Appendices for determination of SIV production by infected activated and resting cells

Cavan Reilly*[1] Steve Wietgrefe[2] Gerald Sedgewick[3] Ashley Haase[2]

Acknowledgements

This work was supported by NIH grants 4R37 AI28246, R01 AI48484, R01 AI56997.

Appendix 1: Likelihood derivation

Likelihood for a virion count

In order to conduct statistical inference for the ratio of θ for activated to resting cells, we will first derive a likelihood for the counts of virions around an infected cell for a given cell type. Let t denote the length of time that a cell was infected prior to being sampled. Break this time up into discrete intervals $[j\delta, (j+1)\delta)$ for $j = 1, \ldots, J$ where $(J+1)\delta = t$. Suppose the quantity of virus produced over each of these intervals is approximated by $y(\delta_j)\delta$, where $\delta_j = j\delta$ (we will eventually take the limit as δ approaches zero). Let u_j represent the number of virions produced in time interval $[\delta_j, \delta_{j+1})$ that survive until the time that the tissue sample is obtained. We will suppose that $u_j|t$ is distributed according to the Binomial distribution with the number of trials given by $y(\delta_j)\delta$ and the success probability is given by $\mathrm{e}^{-(t-\delta_j)/\phi}$. The assumption for the success probability follows from assuming that the lifetimes of the virions follow an exponential law with parameter ϕ since the probability of a virion surviving from time δ_j until a time greater than or equal to t is $\mathrm{e}^{-(t-\delta_j)/\phi}$.

Now suppose x is the number of virions we observe in the vicinity of some cell. The basis for the derivation

[1]Corresponding Author: Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Bldg., MMC 303, 420 Delaware St. SE, Minneapolis, MN 55455-0378. e-mail: cavanr@biostat.umn.edu, Phone: 612-624-9644, FAX: 612-626-0660

[2]Department of Microbiology, School of Medicine, University of Minnesota, 1460 Mayo Bldg., MMC 196, 420 Delaware St. SE, Minneapolis, MN 55455-0378.

[3]Department of Neuroscience, School of Medicine, University of Minnesota, Rm. 6-145 JacH 1216, 321 Church St., Minneapolis, MN 55455-0378.

is that the count we observe in the tissue sample is just the total number of virions that survive from the time interval in which they were produced until the time at which the tissue sample was taken. We obtain the likelihood of the virion count x as it depends on the parameters governing $y(t)$ and the distribution of the lifetimes of the virions as follows:

$$L(x|\beta_0,\beta_1,\tau,\phi) = \frac{\mathrm{d}}{\mathrm{dx}}\mathrm{E}\left[P\left\{\sum_j u_j\delta \le x|t\right\}\right], \tag{1}$$

where the expectation is with respect to the distribution of how long the cell has been infected. But since $u_j|t \sim \mathrm{Bin}(y(\delta_j)\delta, \mathrm{e}^{-(t-\delta_j)/\phi})$ it follows that conditional on t, $\sum_j u_j\delta$ is distributed according to the Binomial distribution with parameters $\sum_j y(\delta_j)\delta$ and $\pi_0(t) = \frac{\sum_j y(\delta_j)\mathrm{e}^{-(t-\delta_j)/\phi}}{\sum_j y(\delta_j)}$. If we then differentiate equation 1 to obtain the mass function of x we obtain

$$\mathrm{E}\left[\binom{\sum_j y(\delta_j)\delta}{x}\pi_0(t)^x(1-\pi_0(t))^{\sum_j y(\delta_j)\delta - x}\right]. \tag{2}$$

