum_{N\subset T} \left| \bigcap_{i\ \in N} X_i \cap \bigcap_{j\ \in M} \overline{X_j} \right|$

Let k be additional sample $\notin M \cup N$,

Let M' denotes $M \cup \{k\}$, N' denotes $N \cup \{k\}$, T' denotes $T \cup \{k\}$

$$V_T^f = \sum_{N \subset T} \left| \bigcap_{i \in N} X_i \cap \bigcap_{j \in M} \overline{X_j} \right| = \sum_{N \subset T} \left(\left| \bigcap_{i \in N} X_i \cap \bigcap_{j \in M} \overline{X_j} \cap X_k \right| + \left| \bigcap_{i \in N} X_i \cap \bigcap_{j \in M} \overline{X_j} \cap \overline{X_k} \right| \right)$$

So, $$\sum_{k \in \overline{T}} V_T^f = \sum_{k \in \overline{T}} \sum_{N \subset T} \left(\left| \bigcap_{i \in N} X_i \cap \bigcap_{j \in M} \overline{X_j} \cap X_k \right| + \left| \bigcap_{i \in N} X_i \cap \bigcap_{j \in M} \overline{X_j} \cap \overline{X_k} \right| \right)$$

Thus for all subsets whose module $= t$ in set S.

$$\sum_{T \subset S} \sum_{k \in \overline{T}} V_T^f = \sum_{k \in \overline{T}} \sum_{T \subset S} \sum_{N \subset T} \left(\left| \bigcap_{i \in N} X_i \cap \bigcap_{j \in M} \overline{X_j} \cap X_k \right| + \left| \bigcap_{i \in N} X_i \cap \bigcap_{j \in M} \overline{X_j} \cap \overline{X_k} \right| \right)$$

$$= \sum_{T' \subset S} \sum_{N' \subset T'} \sum_{k \in N'} \left(\left| \bigcap_{i \in N'} X_i \cap \bigcap_{j \in M} \overline{X_j} \right| \right) + \sum_{T' \subset S} \sum_{M' \subset T'} \sum_{k \in M'} \left(\left| \bigcap_{i \in N} X_i \cap \bigcap_{j \in M'} \overline{X_j} \right| \right)$$

$$= |N'| \sum_{T' \subset S} \sum_{N' \subset T'} \left(\left| \bigcap_{i \in N'} X_i \cap \bigcap_{j \in M} \overline{X_j} \right| \right) + |M'| \sum_{T' \subset S} \sum_{M' \subset T'} \left(\left| \bigcap_{i \in N} X_i \cap \bigcap_{j \in M'} \overline{X_j} \right| \right)$$

Considering $\left| \{T \mid T \subset S\} \right| = \begin{pmatrix} |S| \\ |T| \end{pmatrix} = \begin{pmatrix} s \\ t \end{pmatrix}$, $\sum_{T \subset S} V_T^f = \begin{pmatrix} s \\ t \end{pmatrix} * Mean(V_T^f) = \begin{pmatrix} s \\ t \end{pmatrix} V_t^f$

$$= (n+1) \begin{pmatrix} s \\ t+1 \end{pmatrix} Mean \left(\sum_{T' \subset S} \sum_{N' \subset T'} \left(\left| \bigcap_{i \in N'} X_i \cap \bigcap_{j \in M} \overline{X_j} \right| \right) \right) + (m+1) \begin{pmatrix} s \\ t+1 \end{pmatrix} Mean \left(\sum_{T' \subset S} \sum_{M' \subset T'} \left(\left| \bigcap_{i \in N} X_i \cap \bigcap_{j \in M'} \overline{X_j} \right| \right) \right)$$

$$= (n+1) \begin{pmatrix} s \\ t+1 \end{pmatrix} V_{t+1}^{n+1} + (t-n+1) \begin{pmatrix} s \\ t+1 \end{pmatrix} V_{t+1}^n$$

Similarly, $\sum_{T \subset S} \sum_{k \in \overline{T}} V_T^f = \left| \overline{T} \right| \begin{pmatrix} s \\ t \end{pmatrix} * Mean\left(V_T^f \right) = (s-t) \begin{pmatrix} s \\ t \end{pmatrix} V_t^n$

