Journal of Developmental Origins of Health and Disease (2015), **6**(1), 17–26.

doi:10.1017/S2040174414000476

ORIGINAL ARTICLE

A new, improved and generalizable approach for the analysis of biological data generated by -omic platforms

A. B. Pleasants[1,2], G. C. Wake[2,3], P. R. Shorten[1,2], C. Z. W. Hassell-Sweatman[4], C. A. McLean[4], J. D. Holbrook[5], P. D. Gluckman[2,4,5] and A. M. Sheppard[2,4]*

[1]*Mathematical Biology Department, AgResearch, Hamilton, New Zealand*
[2]*Gravida National Centre for Growth and Development, Auckland, New Zealand*
[3]*Department of Mathematics and Statistics, Massey University, Albany, New Zealand*
[4]*Liggins Institute, University of Auckland, Auckland, New Zealand*
[5]*Singapore Institute for Clinical Sciences, National University of Singapore, Singapore*

The principles embodied by the Developmental Origins of Health and Disease (DOHaD) view of 'life history' trajectory are increasingly underpinned by biological data arising from molecular-based epigenomic and transcriptomic studies. Although a number of 'omic' platforms are now routinely and widely used in biology and medicine, data generation is frequently confounded by a frequency distribution in the measurement error (an inherent feature of the chemistry and physics of the measurement process), which adversely affect the accuracy of estimation and thus, the inference of relationships to other biological measures such as phenotype. Based on empirical derived data, we have previously derived a probability density function to capture such errors and thus improve the confidence of estimation and inference based on such data. Here we use published open source data sets to calculate parameter values relevant to the most widely used epigenomic and transcriptomic technologies Then by using our own data sets, we illustrate the benefits of this approach by specific application, to measurement of DNA methylation in this instance, in cases where levels of methylation at specific genomic sites represents either (1) a response variable or (2) an independent variable. Further, we extend this formulation to consideration of the 'bivariate' case, in which the co-dependency of methylation levels at two distinct genomic sites is tested for biological significance. These tools not only allow greater accuracy of measurement and improved confidence of functional inference, but in the case of epigenomic data at least, also reveal otherwise cryptic information.

Received 11 June 2014; Revised 9 September 2014; Accepted 14 September 2014; First published online 22 October 2014

Key words: epigenomics, methylation, systems biology, transcriptomics

Introduction

The development of various '-omic' platforms has greatly enhanced the quantitative measurement of dynamic biological phenomena. Microarray-based assays of gene expression were the first widely applied -omic platform, but recent advances in deep-sequencing and mass spectrometry technologies now allow for more comprehensive and quantitatively sensitive surveys of gene, protein and epigenetic profiles in cells and tissues. It is generally anticipated that this depth of interrogation will underpin a 'systems' level appreciation of the biological processes. However, as measurement resolution and sensitivity continue to increase greater attention must be directed at the analysis of the data generated if the full benefit of these technical advances is to be realized.

A particular case in point is the accurate measurement of DNA methylation at specific genomic sites. Although it is already recognized that interpretation of such measurements is confounded by the inherent cellular and physiological heterogeneity that exists within complex tissues,[1–3] it is also critical to robustly partition the true biological signals from compounding technical measurement errors arising from the increasingly complex assay processes of sample preparation and physical measurement. In a range of technology platforms (see below), measurement errors display frequency distributions with unusual properties, notably strong 'kurtosis' (highly peaked with fat tails) and a level of 'skewness' arising because empirical methylation measurements are bounded between 0 and 1, which consequentially constrains measurement errors between − 1 and + 1. Thus, if the mean of the methylation measurement is close to zero, the error distribution is positively skewed, while if close to 1 the error distribution is negatively skewed. These features must be considered when deriving an expression for a probability density function, which accurately accounts for the nature of the methylation error distribution, without which accurate estimation and inference (hypothesis testing) is compromised. The commonly observed distributional characteristics can particularly mislead interpretation when the methylation measurement constitutes either the response variable, or the independent variable within a regression analysis.

Our previous analysis of DNA extracted from human umbilical tissue samples using the Sequenom EpiTyper massARRAY platform (www.sequenom.com/), which measures methylation by comparison of the mass of transcription cleavage

*Address for Correspondence: Dr A. M. Sheppard, Liggins Institute, The University of Auckland, Private Bag 92019, Victoria Street West, Auckland 1142, New Zealand.
(E-mail a.sheppard@auckland.ac.nz)

products derived from amplified bisulphite-modified DNA by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry,[4] confirms that error frequency distributions are far from normal. When quantifying levels of methylation, the observed variance in methylation measurements across subjects represents the sum of naturally occurring biological variance and the variability inherent within the measurement procedure itself. The issue of errors in Sequenom measurement protocols has received attention previously[5–7] and have been linked with a range of variables including the type of, and position within the thermal cycler, logistics of robotic handling and chip batch and positional effects. The importance of minimizing such error effects cannot be understated. We have since extended this interrogation to more fully describe the errors implicit in *the process of generating* epigenomic data by this platform.Having optimized the chemistry of the Sequenom assay by multifactorial testing of as many variables as we could practically control, we have made a very large number of replicate measures on the same biological sample to evaluate the deviation in measurement due specifically to sample spotting and the stochastic process MALDI-TOF mass spectrometry detection. We have assumed that the heterogeneity caused solely by the physics of the measurement procedure follows a two-dimensional Poisson process with an exponentially distributed Poisson parameter.[8] In which case, the probability that a particular methylation measurement will contain a given level of absolute deviation will similarly follow an exponential distribution (see Appendix 1). Accounting for the fact that deviations may be positive or negative this gives a Laplace distribution as a suitable description for the deviations.[9] Having empirically estimated the error contribution arising from the assay chemistry and physics of machine measurement, we found that the Laplace distribution required an extension based on Hermite polynomials[10] to properly describe the observed deviations. This extended Laplace distribution allows for more confident inference for Sequenom platform generated data.[11]

