Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 29

Validation in Genomics: CpG Island Methylation Revisited

Mark R. Segal*

*University of California, San Francisco, mark@biostat.ucsf.edu

Validation in Genomics: CpG Island Methylation Revisited*

Mark R. Segal

Abstract

In a recent article in PLoS Genetics, Bock et al., (2006) undertake an extensive computational epigenetics analysis of the ability of DNA sequence-derived features, capturing attributes such as tetramer frequencies, repeats and predicted structure, to predict the methylation status of CpG islands. Their suite of analyses appears highly rigorous with regard to accompanying validation procedures, employing stringent Bonferroni corrections, stratified cross-validation, and follow-up experimental verification. Here, however, we showcase concerns with the validation steps, in part ascribable to the genome scale of the investigation, that serve as a cautionary note and indicate the heightened need for careful selection of analytic and companion validation methods. A series of new analyses of the same CpG island methylation data helps illustrate these issues, not just for this particular study, but also analogous investigations involving high-dimensional predictors with complex between-feature dependencies.

KEYWORDS: multiple testing, cross-validation, local false discovery rate, classification, sequence features

---

*The author wishes to thank Christoph Bock for providing the data. Yuanyuan Xiao and Ru-Fang Yeh provided helpful suggestions.

1 Introduction

It is now widely recognized that the data structures resulting from many contemporary genomics studies, performed utilizing high-throughput technologies, demand novel methods of statistical analysis and attendant validation. This derives, in part, from characteristic attributes of such datasets. They are generally high-dimensional with respect to the number, p, of features (or covariates, variables) examined, and low-dimensional with respect to the number, n, of samples (or units, cases) obtained. Moreover, complex between-feature dependencies are commonplace. To illustrate in the context of microarray gene expression studies, it is typical to have intensity measures on $p > 10^3$ genes (features) but only $n < 10^2$ arrays (samples), with between-gene dependencies arising from pathway and network relationships. Similarly, as will be our focus here, when features are extracted from genome-scale sequence data, we are confronted with such "$p \gg n$" configurations, along with intricate dependencies.

Numerous methodologies have been advanced to analyze such data and to address the companion, often interwoven, task of validation of findings. In particular, the multiple testing issues, engendered by separately associating each feature with some phenotype (outcome, response) of interest, have received considerable attention [2; 3; 4; 5]. These methods confer some level of validation of significant findings in the face of conducting large numbers of hypothesis tests. Conversely, when we employ all features simultaneously, and pursue joint modeling of the phenotype, a variety of custom supervised learning (regression, classification) techniques have been advanced to deal with inherent over-fitting concerns [6; 7; 8; 9]. Verification of the reproducibility of prediction rules found using such methods typically makes recourse to cross-validation [10] or even pre-validation [11]. This is due to the absence of independent test data, or small sample sizes precluding artificial creation thereof although, of course, exceptions exist [12; 13] and may indeed become more common as per assay costs diminish. Appeal to cross-validation is frequently uncritical, perhaps because concerns – primarily pertaining to stability [14; 15] – are not widely appreciated. These concerns are exacerbated in small n settings. Finally, despite the rigors of such data analytic approaches to validation, additional experimental validation is pursued and, in fact, often mandated for genome-scale studies.

In this paper we scrutinize the series of validation steps – both analytic and experimental – conducted in a recent analysis, published in *PLoS Genetics*, on predicting the methylation status of CpG islands [1]. It is important to emphasize from the outset that this study was selected for examination *not*

because validation approaches were either poor or lacking. On the contrary, the authors' diligent use of stringent multiple testing corrections and stratified cross-validation was coupled with experimental verification of findings. Consequently, from a validation perspective, it is safe to assert that this work exceeds much of what appears in the literature. Thus, by showcasing that despite these high standards serious concerns surrounding purportedly validated results exist, we hope to demonstrate that an even higher bar is required when attempting to validate findings in genome-scale investigations.

