Identification of endogenous retrovirus integration sites using a mixture model

Le Bao
Department of Statistics
The Pennsylvania State University

Joint work with David R. Hunter and Mary Poss

Latent Variables 2016 Conference
Department of Statistics
The University of South Carolina

Outline

Endogenous retroviruses variation

Understanding how genome sequences vary among individuals and populations is important because genetic differences can confer phenotypic differences.

We focus on endogenous retroviruses (ERVs) which are derived from infectious retroviruses.

Animals that share an ERV are related because ERVs are inherited along family lineages like any host gene.

Our task: correctly determine the presence/absence of ERVs, referred as "viruses or integration site'' thereafter.

Read count data

Cervid mule deer is a species that lacks a reference genome. Tissues of deer were collected at hunter check stations by state officials in Oregon, Montana and Wyoming.

The next generation sequencing and De Novo clustering are used to analyze the DNA fragments.

The resulting data are an $m \times n$ matrix X, where the (i, j) element X_{ij} gives the count of DNA sequences of host virus i observed from tissues of animal j.

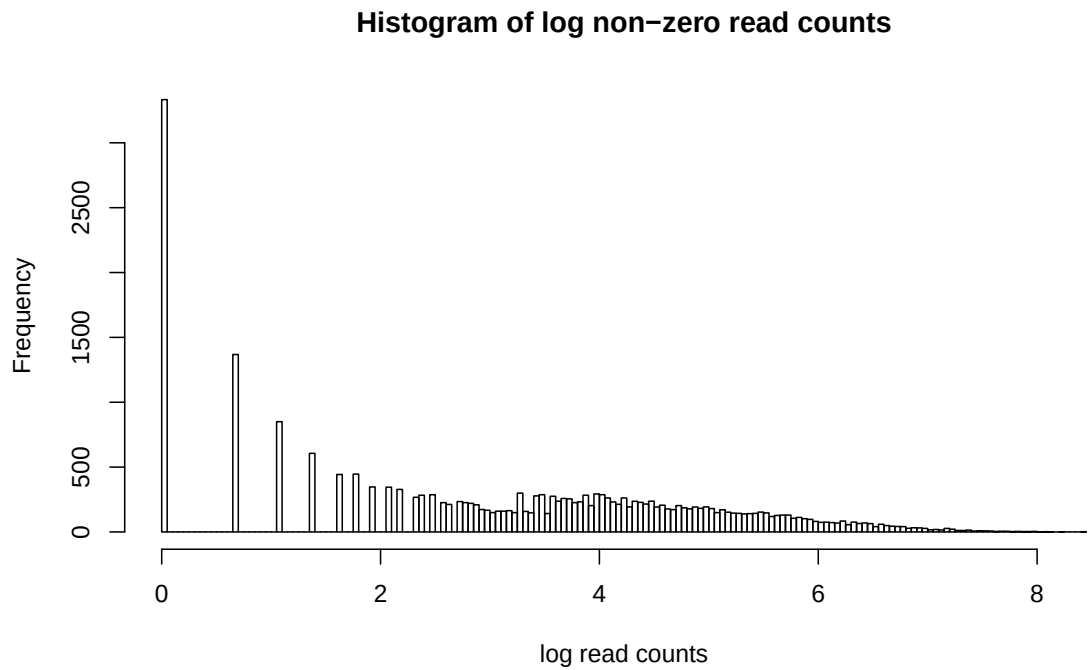
In our example, $m = 2722$ and $n = 77$ are the total numbers of viruses and animals, respectively.

Data summary

Of all the read counts, 82.6% are zero and another 6.3% are between one and ten, and 11.1% are greater than 10.

The mean of all non-zero counts is 98.6.

Large variations are observed across animals and across viruses.



Challenges

Read counts may contain both false positives and false negatives:

- A small number of reads may be attributed to an animal not carrying the virus due to either measurement errors in the high-throughput methods or mis-assignment in the clustering process;
- It is possible to get zero read for an animal actually carrying a particular virus when there are insufficient sequences.

One approach is to set a threshold, and assume that a virus is carried by an animal whenever the corresponding read count is above the threshold.