Hence, $(s-t) \begin{pmatrix} s \\ t \end{pmatrix} V_t^n = (n+1) \begin{pmatrix} s \\ t+1 \end{pmatrix} V_{t+1}^{n+1} + (t-n+1) \begin{pmatrix} s \\ t+1 \end{pmatrix} V_{t+1}^n$

$$\frac{s!}{t!(s-t-1)!} V_t^n = (n+1) \frac{s!}{(t+1)!(s-t-1)!} V_{t+1}^{n+1} + (t-n+1) \frac{s!}{(t+1)!(s-t-1)!} V_{t+1}^n$$

$$V_t^n = \frac{n+1}{t+1} V_{t+1}^{n+1} + \frac{t-n+1}{t+1} V_{t+1}^n$$

Let $A_T^N (N \subseteq T)$ *represents the set of SVs which are exclusive in set* N*, but not in any other sample.*

A_t^n *represents the mean of set* $\{\left| A_T^N \right| \mid N \subseteq T \text{ and } \left|N\right| = n\}$.

Appreally, $A_T^{N1} \cap A_T^{N2} = \varnothing$ *and* $V_T^n = \left|\bigcup \{\left| A_T^N \right| \mid N \subseteq T \text{ and } \left|N\right| = n\}\right|$. *We have*

$V_T^n = sum(\{\left| A_T^N \right| \mid N \subseteq T \text{ and } \left|N\right| = n)$.

$$V_T^n = A_t^n * \left|\{\left| A_T^N \right| \mid N \subseteq T \text{ and } \left|N\right| = n\}\right| = \binom{t}{n} A_t^n$$

Replacing V_t^n *with* A_t^n,

$$\frac{t!}{n!(t-n)!} A_t^n = \frac{n+1}{t+1}\frac{(t+1)!}{(n+1)!(t-n)!} A_{t+1}^{n+1} + \frac{t-n+1}{t+1}\frac{(t+1)!}{n!(t+1-n)!} A_{t+1}^n = \frac{t!}{n!(t-n)!} A_{t+1}^{n+1} + \frac{(t)!}{n!(t-n)!} A_{t+1}^n$$

Hence, $A_t^n = A_{t+1}^{n+1} + A_{t+1}^n$ ①

A_t^n *follows binomial distribution. Thus,*

$$A_{s-x}^n = \sum_{0 \le i \le x} \binom{x}{i} A_s^{n+i}$$

Hence, $V_{s-x}^n = \sum_{0 \le i \le x} \left(\frac{\binom{x}{i}\binom{s-x}{n}}{\binom{s}{n+i}} \right) V_s^{n+i}$ ②

Up-sampling:

let T_t *denotes total number of non−singleton SVs in all samples when sample size* $= t$.

S_t *denotes mean number of non−singleton SVs in each sample when sample size* $= t$.

Obvious,

$$T_t = \sum_{2 \le n \le s} V_s^n$$

$$S_t = \frac{1}{s} \sum_{2 \le n \le s} n V_s^n$$

Considering the equation ②,

$$V_{s-1}^n = \left(\frac{n+1}{s}\right) V_s^{n+1} + \left(\frac{s-n}{s}\right) V_s^n, \text{ thus,}$$

$$\sum_{2\le n\le s-1} V_{s-1}^{n} = \frac{1}{s}\sum_{3\le n\le s} n*V_s^n - \frac{1}{s}\sum_{2\le n\le s-1} n*V_s^n + \sum_{2\le n\le s-1} V_s^n$$

$$\sum_{2\le n\le s-1} V_{s-1}^{n} = \frac{1}{s}\sum_{3\le n\le s} n*V_s^n - \frac{1}{s}\sum_{2\le n\le s} n*V_s^n + \sum_{2\le n\le s} V_s^n$$

$$\sum_{2\le n\le s-1} V_s^n - \sum_{2\le n\le s-1} V_{s-1}^{n} = \frac{2}{s}*V_s^2 = (s-1)A_s^2$$

$$T_s - T_{s-1} = (s-1)A_s^2 = (s-1)V_s^2 / \binom{s}{2}$$

Similarly, considering the equation ①,

$$nV_{s-1}^{n} = \left(\frac{n}{s}\right)(n+1)V_s^{n+1} + \left(1-\frac{n}{s}\right)nV_s^n = nV_s^n - \left(\frac{n}{s}\right)nV_s^n + \left(\frac{n}{s}\right)(n+1)V_s^{n+1}$$

$$nV_{s-1}^{n} = \left(\frac{n+1}{s}\right)(n+1)V_s^{n+1} - \left(\frac{n}{s}\right)nV_s^n + nV_s^n - \frac{1}{s}(n+1)V_s^{n+1}$$