In brief, the bounded probability density we derived is defined thus:

$$f(z) = \frac{p^4 e^{-p|z|}\left[1 + qH_3(|z|)\right]}{2\left\{p^3 - 3qp^2 + 6q - e^{-p}\left(3\left(1-2q\right)p^3 + 6pq + 6q\right)\right\}} \quad (1)$$

where p and q are parameters likely to be population dependent, z is the methylation measurement deviation and $H_3(z)$ is the third-order Hermite polynomial equal to $z^3 - 3z$ (note the use of the absolute value signs). The variables z_i in probability density (equation 1) can be the residuals of a linear ($z_i = y_i - (\mu + \beta x_i)$), or nonlinear ($z_i = (y_i - f(x_i\beta))$) regression model, for observations y_i and parameters μ, β. Because of the presence of the absolute value of the residual in the likelihood, differentiation of the log likelihood presents a technical difficulty. However, the parameters may still be found by maximizing the likelihood using non-gradient methods[12] (using the equations summarized in Appendix 2) and for the case of the Sequenom platform were estimated to be $p = 37.21$; $q = 0.0429$. Subsequent analyses assume this form of the distribution for methylation error measurement.

Problems also arise in performing tests of inference and in obtaining estimates of the standard errors of the regression coefficients. Using bootstrap estimates provided exceptionally poor estimates of the standard errors. Bootstrap estimates are known to perform poorly when applied to skewed distributions of the type of the methylation measurement error distribution.[13] The skewness parameter associated with the third-order Hermite polynomial for this probability density (now defined with absolute values) is defined as $6q$,[14] that is, the skewness parameter for the Sequenom measurements is $6 \times 0.0429 = 0.2574$, representing a difference from the Laplace distribution. Notably, a similar Laplace probability density has also been reported to fit the frequency distribution of transcriptomics data generated by a microarray platform[15] suggesting that error distributions across differing -omics platform may have a common character.

By re-examining available measurements in the public data (summarized in Appendix 3), we now demonstrate that the 'usual' methods of statistical analysis are unsuitable for a wide range of -omic platform assays and illustrate that our bounded probability density accurately describes the error distributions in all of the measurement platforms considered. By the way of illustration, alignment of the measurement error frequency distributions between maximum-likelihood estimated normal (dashed) and extended Laplace (solid) for meDIP data is shown in Fig. 1a. Close examination of the error frequency distributions for other omics platforms indicates that they too display features indicative of non-normality. Further by re-examining our own published data, we illustrate that our analysis approach, applied when methylation is treated as a response variable, performs better than the currently preferred approach, which uses the forgiving β regression, but is based on the untested assumption that the measurement distributions are in fact related to a β distribution.[16] Further we illustrate the application of our approach when methylation measurement is the independent variable, in which case data analysis is complicated by an 'errors-in-variables problem'.[17] Finally, using empirical measurements we extend the utility of this function beyond the 'univariate' case. The dependency parameter of the 'bivariate' distribution we present can be standardized and is shown to provide a better estimate of the relationship between methylation measurements at two CpG sites (i) residing within the same gene target; (ii) within two different gene targets; or (iii) whether temporally linked.

Methods

A range of author processed data was downloaded from the NCBI repository GEO (http://www.ncbi.nlm.nih.gov/gds) for four series (designated as GSE22513, GSE29231, GSE38352, GSE40870) chosen to (i) contain technical replicate data; (ii) for human tissue samples; and (iii) interrogated on different but widely used epigenomic (meDIP and Infinium 450 bead array) and transcriptomic (Illumina HT12 and Affymetrix U133_2)

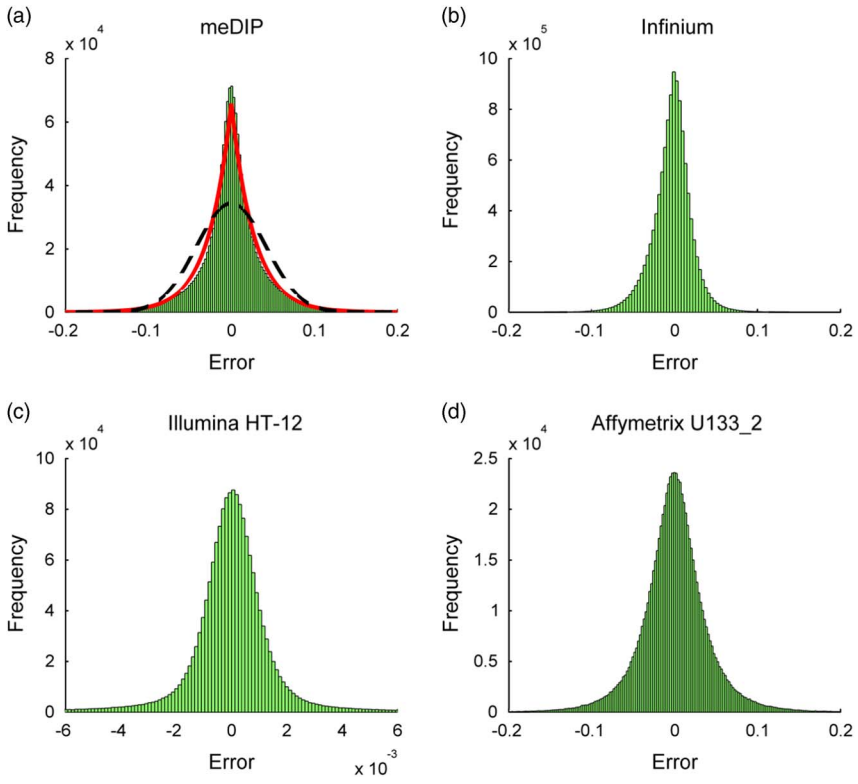

Figure 1. The measurement error frequency distribution for (a) meDIP, (b) Infinium 450, (c) Illumina HT12 and (d) Affymetrix U133_2. Errors were transformed to the interval [− 1,1] by dividing the errors by 10, 1, 10 and 10^4 for meDIP/Agilent array, Infinium 450, Illumina HT12 and Affymetrix U133_2 platforms, respectively. Also shown is the corresponding maximum-likelihood estimated Normal (dashed) and extended Laplace (solid) fit to the data. The measurement error distribution is not consistent with the Normal distribution.

platforms (see Appendix 3). No additional processing was applied to the data before the difference between replicates was calculated for each probe.