Challenges

Read counts may contain both false positives and false negatives:

- A small number of reads may be attributed to an animal not carrying the virus due to either measurement errors in the high-throughput methods or mis-assignment in the clustering process;
- It is possible to get zero read for an animal actually carrying a particular virus when there are insufficient sequences.

One approach is to set a threshold, and assume that a virus is carried by an animal whenever the corresponding read count is above the threshold.

- Uncertainty is not addressed when the count data is transformed to the binary virus status data;
- The virus status implied by small count is sensitive to the threshold, but the choice of threshold value is often arbitrary.
- It fails to account for differences in total read number per animal and per integration site.

Outline

Mixture model – two classes

"True Positive" case $\sim$ Negative Binomial(α_i, r_j):

$$f_{ij}(x; r, \alpha) \stackrel{\text{def}}{=} \binom{x + r_j - 1}{x} \alpha_i^{r_j} (1 - \alpha_i)^x, \quad x = 0, 1, 2, \ldots. \tag{1}$$

The mean and variance are $r_j(1 - \alpha_i)/\alpha_i$ and $r_j(1 - \alpha_i)/\alpha_i^2$, respectively.

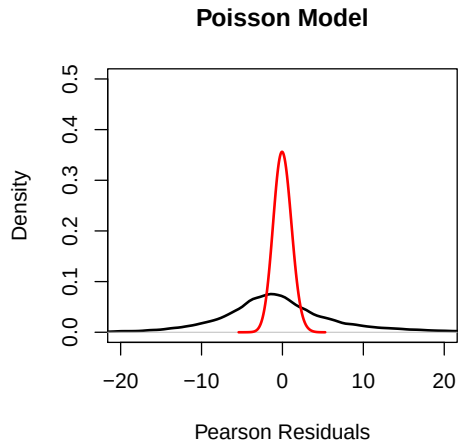
"True Negative" case $\sim$ Negative Binomial$(p_{k(j)}, r_j)$:

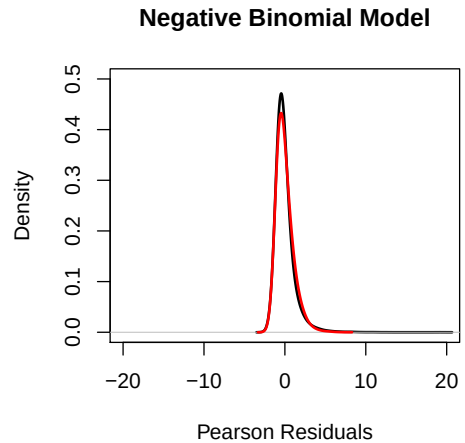
$$g_{ij}(x; r, p) \stackrel{\text{def}}{=} \binom{x + r_j - 1}{x} p_{k(j)}^{r_j} (1 - p_{k(j)})^x, \quad x = 0, 1, 2, \ldots \tag{2}$$

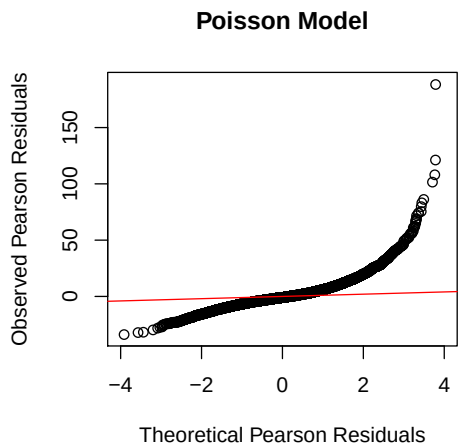
where $k(j)$ is the experiment in which animal j was sequenced.

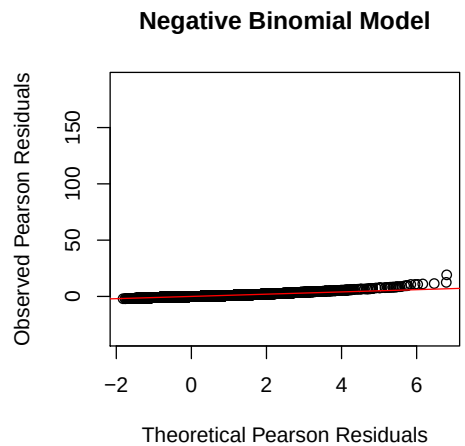
Mixture Model – residual diagnosis

The Negative Binomial distribution is motivated by the strong evidence of overdispersion when the Poisson distribution is used — that is, evidence that the variance of these counts is much larger than the mean.









