Some Regression Models of HIV Integration Targetting

Charles C. Berry

September 7, 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 the a collection of variables previously shown to be correlated with integration targetting, LEDGF responsiveness, and density of GC in genomic regions surrounding a potential site is studied.

Contents

1 Data Used

1.1 Integration Sites

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, 3 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 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	19921	59763
Wang-VSVGgfp-Jurkat-454-Mse	20564	61692

1.2 Variables Used

A collection of variables found to be associated with HIV integration in previous studies is included in this analysis along with a collection of variables giving the proportions of G or C bases in windows of various sizes surrounding the site of interest.

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

uni.100k Unigene Genes within ±50 kilobases

uni.200k Unigene Genes within ±100 kilobases

uni.500k Unigene Genes within ±250 kilobases

uni.1M Unigene Genes within ±500 kilobases

uni.2M Unigene Genes within ±1 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.50k Density of CpG sites within ±25 kilobases

cpg.dens.250k Density of CpG sites within ±125 kilobases

dnaseI.100k Density of DNAse I sites within ±50 kilobases

dnaseI.1M Density of DNAse I sites within ±500 kilobases

ensGene.genes Whether site is in an Ensembl gene

refGene.genes Whether site is in a RefSeq gene

LEDGF Whether a LEDGF response was measured.

site summed over each position in twenty bases of flanking sequence (10 upstream and 10 downstream). In order to avoid overfitting and resubstitution bias in estimates of association with integration, the score based on the sum of all twenty bases was computed using leave-one-out cross-validation

score20bp The loglikelihood for integration versus control

score50bp The loglikelihood as just described, but for 50 bases

score100bp The loglikelihood as just described, but for 100 bases

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

gc50 The proportion of G or C bases within ±25 bases not counting those within ±10 bases

gc100 The proportion of G or C bases within ±50 bases not counting those within ±25 bases

gc250 The proportion of G or C bases within ±125 bases not counting those within ±50 bases

gc500 The proportion of G or C bases within ±250 bases not counting those within ±125 bases

gc1000 The proportion of G or C bases within ±500 bases not counting those within ±250 bases

gc2000 The proportion of G or C bases within ±1000 bases not counting those within ±500 bases

gc5000 The proportion of G or C bases within ±2500 bases not counting those within ±1000 bases

gc10000 The proportion of G or C bases within ±5000 bases not counting those within±2500 bases

gc25000 The proportion of G or C bases within ±12.5 kilobases not counting those within ±5000

gc50000 The proportion of G or C bases within ±25 kilobases not counting those within ±12.5 kilobases

gc100000 The proportion of G or C bases within ±50 kilobases not counting those within ±25 kilobases

gc250000 The proportion of G or C bases within ±125 kilobases not counting those within ±50 kilobases

gc500000 The proportion of G or C bases within ±250 kilobases not counting those within ±125 kilobases

gc1000000 The proportion of G or C bases within ±0.5 megabases not counting those within ±250 kilobases

gc5000000 The proportion of G or C bases within ±2.5 megabases not counting those within ±0.5 megabases

gc10000000 The proportion of G or C bases within ±5 megabases not counting those within ±2.5 megabases

A transformation is applied to all variables other than the GC proportions, the loglikelihood scores, and the gene indicators. The transformation takes ranks of the values in each data set and then scales them to lie in the interval $(-1, 1)$.

2 Predictive Accuracy of Individual Variables

The predictive value of each variable is assessed using the area under the ROC curve.

The following table gives the areas under the ROC curve and its standard error.