We briefly provide some background context in order to detail the motivation behind the Bock et al., [1] analyses. In human DNA, the measured frequency of CpG dinucleotides is very low (< 2 %), and duly labeled as *CG suppression*. Such suppression is characteristic of genomes that use Cytosine methylation and may be related to hypermutability of methylated Cytosines. However, there are exceptions: small regions, typically less than 5,000 base pairs (bp), where CpG frequencies equal or exceed expectations, known as *CpG islands* (CpGIs). These stretches are operationalized in terms of (i) GC content, (ii) ratio of observed to expected (under an independence assumption) numbers of CpG dinucleotides, and (iii) length, although the defining values/thresholds employed can vary. Collectively, CpGIs account for $\approx 1\%$ of the human genome. They are primarily located in the 5' region of expressed genes, with more than 60% of known promoters contained therein. Unlike most CpG dinucleotides, those occurring within CpGIs are usually unmethylated. But, when they are methylated, the associated gene (if any) is permanently silenced. This silencing is transmitted through mitosis and thereby constitutes an epigenetic means of inheritance. Numerous exceptions to the methylation-free state of CpGIs have been documented including instances associated with X-chromosome inactivation, imprinting, senescence and cancer [16; 17; 18]. Particularly in view of the latter association investigation of mechanisms leading to the methylation of select CpGIs is of obvious importance. While little is known in this regard, some recent work implicates local DNA sequence in determining methylation of CpGs [19; 20; 21]. It is these findings that provide the impetus for the comprehensive evaluation of the role of local DNA sequence, and attendant predicted DNA structure, in determining CpGI methylation status undertaken by Bock et al. After a series of analyses, outlined and dissected below, they conclude that certain DNA sequence patterns, specific DNA repeats and a particular DNA structure plays a significant role in predisposing CpGIs for methylation. Our revisiting of these analyses indicate that these findings are overstated.

2 Results

The dataset analyzed in [1], where it is described in detail, was generously provided by Christoph Bock. This data, in turn, builds on a prior comprehensive assessment of CpG island methylation on human Chromosome 21 [22]. Bock et al analyze a sizable subset (132/149) of all the CpGIs identified in [22], restricting to those where definitive methylation categories were obtained. These breakdown as 103 unmethylated (UnM) and 29 methylated cases (M). They then compiled an extensive list of features derived from the DNA sequence of each of the CpGIs as well as surrounding sequence windows. The resulting 1184 features fall into eight biological classes, including DNA sequence properties and patterns (428 features), repeat frequency and distribution (494 features) and predicted DNA structure (28 features). It is worth noting that the first category includes (standardized) frequencies of all possible tetramers (both strand- and non-strand specific), while the last category includes not only predicted structural elements such as rise, twist, tilt and solvent accessible surface area, but also up to fourth moments (kurtoses) thereof. We reevaluate the series of analyses as performed by Bock et al., using this common dataset.

2.1 Univariate Assessments Based on CpGI Features

The first set of analyses consists of performing a battery of Wilcoxon rank sum tests on each feature separately in order to elicit which features differ between the two (M and UnM) groups. For this suite of analyses only feature values for the CpGI itself were used. In order to handle the multiple testing issues spawned by examining such a large number of features Bonferroni corrections were employed and a two-sided significance level threshold of $\alpha = 0.01$ was imposed. This approach is seemingly stringent as both a conservative correction procedure and significance level are used. Some 41 features are deemed significant under this approach [1, Table 1], but only a select few are chosen for follow-up interpretation. In particular, *non-strand-specific CACC/GGTG* is highlighted by virtue of being the sole pattern (among the 41 top features) that is over-represented in the methylated CpGIs. Now, prior to employing Bonferroni correction, numerous features are excluded on the basis of being zero for *most* CpGIs. This filtering is undertaken in order "to simplify the statistical analysis". However, there are no difficulties in computing Wilcoxon statistics for such features and/or effecting subsequent multiplicity corrections. Here, the filtering reduces the number of candidate features from 1184 to 706. Had the filtering not been employed *non-strand-specific CACC/GGTG*, which

ranked 38^{th}, would not have survived the chosen multiple corrections procedure. As the number of features retained is, of course, highly sensitive to the manner whereby "mostly zero" is operationalized, this filtering practice can distort simultaneous inference. It seems further misplaced in settings, such as the present situation, where features are generated in a scattershot fashion; that is in a maximally inclusive manner without prescribing any prior hypotheses or feature importance hierarchies. It has been contended that this filtering does not invalidate simultaneous inference since it is done blind to phenotype (here methylation status). But, this is not the case: any feature that is constant across samples (here CpGIs) is necessarily null.

