Mu integration in the Mouse Genome

Charles C. Berry

November 1, 2007

Abstract
Data from integration sites for the Savilahti Bacterial Mu transposon are studied to determine the combined effects of genomic features on site selection.

1 Data Used

The data used were processed by Nirav Malani of Dr. Bushman’s laboratory using a variety of software programs developed by that lab for managing and annotating integration site data.

The data derive from integration sites based on the Savilahti Bacterial Mu transposon. The following numbers of integration sites and genomic matched random controls are used:

| type       | insertion | match |
|------------|-----------|-------|
|            | 214       | 2140  |

For statistical analysis of Mu integration in mouse cells, it was useful to compare integration site distributions to random expectation. For this, matched random controls were generated. A large set of random sites in the mouse genome was drawn computationally. However recovery of Mu integration sites using restriction enzymes introduces a recovery bias favoring sites near suitable restriction enzyme cleavage sites in the mouse genome. This bias is addressed by the use of matched random controls. Each random site generated in silico was annotated for proximity to restriction enzyme recognition sites. For each experimental site of Mu integration, the distance to the restriction site used for recovery was measured, then ten random sites were drawn that were the same distance from a recognition site for the same enzyme. The statistical analysis preserves the pairing between Mu integration sites and matched random controls. This matching procedure “washes out” recovery biases due to placement of restriction enzyme recognition sites, which otherwise can be severe.

The variables to be used are listed below. The variables used describe genomic features that summarize characteristics of the genomic sequence surrounding the integration (or control) site. Several of the variables depend on gene annotations, the sources of these annotations are:
The data on gene activity was from Affymetrix microarrays querying the activity of mouse genes in MEF cells. The data on sites of histone post-translation modification was from PMID: 17603471.

The variables used and brief abbreviations for them are as follows:

- **ref.100k** † RefSeq Genes within ±50 kilobases
- **ref.200k** † RefSeq Genes within ±100 kilobases
- **ref.500k** † RefSeq Genes within ±250 kilobases
- **ref.1M** † RefSeq Genes within ±500 kilobases
- **ref.2M** † RefSeq Genes within ±1 megabase
- **ref.4M** † RefSeq Genes within ±2 megabase
- **ref.8M** † RefSeq Genes within ±4 megabase
- **ens.100k** † Ensembl Genes within ±50 kilobases
- **ens.200k** † Ensembl Genes within ±100 kilobases
- **ens.500k** † Ensembl Genes within ±250 kilobases
- **ens.1M** † Ensembl Genes within ±500 kilobases
- **ens.2M** † Ensembl Genes within ±1 megabase
- **ens.4M** † Ensembl Genes within ±2 megabase
- **ens.8M** † Ensembl Genes within ±4 megabase
- **low.ex.250k** † Affymetrix probesets achieving the 50\(^{th}\) percentile of expression within ±125 kilobases
- **med.ex.250k** † Affymetrix probesets achieving the 75\(^{th}\) percentile of expression within ±125 kilobases
- **high.ex.250k** † Affymetrix probesets achieving the 87.5\(^{th}\) percentile of expression within ±125 kilobases
- **low.ex.2M** † Affymetrix probesets achieving the 50\(^{th}\) percentile of expression within ±1 megabase
- **med.ex.2M** † Affymetrix probesets achieving the 75\(^{th}\) percentile of expression within ±1 megabase
- **high.ex.2M** † Affymetrix probesets achieving the 87.5\(^{th}\) percentile of expression within ±1 megabase
cpg.dens.1k †† Count of CpG sites within ±1 kilobase

**cpg.dens.5k †† Count of CpG sites within ±5 kilobases**

**cpg.dens.10k †† Count of CpG sites within ±10 kilobases**

**cpg.dens.25k †† Count of CpG sites within ±25 kilobases**

**cpg.dens.50k †† Count of CpG sites within ±50 kilobases**

**cpg.dens.250k †† Count of CpG sites within ±250 kilobases**

**ensGene.genes** Whether site is in an Ensembl gene

**refGene.genes** Whether site is in a RefSeq gene

**score.20** The loglikelihood for integration versus control based on a position weight matrix (PWM) derived from the FASTA sequence of the 20 bases flanking the integration site