Results

The extended Laplace measurement error distribution is generally applicable across -omics platforms

The measurement error frequency distribution was calculated for each data set (Fig. 1) after transformation to the interval [− 1,1] by division of the errors by 10, 1, 10 and 10^4 for the meDIP, Infinium 450, Illumina HT12 and Affymetrix U133_2 platforms, respectively. The Normal distribution does not provide a suitable characterization of the error distribution in any of these cases [based on a Lilliefors test ($P < 0.001$), as illustrated by the dashed curve (Fig. 1a), representing the corresponding maximum-likelihood estimated Normal for the meDIP/Agilent array assay protocol. However, the modified-Laplace probability density (equation 1) was found to fit the measurement error distributions for all four measurement platforms, representing the extended Laplace for the meDIP/Agilent array assay protocol. The estimated parameters that characterize the measurement error distributions for each platform are presented in Table 1.

Table 1. *Parameter values for each of the major -omics platforms*

Platform	*P*	*Q*
meDIP/Agilent array	36.54	0.028
Infinium 450	55.67	0.118
Illumina HT-12	33.11	0.067
Affymetrix U133_2	463	0.033

The estimated parameters that characterize the measurement error distribution was derived by equation 1. Errors were transformed to the interval [− 1,1] by dividing the errors by 10, 1, 10, 10^4 for the meDIP, Infinium 450, Illumina HT12 and Affymetrix U133_2 platforms, respectively.

The extended Laplace regression compares favourably to currently used analytic approaches when methylation represents the response variable (case I)

The difficulties of estimation and inference with CpG methylation measurements have been recognized previously. The most recent literature applies β regressions[16,18,19] to deal with the distributional problems of CpG methylated data and demonstrates better performance than the traditional ordinary least squares approach. A comparison of our own algorithm

https://doi.org/10.1017/S2040174414000476 Published online by Cambridge University Press

Table 2. *Inference for the comparison of methylation levels for CpG sites in the ABCG2 gene a disease sheep reported previously*[21]

CpG site for ABCG2 gene	Significance level for least squares estimation	Significance level for β regression estimation	Significance level for extended Laplace estimation
1	0.02	0.015	0.0001
2	0.01	0.002	0.0006
3	0.01	0.004	0.0006
4	ns	ns	ns
5	ns	ns	0.02
6	ns	ns	ns
7	ns (0.07)	ns (0.09)	ns
8	ns (0.07)	0.04	0.004
9	ns (0.08	ns (0.06)	0.006

(see equation 1) against both ordinary least squares and β regression using the R package Betareg[20] for methylation measurements that we have previously published for the ABCG2 gene[21] is shown in Table 2. In this instance, the methylation status of the individual CpGs in the ABCG2 gene was found to vary as the response variable and align to phenotypic outcome, in sheep following experimental exposure to mycotoxin. Generally, the extended Laplace regression yields higher levels of significance (lower *P*-values) than the other methods, a feature consistent with it being based on a probability density more directly associated with the actual error frequencies, rather than the assumption of Normality or a β distribution for the errors.[12] Note particularly the result with ABCG2 gene CpG7, which is indeterminate under ordinary least squares and β regression analysis, but not seen to be significant under our algorithm. Further, analysis of the ABCG2 gene CpG9 is inconclusive for the comparison between 'clinical' and 'resistant' biological phenotypes with ordinary least squares and β regression, but is found to be highly significant ($P < 0.004$) with our algorithm. Using *in silico* mapping tools we have identified consensus sites for potential transcriptional regulation in the region of CpG9, supporting the notion that this genomic region is indeed significant for the expression of phenotype responses (Zhang *et al.*, unpublished observations).

Application of the extended Laplace when estimation of methylation influences phenotype; the errors-in-variables problem (case II)

It is well known that errors in the independent variables of a regression analysis produce estimators that are biased and inconsistent.[17] A number of general methods have been developed for dealing with errors in the independent variables of a regression, employing simulation where representative samples are drawn from a probability distribution of the errors in the independent variables. Although observed errors in methylation measurements appear to be particularly prone to these problems, here we show that the extended Laplace probability density (equation 1) readily forms a basis for an 'errors-in-variables' or model II regression analysis. Specifically, we have considered the instance in which the degree of methylation at a genomic CpG site early in the life course is thought to have an influence on later phenotypic outcome, by re-analyzing data previously reported that describes a relationship between methylation of the RXRA gene in umbilical cord tissue at birth and the body mass index (BMI) of children at ages 4 and 6 years.[22] We have applied the scoring method described by Carroll *et al.*,[17] to this re-analysis, with random samples drawn from the methylation measurement probability density (equation 1).