The only other features singled out are the two belonging to the predicted DNA structure category. Only one – *predicted average rise* – is mentioned; the other – *predicted roll skewness* – being more challenging to interpret. However, even interpretation of average rise is deferred until univariate analyses employing not just the CpGI feature values, but also values measured at the surrounding windows, are conducted. This has the impact of making "the role of predicted DNA structure even more pronounced". Now *predicted roll skewness* is no longer significant (ranking 360 out of 833 features tested), but *predicted average rise* ranks second and *predicted average twist*, which previously ranked 126 (out of 706) now ranks third. As these findings attract considerable attention [1, Figure 1] and are used to conclude that "methylated CpG islands tend to co-locate with areas of unusual predicted DNA structure", we carefully revisit the underlying data analyses.

2.2 Univariate Assessments Using Features From Surrounding Sequence Windows

In addition to feature values derived from the DNA sequence of the CpGI itself, values were also computed for 10 surrounding windows straddling -20kb to +20kb. Using this expanded data, univariate feature significance was assessed as follows. For each feature, Z, a quadratic regression model was fitted:

$$Z_j = Meth + Posn_j + Meth * Posn_j + Posn_j^2 + Meth * Posn_j^2 \qquad (1)$$

where Z_j denotes the feature value for the j^{th} sequence window, $Meth$ is an indicator variable for whether the corresponding (central) CpGI is M or UnM, and $Posn_j$ codes for the relative position of the j^{th} sequence window ($j = -5, \ldots, -1, 0, 1, \ldots, 5$). Quadratic regression was used to "capture symmetry around the CpG island", although the motivation for desiring such symmetry

is unclear. We note that the flavor of the results that follow is unchanged whether pure (without the *Posn* terms in (1)) or mixed quadratic regression is used. The overall F test for regression, obtainable from the associated analysis of variance (ANOVA) table, yields p-values that are then subjected to the same Bonferroni correction procedure, again using a strict significance level of $\alpha = 0.01$.

But, despite the use of the same multiplicity correction procedure applied to the same features, albeit measured over surrounding sequence windows, the results obtained are radically different. Firstly, now more than a quarter of the features tested (220/833) are declared significant, even under the stringent procedure employed. Multiplicity control using the generally more liberal false discovery rate (FDR) approach [23] brands more than 40% (339/833) of the tested features as significant, once more at $\alpha = 0.01$. Recall that the features themselves were generated in a catch-all fashion, so that the anticipation would be that the majority would be null. Additionally, the p-values themselves are remarkable, with 100 features attaining values $< 10^{-10}$. The top ranked feature – *standard deviation of total length of self-alignments* – which achieves $p < 10^{-51}$ is not appraised. By way of comparison, when using features based solely on the CpGI sequence, *only* the top ranked feature attains a p-value $< 10^{-10}$ and, even then, just barely (2.62×10^{-11}).

It seems puzzling that the inclusion of surrounding sequence windows would sharpen inference to this extent, especially in light of the modest number of methylated CpGIs. What has changed, in addition to using surrounding windows, is the test statistic employed. It is well known that the F statistics used are notoriously non-robust to departures from underlying assumptions and that they can preferentially select for features with limited variation [24; 25]. Several approaches to counter these shortcomings have been proposed. In the $p \gg n$ context, penalization/moderation schemes have been devised that, in part, strive to use between-feature variance components to shrink test statistics [24; 26; 27]. These approaches are applicable when all features are commensurate, for example, expression measures obtained from microarray platforms. However, this is not the case for the sequence-derived features under consideration here. While we did investigate use of the `eBayes/lmFit` functions contained in the Bioconductor [28] `limma` package [29], effected using t-statistics obtained by decomposing into single degree-of-freedom contrasts [30, p153-4] and stratifying on feature class, the resultant attenuation (using default settings) was insufficient.

Using these contrasts proved informative with regard the role of the predicted DNA structural features. Figure 1 displays a volcano plot [31] for the intercept contrast which was by far the most dominant effect, driving signif-

icance for the (overall) F statistics. Such plots are used to emphasize that variability plays a key role in significance as well as mean differences – here methylation effect. The plot has been truncated for display purposes. What is immediately striking is that the highly significant p-values obtained by the showcased features – approximately 10^{-30} and 10^{-25} for *predicted average rise* and *predicted average twist* respectively – correspond to exceedingly small effects. On this basis, it is misplaced to argue that these features have any mechanistic role in CpG island methylation. Since the claimed role for predicted DNA structure was predicated on the importance of these two features, we are forced to view such claims with skepticism, at least based on the data at hand.

