BSC6883: Simulated sequencing

Jeffrey West[1,🐦1]

[1]Integrated Mathematical Oncology Department, H. Lee Moffitt Cancer Center & Research Institute

🐦[1] @mathoncbro

1 A model of neutral evolution

We now re-state the model introduced in ref. 1. Consider a tumor of size N cells at time t, $N(t)$ with cells dividing at a rate of λ. During cell division, somatic mutations occur at rate μ. The average ploidy of a cell (number of chromosome sets) is π. The expected number of new mutations per time interval is given by:

$$\frac{dM}{dt} = \mu\pi\lambda N(t) \tag{1}$$

In other words, we accrue more mutations, M, as mutation rate increases, as division rate increases, or as ploidy increases. Solving this equation requires us knowing the tumor size at each point in time:

$$M(t) = \mu\pi\lambda \int_{t_0}^{t} N(t)dt \tag{2}$$

Next, we introduce a new variable, β, which represents the number of 'effective' cell divisions in which both daughter cells survive. In an exponentially growing population, the tumor size is given by:

$$N(t) = e^{\lambda\beta t} \tag{3}$$

Substituting this into 2 gives the explicit solution:

$$M(t) = \frac{\mu\pi}{\beta}\left(e^{\lambda\beta t} - e^{\lambda\beta t_0}\right) \tag{4}$$

This describes the total number of subclonal mutations that accumulate in a tumor within the time interval $[t_0, t]$.

However, it should be noted that it is difficult (impossible?) to directly measure parameters μ, λ, β in humans. Yet, this model does provide useful information to make clinically-translational predictions.

For a new mutation occuring at time t, when the tumor size is $N(t)$, the allelic frequency is given by:

$$f = \frac{1}{\pi N(t)} = \frac{1}{\pi e^{\lambda\beta t}} \tag{5}$$

If a new mutation arises in a tumor consisting of 100 cells, the cell fraction will be $f = 1/100$ and remain constant over time.

Q1: what assumptions does this make?

In the absence of clonal selection or substantial drift, this allelic frequency will remain constant during expansion. This is because all cells with *and* without the new mutation grow at a constant rate, $\lambda\beta$.

In a neutrally evolving tumor, this implies that we can use frequency f to infer tumor age t, as the two variables are interchangeable.

In a diploid tumor ($\pi = 2$), the maximum expected allelic frequency of clonal variants arising at $t_0 = 0$:

$$f_{\mathtt{max}} = \frac{1}{\pi e^{\lambda\beta t}} = \frac{1}{\pi} = 0.5 \tag{6}$$

Rearranging this equation and plugging back into our explicit function of $M(t)$ gives us a formula relating the expected number of mutations as a function allelic frequency, f:

$$M(f) = \frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{\mathtt{max}}}\right) \tag{7}$$

Importantly, this provides a measureable prediction. Allelic freqency, f, and the distribution $M(f)$ is provided by next-generation sequencing from bulk sequencing of tumor biopsies.

This equations is known as the $1/f$ power-law distribution prediction. Plotting the cumulative number of mutations observed as a function of inverse of allelic frequency is shown in figure 1B. Here, the data matches the model prediction (red line) with good agreement.

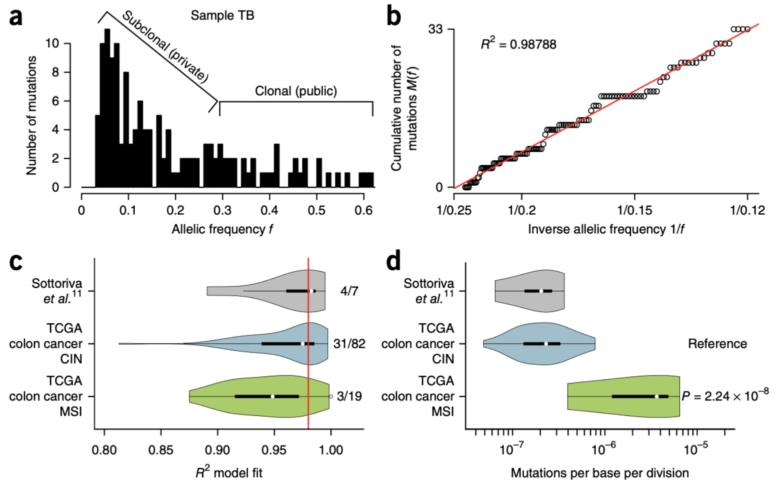

Figure 1 Neutral evolution is common in colon cancer and allows measurement of the mutation rate in each tumor. (**a**) The output of next-generation sequencing, such as whole-exome sequencing, can be summarized as a histogram of mutant allele frequencies, here for sample TB. Considering purity and ploidy, mutations with relatively high frequency (>0.25) are likely to be clonal (public), whereas low-frequency mutations capture the tumor subclonal architecture. (**b**) The same data can be represented as the cumulative distribution *M*(*f*) of subclonal mutations. This distribution was found to be linear with 1/*f*, precisely as predicted by the neutral model. (**c**) The R^2 goodness-of-fit measure for our CRC cohort ($n = 7$) and the TCGA colon cancer cohort ($n = 101$) grouped as having CIN or MSI confirmed that neutral evolution is common (38/108; 35.1% of samples with $R^2 \geq 0.98$). The red line indicates the $R^2 = 0.98$ threshold for discriminating neutral from non-neutral tumors. (**d**) Measurements of the mutation rate showed that the groups with CIN had a median mutation rate of $\mu_e = 2.31 \times 10^{-7}$, whereas tumors with MSI had a 15-fold higher rate (median $\mu_e = 3.65 \times 10^{-6}$; *F* test, $P = 2.24 \times 10^{-8}$), as predicted on the basis of their deficiency in DNA mismatch repair.

Figure 1. Reproduced from ref. 1 (Figure 1).

The slope of this line is equal to the model's prediction for mutation rate per effective cell division:

$$\mu_e = \frac{\mu}{\beta} \tag{8}$$

Q2: what if the model does not match the data?

In the event that the model best fit line does not fit the data with good agreement, one of our modeling assumptions must be incorrect. Here, we assumed exponential growth with a constant growth rate for each new mutation (known as the “neutral” evolution assumption).

If subclonal selection were present, the model would not provide good aggreement (known as the “non-neutral” assumption.

Here, the authors used the R^2 value as a metric for how well the model fit the data. If $R^2 > 0.98$ (an arbitrary cut off!), the authors assume that the neutral evolution hypothesis can not be rejected. If $R^2 \leq 0.98$, then this particular tumor is assumed to be non-neutral.

A significant proportion of tumors were found to be neutrally evolving (figure 1C).

2 Availability of code

Resources for these lecture notes can be found online[2].

References

1. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature genetics* **48**, 238–244 (2016).
2. West, J. BSC-6882 and BSC-6883 lecture notes. https://github.com/jeffreywest/IMO-lecture-notes (2023).

3 Supplementary