**gc20** The proportion of G or C bases within ±10 bases

**gc50** The proportion of G or C bases within ±25 bases

**gc100** The proportion of G or C bases within ±50 bases

**gc250** The proportion of G or C bases within ±125 bases

**gc500** The proportion of G or C bases within ±250 bases

**gc1000** The proportion of G or C bases within ±500 bases

**gc2000** The proportion of G or C bases within ±1000 bases

**gc5000** The proportion of G or C bases within ±2500 bases

**gc10000** The proportion of G or C bases within ±5000 bases

**gc25000** The proportion of G or C bases within ±12.5 kilobases

**gc50000** The proportion of G or C bases within ±25 kilobases

**gc100000** The proportion of G or C bases within ±50 kilobases

**gc250000** The proportion of G or C bases within ±125 kilobases

**gc500000** The proportion of G or C bases within ±250 kilobases

**gc1000000** The proportion of G or C bases within ±0.5 megabases

**gc5000000** The proportion of G or C bases within ±2.5 megabases

**gc10000000** The proportion of G or C bases within ±5 megabases

**MEFK9.1k †** MEFK9 count within ±1k bases
MEFK9.10k † MEFK9 count within ± 10k bases
MEFK4.1k † MEFK4 count within ± 1k bases
MEFK4.10k † MEFK4 count within ± 10k bases
MEFK36.1k † MEFK36 count within ± 1k bases
MEFK36.10k † MEFK36 count within ± 10k bases
MEFK27.1k † MEFK27 count within ± 1k bases
MEFK27.10k † MEFK27 count within ± 10k bases

† — Variables marked with the dagger are ranked and rescaled to range from -1 to 1. This has no effect on calculation of ROC curves, but does affect regression coefficients reported below. †† — Variables marked with a double dagger are counts; these are adjusted to expected counts when the window width is narrower than the intended size (e.g. because the window starts near a chromosome end).

2 Software Used

The software used in preparing this report is the R Language and Environment for Statistical Computing [R Development Core Team, 2007]. This document was prepared using the Sweave report generating function [Leisch, 2002]. Several additional R packages were used and will be cited below in context.

3 Association of Integration with Genomic Features

The relationship between integration frequency and genomic annotation can be conveniently quantified using ROC areas as described in PMID: 17166054. Briefly, ROC areas range between 0 and 1. An ROC area of 0.5 indicates no correlation between integration frequency and genomic annotation, either positive or negative. An ROC area >0.5 indicates a positive correlation, an ROC area of <0.5 indicates a negative correlation. Generation of the ROC areas is accomplished by comparison of Mu integration site distributions to the distributions of matched random controls (MRCs).

The following table show the areas under the ROC curves for each of the genomic features considered.

| ROC.areas |    |
|-----------|----|
| gc20      | 0.746 |
| gc50      | 0.758 |
| gc100     | 0.762 |
| gc250     | 0.777 |
| Feature         | Value  |
|-----------------|--------|
| gc500           | 0.763  |
| gc1000          | 0.742  |
| gc2000          | 0.736  |
| gc5000          | 0.719  |
| gc10000         | 0.720  |
| gc25000         | 0.722  |
| gc50000         | 0.725  |
| gc100000        | 0.726  |
| gc250000        | 0.718  |
| gc500000        | 0.706  |
| gc1000000       | 0.710  |
| gc5000000       | 0.671  |
| gc10000000      | 0.634  |
| ref.100k        | 0.669  |
| ref.200k        | 0.695  |
| ref.500k        | 0.687  |
| ref.1M          | 0.677  |
| ref.2M          | 0.659  |
| ref.4M          | 0.653  |
| ref.8M          | 0.620  |
| ens.100k        | 0.654  |
| ens.200k        | 0.687  |
| ens.500k        | 0.683  |
| ens.1M          | 0.678  |
| ens.2M          | 0.654  |
| ens.4M          | 0.645  |
| ens.8M          | 0.620  |
| low.ex.250k     | 0.635  |
| med.ex.250k     | 0.610  |
| high.ex.250k    | 0.549  |
| low.ex.2M       | 0.679  |
| med.ex.2M       | 0.671  |
| high.ex.2M      | 0.624  |
| cpg.dens.1k     | 0.555  |
| cpg.dens.5k     | 0.575  |
| cpg.dens.10k    | 0.614  |
| cpg.dens.25k    | 0.680  |
| cpg.dens.50k    | 0.702  |
| cpg.dens.250k   | 0.718  |
| score.20        | 0.928  |
| ensGene.genes   | 0.570  |
| refGene.genes   | 0.576  |
| MEFK9.1k        | 0.584  |
| MEFK9.10k       | 0.622  |
| MEFK4.1k        | 0.637  |
| MEFK4.10k       | 0.679  |
4 Regression Models for Integration Preference

While there are a number of features that are correlated with integration preference, it is not yet clear how these features work together or whether there are some features that are merely associated with other features that more closely represent processes influencing integration.

To address this, conditional logit regression models are used that incorporate the candidate variables in combination to discriminate between sites and their matched random controls.

Since there are limited numbers of integration sites, using all variables simultaneously in a logistic regression is not likely to be helpful as the estimates obtained will have large variation. To cope with this Bayes Model Averaging as implemented in the BMA package [Raftery et al., 2006] to find posterior probabilities of the regressors, and posterior means and standard errors of the regression coefficients.

