1 Lines in solution space for a given VAF-distribution

In simulations of the VAF distribution of neutrally evolving populations that reproduce according to a moran-process, we found that there is a striking similarity between down-sampled VAFs for certain combinations of time, population size and mutation rate. More specifically, we found that, if you start with a population size N_1, a time measured in reproduction events t_1, and a mutation rate μ_1, you can calculate t_2, μ_2 given a N_2 where both combinations lead to almost identical VAFs:

$$t_2 = t_1 \cdot (N_2/N_1)^2 \ , \ \mu_2 = \mu_1 \cdot (N_1/N_2) \tag{1.0.1}$$

You can also look at this from the perspective of a variable division rate ρ with a fixed time τ. This would mean that $t = N \cdot \tau \cdot \rho$. We could then calculate ρ_2 as :

$$N_2 \cdot \tau \cdot \rho_2 = N_1 \cdot \tau \cdot \rho_1 \cdot (N_2/N_1)^2 \implies \rho_2 = \rho_1 \cdot (N_2/N_1)$$

Note that this does not appear to hold true for very small values of t (i.e. $t < N$). In the context of (blood) stem cells, we can then further split this up into an symmetric division rate ρ and an asymmetric division rate φ. Since asymmetric division do not propagate the distribution along the prevalences, but add a few mutations to an individual, it effectively increases only the mutation rate:

$$\mu_{eff} = \mu \cdot \frac{\rho + \varphi/2}{\rho} \tag{1.0.2}$$

If we further re-parametrize this to a general division rate $r = \rho + \varphi$ and a chance for asymmetric division $p = \frac{\varphi}{r}$, then:

$$\rho = r \cdot (1 - p)$$

In other words, in this parametrization, if we keep the r fixed, ρ is reduced when increasing p (and the same goes for t). In either parametrization, this means that we actually can project what happens for a chosen p or φ based on a simulation with no asymmetric divisions, by only changing μ and ρ.

There are two possible arguments/intuitive explanations why the $1/N^2$ relationship between t and N might be true that come to my mind.
The first argument is that fixation time of a single mutation scales with N^2 as well. So it would be intuitive for mutations to wander across the prevalences with a similar speed as well.
The other explanation concerns the equations for a single step of the moran process directly:

$$C_k = \frac{k}{N} \cdot (1 - \frac{k}{N}) \cdot M_k = \frac{Nk - k^2}{N^2} \cdot M_k$$

This relates to the number of mutations M_k with a certain prevalence k as follows:

$$\frac{dM_k}{dt} = C_{k-1} + C_{k+1} - 2C_k$$

For the low prevalences (k«N), which are the ones we are interested in and which are the easiest to analyze, we can estimate this as:

$$C_k \sim \frac{k}{N} \cdot M_k$$

If we put this into the entire change of M_k:

$$\frac{dM_k}{dt} = C_{k-1} + C_{k+1} - 2C_k \sim ((k-1) \cdot M_{k-1} + (k+1) \cdot M_{k+1} - 2k \cdot M_k) \cdot \frac{1}{N}$$

So, a single step changes all low-prevalence M_k with a magnitude scaling with $1/N$. However, you have to keep in mind that the number of prevalences also increase with N directly. So moving a single prevalence becomes a smaller and smaller effective distance for increasing N. Therefore, in total, a single step should expect to move the distribution with a speed scaling with $1/N^2$.
This also explains why it doesn't work for very low times: If only a small amount of time has passed, the main change to the distribution comes from the new mutations turning up at prevalence 1. This scales with μ directly, with no influence from N.
Overall, this is no proof, but I think we have a decent reason to suspect this relationship to be true.