Finally we take the limit in equation 2 as δ goes to zero to obtain the likelihood for x

$$L(x|\beta_0,\beta_1,\tau,\phi) = \mathrm{E}\left[\binom{\int_0^t y(s)\,ds}{x}\pi(t)^x(1-\pi(t))^{\int_0^t y(s)\,ds - x}\right], \tag{3}$$

where $\pi(t) = \frac{\int_0^t y(s)\mathrm{e}^{-(t-s)/\phi}\,ds}{\int_0^t y(s)\,ds}$. Given the previous expression for $y(t)$ we obtain the following expressions

$$\int_0^t y(s)\,ds = \beta_0\left[\frac{\mathrm{e}^{\beta_1\min(t,\tau)}-1}{\beta_1} + \mathrm{e}^{\beta_1\tau}(t-\tau)1_{\{\tau\le t\}}\right]$$

and

$$\int_0^t y(s)\mathrm{e}^{-(t-s)/\phi}\,ds = \beta_0\left[\frac{\mathrm{e}^{-t/\phi}(\mathrm{e}^{(\beta_1+1/\phi)\min(t,\tau)}-1)}{\beta_1+1/\phi} + \phi\mathrm{e}^{\beta_1\tau}(1-\mathrm{e}^{(\tau-t)/\phi})1_{\{\tau\le t\}}\right].$$

Hence we can evaluate the likelihood by integrating the expression in the expectation sign in equation 3 against the probability density of the t variables (this density is derived in Appendix 2). While the expectation can not be done analytically, we can numerically evaluate the likelihood (here we use 10 point Gaussian quadrature), and this is sufficient for likelihood based inference.

Joint likelihood for all cell counts

Supposing that the counts are independent of one another for distinct cells and recalling that we use $\beta_{0\mathrm{a}}$ to represent the intercept for activated cells and $\beta_{0\mathrm{r}}$ for the intercept for resting cells (and use similar notation

for the other parameters and the counts), then the joint likelihood for all parameters can be written

$$L(x_{i\mathrm{a}}, x_{i\mathrm{r}}, i = 1, \ldots, 100 | \beta_{0\mathrm{a}}, \beta_{0\mathrm{r}}, \beta_{1\mathrm{a}}, \beta_{1\mathrm{r}}, \tau_{\mathrm{a}}, \tau_{\mathrm{r}}, \phi) = \left[\prod_i L(x_{i\mathrm{a}} | \beta_{0\mathrm{a}}, \beta_{1\mathrm{a}}, \tau_{\mathrm{a}}, \phi_{\mathrm{a}})\right] \left[\prod_i L(x_{i\mathrm{r}} | \beta_{0\mathrm{r}}, \beta_{1\mathrm{r}}, \tau_{\mathrm{r}}, \phi_{\mathrm{r}})\right].$$

Appendix 2: The probability distribution of the time since infected

Let z denote the time at which a cell was infected (measured from the time since infection), and let t denote the length of time that this cell was infected before the cell dies (so the cell dies at time $z + t$ measured from the time of infection). Let T denote the time at which the animal was euthanized. Suppose that z is uniformly distributed over the interval $[0, T]$ (since the animal was euthanized on the 12^{th} day, $T = 12$). This is a conservative assumption from the perspective of determining the relative quantity of virus production attributable to resting cells since assuming that the likelihood of a cell getting infected increases over the initial course of infection implies that the infected life times of all cells is shorter than our assumption which uses the uniform distribution. But since resting cells are productively infected for a longer period than activated cells, an assumption which shortens the lifetimes of infected cells has a greater impact on reducing the implied lifetime of resting cells, and this would lead to the conclusion that resting cells produce more virus. In addition, we also set $T = 6$ since its possible that virus doesn't gain entry to the lymph node until up to a week after initial infection. The results of setting $T = 6$ were very similar to the results reported here and not discussed further.