Mixture Model – animal-specific parameter r_j

"True Positive" case $\sim$ Negative Binomial(α_i, r_j),
mean $= r_j(1-\alpha_i)/\alpha_i$.

"True Negative" case $\sim$ Negative Binomial$(p_{k(j)}, r_j)$,
mean $= r_j(1-p_{k(j)})/p_{k(j)}$.

Both negative binomial distributions can be interpreted as a sum of r_j independent geometric distributions.

The quality and quantity of animal sample varies, which is reflected by r_j. The TP counts and the TF counts are affected in the same way.

Mixture Model – α_i and $p_{k(j)}$

"True Positive" case $\sim$ Negative Binomial(α_i, r_j),
mean $= r_j(1-\alpha_i)/\alpha_i$.

$1-\alpha_i$ approximates the enrichment of virus i.

"True Negative" case $\sim$ Negative Binomial$(p_{k(j)}, r_j)$,
mean $= r_j(1-p_{k(j)})/p_{k(j)}$.

Counts may be considered to be "background noise" and therefore likely to depend on the particular experiment but not the virus in question.

We allow the true negative count distribution to depend on $p_{k(j)}$, where $k(j)$ denotes the experiment number of animal j.

Mixture Model – α_i and $p_{k(j)}$

"True Positive" case $\sim$ Negative Binomial(α_i, r_j),
mean $= r_j(1-\alpha_i)/\alpha_i$.

$1-\alpha_i$ approximates the enrichment of virus i.

"True Negative" case $\sim$ Negative Binomial$(p_{k(j)}, r_j)$,
mean $= r_j(1-p_{k(j)})/p_{k(j)}$.

Counts may be considered to be "background noise" and therefore likely to depend on the particular experiment but not the virus in question.

We allow the true negative count distribution to depend on $p_{k(j)}$, where $k(j)$ denotes the experiment number of animal j.

The estimates of the α_i parameters are all less than 0.503, and the estimates of the $p_k(j)$ parameters are 0.979, 0.963, and 0.981.

It implies that $E(X_{ij}|j \text{ contains } i) > E(X_{ij}|j \text{ does not contain } i)$ for all i and j, even though we do not enforce this constraint in the model.

Mixture Model – mixing probabilities π's

Let π_{ij} represents the a priori probability that animal j carries virus i. The full likelihood of our mixture model becomes

$$L(\pi, r, \alpha, p) = \prod_i \prod_j [\pi_{ij} f_{ij}(x_{ij}; r, \alpha) + (1 - \pi_{ij}) g_{ij}(x_{ij}; r, p)], \qquad (3)$$

Occasionally, samples from the same animal could be run in different experiments. In such cases, we revise the likelihood:

$$\prod_i [\pi_{ij} \prod_{j' \in S_j} f_{ij'}(x_{ij'}; r, \alpha) + (1 - \pi_{ij}) \prod_{j' \in S_j} g_{ij'}(x_{ij'}; r, p)], \qquad (4)$$

where $S_j = \{j' : j \text{ and } j' \text{ are the same animal}\}$.

Mixture Model – parameter estimation and model selection

Estimation of the model parameters was accomplished using maximum likelihood via a straightforward Expectation-Conditional Maximization (ECM) algorithm (Meng and Robin, 1993).

The estimation could be done within 10 minutes for each model we considered: (1) $\pi_{ij} = \pi$ for all i and j; (2) $\pi_{ij} = \pi_i$ for all j; and (3) $\pi_{ij} = \pi_j$ for all i.

Model (2) was chosen based on Bayesian Information Criterion (BIC).

Outline

Model evaluation via replicated animals

We treat replicated animals as independent samples, and calculate the proportion of consistent virus statuses across replicates.