2 The Experiment

Let's now pretend we have a specific experimentally derived VAF, we know the sample size k and the true mutation rate μ, and we want to find out the true population size N, the true number of symmetric replication events t as well as the true percentage of asymmetric divisions p. I use this parametrization because I consider it easier to calculate with, but the approach works just as well with a different parametrization.
First, we pretend that $p = 0$ and that $N = k$. This allows us to run an optimization approach only over t. We call the resulting 'false' time t_f, and we get a 'false' mutation rate μ_f as well. Then, we know what the 'true' effective mutation rate for our population is based on equation 1.0.2 :

$$\mu_{eff}(p) = \mu \cdot \frac{\rho + \varphi/2}{\rho} = \mu \cdot \frac{1 - p/2}{1 - p}$$

Let's now change equation 1.0.1 to be based on a known μ instead of a known N:

$$N_2 = N_1 \cdot \frac{\mu_1}{\mu_2} \text{ , } t_2 = t_1 \cdot (\frac{\mu_1}{\mu_2})^2 \tag{2.0.1}$$

We can then use the relative magnitudes of the measured μ_f and the 'known' μ_{eff} to get a line of possible N's and t's based on a given p. We insert into equation 2.0.1 as follows: $N_2 = N(p)$, $N_1 = k$, $\mu_2 = \mu_{eff}(p) = \mu \cdot \frac{1-p/2}{1-p}$, $\mu_1 = \mu_f$, $t_2 = t(p)$, $t_1 = t_f$:

$$N(p) = k \cdot \frac{\mu_f}{\mu} \cdot \frac{1-p}{1-p/2}$$

$$t(p) = t_f \cdot (\frac{\mu_f}{\mu})^2 \cdot (\frac{1-p}{1-p/2})^2$$

If N is known but p is not, we get:

$$p = \frac{k \cdot \mu_f - N \cdot \mu}{k \cdot \mu_f - N/2 \cdot \mu}$$

If we want to know the true number of asymmetric AND symmetric replication events τ, it is $\tau = t_f \cdot \frac{1}{1-p}$.

In conclusion, we need to know one of either p, N or t to find out the other two.

3 Looking at cell burden

By looking at the mutational burden, we can derive the rate r_{cell} by which the cell has divided if we know the mutation rate μ and the real time that has passed τ: If a cell has a burden of ~ 100, we know the mutation rate is 1.3, and 10 years have passed, then $r_{cell} \approx 1000/(1.3 \cdot 10) \approx 7.7$ per year. What does this tell us about the aforementioned $r = \rho + \phi$?
For this purpose, imagine a population of only 2 cells and that there is no chance to randomly "overwrite" yourself during symmetric cell divisions(this makes the difference more stark between them, but does fundamentally change the principles). First, we assume $\phi = 1$ and $\rho = 0$, i.e. cells only have asymmetric cell division and hence "overwrite" themselves. If you wait $\tau = 100$ units of time, both cells will have $\phi \cdot \tau = 100$ mutations.
Now, let's assume the opposite: $\rho = 1$, $\phi = 0$. Now it gets more interesting. Since in this case cell A is overwriting cell B or vice versa, in every replication event, both cells will gain a mutation! This means that in every time step, both cells gain 2 mutations, as opposed to 1, when they only replace themselves. Hence, both cells will have $\rho \cdot 2 \cdot \tau = 200$ mutations.
If we know see 100 mutations in an experiment, even if we know μ and τ, there is still a factor of 2 by which the true r might differ from all the r_{cell}.
Even worse, the variable we're actually interested in is ρ because this variable controls the shape of the VAF-distribution, ϕ has no influence on that. ρ can still be anything between 0 and $r_{cell}/2$ solely based on the mutation burden. For $\rho = r = r_{cell}/2$, it corresponds simply to the $p = 0$ estimate of the ABC for N, and $\rho \to 0$ simply leads to $N \to 0$ as well, so there is no added information compared to what the ABC finds when analyzing the VAF.

With perfect strand segregation, this might be solvable since asymmetric cell division would have no effect on the distribution, but as soon as it becomes non-perfect, it's unsolvable again and just changes the factor $\frac{1-p}{1-p/2}$ to $\frac{1-p}{1-p(1+s)/2}$, with s being the likelihood of strand segregation.