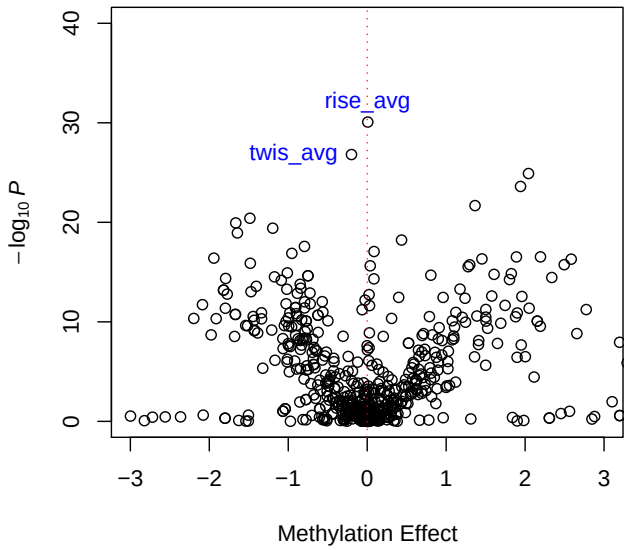

Figure 1: Volcano plot based on t-statistic contrasts highlighting the predicted DNA structural features emphasized by Bock et al., [1].

To address the first (inter-related) issue of excessive significance declarations even under stringent Bonferroni adjustment we showcase an alternate strategy based on utilizing a more appropriate null distribution. However repair of, or alternatives to, the chosen F statistics remains the central concern. In a series of recent papers Efron advances the idea of using an *empiric null* distribution in large-scale (i.e., many features) hypothesis testing situations

[32; 33; 34]. These empiric nulls may be consequentially different from the theoretic null appropriate for individual level (i.e., single feature) testing. The corresponding estimation scheme assumes that the bulk of the features are indeed null which, as argued above, we believe to be the case here. Having obtained estimates of the empiric null and the *mixture distribution* describing all (null and non-null) features, it is straightforward to estimate the non-null distribution along with the corresponding *local false discovery rate*. It is important to note that (i) the distinction between empiric and theoretic null distributions transcends multiple testing concerns, and (ii) permutation approaches serve to refine the theoretic null rather than capturing the (appropriate) empiric null.

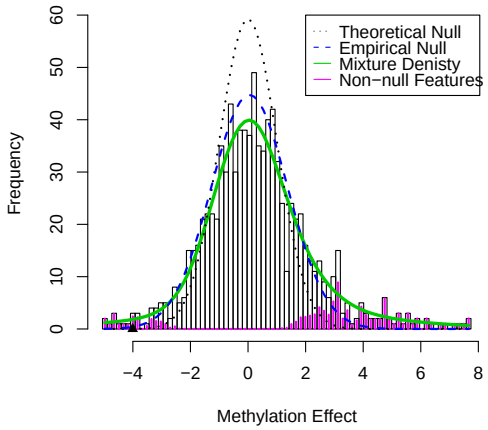

Figure 2: Comparison of theoretic and empirical null distributions. The histogram depicts t-statistic contrast values; see text.

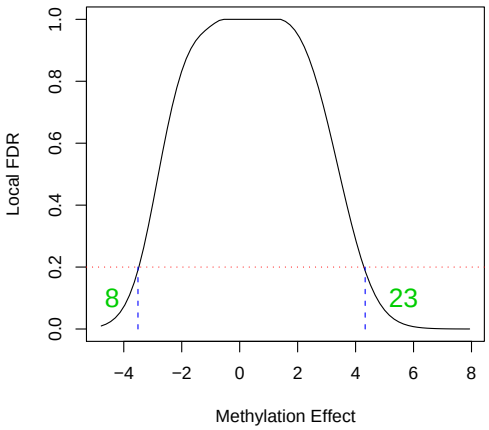

Figure 3: Local false discovery rates. Thresholding at the recommended level of 0.2 yields a total of 31 significant features.

Implementing Efron's procedure using the R (`http://www.r-project.org`) package `locfdr` gives the results depicted in Figures 2 and 3. From Figure 2 it is apparent that the theoretic null density, effectively a standard Gaussian, is appreciably under-dispersed compared to the empirically estimated null density. This is likely attributable to strong between feature correlations which are to be anticipated given the inter-relatedness between many features and the manner in which they are derived. The associated mixture density provides a good fit to the histogram of t-statistic contrasts that capture (intercept) methylation effects. In Figure 3 we have thresholded the resultant local

false discovery rates at the recommended [33] 0.2 level. This yields a total of 8 + 23 = 31 significant features, in stark contrast to the 220 or 339 as obtained by the multiplicity control schemes employed by Bock et al. It is important to reiterate that even among these 31 are features whose significance results from breakdowns of the underlying F statistics.