	Avr.area	Avr.stderr	Mse.area	Mse.stderr
ace.100k	0.784	0.002	0.799	0.002
ace.200k	0.778	0.002	0.795	0.002
ace.500k	0.767	0.002	0.784	0.002
ace.1M	0.756	0.002	0.772	0.002
ace.2M	0.738	0.002	0.756	0.002
uni.100k	0.757	0.002	0.770	0.002

uni.200k	0.759	0.002	0.773	0.002
uni.500k	0.749	0.002	0.767	0.002
uni.1M	0.741	0.002	0.759	0.002
uni.2M	0.728	0.002	0.749	0.002
low.ex.250k	0.780	0.002	0.792	0.002
med.ex.250k	0.770	0.002	0.779	0.002
high.ex.250k	0.725	0.002	0.731	0.002
low.ex.2M	0.741	0.002	0.757	0.002
med.ex.2M	0.738	0.002	0.755	0.002
high.ex.2M	0.726	0.002	0.742	0.002
cpg.dens.50k	0.678	0.002	0.696	0.002
cpg.dens.250k	0.729	0.002	0.750	0.002
dnaseI.100k	0.744	0.002	0.762	0.002
dnaseI.1M	0.750	0.002	0.769	0.002
LEDGF	0.511	0.001	0.512	0.001
ensGene.genes	0.702	0.002	0.705	0.002
refGene.genes	0.693	0.002	0.697	0.002
score20bp	0.825	0.002	0.820	0.002
score50bp	0.838	0.002	0.836	0.002
score100bp	0.837	0.002	0.839	0.002
gc20	0.499	0.003	0.530	0.002
gc50	0.457	0.003	0.491	0.003
gc100	0.466	0.003	0.508	0.003
gc250	0.471	0.003	0.515	0.003
gc500	0.484	0.003	0.525	0.003
gc1000	0.494	0.003	0.530	0.003
gc2000	0.507	0.003	0.535	0.003
gc5000	0.539	0.003	0.566	0.003
gc10000	0.564	0.003	0.595	0.003
gc25000	0.596	0.003	0.625	0.003
gc50000	0.620	0.003	0.649	0.003
gc100000	0.653	0.003	0.682	0.002
gc250000	0.697	0.002	0.722	0.002
gc500000	0.709	0.002	0.734	0.002
gc1000000	0.709	0.002	0.731	0.002
gc5000000	0.685	0.003	0.708	0.002
gc10000000	0.653	0.003	0.668	0.003

As is evident, the local sequence variables have the strongest associations (in the sense of the largest departures of the ROC curve areas from 0.50), followed by variables that reflect the density of expressed genes or just genes, then the density of DNAse I sites, then CpG density and GC proportion at wider window widths or location in a gene, followed by the remaining variables.

Since these variables tend to be correlated with others in the set, it is helpful to determine whether some of these associations are merely the result of *confounding*, i.e. that some variables appear to be associated with integration

merely because they are correlated with variables that bear a more proximal relation to the integration process. To determine whether this may be the case, we use conditional logit regression as implemented in the the R `survival` library [Therneau and Lumley, 2006].

An examination of the GC proportions will appear first followed by consideration of the effects of the other variables.

3 Regression of GC proportions

Below is a table of results from conditional logit regression using the GC proportions. All variables are included in a single regression model for each dataset. A penalized loglikelihood approach ala ridge regression was used to protect against unstable results due to the large number of correlated variables being used.

Preliminary 10-fold cross-validation showed the highest log-likelihoods were obtained when a penalty was chosen that resulted in approximately 17 effective degrees of freedom — and that value was used, but any choice approaching the number of regressors seemed to work well. Likewise, the cross-validated ROC curve areas based on the fitted log-odds for these datasets were highest for effective degrees of freedom approaching the number of regressors. Likely, the large number of observations accounts for the stability of these results with little or no penalization required.

	Avr.coef	Avr.stderr	Avr.p.value	Mse.coef	Mse.stderr	Mse.p.value
gc20	0.854	0.086	2.0e-23	0.752	0.085	8.5e-19
gc50	-1.557	0.102	9.6e-53	-1.221	0.101	1.6e-33
gc100	-0.907	0.126	5.1e-13	-0.529	0.124	2.2e-05
gc250	-1.350	0.166	4.5e-16	-0.508	0.162	1.8e-03
gc500	-1.371	0.190	5.5e-13	-0.882	0.186	2.2e-06
gc1000	-2.742	0.217	1.7e-36	-2.200	0.211	1.7e-25
gc2000	-3.983	0.255	3.4e-55	-4.318	0.248	5.5e-68
gc5000	-3.817	0.319	4.4e-33	-4.165	0.311	6.2e-41
gc10000	-2.436	0.340	7.9e-13	-2.636	0.338	6.8e-15
gc25000	-2.535	0.444	1.2e-08	-2.936	0.443	3.4e-11
gc50000	-0.888	0.502	7.7e-02	-0.911	0.498	6.7e-02
gc100000	2.035	0.536	1.5e-04	2.924	0.535	4.6e-08
gc250000	12.876	0.579	1.4e-109	11.035	0.574	1.9e-82
gc500000	8.905	0.578	1.6e-53	10.660	0.581	4.0e-75
gc1000000	9.157	0.559	2.0e-60	7.917	0.550	6.4e-47
gc5000000	3.559	0.601	3.1e-09	4.231	0.592	8.6e-13
gc10000000	-2.554	0.480	1.0e-07	-1.992	0.475	2.8e-05

