Model and Methods for the `cellTypeCompositions` Package

Charles Berry

June 7, 2022

The package implements a Bayesian hierarchical model that captures key features of the cell types produced by clones resulting from gene therapy using constructs that integrate into the host genome. Retroviral gene therapy for some diseases seeks to provide a functional gene to correct Hematopoetic Stem Cell (HSCs). Monitoring of the populations of cells over time can help establish an understanding the success or failure of any gene therapy construct and may give early indications of the likelihood of long term success. In humans undergoing gene therapy, the integration site (IS) locations almost always identify unique integration events, so that all cells sharing an IS location are clones. Thus, the IS locations serve as molecular tags to enable study of gene corrected cell populations in human patients [1, 6, 3, 5]. IS can be isolated using random shearing of the host genome followed by LAM-PCR. Within a single specimen, cell type fraction, and replicate shearing, the unique breakpoints correspond to unique cells with high probability unless the number of cells is so large that cells coincidentally share the same breakpoint [2, 8]. (Even in the latter case, the number of breakpoints is predictably related to the number of cells.)

The essential features of the model for such data are given in the plate diagram in Figure 1. The model specifies two Dirichlet Process priors — one for compositional proportions and one for abundances. A plate graph for a genric Dirichlet Process prior is given in Figure 2. The details of the model and the sampler are laid out in the next section. In particular, Section 2.1 sets out the model, while Section 2.2 lays out the updates and 2.3 give details of an auxilary variable gambit that aids one of the updates.

1 Overview

In this document, the notation uses dots and plus signs in subscripts to indicate vectors composed of all elements using the subscript a dot occupies or summation over all such elements of the subscript occupied by a plus sign. So, $X_{i\bullet} = (X_{i1}, X_{i2}, \dots, X_{iJ})$ and $X_{i+} = (X_{i1} + X_{i2} + \cdots + X_{iJ})$.

As shown schematically in Figures 1 and 2 there is an **Initial Sample** whose cells are from a specimen drawn from a patient at one time. That sample is represented by a $N \times J$ matrix of cell type frequencies for N integration sites and J cell types. A row of that matrix is denoted by $X_{i\bullet}$. The vector of cell type frequencies, $X_{i\bullet}$, depends on a composition specifying cell type proportions ($\eta_{Z_i^{(\eta)}\bullet}$) and an abundance ($\lambda_{Z_i^{(\lambda)}}$). ($Z_i^{(\eta)}$ indexes the element of η associated with IS i and similarly for $Z_i^{(\lambda)}$ and λ.) Each of these is drawn from a Dirichlet Process Prior. The cell type frequencies are subject to filtering in which some cells may be lost (resulting in $\tilde{X}_{i\bullet}$ as governed by a vector of proportions, Υ), leading to a **Secondary Sample**. The Secondary Sample is sorted with possible sorting errors (resulting in $Y_{i\bullet}$ as governed by the transition matrix, Ξ). The **Sequenced Sample** is subject to post-sort loss (resulting in $W_{i\bullet}$ governed by the vector of proportions Ψ). Finally, those $W_{i\bullet}$ for which at lesst one element is non-zero are observed as $W_{i\bullet}^{(+)}$ and the remainder ($W_{i\bullet}^{(-)}$) are not. The data analyst is only cognizant of the $W_{i\bullet}^{(+)}$.

2 Updates

2.1 Priors and Likelihood

Two Dirichlet Process priors are invoked in the model: one for the abundance parameter, λ, and one for the composition, η.

The concentration parameters for the stick-breaking component of the priors have gamma distributions, with fixed parameters for the shape, a, and the rate, b.

$$\alpha_\lambda \sim Ga(a_\lambda, b_\lambda)$$

$$\alpha_\eta \sim Ga(a_\eta, b_\eta)$$

The prior for λ_{i}, has $Ga(k, \beta)$ as its base distribution and $\delta(\theta)$ is the distribution that contentrates all of its mass at θ :

$$\lambda_i|\lambda_1, \lambda_2, \ldots, \lambda_{i-1} \sim \frac{1}{i+\alpha_\lambda-1}\sum_{j=1}^{i-1}\delta(\lambda_j) + \frac{\alpha_\lambda}{i+\alpha_\lambda-1}Ga(\kappa, \beta)$$