The correlations between degrees of methylation of each of the six adjacent CpG sites in this analysis was high, so it would be expected that the relationships with children's BMI would be similar for each CpG site. Compared with the 'traditional' least squares estimates (model I) calculated from the same data set (Table 3), there are two essential points of importance to be recognized. First, the errors-in-variables regression coefficients that incorporate the error probability density (equation 1) are much more consistent with each other. Note, the least squares estimates do not show this relationship (and two estimators (CpG 1 and 3 in the 6-year data show reversed slopes) in keeping with the known lack of consistency of least squares estimators in this situation.[17] Second, the standard error of the errors-in-variables estimators is considerably lower, and the significant levels are better identified with the error distribution based on equation 1.

Extension to a bivariate methylation measurement error distribution

Of particular biological interest is determining the potential relationship between methylation measurements made at different CpG sites, either (a) within a given gene, (b) at CpG sites in different genes, (c) at the same CpG site in a gene but at different time points and (d) at the same CpG site in a gene but in different tissues. The usual statistic for this purpose is to calculate the correlation between variables. However, the high frequency of significance deviations that characterize methylation measurements may be misleading for the usual correlation calculation, either parametric or non-parametric, because these calculations do not allow for the probabilities or expected frequencies of large measurement deviations. Fortunately, there are a number of ways that the univariate methylation measurement probability density (equation 1) could be extended to two dimensions to accommodate these questions. The simplest approach adopted here is to build on conditional densities using the identity

$$P[z, y] = P[z\,|\,y]P[y]$$

where z, y denote methylation levels. A suitable construction must be found for the conditional probability density $P[z|y]$, which is the probability that methylation z will be measured on CpG1 given that methylation y has already been observed on CpG2. There are a variety of ways that this might be done, and

https://doi.org/10.1017/S2040174414000476 Published online by Cambridge University Press

Table 3. *The model II regression of BMI on the level of methylation at each of the measured CpG sites in the RXRA gene*

Age	CpG	Intercept model I	Slope model I	Significance of slope model I	Intercept model II	Slope model II	Significance of slope model II
4 years	1	16.00 ± 0.219	0.116 ± 0.406	ns	15.90 ± 0.063	0.251 ± 0.130	ns
	2	15.91 ± 0.159	0.506 ± 0.433	ns	15.85 ± 0.051	0.551 ± 0.141	$P < 0.01$
	3	15.87 ± 0.234	0.174 ± 0.446	ns	15.87 ± 0.071	0.302 ± 0.151	ns
	4	15.74 ± 0.291	0.491 ± 0.435	ns	15.86 ± 0.093	0.244 ± 0.158	ns
	5	15.88 ± 0.269	0.283 ± 0.423	ns	15.98 ± 0.083	0.056 ± 0.150	ns
	6	15.43 ± 0.940	0.940 ± 0.405	$P < 0.05$	15.58 ± 0.088	0.665 ± 0.134	$P < 0.001$
6 years	1	16.23 ± 0.360	− 0.186 ± 0.651	ns	15.86 ± 0.085	0.356 ± 0.172	$P < 0.05$
	2	16.00 ± 0.263	0.459 ± 0.713	ns	15.78 ± 0.057	0.798 ± 0.187	$P < 0.001$
	3	16.29 ± 0.389	− 0.336 ± 0.735	ns	15.75 ± 0.114	0.440 ± 0.192	$P < 0.05$
	4	15.44 ± 0.459	1.103 ± 0.682	ns	15.62 ± 0.114	0.674 ± 0.195	$P < 0.01$
	5	15.21 ± 0.444	1.535 ± 0.690	$P < 0.05$	15.62 ± 0.115	0.670 ± 0.193	$P < 0.01$
	6	15.05 ± 0.444	1.636 ± 0.635	$P < 0.05$	15.41 ± 0.108	0.935 ± 0.170	$P < 0.0001$

BMI = body mass index.

This analysis removes the outlier. The model I (i.e. standard least squares) regression estimates are included for comparison. Intercepts and slopes are reported ± standard errors.

the usefulness of any construction must be judged in application. The simplest modification adopted here is to assume a linear relationship between the CpG measurements. Thus,

$$P[z,y] = Qp^2 e^{-p(|z|+|y|+\theta zy)}\left[1+q(H_3(|z|)+H_3(|y|))\right] \quad (2)$$

where Q is the normalizing constant and the cross-product in the Hermite polynomials is ignored as being of order q^2. Note that the products of the two methylation measurements in the exponential of probability density (equation 2) are not taken as absolute values. The dependency parameter in the bivariate probability density (equation 2) is θ.

The normalizing constant Q is given by

$$Q^{-1} = \int_{-1}^{1}\int_{-1}^{1} p^2 e^{-p(|z|+|y|+\theta zy)}\left[1+q(H_3(|z|)+H_3(|y|))\right]\mathrm{dzdy} \quad (3)$$

This double integral cannot be evaluated explicitly and must be solved numerically, although this equation (equation 3) may be manipulated so that only a single integral must be performed numerically to calculate Q. These calculations are given in Appendix 4. The log likelihood based on the bivariate probability density (equation 2) can be maximized for the parameter θ using numerical optimization methods.

