

File S1

Support for Assumption of 2 to 4 Germ-Line Progenitor Nuclei

There is some contention in the literature regarding the number of progenitor nuclei that found the germline. Estimates range from 2 through more than 8. We have tested these postulates by comparing the clonal composition of germ lines in our observed data to model results while varying several parameters in the model. The term “clone” is used here to denote all the descendants of one replication of an alkylated strand. Clonal composition is defined as the number of different clones represented in the germ line. To make this comparison we first eliminated several experimental lines that had resulted in overlapping clones and were therefore ambiguous. (See Table 6 in the body of the paper – the excluded lines are denoted in bold face.) We summed observed results of data from chromosome 2 and 3 three. We removed 12% from the observed number of monoclonals to account for complete mutations as described in the article. [See discussion of the contribution of completes in the section entitled *Estimating the frequency of complete (whole body) mutations*].

As a result of stochastic events leading up to the selection of germ-line-progenitor nuclei, our model predicts a distribution in the ratio of mono- and polyclonal germ lines. By running the model repeatedly while varying the degree of randomness with which the nuclei are selected from the syncytium, and by varying the number of progenitor nuclei, we are able to confirm that the best fit of the model parameters to our experimental data lies between two and four progenitor nuclei.

We vary the degree of randomness with which nuclei are selected by varying the size of the pool of nuclei from which progenitors are drawn from 20 through 50 (the smaller the pool, the more closely related the nuclei). Varying the pool size from 20 to 50 nuclei had little effect. By varying the number of progenitor nuclei from 1 through 8 we find that the model configured for one progenitor nucleus predicts more monoclonals and fewer polyclonals than observed. Models configured for four through eight progenitor nuclei predict too few monoclonals, and an excess of polyclonals beyond tri-clonal. The model configured for three progenitor nuclei fits the observed data for every clonal count with the exception of five: the model predicts that there will be no nuclei with a clonal count of five while the observed data contain a single instance. This one instance may be an experimental artifact, or the result the inclusion of an alkylated strand that then amplified after germ line selection, the effect of which is not taken into account by the model. Regardless, a complete fit of the model parameters to the observed data is obtained by the reasonable assumption that the observed data result from germ lines descended from a mix of two, three or four progenitor nuclei.

File S1

Table S1 Comparison of model predictions to observed clonal counts assuming one through eight progenitor nuclei.

		Number of Progenitor Nuclei																
			1		2		3		4		5		6		7		8	
Pool	20	Observed	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Clonal Counts	1	71.8%	97.3%	100.0%	59.8%	79.7%	45.7%	72.4%	41.7%	66.9%	38.1%	60.2%	36.9%	60.2%	34.9%	57.1%	32.9%	54.3%
	2	20.9%	0.0%	2.7%	19.0%	39.2%	20.1%	47.2%	24.0%	47.2%	25.0%	44.1%	25.4%	43.7%	26.2%	43.7%	25.4%	43.7%
	3	6.4%			0.0%	4.3%	2.0%	13.0%	2.8%	15.9%	1.6%	15.5%	2.4%	15.4%	2.8%	14.3%	2.8%	12.8%
	4	0.5%					0.0%	3.9%	0.4%	6.6%	1.2%	9.1%	1.2%	10.0%	1.2%	10.4%	1.6%	10.8%
	5	0.5%					0.0%	0.4%	0.0%	3.5%	0.0%	4.8%	0.4%	7.5%	0.4%	7.5%	1.2%	9.4%
	6	0.0%									0.0%	2.4%	0.0%	3.2%	0.0%	4.8%	0.0%	6.0%
	7	0.0%											0.0%	1.6%	0.0%	2.4%	0.0%	2.8%
	8	0.0%											0.0%	0.4%	0.0%	0.8%	0.0%	0.8%
	9	0.0%													0.0%	0.4%	0.0%	0.4%
Pool	50	Observed	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Clonal Counts	1	71.8%	96.8%	100.0%	59.1%	80.6%	46.8%	72.4%	41.2%	64.8%	37.6%	62.0%	36.8%	57.6%	35.2%	56.8%	34.4%	55.6%
	2	20.9%	0.0%	3.2%	16.7%	39.7%	21.2%	44.8%	24.4%	44.4%	25.2%	44.8%	25.6%	44.8%	26.0%	44.0%	27.6%	43.6%
	3	6.4%			0.0%	4.3%	2.4%	13.2%	2.4%	15.2%	3.1%	15.6%	3.2%	14.8%	3.2%	14.0%	3.2%	14.0%
	4	0.5%					0.0%	4.4%	0.0%	7.6%	0.8%	10.8%	1.2%	10.4%	1.6%	11.2%	2.0%	10.4%
	5	0.5%							0.0%	3.5%	0.0%	5.2%	0.0%	6.8%	0.4%	8.4%	0.4%	10.0%
	6	0.0%									0.0%	2.4%	0.0%	3.6%	0.0%	4.4%	0.0%	5.6%
	7	0.0%											0.0%	1.2%	0.0%	2.4%	0.0%	2.8%
	8	0.0%													0.0%	0.8%	0.0%	0.8%
	9	0.0%																

Table S1 compares the observed ratio of germ lines with those predicted by our model. The upper half of the table shows model results when germ line precursors are drawn at random from 20 adjacent nuclei; the lower half shows results when precursors are drawn from a pool 50. The larger pool size simulates a more random selection of nuclei from the syncytium. Row headings indicate clonal count of one through nine. The observed data appear in the third column. Column headings indicate the number of progenitor nuclei modeled, with the minimum and maximum model results obtained for each out of a thousand independent simulations for each pair of variables. Model predictions with a minimum and maximum of zero are left empty. Cells highlighted in green indicate when the observed values fall between the minimum and maximum model predictions, including those cases when the observed results were zero.