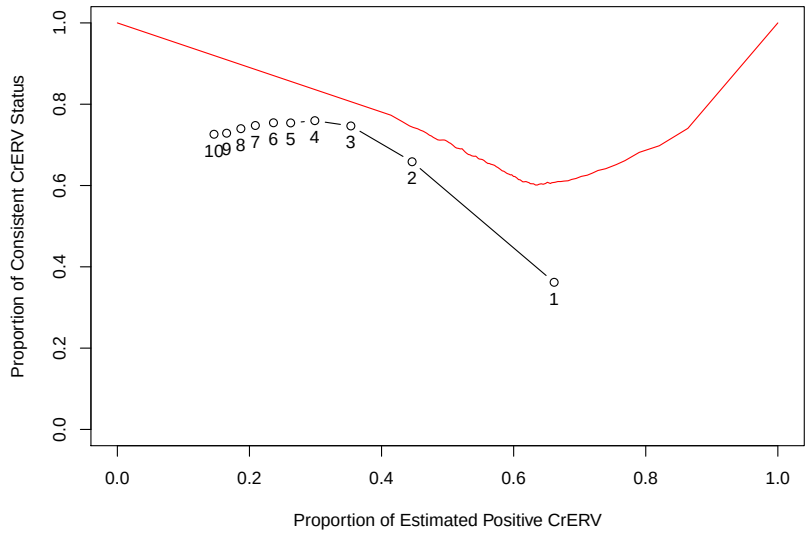


Figure: The proportion of consistent Y_{ij} across replicates v.s. the proportion of $Y_{ij} = 1$: black line uses various threshold values on read counts, red line uses various threshold values on estimated posterior probabilities in the mixture model.

Model evaluation via polymerase chain reaction (PCR)

Directly visualizing the DNA fragment amplified at a particular integration site using PCR.

We have obtained the true status of 6 insertion sites in 32 unique animals. Some of these animals occur in more than one batch, so we have a total of 45 samples to consider, comprising a total of $6 \times 45 = 270$ probability assignments to the "present" mixture component.

47 of the 270 probability assignments correspond to truly present integration sites, whereas the remaining 223 correspond to absent sites.

Model evaluation via polymerase chain reaction (PCR)

Directly visualizing the DNA fragment amplified at a particular integration site using PCR.

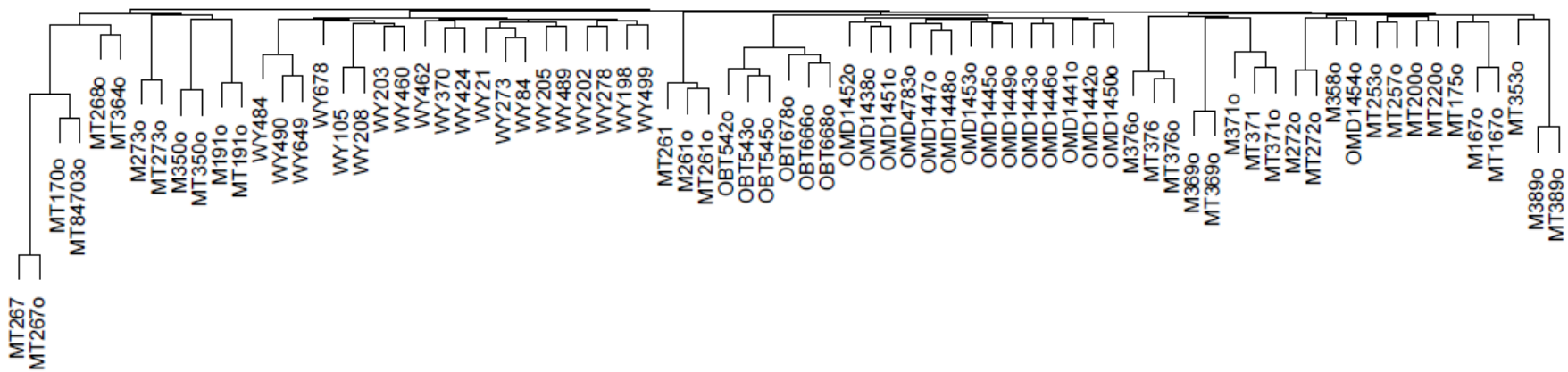
We have obtained the true status of 6 insertion sites in 32 unique animals. Some of these animals occur in more than one batch, so we have a total of 45 samples to consider, comprising a total of $6 \times 45 = 270$ probability assignments to the "present" mixture component.

47 of the 270 probability assignments correspond to truly present integration sites, whereas the remaining 223 correspond to absent sites.