To interpret the dependency parameter as a measure of the relationship between the methylation of two CpG sites, the value needs to be standardizd, much as the product moment correlation coefficient is a standardized covariance. In this case, the standardization is carried out by estimating the dependency parameter for each set of methylation measurements linked *with itself*, and then dividing the bivariate dependency parameter by the maximum of these two estimates. That is, find the maximum θ estimate from calculations of $P[z, z]$ or $P[y, y]$, then divide the dependency parameter calculated for $P[z, y]$ by this maximum parameter. The necessity of calculating the relationship between the methylations of two CpG sites in this way can be seen in a comparison of multiple methylation measurements made on the same CpG site. Because of the discrepancies induced by the measurement procedure there is a notable frequency of high errors in the multiple measurements. To illustrate this, we present a comparison of methylation values at the RXR gene promoter CpG sites (Table 4), in umbilical cord[22] and subject matched postnatal buccal swab tissue (unpublished data). Note, for example, the product moment correlation of these multiple tissue measurements for CpG1 is 0.03 (low). However, the standardized dependency parameter for this CpG calculated from the bivariate probability density (equation 2) is –0.43 with 95% confidence interval (−0.28 to − 0.64) showing that, as would be expected the multiple measurements are actually relatively strongly related. This improvement in detecting relationships among the methylation status of the CpG sites is because the bivariate probability density (equation 2) takes into account the frequency of large errors expected in duplicate measurements. In this instance, comparison of multiple measurements at the same CpG sites suggests the following biological behaviour through the early life of the organism, which informs a discussion about the comparative impact of foetal programming and postnatal modification on epigenomic profiles. Specifically, at all of the measured CpG sites in RXR estimated methylation is higher in the cord at birth than in postnatal buccal samples. Further, the order of decrease in estimated methylation between subjects was preserved for most CpGs, with the specific exception of CpG4.

Discussion

Statistical procedures based on the Normal probability density are widely used in biology. At the level of the organism, the characteristics of interest are generally affected by the combination of

https://doi.org/10.1017/S2040174414000476 Published online by Cambridge University Press

Table 4. *The scaled dependency coefficient and its statistical significance for cord and buccal methylation measurements in the RXRA gene*

CpG	Standardized dependency parameter θ	Significance of dependency	95% confidence interval for the dependency parameter θ	Product moment correlation
RXRA_int1_2_CpG_1	− 0.43	$P < 0.01$	− 0.28 to − 0.64	0.03
RXRA_int1_2_CpG_4	0.10	ns		0.01
RXRA_int1_2_CpG_5	− 0.93	$P < 0.01$	− 0.91 to − 0.96	0.19
RXRA_int1_2_CpG_7	0.43	$P < 0.05$	0.26 to 0.56	0.03
RXRA_int1_2_CpG_9	− 0.91	$P < 0.01$	− 0.78 to − 0.94	0.12

many processes and the central limit theorem means that the frequency of these characteristics tends towards a Normal distribution, although the rate of this convergence can be slow. However, at the molecular level nonlinear processes may dominate, and become a factor especially if small amounts of biological material are being measured. Under these circumstances of nonlinear relationships, the central limit theorem may not apply and the basic probability distributions of the molecular measurements may not be Normal, or even close to Normal.[12] In these circumstances, basing estimation and inference on statistical procedures that assume the error distribution is Normal may be misleading, as we have sought to illustrate. Notably, important aspects of the biology of these processes may be overlooked, lost within the observation errors and not uncovered because the nature of the errors is not accounted for correctly.

These problems appear to manifest in the measurement of both epigenomics[1–3] and transcriptomics[15] data generated by widely used and diverse technology platforms. The extended Laplace approach based on a detailed empirical knowledge of the error frequency distributions inherent in the measurements represents a first attempt to deal with these issues. Although not necessarily optimal, the current study provides a generic approach for investigating and formulating suitable statistical methods for estimation and inference when the assumption of Normal errors fails.

The suggested steps are shown below:

- Perform a factorial experiment (changing the protocols under the control of the experimenter) to optimize sample preparation and make measurements in a designed way. Analyse these data to partition the sources of experimental variation and use this analysis to define a protocol that minimizes variation in the measurements due to experimental procedure.
- With this optimized assay, perform a number of repeated measures of the same sample to obtain a measurement error frequency distribution. Assess whether this distribution satisfies requirements to be Gaussian or another known frequency distribution. In particular, consider the impact of other factors (e.g. the requirement that the error distribution be bounded) on the chosen representation for this distribution.
- If the error frequency distribution does not satisfy the criteria for known families of probability distributions, construct a suitable representation (e.g. using Edgeworth expansions or Gram–Charlier series[23]).
- When a suitable representation of the error distribution is found derive methods of estimation and inference based on, for example, maximum-likelihood methodology.[9] Test the efficacy of these methods using simulated data to quantify the improvement in estimation and inference. If these tests are satisfactory then analyse the data accordingly.

The most dramatic improvements in the application of the derived methylation measurement probability density have been in the errors-in-variables case, or model II regressions. When the CpG site methylation affects some phenotypic outcome the derived methylation measurement probability density can be used with errors-in-variables methods such as the scoring algorithm to considerably improve both estimation and inference. The clarity brought to hypothesis testing using this methodology is notable. Our data suggests that this issue is not restricted either to epigenetic analysis or to the methodologies of epigenetic analysis but is likely to be a broadly based issue in systems biology.

Acknowledgements

The authors extend their gratitude to Prof. Terry Speed (WEHI Melbourne, Australia) and members of the Developmental Epigenetics Group (Liggins Institute, University of Auckland) for critical reading and valuable feedback on the manuscript.

Financial Suppoort

This work was funded by New Zealand Government contract UOAX0808 and Gravida National Centre for Growth and Development, NZ.

Conflicts of Interest

None.