The prior for $\eta_{\mathrm{i}\bullet}$, has the Dirichlet distribution $Dir(d\mathbf{1})$ (typically with d being one and $\mathbf{1}$ being the unit vector) as its base distribution

$$\eta_{i\bullet}|\eta_{1\bullet}, \eta_{2\bullet}, \ldots, \eta_{i-1,\bullet} \sim \frac{1}{i+\alpha_\eta-1}\sum_{j=1}^{i-1}\delta(\eta_{j\bullet}) + \frac{\alpha_\eta}{n+\alpha_\eta-1}Dir(d\mathbf{1})$$

For the purpose of Gibbs sampling, the conditional priors for λ_{i} and $\eta_{\mathrm{i}\bullet}$ can be derived by taking the i^{th} element as the last of n elements observed as noted by Neal [7] :

$$\lambda_i|\lambda_{-i} \sim \frac{1}{n+\alpha_\lambda-1}\sum_{i\neq j}\delta(\lambda_j) + \frac{\alpha_\lambda}{n+\alpha_\lambda-1}Ga(\kappa, \beta)$$

$$\eta_{i\bullet}|\eta_{-i\bullet} \sim \frac{1}{n+\alpha_\eta-1}\sum_{i\neq j}\delta(\eta_{j\bullet}) + \frac{\alpha_\eta}{n+\alpha_\eta-1}Dir(d\mathbf{1})$$

where $\lambda_{-i} = \{\lambda_j : j \neq i\}$ and $\eta_{-i\bullet} = \{\eta_{j\bullet} : j \neq i\}$.

The cells of each type for observation i follow a Poisson distribution:

$$X_{ij}|\eta_{ij}, \lambda_i \sim Pois(\lambda_i\eta_{ij})$$

Those cells are sorted and subsampled. The counts of cells omitted and retained after under subsampling are given by C_i and $Y_{i\bullet}$.

$$(C_i, Y_{i\bullet})|X_{i\bullet} \sim \sum_{j=1}^{J} Mn(X_{ij}, \widetilde{\Omega}_{j\bullet})$$

$$\widetilde{\Omega}_{j\bullet} = (1 - \Omega_{j+}, \Omega_{j\bullet})$$

The sum of each row of Ω, Ω_{j+}, gives the probability of retaining a cell of type j, and the vector $\frac{1}{\Omega_{j+}}\Omega_{j\bullet}$ give the probabilities that retained cells are sorted into tubes intended for the respective cell types.

In practice, the analyst only sees the vector of counts for integration site i, $Y_{i\bullet}$, if its sum is non-zero, i.e. the integration site is detected. The subset of detected sites is given by $W_{1\bullet}, \ldots, W_{n\bullet}$

$$W_{i\bullet} = Y_{i'\bullet} \quad \text{where}\, i' = \min\left\{ t : i = \sum_{j=1}^{t} \min(1, Y_{j+}) \right\}$$

The probability law for $W_{i\bullet}$ given $\eta_{i'\bullet}$ and $\lambda_{i'}$ turns out to be the product of Poisson laws truncated when all counts are zero.

$$W_{i\bullet} | \eta_{i'\bullet}, \lambda_{i'} \sim PoisPos(\lambda_{i'} \eta_{i'\bullet} \Omega)$$

where $PoisPos(\cdot)$ is the distribution of a product of independent Poisson variables conditioned on having at least one non-zero value. This distribution is easily shown to be the product of a multinomial and a zero truncated Poisson whose mass function is:

$$dPoisPos(W_{i\bullet}; \rho_\bullet) = dTrPois(W_{i+}; \rho_+) \cdot dMn\left(W_{i\bullet}; W_{i+}, \frac{1}{\rho_+} \rho_\bullet \right)$$

with $dTrPois(\,)$ as the zero-truncated Poisson mass and $dMn(\,)$ as the multinomial mass function.

In what follows, i will be equated with i' for notational convenience (as might happen if all truncated observations had $i' > n$).

2.2 Gibbs Updates

The update strategy is much like that of algorithm 8 of Neal [7] : a scan for $i = 1 : n$ is used to sample a parameter vector for each observation and then each unique parameter vector is sampled conditional on the data elements that depend on it using the base distribution as its prior. The scan-update process yields a draw from the posterior and is repeated to estmate the posterior distribution.