The regression coefficients are partial derivatives of the log-intensity of integration with respect to the GC proportion variables. Thus, each coefficient measures the effect of the corresponding variable when all other variables are held constant. Since the GC proportions are highly correlated, the pattern of

associations seen between them and integration targetting is not expected to be mirrored in the regression coefficients.

For both datasets, the same pattern of coefficients is seen. At the narrowest width (20 bases), higher GC proportion favors integration, but for somewhat wider widths (50 – 100000 bases) higher AT proportion favors integration, then at still higher widths (250 kilobases – 5 megabases) higher GC proportion favors integration, and at the highest width (10 megabases) hihger AT proportion favors integration.

All in all, how predictive are the GC proportions? The areas under the ROC curves based on the fitted log-odds for these datasets are 0.788 (AVR) and 0.782 (MSE). Comparing this to the ROC curve areas above for single variables, the 'combined GC' values are competitive with the gene density variables.

4 Regression of Other Variables

Here are the results for regressing integration siting on the other variables. Since the loglikelihood scores pertain to overlapping regions, the score for the 20bp region is subtracted from thatfor the 50bp region, and that for the 50bp region is subtracted from that for the 100bp region.

The variables other than the loglikelihood scores were penalized ala ridge regression. Preliminary 10-fold cross-validation showed the highest log-likelihoods were obtained when a penalty was chosen that resulted in approximately 20 effective degrees of freedom — and that value was used, but any choice approaching the number of regressors seemed to work well. Likewise, the cross-validated ROC curve areas based on the fitted log-odds for these datasets were highest for effective degrees of freedom approaching the number of regressors. Likely, the large number of observations accounts for the stability of these results with little or no penalization required.

	Avr.coef	Avr.stderr	Avr.p.value	Mse.coef	Mse.stderr	Mse.p.value
score20bp	0.9631	0.012	0.0e+00	0.927	0.012	0.0e+00
score50bp - score20bp	0.9367	0.027	5.9e-270	1.005	0.028	5.9e-281
score100bp - score50bp	0.7989	0.034	6.8e-125	0.608	0.038	3.5e-56
ace.100k	0.6966	0.075	9.0e-21	0.675	0.070	9.9e-22
ace.200k	-0.3756	0.097	1.1e-04	-0.233	0.093	1.2e-02
ace.500k	-0.0043	0.104	9.7e-01	-0.085	0.099	3.9e-01
ace.1M	-0.0247	0.110	8.2e-01	0.051	0.105	6.3e-01
ace.2M	-0.0996	0.110	3.6e-01	-0.187	0.104	7.1e-02
uni.100k	0.3992	0.064	3.8e-10	0.293	0.060	1.1e-06
uni.200k	-0.1610	0.082	5.0e-02	-0.113	0.078	1.5e-01
uni.500k	0.0033	0.092	9.7e-01	-0.038	0.088	6.7e-01
uni.1M	-0.1219	0.102	2.3e-01	-0.130	0.098	1.8e-01
uni.2M	-0.0340	0.093	7.1e-01	0.045	0.089	6.1e-01
low.ex.250k	0.4970	0.064	8.3e-15	0.583	0.061	1.7e-21
med.ex.250k	0.4920	0.065	3.4e-14	0.429	0.062	5.2e-12

