Association of ENCODE Features on HIV Integration Targetting

Charles C. Berry

September 12, 2006

Abstract

Gary Wang has developed two datasets using 454 sequencing to isolate tens of thousands of HIV integration sites in human cells. Here the relation of a collection of variables annotated by the ENCODE consortium [Consortium, 2004] annotation are studied.

Contents

1 Data Used

1.1 Integration Sites

The data used for this report are a collection of HIV integration sites developed by Gary Wang. The full set of integration sites was used to develop a prediction rule for HIV integration intensity [Berry, 2006]. Here the subset of integration sites that fall into regions annotated by the ENCODE consortiums pilot phase are studied.

Integration sites are recovered via 454 sequencing from a pool of integration sites obtained by infecting a human cell line (Jurkat). Two methods are used to recover the sites. In the **Wang-VSVGgfp-Jurkat-454-Avr** subset, a cocktail of enzymes (AvrII,SpeI,NheI) is used to digest DNA and sequencing is from the HIV U3 end,while in the **Wang-VSVGgfp-Jurkat-454-Mse** subset, the MSEI enzyme is used to digest DNA and sequencing is from the HIV U5 end.

These results are screened to delete duplicate reports of the same integration event.

For each integration event, 10 control sites are sampled from among those genomic sites that are the same distance (in the direction of sequencing) from the nearest restriction site as the integration site.

Some few sites are not assigned chromosomal position as the BLAT alignments match genomic segments whose position is unknown. These are omitted from the analysis.

The available data for analysis of ENCODE annotations broken down by subset and type of site (actual insertion or matched genomic control) is as follows:

```
                                type
  setName                     insertion match
  Wang-VSVGgfp-Jurkat-454-Avr       416  4160
  Wang-VSVGgfp-Jurkat-454-Mse       450  4500
```

1.2 Variables Used

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

mvfit Fitted log relative intensity of integration based on a the analysis in [Berry, 2006]

AffyChIpHl60PvalStrictH3K9K14DHr00 Refer to `http://hgw4.cse.ucsc.edu/`

AffyChIpHl60PvalStrictPol2Hr32 Refer to `http://hgw4.cse.ucsc.edu/`

AffyRnaGm06990Signal Refer to `http://hgw4.cse.ucsc.edu/`

RegulomeBaseCACO2 Refer to `http://hgw4.cse.ucsc.edu/`

RegulomeDnaseGM06990Sens Refer to `http://hgw4.cse.ucsc.edu/`

SangerChipH3ac Refer to `http://hgw4.cse.ucsc.edu/`

SangerChipH3K4me1 Refer to `http://hgw4.cse.ucsc.edu/`

SangerChipH3K4me2 Refer to `http://hgw4.cse.ucsc.edu/`

SangerChipH3K4me3 Refer to `http://hgw4.cse.ucsc.edu/`

SangerChipH4ac Refer to `http://hgw4.cse.ucsc.edu/`

StanfordChipSmoothedJurkatSp3 Refer to `http://hgw4.cse.ucsc.edu/`

StanfordChipSmoothedK562Sp1 Refer to `http://hgw4.cse.ucsc.edu/`

StanfordMethCRL1690 Refer to `http://hgw4.cse.ucsc.edu/`

UcsdChipH3K27me3 Refer to `http://hgw4.cse.ucsc.edu/`

UncFaireSignal Refer to `http://hgw4.cse.ucsc.edu/`

YaleChipPvalFos Refer to `http://hgw4.cse.ucsc.edu/`

2 Statistical Analysis

The table below summarizes the associations of ENCODE variables with HIV integration. The associations of the ENCODE features are assessed by calculating the area under ROC curve for each (column `alone`) and by fitting each in a model in which the prediction given by `mvfit` is also included. A transformation is applied to the ENCODE variables. The transformation takes ranks of the values and scales them to lie in the interval $(-1 \leq 1)$. The fitted models are summarized in two ways: First, the increment in the area under the ROC curve is calculated comparing the model that includes both one ENCODE feature and `mvfit` to the ROC curve based on only `mvfit` (column `improve`); and second, by presenting the regression coefficient (column `coef`), its standard error (column `se(coef)`), and its p-value (column `p`). The fitting is carried out using the `clogit` funciton of the `survival` library [Therneau and Lumley, 2006].

	alone	improve	coef	se(coef)	p
AffyChIpHl60PvalStrictH3K9K14DHr00	0.64	0.00079	0.381	0.13	4.7e-03
AffyChIpHl60PvalStrictPol2Hr32	0.64	0.00079	0.148	0.13	2.6e-01
AffyRnaGm06990Signal	0.66	0.00344	0.164	0.13	2.0e-01
RegulomeBaseCACO2	0.38	0.00000	0.035	0.26	8.9e-01
RegulomeDnaseGM06990Sens	0.58	-0.00034	0.122	0.15	4.2e-01
SangerChipH3ac	0.66	-0.00023	0.214	0.14	1.3e-01
SangerChipH3K4me1	0.65	-0.00066	0.237	0.13	7.5e-02
SangerChipH3K4me2	0.62	0.00000	0.246	0.13	5.4e-02
SangerChipH3K4me3	0.57	-0.00090	0.238	0.12	4.7e-02
SangerChipH4ac	0.73	0.00178	0.759	0.16	1.3e-06
StanfordChipSmoothedJurkatSp3	0.59	0.00266	0.271	0.20	1.8e-01
StanfordChipSmoothedK562Sp1	0.59	-0.00089	0.095	0.23	6.7e-01
StanfordMethCRL1690	0.40	-0.00047	-0.334	0.19	7.1e-02
UcsdChipH3K27me3	0.36	0.00112	-0.213	0.19	2.6e-01
UncFaireSignal	0.60	0.00029	0.125	0.11	2.5e-01
YaleChipPvalFos	0.43	0.00000	0.088	0.16	5.9e-01

References

[Berry, 2006] Berry, C. C. (2006). Some regression models of hiv integration targetting. Supplement to paper of Wang et al.

[Consortium, 2004] Consortium, E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640.

[Therneau and Lumley, 2006] Therneau, T. and Lumley, T. (2006). *survival: Survival analysis, including penalised likelihood. S original by Terry Therneau and ported by Thomas Lumley.* R package version 2.24.