However, there are two sets of parameters having Dirichlet Process priors, and the updates to one must condition on the other. So some method of initializing the scan is needed. Also, in algorithm 8, one or more samples drawn from the prior base distribution in each step, i, of the scan compete with samples drawn earlier. Using a diffuse base distribution for λ poses a problem in that only values with small posterior density may be drawn. This could require excessively long MCMC burn-ins to converge to the stationary distribution. Fortunately, efficent numerical integration of the posterior with respect to $\lambda_i | W_{i\bullet}, \eta_{i\bullet}$ is feasible as is sampling from it when its factor in the likelihood depends on only one value of $W_{i\bullet}$. This ensures good choices for the values of λ selected during the scan. Even though the base prior is not conjugate for λ, the computability of the integral and the posterior draws allows λ_{i} to sampled as in Algorithm 1 of Neal [7] during the scans. The later updates are not affected by this problem as they use Gibbs samples that are initialized with the earlier values.

In this section, the unique values of $\eta_{i\bullet}$, $i = 1 : n$, must be referenced. The notation $\eta_{c_h\bullet}$ will be used to refer to the h^{th} such value, and when an update to that value is performed (in step 2), it is implied that the update applies to all parameters that shared that value after step 1. Similarly, this applies to λ_i, $i = 1 : n$. A single cycle of updates proceeds as follows

1. for each $i = 1 : n$
    - sample $\eta_{i\bullet}|W_{i\bullet}, \lambda_i, \eta_{-i\bullet}$ (on the first pass sample $\eta_{i\bullet}|W_{i\bullet}, \{\eta_{j\bullet} : j < i\}$ using just the multinomial factor of the likelihood)
    - sample $\lambda_i|W_{i\bullet}, \eta_{i\bullet}, \lambda_{-i}$ (on the first pass let $\lambda_{-i} = \{\lambda_j : j < i\}$)
2. (possibly) repeat
    - let $c_1, \ldots, c_H$ index the H unique values of $\eta_{i\bullet}$
    - for (h=1:H) sample $\eta_{c_h\bullet}|\{W_{i\bullet} : \eta_{i\bullet} = \eta_{c_h\bullet}\}, \{\lambda_i : \eta_{i\bullet} = \eta_{c_h\bullet}\}$
    - let $s_1, \ldots, s_K$ index the K unique value of λ_i
    - for (k=1:K) sample $\lambda_{s_k}|\{W_{i\bullet} : \lambda_i = \lambda_{s_k}\}, \{\eta_{i\bullet} : \lambda_i = \lambda_{s_k}\}$
3. sample $\alpha^{(\eta)}$
4. sample $\alpha^{(\lambda)}$

Steps 1–2 may be repeated more than once before proceeding to steps 3–4, and step 2 may be repeated several times for each time step 1 is executed. Typically, steps 1–2 will be repeated many times to *burn-in* the sampler without recording the results, then many replications of 1–4 will be recorded possibly *thinning* the output by discarding several cycles of results for every one that is saved.

In step 2, the update for $\eta_{c_h\bullet}|\{W_{i\bullet} : \eta_{i\bullet} = \eta_{c_h\bullet}\}, \{\lambda_i : \eta_{i\bullet} = \eta_{c_h\bullet}\}$ is given here.

$\eta_{c_h\bullet} \sim Dir(X + d\mathbf{1})$ where $\mathbf{1}$ is a unit vector of length J. The vector, $X = X^{(+)} + X^{(-)}$, depends on samples as follows:

$$\rho_{i\bullet} = \eta_{i\bullet}\Omega$$

$$C_i \sim Nb\left(W_{i+}, 1 - \exp(-\lambda_i \rho_{i+})\right)$$

$$T_{\bullet j} \sim Mn\left(\sum_{i:\eta_{i\bullet}=\eta_{c_h\bullet}} (W_{ij}), \Omega_{\bullet j}/\Omega_{+j}\right)$$

$$X^{(+)} = T\mathbf{1}'$$

The elements of $X^{(-)}$ are sampled as

$$X_j^{(-)} \sim Pois(\sum_{i:\eta_{i\bullet}=\eta_{c_h\bullet}} (C_i + 1)\lambda_i \eta_{ij}(1 - \Omega_{j+}))$$