high.ex.250k	0.1838	0.046	7.0e-05	0.159	0.045	4.5e-04
low.ex.2M	0.3256	0.094	5.3e-04	0.213	0.088	1.6e-02
med.ex.2M	-0.0176	0.097	8.6e-01	-0.018	0.091	8.5e-01
high.ex.2M	0.0169	0.064	7.9e-01	0.158	0.062	1.2e-02
cpg.dens.50k	-0.1128	0.039	4.2e-03	-0.209	0.038	3.8e-08
cpg.dens.250k	-0.1168	0.053	2.7e-02	-0.052	0.050	3.0e-01
dnaseI.100k	0.3558	0.041	2.2e-18	0.454	0.039	7.3e-31
dnaseI.1M	0.5598	0.060	9.5e-21	0.357	0.057	3.5e-10
LEDGF	0.4134	0.096	1.7e-05	0.448	0.090	6.9e-07
ensGene.genes	0.5893	0.051	2.0e-31	0.568	0.049	2.6e-31
refGene.genes	0.3715	0.049	3.8e-14	0.410	0.047	5.2e-18

5 All Regressor Variables

Here is the table of results when all regressor variables are used together. Again, a penalized loglikelihood was used with the shrinkage parameters set at the same values as in the regression above.

	Avr.coef	Avr.stderr	Avr.p.value	Mse.coef	Mse.stderr	Mse.p.value
score20bp	0.9622	0.013	0.0e+00	0.9393	0.013	0.0e+00
score50bp - score20bp	0.9679	0.033	9.1e-190	0.9234	0.030	1.6e-204
score100bp - score50bp	0.7677	0.044	1.8e-67	0.7482	0.042	9.6e-71
ace.100k	0.5710	0.066	6.1e-18	0.5449	0.064	1.2e-17
ace.200k	-0.2192	0.081	6.7e-03	-0.1347	0.078	8.5e-02
ace.500k	0.0514	0.085	5.4e-01	-0.0513	0.081	5.3e-01
ace.1M	0.0021	0.087	9.8e-01	0.0560	0.084	5.0e-01
ace.2M	-0.3325	0.089	1.8e-04	-0.3496	0.085	3.9e-05
uni.100k	0.3540	0.058	1.4e-09	0.2705	0.056	1.5e-06
uni.200k	-0.1009	0.072	1.6e-01	-0.0478	0.069	4.9e-01
uni.500k	0.0678	0.077	3.8e-01	0.0514	0.075	4.9e-01
uni.1M	-0.0331	0.083	6.9e-01	-0.0019	0.080	9.8e-01
uni.2M	-0.2938	0.078	1.7e-04	-0.2441	0.075	1.2e-03
low.ex.250k	0.4782	0.061	4.2e-15	0.5289	0.059	2.7e-19
med.ex.250k	0.4663	0.062	4.8e-14	0.4014	0.060	2.0e-11
high.ex.250k	0.2044	0.046	1.0e-05	0.2206	0.046	1.5e-06
low.ex.2M	0.1353	0.080	9.0e-02	0.0813	0.077	2.9e-01
med.ex.2M	0.1197	0.081	1.4e-01	0.1096	0.078	1.6e-01
high.ex.2M	0.1130	0.060	5.8e-02	0.1913	0.058	9.6e-04
cpg.dens.50k	0.0888	0.043	3.7e-02	0.0194	0.041	6.4e-01
cpg.dens.250k	-0.4583	0.056	2.9e-16	-0.3084	0.054	9.1e-09
dnaseI.100k	0.5394	0.042	1.7e-38	0.6280	0.040	1.1e-54
dnaseI.1M	0.1703	0.061	5.0e-03	0.0940	0.058	1.1e-01
LEDGF	0.4197	0.101	3.5e-05	0.4793	0.097	7.2e-07
ensGene.genes	0.4931	0.050	1.4e-22	0.4248	0.049	4.5e-18
refGene.genes	0.3057	0.049	5.1e-10	0.3585	0.048	7.8e-14

```
gc20          0.8142    0.132    7.2e-10  -0.7832    0.131    2.4e-09
gc50          1.0856    0.175    5.5e-10  -0.5758    0.156    2.2e-04
gc100         1.1280    0.221    3.5e-07  -0.2906    0.190    1.3e-01
gc250        -1.0933    0.257    2.1e-05  -0.6600    0.247    7.6e-03
gc500        -1.0678    0.293    2.7e-04  -0.4015    0.286    1.6e-01
gc1000       -3.0843    0.337    5.4e-20  -2.8267    0.324    3.0e-18
gc2000       -3.7401    0.397    4.1e-21  -3.8597    0.380    3.4e-24
gc5000       -4.2306    0.506    5.9e-17  -4.5470    0.489    1.4e-20
gc10000      -1.6375    0.548    2.8e-03  -1.2920    0.542    1.7e-02
gc25000      -2.3686    0.752    1.6e-03  -2.2701    0.737    2.1e-03
gc50000      -0.1027    0.897    9.1e-01  -0.4486    0.878    6.1e-01
gc100000     -2.1607    1.002    3.1e-02  -0.3436    0.989    7.3e-01
gc250000      8.2597    1.093    4.2e-14   4.4049    1.079    4.5e-05
gc500000      3.5151    1.130    1.9e-03   6.2932    1.118    1.8e-08
gc1000000     5.6719    1.061    9.1e-08   2.0019    1.047    5.6e-02
gc5000000     7.8248    1.083    4.9e-13   8.3770    1.093    1.8e-14
gc10000000   -0.4855    0.822    5.5e-01   1.7229    0.825    3.7e-02
```