The following table shows the BMA posterior probabilities (times 100), BMA regression coefficients, and BMA standard errors in its first three columns. The last two columns show the logistic regression coefficients and their standard errors when each feature is separately regressed on site/MRC status.

|        | Post Prob | Post Mean | Post SD | solo.Coef | solo.SD |
|--------|-----------|-----------|---------|-----------|---------|
| gc20   | 100.0     | -5.729    | 1.508   | 7.528     | 0.654   |
| gc50   | 8.7       | 0.337     | 1.259   | 9.442     | 0.798   |
| gc100  | 0.0       | 0.000     | 0.000   | 11.131    | 0.917   |
| gc250  | 29.4      | 1.586     | 2.720   | 12.455    | 1.003   |
| gc500  | 0.0       | 0.000     | 0.000   | 13.094    | 1.075   |
| gc1000 | 4.7       | 0.284     | 1.458   | 13.612    | 1.179   |
| gc2000 | 50.2      | 6.740     | 7.942   | 14.803    | 1.305   |
| gc5000 | 36.0      | -7.781    | 11.502  | 15.240    | 1.456   |
| gc10000| 0.0       | 0.000     | 0.000   | 16.082    | 1.578   |
| gc25000| 10.0      | 1.438     | 5.064   | 17.195    | 1.666   |
| gc50000| 50.5      | 7.791     | 9.131   | 18.277    | 1.761   |
| gc100000| 0.0     | 0.000     | 0.000   | 18.876    | 1.855   |
| gc250000| 1.3     | 0.028     | 1.305   | 19.160    | 1.953   |
| gc500000| 2.6     | 0.202     | 1.662   | 19.723    | 2.052   |
| gc1000000| 0.0    | 0.000     | 0.000   | 20.606    | 2.168   |
| gc5000000| 0.4    | -0.016    | 0.462   | 19.829    | 2.541   |
| gc10000000| 1.3   | -0.087    | 1.020   | 17.193    | 2.827   |
| ref.100k| 20.4     | 0.124     | 0.272   | 1.228     | 0.140   |
| ref.200k| 52.1     | 0.368     | 0.405   | 1.344     | 0.144   |
Evidently, the local sequence ("score.20"), local GC content ("gc20"), and histone methylation MEFK36 in a 2 kilobase window ("MEFK36.1k") are the dominant effects in determining integration preference. However, there are a number of variables with posterior probabilities higher than 25 percent. Deciding among those variables which should be counted as affecting target preference and which should be dismissed is a challenge that will likely require a larger dataset to settle with certainty.

There is some evidence for effects near CpG islands (‘near’ as in ± 1 kilobase); the posterior probability for it is not nearly high enough to claim an effect. In a regression model with just “score.20”, the coefficient is reduced, but the effect
is still nominally significant. This is also true for it in a regression model with just “cpg.dens.25k”. When both “score.20” and “cpg.dens.25k” are used together, the effect of “cpg.dens.1k” is markedly reduced and no longer near conventional levels of statistical significance as shown here:

|        | coef  | se(coef) | z     | p   |
|--------|-------|----------|-------|-----|
| score.20 | 0.956 | 0.070    | 13.684| 0.00|
| cpg.dens.25k | 0.419 | 0.102    | 4.117 | 0.00|
| cpg.dens.1k  | 0.552 | 0.509    | 1.084 | 0.28|

Also “gc250”, which has posterior probability greater than 25 percent markedly reduces its effect as shown in a regression of “gc250” and “cpg.dens.1k” here:

|        | coef  | se(coef) | z     | p     |
|--------|-------|----------|-------|-------|
| cpg.dens.1k  | 0.677 | 0.379    | 1.787 | 0.074|
| gc250      | 11.655| 1.098    | 10.610| 0.000|

So CpG count within ± 1 kilobase seems unlikely to have a big effect, but it cannot be entirely dismissed. Similarly CpG island count in a wider window (± 25 kilobases) may have an effect.

The effect of “gc20” is interesting. Here are the correlations among the variables for which the posterior probability near 100.0 percent:

|        | score.20  | gc20  | MEFK36.1k |
|--------|-----------|-------|-----------|
| score.20 | 1.000     | 0.570 | 0.112     |
| gc20    | 0.570     | 1.000 | 0.073     |
| MEFK36.1k | 0.112    | 0.073 | 1.000     |

As is evident and expected, “gc20” is correlated with “score.20”. Likely, there is a local sequence effect that the PWM for the 20 flanking bases is not quite capturing. The PWM is calculated without consideration of the effects of any other variables, so performing a calculation that accounts for these other variables might show still stronger effects of the local sequence. However, this would probably require more data than available here.

5 References

1. Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. Compstat, pages 575-580.
2. R Development Core Team (2007). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
3. Raftery, A., Hoeting, J., Volinsky, C., Painter, I., and Yeung, K. Y. (2006). BMA: Bayesian Model Averaging. R package version 3.03.