In step 2, the update for $\lambda_{c_k}|\{W_{i\bullet} : \lambda_i = \lambda_{c_k}\}, \{\eta_{i\bullet} : \lambda_i = \lambda_{c_k}\}$ is given here using the updated $\eta_{i\bullet}, i = 1 : n$.

$$\rho_{i\bullet} = \eta_{i\bullet}\Omega$$

$$D_i \sim Nb\left(1, 1 - \exp(-\lambda_i \rho_{i+})\right)$$

$$\lambda_{c_k} \sim Ga\left(\kappa + \sum_{i:\lambda_i=\lambda_{c_k}} W_{i+},\ \beta + \sum_{i:\lambda_i=\lambda_{c_k}} \rho_{i+}(D_i + 1)\right)$$

Steps 3 and 4 use the updates of Escobar and West [4].

2.3 Posterior Integration and Sampling for λ_i

As mentioned earlier, sampling fresh values of λ_i from the prior will often yield values with very low likelihoods. Here, the integral of and sample from the posterior is developed.

The factor of the posterior containing a unique λ_i is

$$\pi_\lambda(\lambda_i) = \frac{\beta^\kappa \lambda_i^{(\kappa-1)} \exp(-\beta\lambda_i)}{\Gamma(\kappa)} \frac{\exp(-\rho_{i+}\lambda_i)(\rho_{i+}\lambda_i)^{W_{i+}}}{W_{i+}!(1 - \exp(-\rho_{i+}\lambda_i))}$$

To integrate this term, a discrete auxiliary variable is introduced that simplifies integration after which the auxiliary variable is eliminated by summing over all of its values.

Multiplication by a geometric variable, t, with failure probability $\exp(-\rho_{i+}\lambda_i)$ (and hence mass of $\exp(-t\rho_{i+}\lambda_i)(1 - \exp(-\rho_{i+}\lambda_i))$

gives the joint density of λ_{i} and t proportional to

$$f(\lambda_i, t) = \exp(-(\beta + \rho_{i+}(t+1))\lambda_i)\lambda_i^{(\kappa+W_{i+}-1)}$$

which is the kernel of a gamma density for fixed t. The integral with respect to λ_{i} is just the reciprocal of the normalizing constant of that gamma density, viz.

$$g(t) = \int_0^\infty f(\lambda_i, t)\partial\lambda_i = \frac{\Gamma(\kappa + W_{i+})}{(\beta + \rho(t+1))^{\kappa+W_{i+}}}$$

So, the integral of the factor is

$$\int_0^\infty \pi_\lambda(\lambda_i)\partial\lambda_i = C\sum_{t=0}^{\infty} g(t)$$

The terms in $g(t)$ are decreasing in t, but for small values of $\kappa + W_{i+}$ do not decrease rapidly enough to allow just a few terms to be summed. However, the indefinite integral of $g(t)$ is easily found and the approximation

$$g(t) \approx \int_{t-\frac{1}{2}}^{t+\frac{1}{2}} g(t)\partial t$$

becomes better as t increases, and taking

$$\sum_{t=0}^{\infty} g(t) \approx \sum_{t=0}^{m} g(t) + \int_{m+\frac{1}{2}}^{\infty} g(t)\partial t$$

yields an adequate approximation of $\int_0^\infty \pi_\lambda(\lambda_i)\partial\lambda_i$ for reasonably small values of m.

Posterior samples for λ_i can be drawn from the gamma distribution with kernel $f(\lambda_i, t)$ by conditioning on a value of t drawn from its marginal distribution. Samples from the marginal distribution of t can be had by rejection sampling with proposals from the inverse CDF of a continuous distribution proportional to $g(t)$, $-\frac{1}{2} < t < \infty$, rounded to the nearest integer. If a uniform draw on $(0, 1)$ exceeds $g(t)/\int_{t-\frac{1}{2}}^{t+\frac{1}{2}} g(x)\partial x$ the proposal is rejected.