How much improvment does the additional of the GC proportion variables make to the fit? This can be answered in several ways. First, consider the increase in the loglikelihood reflected by the analysis of deviance table (deviance $= -2\times$ log-likelihood). The deviance for a (saturated) model that fits perfectly is zero, so one can speak of proportional reductions in the deviance. Here is the analysis of deviance table for the **AVR** data.

```
Analysis of Deviance Table

Model 1: Only variables besides GC proportions
Model 2: All Variables
  Resid. Df Resid. Dev    Df Deviance P(>|Chi|)
1     79658    19286.9
2     79641    17257.5    17   2029.5       0.0
```

As is evident there is a 10.5 percent reduction in the deviance due to the addition of the GC proporitons.

Here is the analysis of deviance for the **MSE** data.

```
Analysis of Deviance Table

Model 1: Only variables besides GC proportions
Model 2: All Variables
  Resid. Df Resid. Dev    Df Deviance P(>|Chi|)
1     82230    20505.8
2     82213    18090.5    17   2415.2       0.0
```

As is evident there is a 11.8 percent reduction in the deviance due to the addition of the GC proporitons.

Likewise, one can consider the improvement in the area under the ROC curve.

Here are the ROC curve areas for the **AVR** data.

	AVR ROC areas
Only variables besides GC proportions	0.920
All Variables	0.929
Improvement	0.009
Percent of Maximum Improvement	11.130

Here are the ROC curve areas for the **MSE** data.

	MSE ROC areas
Only variables besides GC proportions	0.910
All Variables	0.922
Improvement	0.011
Percent of Maximum Improvement	12.451

So, by either metric — reduction in deviance or increase in area under the ROC curve — a modest improvement is due to GC proportions.

6 AVR vs MSE

The results for **AVR** and **MSE** seem a bit different. Here is a table that compares the regression coefficients for the model that uses all of the variables:

	Avr.coef	Mse.coef	AVR-MSE	std.err	t-stat	p-value
score20bp	0.9622	0.9393	0.023	0.018	1.271	2.0e-01
score50bp - score20bp	0.9679	0.9234	0.045	0.045	0.996	3.2e-01
score100bp - score50bp	0.7677	0.7482	0.019	0.061	0.319	7.5e-01
ace.100k	0.5710	0.5449	0.026	0.092	0.284	7.8e-01
ace.200k	-0.2192	-0.1347	-0.085	0.113	-0.751	4.5e-01
ace.500k	0.0514	-0.0513	0.103	0.117	0.878	3.8e-01
ace.1M	0.0021	0.0560	-0.054	0.121	-0.445	6.6e-01
ace.2M	-0.3325	-0.3496	0.017	0.123	0.139	8.9e-01
uni.100k	0.3540	0.2705	0.084	0.081	1.031	3.0e-01
uni.200k	-0.1009	-0.0478	-0.053	0.099	-0.535	5.9e-01
uni.500k	0.0678	0.0514	0.016	0.107	0.153	8.8e-01
uni.1M	-0.0331	-0.0019	-0.031	0.115	-0.271	7.9e-01
uni.2M	-0.2938	-0.2441	-0.050	0.109	-0.457	6.5e-01
low.ex.250k	0.4782	0.5289	-0.051	0.085	-0.598	5.5e-01
med.ex.250k	0.4663	0.4014	0.065	0.086	0.754	4.5e-01
high.ex.250k	0.2044	0.2206	-0.016	0.065	-0.249	8.0e-01
low.ex.2M	0.1353	0.0813	0.054	0.111	0.488	6.3e-01
med.ex.2M	0.1197	0.1096	0.010	0.112	0.090	9.3e-01
high.ex.2M	0.1130	0.1913	-0.078	0.083	-0.942	3.5e-01

cpg.dens.50k	0.0888	0.0194	0.069	0.059	1.167	2.4e-01
cpg.dens.250k	-0.4583	-0.3084	-0.150	0.078	-1.933	5.3e-02
dnaseI.100k	0.5394	0.6280	-0.089	0.058	-1.530	1.3e-01
dnaseI.1M	0.1703	0.0940	0.076	0.084	0.906	3.6e-01
LEDGF	0.4197	0.4793	-0.060	0.140	-0.426	6.7e-01
ensGene.genes	0.4931	0.4248	0.068	0.070	0.971	3.3e-01
refGene.genes	0.3057	0.3585	-0.053	0.069	-0.769	4.4e-01
gc20	0.8142	-0.7832	1.597	0.186	8.579	9.6e-18
gc50	1.0856	-0.5758	1.661	0.234	7.094	1.3e-12
gc100	1.1280	-0.2906	1.419	0.292	4.863	1.2e-06
gc250	-1.0933	-0.6600	-0.433	0.357	-1.215	2.2e-01
gc500	-1.0678	-0.4015	-0.666	0.409	-1.627	1.0e-01
gc1000	-3.0843	-2.8267	-0.258	0.468	-0.551	5.8e-01
gc2000	-3.7401	-3.8597	0.120	0.550	0.218	8.3e-01
gc5000	-4.2306	-4.5470	0.316	0.703	0.450	6.5e-01
gc10000	-1.6375	-1.2920	-0.345	0.771	-0.448	6.5e-01
gc25000	-2.3686	-2.2701	-0.099	1.053	-0.094	9.3e-01
gc50000	-0.1027	-0.4486	0.346	1.255	0.276	7.8e-01
gc100000	-2.1607	-0.3436	-1.817	1.408	-1.291	2.0e-01
gc250000	8.2597	4.4049	3.855	1.536	2.509	1.2e-02
gc500000	3.5151	6.2932	-2.778	1.590	-1.748	8.1e-02
gc1000000	5.6719	2.0019	3.670	1.491	2.462	1.4e-02
gc5000000	7.8248	8.3770	-0.552	1.538	-0.359	7.2e-01
gc10000000	-0.4855	1.7229	-2.208	1.165	-1.896	5.8e-02

For the most part the coefficients agree well, but `gc20, gc50` and `gc100` show striking differences.

7 Combined vs Separate Regressions

To make it easier to compare the regression results using all variables in a single regression in 5 to those in using all but the GC variables in 4 and just the GC variables in 3, here is a table in which the coefficients in 4 and 3 (labelled with the suffix `sepGC`) are subtracted from those in 5 (labelled with the suffix `all` while the differences are labelled with the suffix `Diff`).