2.3 Classification Analyses

All previous analyses have been inherently univariate: features are evaluated and ranked according to their individual significance. Bock et al., invoke a number of motivating reasons for going beyond these univariate approaches and using machine learning / classification techniques to develop and assess multi-feature models for predicting methylation status which, as previously, is treated as binary (M or UnM). The forefront classifier employed is a support vector machine (SVM) with linear kernel and default tuning parameters. This classifier is applied separately to each of the eight biological feature categories, as well as various combinations thereof, including the totality of features [1, Table 2]. Once again seemingly stringent methods for evaluating classifier performance are used, namely, stratified cross-validation (CV). Based on the resultant (cross-validated) confusion matrix – the 2×2 cross-categorization of actual versus predicted methylation status – interpretative emphasis is placed largely on the derived "correlation" summary. This summary, which is symmetric in false positives (FPs) and false negatives (FNs), is equivalent to the χ^2_1 statistic for testing independence in the two-way table. We revisit these analyses, focusing on the use of all features, and reveal that the combination of small n (132 CpGIs), class imbalance ($n_M = 29$ methylated CpGIs), and reliance on correlation summaries can conspire to produce misleading results.

The trivial model that classifies all CpGIs as UnM has accuracy $n_{UnM}/n = 103/132 = 0.78$. Baldi et al., [35] brand such models as "highly non-informative and useless." Paradoxically, this characterization reveals some utility for these constant prediction models: as null models / baseline procedures from which prediction gains of more sophisticated models can and should be judged, as we now illustrate. Below we consider a more refined baseline model. In evaluating classifier accuracy Bock et al., use 10-fold CV. Each (withheld) fold constitutes one tenth of the data and so, on average, will contain 10 UnM and 3 M CpGIs. An exact binomial 95% confidence interval around accuracy (success probability) $p = 10/13 \approx 0.78$ has upper confidence limit 0.95. This exceeds the achieved accuracies of all classifiers examined. So, from a classical hypothesis testing standpoint, we would conclude that the use of feature-based

classifiers does not significantly improve on a null (constant prediction) model. Underscoring this result is variability attributable to small n, n_M.

The same idea can be alternatively depicted by scrutinizing the stratified CV accuracies for the SVM. Figure 4 displays 10-fold CV accuracies for each of the 20 strata. The variability of these accuracies is evident. Furthermore, while the SVM average accuracy (86%; green) exceeds the trivial model accuracy (78%; red), the trivial model accuracy (roughly) equals or exceeds the SVM lower quartile for 17 out of 20 strata. Conversely, the abovementioned binomial 95% upper confidence limit (95%; blue) exceeds the SVM upper quartile for 12 out of 20 strata. Accordingly, it is possible to view the predictive content of the collection of sequence-derived features as modest at best – especially since comparisons here are against the trivial ("useless") model.

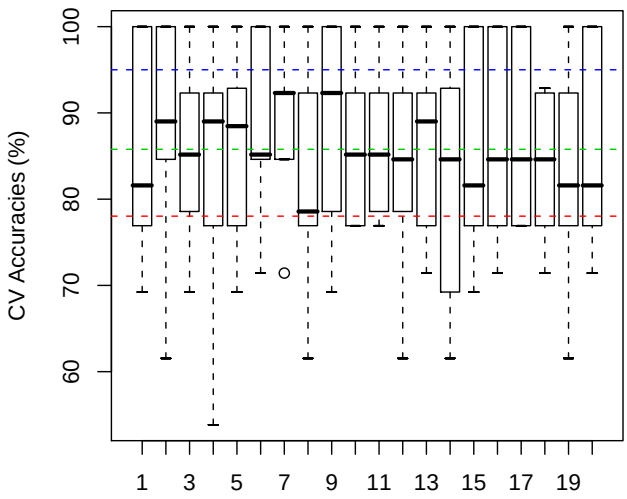

Figure 4: Cross-validated accuracies for the linear SVM. Each of the 20 boxplots corresponds to a 10-fold CV run. The dashed horizontal lines are overall averages as follows: green – SVM; red – trivial (constant) classifier; blue – upper 95% confidence limit for the trivial classifier.