In practice, the approximation is quite poor for small values of t, and the fraction rejected is large. A modification of this scheme is sampling from $m+2$ masses proportional to $g(0), g(1), \ldots, g(m)$ and $\int_{m+\frac{1}{2}}^{\infty} g(x)\partial x$. If any of the values $t = 0, \ldots, m$ is selected, it is accepted. If the last mass is selected, a proposal is drawn from a density proportional to $g(t)$, $m + \frac{1}{2} < t < \infty$ rounded to the nearest integer, and a rejection trial performed as above. If rejection occurs, another draw from the $m + 2$ masses is attempted. Even for values as small as $m = 5$ rejections are rare.

3 plate graph

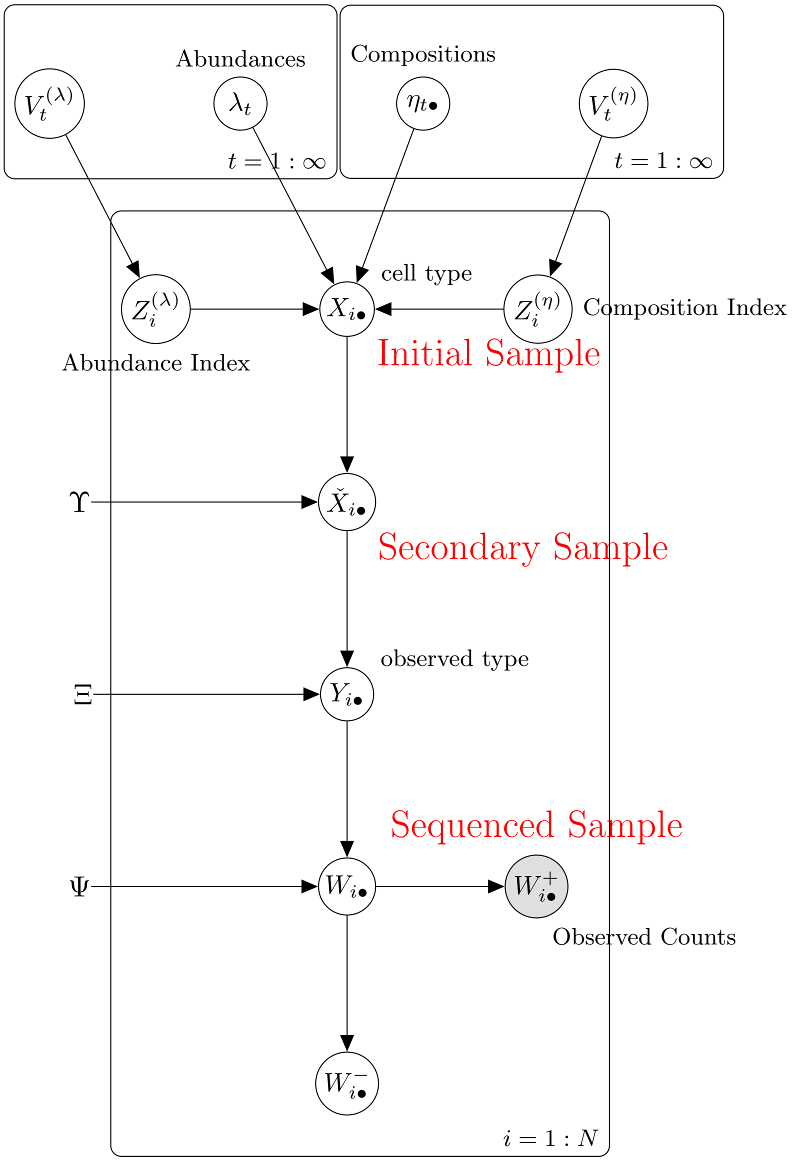

Figure 1: Hierarchical Model of Cell Type Composition. The boxes or 'plates' show replication according to the schema in the lower right corner of each plate with '$1 : N$' indicating the sequence of all integers from 1 to N. The shaded circle shows the observed data, while unshaded circles show latent, unobserved variables. Symbols without circles are fixed constants. The arrows indicate the dependencies of the nodes on one another.