Model	AUC	Mean "False Positives"	Mean "False Negatives"
Read counts only (model-free)	0.957	N/A	N/A
NB-NB, independent samples	0.963	0.100	0.050
NB-NB, replicates recognized	0.975	0.092	0.030

Model evaluation via hierarchical clustering

We first measure the similarities between each pair of samples. Starting from each sample being treated as its own cluster, an agglomerative method merges small clusters into bigger ones sequentially.



Animals from Oregon (OR), Montana (MD) and Wyoming (WY) are separated; and replicated animals are clustered together.

Model evaluation via geographic locations

Treating virus statuses of each animal as a point in m-dimensional space, we may perform principal components analysis (PCA) and visualize the first two principal components.

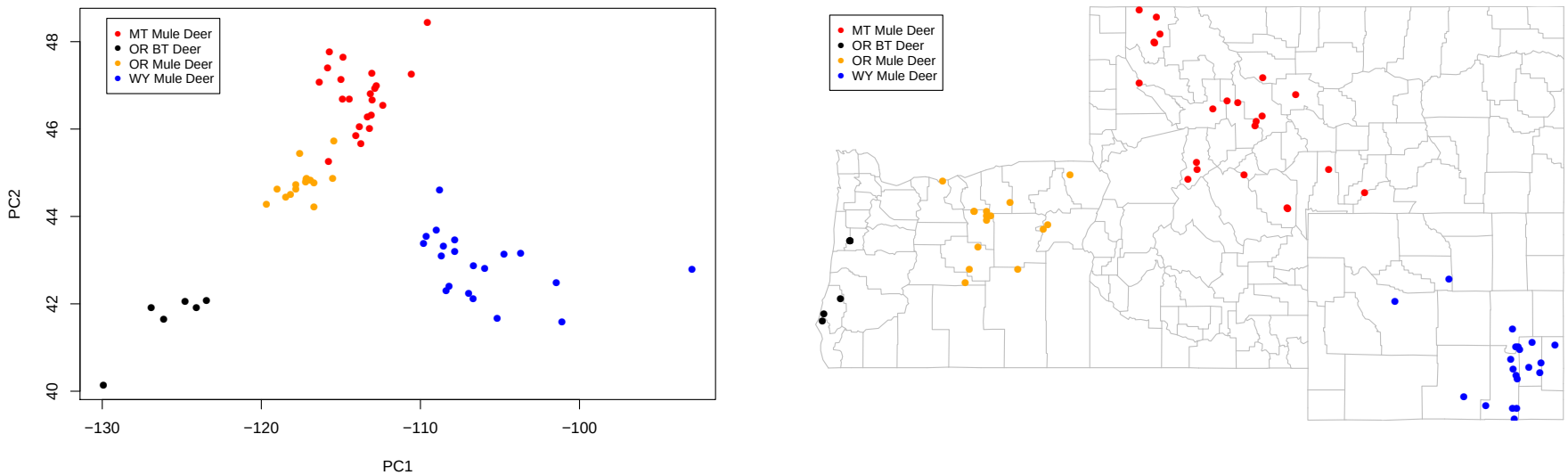


Figure: PC scores and geographic locations

Summary

- Embracing statistical model with scientific belief $\rightarrow$ two-component mixture model.
- Model diagnosis $\rightarrow$ Negative Binomial distributions.
- Model evaluation $\rightarrow$ replicated samples, PCR analysis, geographic locations.

Summary

- The mixture model offers a more flexible and accurate way of identifying the virus statuses from count data.
- It advances the commonly used mixture models for count data such as zero-inflated Poisson.
- It allows seamless integration of data from multiple experiments which is a desirable feature as sequencing technology advances rapidly.

Future Work

We have assumed that the virus status are conditionally independent given the read counts and other parameters.

We plan to introduce the row dependence and column dependence by using the separable covariance model.

Let $W_{ij} = \text{probit} P(Y_{ij} = 1) = \text{probit} \pi_{ij}$,

$$Cov[vec(W)] = \Sigma_2 \bigotimes \Sigma_1, \tag{5}$$

where Σ_1 and Σ_2 represent covariances among rows and columns of W, respectively.

(Dawid, 1981; Hoff, 2011)

Acknowledgement

Thanks all for your coming!

David Hunter
Department of Statistics

Mary Poss
Department of Biology