As mentioned, the correlation summary of classifier performance is symmetric in FPs and FNs. This implicitly imparts a (differential) case-weighting in situations as here, where we have class imbalance, $n_M \neq n_{UnM}$. However,

this weighting was not employed in applying the classifier itself. If correlation summaries are to serve as the primary assessment criterion then a corresponding loss function / weighting scheme ought be used in obtaining predictions. Rather than prescribe a specific set of weights we employ receiver operator characteristic (ROC) curves to evaluate prediction performance which (implicitly) span wide-ranging weights. Further, we present results corresponding to use of boosting [36; 37], although similar findings pertain to other classifiers examined including SVMs and random forests [38]. For comparison purposes we use as a baseline classifier one based on only three crude attributes of a CpGI: length, chromosomal position, and category. Here CpGI category corresponds to the four-level factor as defined, and manually assigned, according to the region where the CpGI is located: promoter, intragenic, gene-terminal, and intergenic [1, p0247].

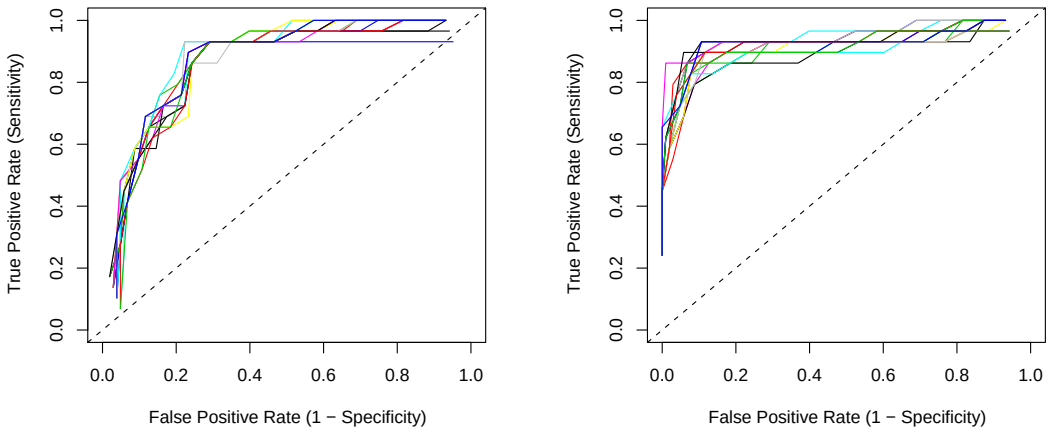

Figure 5: Receiver operator curves (ROCs) of cross-validated boosting-based classification of CpGIs for baseline (left panel) and all feature (right panel) models. The individual traces correspond to differing strata.

The ROC curves for the two scenarios are qualitatively similar, with respective areas under the curve (AUCs) of 0.86 (baseline model) and 0.92 (all features). Now, of course, CpGI category contains sequence-level information as reflected, for example, by the motifs characterizing transcription factor binding sites and exon-intron boundaries. Indeed, Bock et al., indicate that the methylation rates over the four levels are 2.5%, 50%, 83% and 45% respectively. So, it is purposeful to determine if the battery of sequence-derived features provides *additional* predictive information beyond that afforded by

the baseline model and, if so, which features are furnishing this improvement. To address this issue we obtained predictions from the baseline model, and assessed whether residuals therefrom were associated with features using L_1 penalized regression [39; 40], this approach having proved useful for feature selection in other $p \gg n$ settings [41; 42; 9; 43]. Depending on the selection criteria employed, we find contributions from some 10 - 20 features, consistent with the univariate findings and the above difference in AUCs. However, absent from these feature sets are the previously highlighted predicted DNA structural features. So, in summary, in contrast to the classifier comparisons detailed by Bock et al., for which we believe there was an under-appreciation of cross-validation variability and an over-reliance on correlation measures, analyses using ROC-based measures and two levels of rudimentary baseline models, reveal limited additional predictive content for the sequence-derived features.