As previously discussed we suppose that the t are random variables distributed according to an exponential distribution with parameter λ. We now derive the distribution of the random variable $T - z$ conditional on $z + t > T$ and $z < T$ (i.e. the length of time the cell has been infected given that we sampled the tissue at a time the cell was productively infected). Now, by the definition of conditional probability

$$p(T - z < y | z + t > T \text{ and } z < T) = \frac{p(T - z < y \text{ and } z + t > T \text{ and } z < T)}{p(z + t > T \text{ and } z < T)} \tag{4}$$

and if we denote $K = \left(\int_0^T \mathrm{e}^{-s/\lambda}\, ds\right)^{-1}$ (so that $K = \frac{1}{\lambda(1-\mathrm{e}^{-T/\lambda})}$) then

$$\begin{aligned} p(T - z < y \text{ and } z + t > T \text{ and } z < T) &= \int_{T-z<y \text{ and } z+t>T \text{ and } z<T} \frac{1}{T} K \mathrm{e}^{-t/\lambda}\, dt\, dz \\ &= \frac{K}{T} \int_{T-y}^{T} \int_{T-z}^{T} \mathrm{e}^{-t/\lambda}\, dt\, dz \\ &= \frac{\lambda K}{T}\left[\lambda(1 - \mathrm{e}^{-y/\lambda}) - y\mathrm{e}^{-T/\lambda}\right]. \end{aligned} \tag{5}$$

In a similar fashion we obtain

$$p(z + t > T \text{ and } z < T) = \frac{\lambda K}{T}\left[\lambda(1 - \mathrm{e}^{-T/\lambda}) - T\mathrm{e}^{-T/\lambda}\right], \tag{6}$$

so that using equations 5 and 6 in equation 4

$$p(T - z < y | z + t > T \text{ and } z < T) = \frac{\lambda(1 - \mathrm{e}^{-y/\lambda}) - y\mathrm{e}^{-T/\lambda}}{\lambda(1 - \mathrm{e}^{-T/\lambda}) - T\mathrm{e}^{-T/\lambda}}.$$

Furthermore we obtain the density of these variables by differentiating with respect to y to obtain $C(\mathrm{e}^{-y/\lambda} - \mathrm{e}^{-T/\lambda})$ where $C = \left[\lambda(1 - \mathrm{e}^{-T/\lambda}) - T\mathrm{e}^{-T/\lambda}\right]^{-1}$.

Appendix 3: Details of statistical inference

We suppose that all parameters are independent in the prior distribution. Since all of the parameters in our model are necessarily non-negative, for all parameters except the τ parameters and the half-lifes of infected cells we adopt the conventional non-informative prior for each of these parameters which supposes that the prior distribution is proportional to the inverse of the parameter (1). This is equivalent to supposing that the prior is flat on the logarithmic scale. For the τ parameters we use a log normal prior where the logarithm of τ is assumed to follow a normal distribution with mean 2.3 and variance 1 (so 95% of the prior mass is put on the interval [0, 52], and this constitutes a very vague prior for this parameter). We suppose that the logarithm of the half-life of a resting cell is normally distributed either with mean 2.64 and a standard deviation of 0.05 (when we set this parameter to be roughly 14) or 1.39 with a standard deviation of 0.1 (when we set this parameter to be roughly 4). For the activated cells, we suppose that the logarithm of the

half-life is normally distributed with mean 0.41 and standard deviation 0.1. The priors for the half-lifes are rather informative, but they are based on estimates available in the literature.

We use a simulation based approach to characterize the joint posterior distribution of all parameters (see 1 for greater detail on this approach to statistical inference). In detail, we first explored the posterior distribution by finding posterior modes using simulated annealing (using the implementation in reference 2). We used an initial computational temperature of 100 and reduced the temperature by 10% every 500 iterations. We used many different starting values since it is not obvious that all parameters are well identified given the manner in which the likelihood is evaluated. This indicated that there was a single well defined mode since the algorithm always converged to the same point in parameter space. We then numerically approximated the Hessian matrix of the log-likelihood at this mode (using Neville's algorithm as implemented in reference 2) and used the negative inverse of this matrix as our covariance matrix of a normal distribution to use as the jumping kernel in the Metropolis algorithm (we scaled this matrix by 0.8 to obtain a Markov chain that accepted 24% of the proposed moves). We then used the Metropolis algorithm to draw samples from the joint posterior distribution of all the parameters. Finally, we used the simulated values of the parameters to compute simulated values for θ for each cell type and then took the ratio of these quantities to obtain samples from the posterior distribution of the ratio of the quantity of virus produced by activated cells to the quantity of virus produced by resting cells.