$$\sum_{2\le n\le s-1} n*V_{s-1}^{n} = sV_s^s - \frac{2}{s}2V_s^2 + \sum_{2\le n\le s-1} n*V_s^n - \frac{1}{s}\sum_{3\le n\le s} n*V_s^n$$

$$\sum_{2\le n\le s-1} n*V_{s-1}^{n} = -\frac{2}{s}V_s^2 + \sum_{2\le n\le s} n*V_s^n - \frac{1}{s}\sum_{2\le n\le s} n*V_s^n$$

$$\sum_{2\le n\le s-1} n*V_{s-1}^{n} = -\frac{2}{s}V_s^2 + \frac{s-1}{s}\sum_{2\le n\le s} n*V_s^n$$

$$\frac{1}{s-1}\sum_{2\le n\le s-1} n*V_{s-1}^{n} = -\frac{2}{s(s-1)}V_s^2 + \frac{1}{s}\sum_{2\le n\le s} n*V_s^n$$

$$\frac{1}{s}\sum_{2\le n\le s} n*V_s^n - \frac{1}{s-1}\sum_{2\le n\le s-1} n*V_{s-1}^{n} = A_s^2$$

$$\textit{Hence, } S_s - S_{s-1} = A_s^2 = V_s^2 / \binom{s}{2} = \frac{T_s - T_{s-1}}{s-1}$$

Upsampling-projection:

Define the increment of T_s:

$$I_s = T_{s+1} - T_s$$

let p denotes the possibility of an SV in an arbitrary sample X to be find in another sample.
If the sampling is fully random, we believe it should be same in all samples.
Thus, the change it only be captured twice is $p(1-p)^s$ *when sample set is* $s+2$
We have $A_{s+2}^2 = \sum_{f \in X} p(1-p)^s$

Thus,

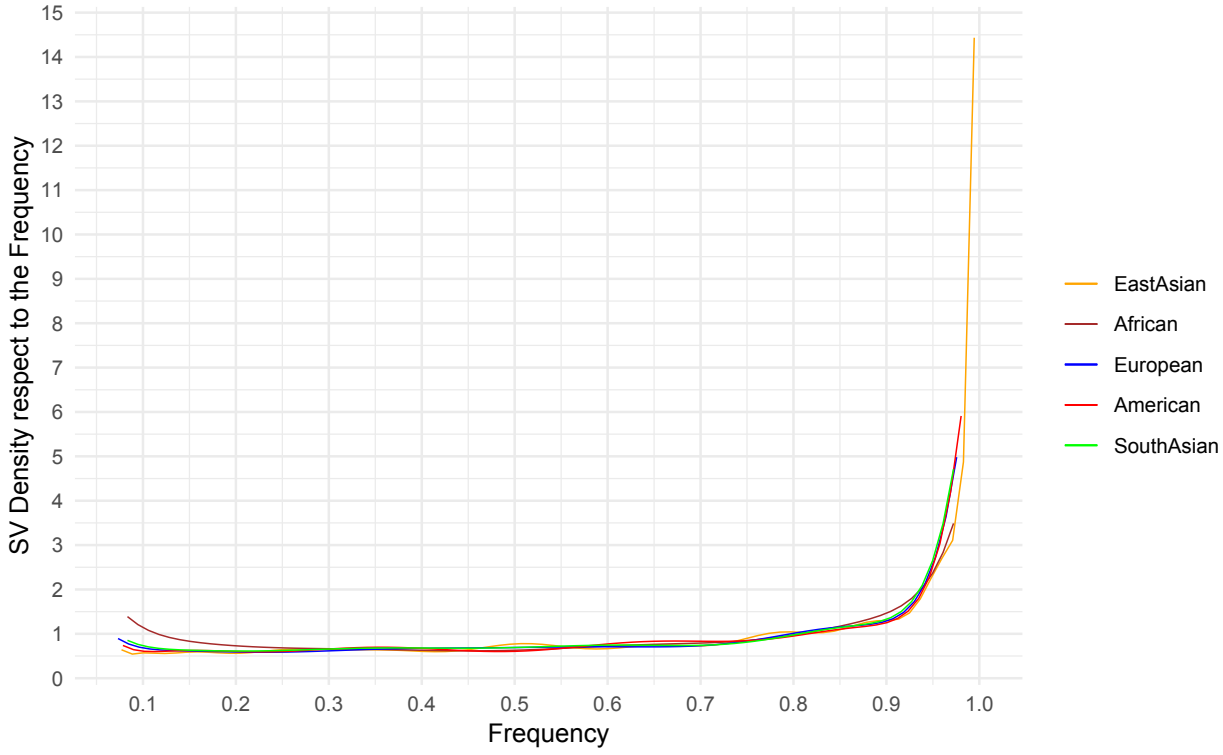
$$I_{s+2} = (s+1)\sum_{p \in X} p(1-p)^s$$