References

1. Talens RP, Boomsma DI, Tobi EW, *et al*. Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology. *FASEB J*. 2010; 24(9), 3135–3144.
2. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Gen*. 2010; 11, 191–203.
3. Gervin K, Hammero M, Akselsen H, *et al*. Extensive variation and low heritability of DNA methylation identified in a twin study. *Genome Res*. 2011; 21, 1813–1821.

https://doi.org/10.1017/S2040174414000476 Published online by Cambridge University Press

4. Ehrich M, Nelson MR, Stanssens P, *et al*. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proc Nat Acad Sci*. 2005; 102, 15785–15790.
5. Warnecke PM, Stirzaker C, Melki JR, *et al*. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Res*. 2007; 25, 4422–4426.
6. Warnecke PM, Stirzaker C, Song J, *et al*. Identification and resolution of artifacts in bisulfite sequencing. *Methods*. 2002; 27, 101–107.
7. Coolen MW, Statham AL, Gardiner-Garden M, Clark SJ. Genomic profiling of CpG methylation and allelic specificity using quantitative high-throughput mass spectrometry: critical evaluation and improvements. *Nucleic Acids Res*. 2007; 35, e119.
8. Gallant AR, Tauchen G. Semi-nonparametric estimation of conditionally constrained heterogeneous processes: asset pricing applications. *Econometrica*. 1989; 57, 1091–1120.
9. Pawitan Y. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, 2001. OUP: Oxford, 528pp.
10. Buckland ST. Fitting density functions with polynomials. *J App Stats*. 1992; 41, 63–67.
11. Hassell-Sweatman CZW, Wake GC, Pleasants AB, McLean CA, Sheppard AM. Linear models with response functions based on the Laplace distribution: statistical formulae and their application to epigenomics. *ISRN Prob and Stats*. 2014; 2013, 1–22.
12. Freund JE, Walpole RE *Mathematical Statistics*, 3rd edn, 1992. Prentice Hall: New Jersey, 547pp.
13. Porter PS, Rao ST, Ku J-Y, Poirot RL, Dakins M. Small sample properties of non-parametric bootstrap t confidence intervals. *J Air Waste Manag Assoc*. 1997; 47, 1197–1203.
14. Jondeau E, Poon S-H, Rockinger M. *Financial Modelling Under Non–Gaussian Distributions*, 2000. Springer-Verlag: London, 539 pp.
15. Purdom E, Holmes SP. Error distribution for gene expression data. *Stat Appl Genet Mol Biol* 2005; 4, 1–33.
16. Seow WJ, Pesatori AC, Dimont E, *et al*. Urinary benzene biomarkers and DNA methylation in Bulgarian petrochemical workers: study findings and comparison of linear and beta regression models. *PLoS One*. 2012; 7, e50471.
17. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective. Monographs on Statistics and Applied Probability*, 2nd edn. 2006. Chapman and Hall/CRC Press: Florida, 488pp.
18. Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. *J App Stats*. 2004; 31, 799–815.
19. Hebestreit K, Dugas M, Klein HU. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*. 2013; 29, 1647–1653.
20. R Core Team. *R: A Language and Environment for Statistical Computing*, 2013. R Foundation for Statistical Computing: Vienna, Austria, http://www.R-project.org/.
21. Babu K, Zhang J, Moloney S, *et al*. Epigenetic regulation of ABCG2 gene is associated with susceptibility to xenobiotic exposure. *J Proteomics*. 2012; 75, 3410–3418.
22. Godfrey KM, Sheppard A, Gluckman P, *et al*. Epigenetic gene promoter methylation at birth is associated with child's later adiposity. *Diabetes*. 2011; 60, 1528–1534.
23. Kolassa JE. *Series Approximation Methods in Statistics. Lecture Notes in Statistics*, 2006. Springer Science: New York, 218pp.

Appendix 1

The theory summarized here is taken from Hassell Sweatman *et al.*,[11] applicable to our case in which methylation proportion is treated as the response variable and multiple explanatory variables are considered. A linear model is assumed. Specifically, let $y \in R^n$ be a vector of response variables, let $\boldsymbol{X}$ be a real-valued $n \times m$ matrix, where $x_{i1} = 1$, $i = 1, \ldots, n$. In practice, we usually have n, the number of data points, much larger than m the number of coefficients. Let $\beta \in R^m$ be a vector of coefficients for our linear model and assume that $E(\boldsymbol{y}) = \boldsymbol{X}\beta$. The goal is to estimate the components of β by ML principles, and to determine their standard errors, given $\boldsymbol{y}$, $\boldsymbol{X}$ and the response variable distribution (1). Let $\boldsymbol{z} = \boldsymbol{y} - X\beta$ be the error vector. We assume that the errors are independent. The response variable distribution is modelled by (1). Then the log likelihood (disregarding the constant term) is

$$\begin{aligned} l(\beta) &= \ln(f(z(\beta))) \\ &= \sum\nolimits_{i=1}^{n} \left[-p|z_i| + \ln\left[1 + q\left(|z_i|^3 - 3|z_i|\right)\right]\right] \end{aligned} \tag{1a}$$

The inclusion of the modulus (absolute value) function in the perturbed Laplace probability density function (equation 1) is the cause of abrupt changes in the gradient of the log-likelihood function. Although gradient-based methods of ML parameter estimation fail, such estimation may be done by non-gradient methods such as the simplex method or simulated annealing.

The abrupt changes in gradient must be taken into account when calculating the standard errors of the parameters. The fact that $l(\beta)$ is not differentiable in the classical sense at a local maximum means that the assumptions made in the derivation of the usual classical formulae for the information matrix, the expected value of the Hessian of the log-likelihood function and the variance–covariance matrix for the model coefficients β_j, $j = 1, \ldots, m$, are not met. For C^2 probability density functions, these formulae are derived using Taylor series. In section 4 in Hassell Sweatman *et al.*[11] using generalized functions, alternative expressions for these quantities are found, assuming truncated and/or perturbed Laplace response functions which are C^3 where the modulus function is non-zero. In section 5 in Hassell Sweatman *et al.*,[11] these expressions are used to prove the asymptotic convergence of our MLE to a random variable with a normal distribution.