Three Markov chains were run in parallel to monitor the convergence of the Metropolis algorithm, and the $\sqrt{\hat{R}}$ statistic (see 3) indicated that the chain converged for the ratio of the θ's for the 2 cell types since the value of this statistic was less than 1.01 (5000 iterations per chain were necessary to obtain convergence when the prior mean of the half-life of resting cells was 4 and 20,000 iterations per chain when it was 14, and the first half of the iterations were discarded for burn-in purposes). While the chains converged quickly for the ratio $\theta_{\mathrm{a}}/\theta_{\mathrm{r}}$, many samples were necessary to obtain convergence for τ_{a} and τ_{r}. The initial values where obtained by first simulating normal deviates with mean given by the logarithm of the posterior mode and covariance given by the negative inverse of the Hessian matrix of the log-likelihood evaluated at the posterior mode, then exponentiating these values. The C++ code that implements this analysis

can be obtained from `www.biostat.umn.edu/~cavanr/virusProd.c`, the data set can be obtained from `www.biostat.umn.edu/~cavanr/virusProd.txt`, and a file that holds the posterior modes and initial values used in the Metropolis algorithm can be found at `www.biostat.umn.edu/~cavanr/virusProdVals.txt`.

Appendix 4: Derivation of the correction factor

Here we derive the correction necessary due to activated cells having a larger diameter than the tissue section. Suppose that cells are spheres in 3 dimensions and virions bud off of the cell in a spherically symmetric fashion. Then, surrounding each cell nucleus is a cloud of virions that itself is a sphere. Let a represent the radius of a cell, b represent the radius of the spherical cloud of virions associated with a productively infected cell, and let c represent half the thickness of the tissue section. The volume of the cloud of virus is the volume of a sphere with diameter b less the volume of a sphere with diameter a, $\frac{4\pi}{3}(b^3 - a^3)$, and of this we can't compute the top and bottom part of the sphere. Note that since the cell diameter exceeds the thickness of the tissue section, the region that we can not count has the shape of the top of a sphere with a top of a smaller sphere removed. If we work in cylindrical coordinates, the region that we can not count is $\{(r, \theta, z) : a^2 \leq r^2 + z^2 \leq b^2, c \leq |z| \leq a\}$, and the volume of this region is just

$$2\int_0^{2\pi}\int_c^b\int_0^{\sqrt{b^2-z^2}} r\,dr\,dz\,d\theta - 2\int_0^{2\pi}\int_c^a\int_0^{\sqrt{a^2-z^2}} r\,dr\,dz\,d\theta = 2\pi\left[\frac{2}{3}(b^3 - a^3) - c(b^2 - a^2)\right].$$

Then taking the ratio of the volume of cloud of virus to the volume of the cloud of virus less the volume of the region we can't count, we find the factor by which we should increase all of our counts is

$$\left(1 - \left[1 - \frac{3c}{2}\frac{a+b}{a^2+ab+b^2}\right]\right)^{-1}.$$

Since we have $a = 6.1$, $b = 9.1$ and $c = 4$ we find that the correction factor is 1.9. Similar results are obtained if one simulates the location of cell centers in the microscopic section, then simulates the locations of virions around these centers and determines what proportion of the virions reside in the microscopic section. This latter analysis indicates that ignoring the slight undercount for resting cells arising from some of their associated virions not being in the microscopic section has only a slight effect on the results.

Acknowledgements

This work was supported by NIH grants 4R37 AI28246, R01 AI48484, R01 AI56997.

References

1. Gelman A, Carlin JB, Stern HS, et al. *Bayesian Data Analysis.* London: Chapman & Hall, 1995.

2. Press WH, Teukolsky SA, Vetterling WT, Flannery BP *Numerical Recipes in C*, Cambridge University Press, 1992.

3. Gelman A, Rubin, DB Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* 1992;7:457–511.