	Avr.all	Mse.all	Avr.sepGC	Mse.sepGC	Avr.Diff	Mse.Diff
score20bp	0.962	0.939	0.963	0.927	-0.001	0.013
score50bp - score20bp	0.968	0.923	0.937	1.005	0.031	-0.082
score100bp - score50bp	0.768	0.748	0.799	0.608	-0.031	0.140
ace.100k	0.571	0.545	0.697	0.675	-0.126	-0.130
ace.200k	-0.219	-0.135	-0.376	-0.233	0.156	0.099
ace.500k	0.051	-0.051	-0.004	-0.085	0.056	0.034
ace.1M	0.002	0.056	-0.025	0.051	0.027	0.006
ace.2M	-0.333	-0.350	-0.100	-0.187	-0.233	-0.162
uni.100k	0.354	0.270	0.399	0.293	-0.045	-0.023

uni.200k	-0.101	-0.048	-0.161	-0.113	0.060	0.066
uni.500k	0.068	0.051	0.003	-0.038	0.065	0.089
uni.1M	-0.033	-0.002	-0.122	-0.130	0.089	0.128
uni.2M	-0.294	-0.244	-0.034	0.045	-0.260	-0.290
low.ex.250k	0.478	0.529	0.497	0.583	-0.019	-0.054
med.ex.250k	0.466	0.401	0.492	0.429	-0.026	-0.027
high.ex.250k	0.204	0.221	0.184	0.159	0.021	0.061
low.ex.2M	0.135	0.081	0.326	0.213	-0.190	-0.131
med.ex.2M	0.120	0.110	-0.018	-0.018	0.137	0.127
high.ex.2M	0.113	0.191	0.017	0.158	0.096	0.034
cpg.dens.50k	0.089	0.019	-0.113	-0.209	0.202	0.228
cpg.dens.250k	-0.458	-0.308	-0.117	-0.052	-0.342	-0.256
dnaseI.100k	0.539	0.628	0.356	0.454	0.184	0.174
dnaseI.1M	0.170	0.094	0.560	0.357	-0.389	-0.263
LEDGF	0.420	0.479	0.413	0.448	0.006	0.031
ensGene.genes	0.493	0.425	0.589	0.568	-0.096	-0.143
refGene.genes	0.306	0.358	0.372	0.410	-0.066	-0.052
gc20	0.814	-0.783	0.854	0.752	-0.040	-1.535
gc50	1.086	-0.576	-1.557	-1.221	2.642	0.645
gc100	1.128	-0.291	-0.907	-0.529	2.035	0.238
gc250	-1.093	-0.660	-1.350	-0.508	0.257	-0.152
gc500	-1.068	-0.401	-1.371	-0.882	0.304	0.481
gc1000	-3.084	-2.827	-2.742	-2.200	-0.342	-0.627
gc2000	-3.740	-3.860	-3.983	-4.318	0.243	0.458
gc5000	-4.231	-4.547	-3.817	-4.165	-0.413	-0.382
gc10000	-1.637	-1.292	-2.436	-2.636	0.799	1.344
gc25000	-2.369	-2.270	-2.535	-2.936	0.166	0.666
gc50000	-0.103	-0.449	-0.888	-0.911	0.785	0.462
gc100000	-2.161	-0.344	2.035	2.924	-4.196	-3.267
gc250000	8.260	4.405	12.876	11.035	-4.617	-6.630
gc500000	3.515	6.293	8.905	10.660	-5.390	-4.366
gc1000000	5.672	2.002	9.157	7.917	-3.486	-5.916
gc5000000	7.825	8.377	3.559	4.231	4.266	4.146
gc10000000	-0.486	1.723	-2.554	-1.992	2.069	3.714

Many of the GC coefficients are greatly reduced in magnitude when the other variables are included in the regression, which suggests that the effects thatt hey represent are partially accounted for by other variables. Also, the signs of the coefficients for `gc50` and `gc100` change for the **AVR** data, the signs for `gc20` change for both datasets (but in opposite directions), and the point at which a long run of negative effects changes to positive effects shifts from between `gc50000` and `gc100000` to between `gc100000` and `gc250000`.

The effects of `ace.2M`, `uni.2M`, `low.ex.2M`, `dnaseI.1M` are greatly reduced and `reGene.genes` and `ensGene.genes` are somewhat reduced when the GC variables are included, but the effects of `dnaseI.100k` become stronger.

References

[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.