To derive these expressions, generalized functions have been used:

$$\operatorname{sgn}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

and $\delta(x)$, which is the δ function that is zero except at one point $x = 0$ and where $\int_{-\infty}^{\infty} \delta(x)dx = 1$. These expressions and the modulus function are connected by

$$\frac{d}{dx}|x| = \operatorname{sgn}(x)$$
$$\frac{d}{dx}\operatorname{sgn}(x) = 2\delta(x)$$

where the differentiation is taken in a generalized sense.

https://doi.org/10.1017/S2040174414000476 Published online by Cambridge University Press

With respect to the model parameters, let $E(H)$ denote the expected value of the generalized Hessian of the log-likelihood function, let J denote the information matrix and let V denote the variance–covariance matrix. It is proved[11] that

$$E(H) = \varsigma X^T X$$
$$J = \upsilon X^T X$$

and

$$V^{-1} = \frac{\varsigma^2}{\nu} X^T X$$

where $\varsigma = E\left(\frac{\partial^2 l}{\partial z_i^2}\right)$ and $\nu = E\left(\frac{\partial l}{\partial z_i}\right)^2$ (these quantities do not depend on the index i) and we assume that $\boldsymbol{X}$ has full rank m. It is shown that the usual classical relation for smooth log-likelihood functions, namely $V = J^{-1} = [-E(H)]^{-1}$ does not hold, although it is a good approximation for large p and small q. For example, when $q = 0$ it turns out that $\upsilon = p^2$ and $\varsigma = -\frac{p^2}{1-e^{-p}}$. To compare a model with M coefficients to a lesser model with P coefficients, let $\lambda = \frac{L(\beta_1, \ldots \beta_M)}{L(\beta_1, \ldots, \beta_P)}$ be the likelihood ratio. The generalized log-likelihood ratio statistic is $D_{\text{gen}} = -2\frac{\varsigma}{\nu}\ln\lambda$, distributed approximately as $x^2(M-P)$ if the lesser model gives a good description of the data.For example, for the case $m = 2$ corresponding to one explanatory variable, the log likelihood (disregarding the constant term) is

$$l(\beta_1, \beta_2) = \sum\nolimits_{i=1}^{n} \left[-p|z_i| + \ln\left[1 + q\left(|z_i|^3 - 3|z_i|\right)\right]\right] \quad (1a)$$

where setting $x_i = x_{i2}$

$$z_i = y_i - (\beta_1 + \beta_2 x_i)$$

Since the log-likelihood function $l(\beta_1,\beta_2)$ in (1a) involves the modulus of the deviations z_i then $l(\beta_1,\beta_2)$ is not differentiable with respect to its parameters when any $z_i = 0$. The nature of the maximization means that the maximum occurs on a ridge in (β_1, β_2, l) space, defined by some $z_i = 0$. Assuming that not all the x_i are equal (so that $\boldsymbol{X}$ has full rank 2), a maximum will occur at the intersection of two distinct ridges defined by $z_i = z_j = 0$, for some $i \neq j$.

Generalized differentiation yields

$$\frac{\partial l}{\partial z_i} = -p\ \mathrm{sgn}(z_i) + \frac{\left(3q|z_i|^2 - 3q\right).\mathrm{sgn}(z_i)}{1 + q|z_i|^3 - 3q|z_i|}$$

and

$$\frac{\partial^2 l}{\partial z_i^2} = \delta(z_i)\left(-2p - 6q\right) + \frac{(-3q)\left(q|z_i|^4 - 2|z_i| + 3q\right)(\mathrm{sgn}(z_i))^2}{\left(1 + q|z_i|^3 - 3|z_i|\right)^2}.$$

Integrating over the error space $[-1,1]^n$ with respect to (1) yields expected values.

Appendix 2

The log likelihood (disregarding the constant term) is

$$f(\alpha, \beta) = \sum\nolimits_{i=1}^{n} \left[-p|z_i| + \ln\left[1 + q\left(|z_i|^3 - 3|z_i|\right)\right]\right] \quad (2a)$$

where

$$z_i = y_i - (\alpha + \beta x_i)$$

Since the log-likelihood function $f(\alpha, \beta)$ in equation 2a involves the modulus of the deviations z_i then $f(\alpha, \beta)$ is not differentiable with respect to the parameters α and β when $z_i = 0$. The nature of the maximization means that this is very likely to occur for some z_i and we have found that this often occurs in practice. That is, the maximum will occur at the intersection of two ridges in the (α, β, f) space at which $z_i = z_j = 0$. Accordingly, maximization based on derivative-free methods is necessary, for example, simplex optimization or simulated annealing.

Notwithstanding this difficulty, the first and second derivatives of the log-likelihood $f(\alpha, \beta)$ with respect to the parameters α and β, needed for calculation of the standard errors, can be determined using generalized function theory:

$$\frac{\partial f}{\partial \alpha} = p\sum_{i=1}^{n} \mathrm{sgn}(z_i) + \sum_{i=1}^{n} \frac{\left(3q|z_i|^2 - 3q\right).\mathrm{sgn}(z_i)}{1 + q|z_i|^3 - 3q|z_i|}$$

$$\frac{\partial f}{\partial \beta} = p\sum_{i=1}^{n} x_i\ \mathrm{sgn}(z_i) + \sum_{i=1}^{n} \frac{x_i\left(3q|z_i|^2 - 3q\right).\mathrm{sgn}(z_i)}{1 + q|z_i|^3 - 3q|z_i|} \quad (2b)$$