2.4 Experimental Validation

Experimental validation of predictions confers compelling verification. While such validation is claimed by Bock et al., again further analysis indicates that matters are not so clear-cut. The manner in which Bock et al., effect such experimental validation is as follows. Firstly, using the linear SVM trained on the full ($n = 132$ CpGIs) dataset and all features, predictions of methylation status were obtained for *all* Chromosome 21 CpGIs that were not included in the original dataset. Next, a selection of 8 CpGIs predicted as UnM and 4 predicted as M was made. Then, using bisulphite sequencing, the methylation status of these 12 CpGIs was experimentally determined. The results of this program where that predictions were correct in 10 out of 11 cases, with this finding accorded a p-value < 0.01. One case was not included since it exhibited incomplete methylation. However, we note that if previously used methylation criteria [1, p0250] were employed this case would have resulted in an additional error, consequential given the small sample size.

We now review all steps of this experimental validation. In obtaining the pool of all Chromosome 21 CpGIs a relaxed definition of a CpGI is used. From the 12 selections it is possible to infer that the minimum required length of a CpGI was reduced to 200 bp as opposed to the previously stipulated 400. The motivation for this relaxation is seemingly to expand the pool available for subsequent selection and evaluation. Otherwise, there is little purpose in validating predictions obtained under conditions that differ from those used in model development. But, even under the original CpGI definition (length

$>$ 400bp) there are approximately 90 CpGIs available (as determined using the EMBOSS-3.0.0 tool `newcpgreport`), so choosing 8 predicted UnM and 4 predicted M is readily achievable. Under the relaxed definition there are approximately 845 CpGIs available, which begs the question as to how 12 were selected. As a notable aside, more than 10% of the CpGIs constituting the original ($n = 132$) training data do *not* conform to the stated definition of a CpGI.

Finally, the cited p-value is seemingly based on binomial testing against a null success proportion of $p_0 = 0.5$. But, given the manner in which the sampling of CpGIs was performed, a null value of $p_0 = 8/12 \approx 0.67$ is appropriate. Observing 10/11 successes under this null gives a p-value > 0.1. Similarly, framing the significance assessment in terms of the 2×2 cross-tabulation of true and false, positives and negatives, produces equivocal results.

3 Discussion

The analyses of Bock et al., [1], addressing CpG island methylation and its relationship to sequence derived features, were coupled to a seemingly rigorous validation process. Thus, the single feature, two-sample (M vs UnM) comparisons used Bonferroni correction with a stringent significance threshold to account for multiple testing concerns; the classification analyses (multi-feature discrimination between M and UnM CpGIs) used stratified cross-validation to obtain unbiased estimates of classifier performance; and follow-up experimental validation of model predictions was undertaken. However, in each validation phase, approaches contributing to potentially misleading results were employed, and/or consequential issues were neglected.

For the two sample testing these facets included filtering of low-variation features, choice of highly non-robust test statistics (F from quadratic regression), and the impact of between-feature dependence on the (theoretical) null distribution. Furthermore, results obtained through use of the F statistic are selectively interpreted, with emphasis accorded to predicted DNA structural features while others, that are highly significant yet problematic to understand, are disregarded. It should be acknowledged that such selectivity is common practice. In the classification context, an over-reliance on correlation summaries, with their attendant symmetry with respect to false positives and false negatives in an asymmetric (class imbalanced) setting, and an underappreciation of cross-validation variability in small sample settings, leads to an inflated assessment of the predictive content of the ensemble of sequence-derived features. The experimental validation is based on a small number

of test CpG islands, that are constitutively different from those used in developing models and predictions, with subsequent evaluation of significance of predictions using an inappropriate referent. Compounding these primarily methodological concerns are the facts that more than 10% of the CpGIs in the (training) dataset do not conform with the stated operating definition and the threshold for determining methylation is not consistently applied.

In summary, then, in order to affirm the conclusions of Bock et al., as to the role of certain DNA sequence patterns, specific DNA repeats and a particular DNA structure in methylation of CpGIs, additional data and alternative analysis approaches seem warranted. In highlighting the difficulties encountered by conscientious, yet perhaps rote, approaches to validation in genome-scale settings we hope to alert researchers to the need for heightened care in choice and application of analytic and companion verification methods.

References

[1] Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J. (2006). CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genetics* 2(3):e26.

[2] Dudoit S, Shaffer JP, Boldrick JC. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* 18:71–103.

[3] Storey JD, Tibshirani R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences* 100: 9440–9445.

[4] Storey JD, Taylor JE, Siegmund D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B* 66: 187-205.

[5] Pollard KS, Dudoit S, van der Laan MJ. (2005). Multiple testing procedures: R multtest package and applications to genomics. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit (Editors)). New York, NY: Springer, pp. 251-272.