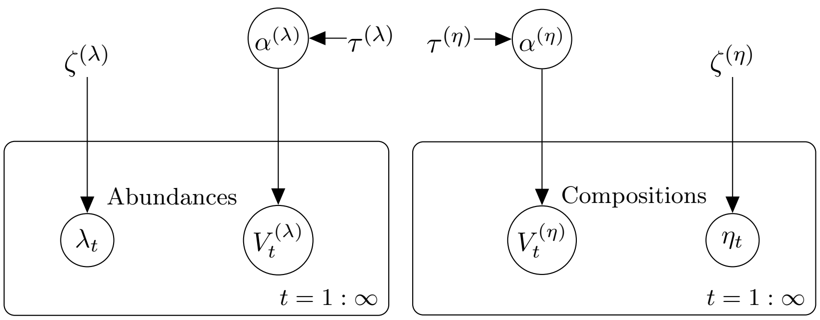

Figure 2: Dirichlet Process Priors for λ and η. Each 'plate' shows replication over an infinite number of draws. The base distribution for λ has a parameter, $\zeta^{(\lambda)}$. The stick-breaking component, $V^{(\lambda)}$, is based on draws from a beta distribution with parameters 1 and $\alpha^{(\lambda)}$. The latter parameter is drawn from a gamma distribution with shape and rate given by the two elements of the vector $\tau^{(\lambda)}$, which are usually chosen to be small to yield a diffuse prior. The mass for the t^{th} component of λ_t is $V_t^{(\lambda)} \prod_{k<t}(1 - V_k^{(\lambda)})$. The setup is analogous for η. Symbols without circles are fixed constants. The arrows indicate the dependencies of the nodes on one another.

References

[1] A. Aiuti, L. Biasco, S. Scaramuzza, F. Ferrua, M. P. Cicalese, C. Baricordi, F. Dionisio, A. Calabria, S. Giannelli, M. C. Castiello, M. Bosticardo, C. Evangelio, A. Assanelli, M. Casiraghi, S. Di Nunzio, L. Callegaro, C. Benati, P. Rizzardi, D. Pellin, C. Di Serio, M. Schmidt, C. Von Kalle, J. Gardner, N. Mehta, V. Neduva, D. J. Dow, A. Galy, R. Miniero, A. Finocchi, A. Metin, P. P. Banerjee, J. S. Orange, S. Galimberti, M. G. Valsecchi, A. Biffi, E. Montini, A. Villa, F. Ciceri, M. G. Roncarolo, and L. Naldini. Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science*, 341(6148):1233151, Aug 2013.

[2] C. C. Berry, C. Nobles, E. Six, Y. Wu, N. Malani, E. Sherman, A. Dryga, J. K. Everett, F. Male, A. Bailey, K. Bittinger, M. J. Drake, L. Caccavelli, P. Bates, S. Hacein-Bey-Abina, M. Cavazzana, and F. D. Bushman. INSPIIRED: Quantification and Visualization Tools for Analyzing Integration Site Distributions. *Mol Ther Methods Clin Dev*, 4:17–26, Mar 2017.

[3] Charles C Berry, Nicolas A Gillet, Anat Melamed, Niall Gormley, Charles RM Bangham, and Frederic D Bushman. Estimating abundances of retroviral insertion sites from dna fragment length data. *Bioinformatics*, 28(6):755–762, 2012.

[4] Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

[5] N. A. Gillet, N. Malani, A. Melamed, N. Gormley, R. Carter, D. Bentley, C. Berry, F. D. Bushman, G. P. Taylor, and C. R. Bangham. The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. *Blood*, 117:3113–3122, Mar 2011.

[6] Salima Hacein-Bey Abina, H. Bobby Gaspar, 24 others, Charles Berry, 2 others, Alain Fischer, Adrian J. Thrasher, Anne Galy, and Marina Cavazzana. Outcomes following gene therapy in patients with severe wiskott-aldrich syndrome. *JAMA*, 313(15):1550–1563, 2015.

[7] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[8] E. Sherman, C. Nobles, C. C. Berry, E. Six, Y. Wu, A. Dryga, N. Malani, F. Male, S. Reddy, A. Bailey, K. Bittinger, J. K. Everett, L. Caccavelli, M. J. Drake, P. Bates, S. Hacein-Bey-Abina, M. Cavazzana, and F. D. Bushman. INSPIIRED: A Pipeline for Quantitative Analysis of Sites of New DNA Integration in Cellular Genomes. *Mol Ther Methods Clin Dev*, 4:39–49, Mar 2017.