We let d_p *to represent the density of sv at frequency* $= p$,

Thus,

$$Exp(I_{s+2}) = Exp((s+1)\sum_{p \in X} p(1-p)^s) \approx (s+1)\int_0^1 d_p p(1-p)^s \, dp$$

We let s to be sufficiently big (>30) as a precondition:



Based on estimated distribution of SV frequencies for individual of current data, we noticed that two peaks at both ends of the graph, which represent reference as major allele and minor allele and are believed to be somewhat symmetric.

We use φ to represent the upper boundary of those rare SVs' frequencies. Similarly, we also noticed that frequencies is relatively flat from φ to 0.6, thus the density between φ-0.6 can be treated as a constant, we use D to represent.

Based on what we observed:

$$(s+1)\int_0^1 d_p p(1-p)^s\,dp$$
$$=(s+1)\int_0^{\varphi} d_p p(1-p)^s\,dp+(s+1)\int_0^1 d_p\,p(1-p)^s\,dp$$
$$=(s+1)\int_0^{\varphi} d_p p(1-p)^s\,dp+(s+1)\int_{\varphi}^1 D\,p(1-p)^s\,dp+(s+1)\int_{0.4}^1 (d_p-D)p(1-p)^s\,dp$$

for p is sufficiently big(>0.4),
coefficient $=(s+1)(1-p)^s<30*0.6\,\hat{}\,30=2.210739*10^{-7}$,
Which is small enough compared to SV number$(estimate\ to\ be<4000)$ *to be ingored*.
So,
$$it\approx(s+1)\int_0^{\varphi}(d_p-D)p(1-p)^s\,dp+(s+1)D\int_{\varphi}^1 p(1-p)^s\,dp$$

for $\int p(1-p)^s\,dp$, *let q denote* $1-p$
$$=\int(1-q)q^s d(1-q)$$
$$=-\int(1-q)q^s\,dq$$
$$=\frac{q_i^{s+1}}{s+1}-\frac{q_i^{s+2}}{s+2}$$

$$Thus\ (s+1)\int_{\varphi}^1 p(1-p)^s\,dp=(s+1)\left(\frac{q_i^{s+1}}{s+1}-\frac{q_i^{s+2}}{s+2}\Big|_0^{1-\varphi}\right)=\frac{1+(s+1)\varphi}{s+2}(1-\varphi)^{s+1}\approx\frac{1}{s+2}(1-\varphi)^{s+1}$$

$$when\ \varphi<<\frac{1}{s+1},\ it\approx\frac{1}{s+2}(1-\varphi)^{s+1}$$ ③

Similarly,We let ε *represents the peak closest to* $\frac{1}{s+2}$, *so*
$$(s+1)\int_0^{\varphi}(d_p-D)p(1-p)^s\,dp$$
$$\approx(1-\varepsilon)^s*(s+1)\int_0^{\varphi}(d_p-D)p\,dp$$
$\int_0^{\varphi}(d_p-D)p\,dp$ *is a constant, we use R to represent*.
Thus, $it\approx(1-\varepsilon)^s*R$ ④

Hence, combining ③ and ④,

$$Exp(I_{s+2}) \approx R*(s+1)(1-\varepsilon)^{s} + \frac{D}{s+2}\left(1-\varphi\right)^{s+1}$$

$$Hence,\ \ Exp(I_s) \approx R*(s+1)(1-\varepsilon)^{s-2} + \frac{\mathbf{D}}{s}(1-\varphi)^{s-1} \qquad ⑤$$

We fit last 10 increments on ⑤ to estimate constants B and D, which is used to project total number of non-singleton SVs on larger sample size.

Reference

1. Crowley, P. (2003). Resampling Methods for Computation-Intensive Data Analysis in Ecology and Evolution. Annual Review of Ecology and Systematics. 23. 405-447. 10.1146/annurev.es.23.110192.002201.