[6] West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Marks JR, Nevins JR. (2001). Predicting the clinical status of human

breast cancer utilizing gene expression profiles. *Proceedings of the National Academy of Sciences* 98: 11462–11467.

[7] Tibshirani RJ, Hastie TJ, Narasimhan B, Chu G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* 99: 6567–6572.

[8] Zhu J, Hastie T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* 5: 427–444.

[9] Segal MR. (2006). Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. *Biostatistics* 7: 268–285.

[10] Hastie TJ, Tibshirani RJ, Friedman JH. (2001). *The Elements of Statistical Learning*. New York, NY: Springer.

[11] Tibshirani RJ, Efron B. (2002). Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology* 1(1): Article 1, 2002.

[12] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537.

[13] Rosenwald A, Wright G, Chan WC, et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* 346: 1937–1947.

[14] Efron B, Tibshirani RJ (1997) Improvements on cross-validation: The .632+ booststrap method. *J Amer Stat Assoc* 92: 548–560.

[15] Efron B. (2004) The estimation of prediction error: Covariance penalties and cross-validation. *J Amer Stat Assoc* 99: 619–642.

[16] Heard E. (2004). Recent advances in X-chromosome inactivation. *Curr Opin Cell Biol* 16: 247255.

[17] Reik W, Santos F, Dean W. (2003). Mammalian epigenomics: Reprogramming the genome for development and therapy. *Theriogenology* 59: 2132.

[18] Feinberg AP, Tycko B. (2004). The history of cancer epigenetics. *Nat Rev Cancer* 4: 143153.

[19] Bhasin M, Zhang H, Reinherz EL, Reche PA. (2005). Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett* 579: 43024308.

[20] Handa V, Jeltsch A. (2005). Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. *J Mol Biol* 348: 11031112.

[21] Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. (2003). Predicting aberrant CpG island methylation. *Proc Natl Acad Sci* 100: 1225312258.

[22] Yamada Y, Watanabe H, Miura F, Soejima H, Uchiyama M, et al. (2004). A comprehensive analysis of allelic methylation status of CpG islands on human Chromosome 21q. *Genome Res* 14: 247266.

[23] Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Royal Stat Soc. Series B* 57: 289–300.

[24] Tusher VG, Tibshirani RJ, Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 98:5116-21.

[25] Storey JD, Dai JY, Leek JT. (2005). The optimal discovery procedure for large-scale significance sesting, with applications to comparative microarray experiments. `http://www.bepress.com/uwbiostat/paper260`

[26] Smyth GK, Yang YH, Speed TP. (2003). Statistical issues in microarray data analysis. In: *Functional Genomics: Methods and Protocols*, M. J. Brownstein and A. B. Khodursky (eds.), *Methods in Molecular Biology* 224: 111-136. Totowa, NJ: Humana Press.

[27] Wu B. (2005). Differential gene expression detection and sample classification using penalized linear regression models. *Bioinformatics* 22: 472–476.

[28] Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer.

[29] Smyth GK. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3: Article 3.

[30] Venables WN, Ripley BD. (1999). *Modern Applied Statistics with S-PLUS*. New York: Springer.

[31] Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nature Genetics* 29: 389–395.

[32] Efron B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J Amer Stat Assoc* 99: 96–104.

[33] Efron B. (2005). Local false discovery rates. Technical Report, Department of Statistics, Stanford University.

[34] Efron B. (2006). Correlation and large-scale simultaneous significance testing. Technical Report, Department of Statistics, Stanford University.

[35] Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16: 412–424.

[36] Freund Y, Schapire RE. (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*. San Francisco: Morgan Kauffman, 148–156.

[37] Friedman J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29:1189–1232.

[38] Breiman L. (2001). Random forests. *Machine Learning* 45: 5–32.

[39] Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B.* 58: 267–288.

[40] Efron B, Hastie TJ, Johnstone I, Tibshirani RJ. (2004). Least angle regression. *Annals of Statistics* 32: 407–451.

[41] Segal MR, Dahlquist KD, Conklin BR. (2003). Regression approaches for microarray data analysis. *Journal of Computational Biology* 10: 961–980.

[42] Gui J, Li H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* 21: 3001–3008.

[43] Park M-Y, Hastie T, Tibshirani R. (2006). Averaged gene expressions for regression. *Biostatistics*, In press.