The standard errors are given by the inverse of the variance–covariance matrix:

$$V = -\begin{bmatrix} \frac{\partial^2 f}{\partial \alpha^2} & \frac{\partial^2 f}{\partial \alpha \partial \beta} \\ \frac{\partial^2 f}{\partial \alpha \partial \beta} & \frac{\partial^2 f}{\partial \beta^2} \end{bmatrix}^{-1}$$

$$\frac{\partial^2 f}{\partial \alpha^2} = 3pq\sum_{i=1}^{n} \frac{\left(q|z_i|^4 - 2|z_i| + 3q\right)(\mathrm{sgn}(z_i))^2}{\left(1 + q|z_i|^3 - 3|z_i|\right)^2} - \left(2p + 3q\right)\sum_{i=1}^{n} \delta(z_i)$$

$$\frac{\partial^2 f}{\partial \alpha \partial \beta} = 3pq\sum_{i=1}^{n} \frac{x_i\left(q|z_i|^4 - 2|z_i| + 3q\right)(\mathrm{sgn}(z_i))^2}{\left(1 + q|z_i|^3 - 3|z_i|\right)^2} - \left(2p + 3q\right)\sum_{i=1}^{n} z_i\delta(z_i)$$

$$\frac{\partial^2 f}{\partial \beta^2} = 3pq\sum_{i=1}^{n} \frac{z_i^2\left(q|z_i|^4 - 2|z_i| + 3q\right)(\mathrm{sgn}(z_i))^2}{\left(1 + q|z_i|^3 - 3|z_i|\right)^2} - \left(2p + 3q\right)\sum_{i=1}^{n} z_i^2\delta(z_i) \quad (2c)$$

To derive these formulae generalized functions have been used:

$$\mathrm{sgn}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

https://doi.org/10.1017/S2040174414000476 Published online by Cambridge University Press

and $\delta(x)$ which is the δ function that is zero except at one point $x = 0$ and where $\int_{-\infty}^{\infty} \delta(x)dx = 1$. These expressions and the modulus function are connected by

$$\frac{d}{dx}|x| = \mathrm{sgn}(x)$$
$$\frac{d}{dx}\mathrm{sgn}(x) = 2\delta(x)$$

where the differentiation is taken in a generalized sense.

Since $\mathrm{sgn}(0) = 0$ and $\delta(z_i) = 0$ for $z_i \neq 0$ only one (but not both) of the terms on the RHS of equation (3a) is non-zero for each z_i. Thus, the calculation of the information matrix V for the standard errors has to be made in the generalized sense if one of the z_is is zero, as it generally appears to be in practice. Suppose that this happens at the k^{th} z_i. Then

$$V = -\begin{bmatrix} \frac{\partial^2 f}{\partial \alpha^2} & \frac{\partial^2 f}{\partial \alpha \partial \beta} \\ \frac{\partial^2 f}{\partial \alpha \partial \beta} & \frac{\partial^2 f}{\partial \beta^2} \end{bmatrix}^{-1}$$

The matrix below is a term in V when $z_k = 0$,

$$-(2p+3q)\begin{bmatrix} 1 & z_k \\ z_k & z_k^2 \end{bmatrix}\delta(z_k) = (2p+3q)\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\delta(z_k)$$

where we have used $x\delta(x) = 0$, $x^2\delta(x) = 0$. This enables calculation of the information matrix V.

Appendix 3

A summary of the features of the author processed data downloaded from the NCBI repository GEO (http://www.ncbi.nlm.nih.gov/gds).

GEOs series	Data type	Data platform	Sample type	No. of samples and replicates	Pubmed
GSE22513	Transcriptome	Affymetrix U133_2	Human breast needle tumour biopsy	14 samples, run in duplicate	21633166 20068102 20878462
GSE29231	Transcriptome	Illumina HT-12	Human visceral adipose biopsy	6 samples, run in 4 replicates	23308243
GSE38352	DNA methylation	MeDIP-Chip, Agilent array	Human post-mortem hippocampus	17 samples, 9 run in duplicate, 8 in triplicate	23045659
GSE40870	DNA methylation	Illumina Infinium s450K	Human bone marrow derived, AML cell lines with treatment	24 samples, run in duplicate	23297133

Appendix 4

Consider the bivariate methylation measurement probability density:

$$P[z,y] = Qp^2 e^{-p(|z|+|y|+\theta zy)}\left[1+q(H_3(|z|)+H_3(|y|))\right]$$

where Q is a normalizing constant and H_3 represents the third-order Hermite polynomial. The cross-product in the Hermite polynomials is ignored as being of order $q^2 << 1$. The dependency parameter in the bivariate probability densityis θ.

The normalizing constant Q is given by

$$Q^{-1} = \int_{-1}^{1}\int_{-1}^{1} p^2 e^{-p(|z|+|y|+\theta zy)}\left[1+q(H_3(|z|)+H_3(|y|))\right]dzdy \quad (4a)$$

Noting that the integral (4a) over the quadrant ($z = 0\ 1, y = 0\ 1$) equals the integral (4a) over the quadrant ($z = -1\ 0, y = -1\ 0$), and similarly the integral (4a) over the quadrant ($z = -1\ 0$, $y = 0\ 1$) equals the integral (4a) over the quadrant ($z = 0\ 1$, $y = -1\ 0$), then

$$Q^{-1} = \frac{2}{p^4}\int_0^1 \frac{e^{-py}}{(1+\vartheta y)^4}\left[S_1(y) - e^{-p(1+\theta y)}S_2(y)\right]dy + \frac{2}{p^4}\int_0^1 \frac{e^{-py}}{(1-\vartheta y)^4}\left[S_3(y) - e^{-p(1-\theta y)}S_4(y)\right]dy \quad (4b)$$

where

$$S_1(y) = (1+\theta y)^3\left[p^3\left(1+q\left(y^2-3y\right)\right)\right] - 3qp^2(1+\theta y)^2 + 6q$$

$$S_2(y) = p^3(1+\theta y)^3\left(1-2q\right) + 6pq(1+\theta y) + 6q + p^3 q(1+\theta y)^3\left(y^3-2y\right)$$

$$S_3(y) = S_1(y) \text{ with } \theta \to -\theta$$
$$S_4(y) = S_2(y) \text{ with } \theta \to -\theta$$

The bivariate probability density (eqn 2) with the normalizing constant given by equation (4b) can be formulated as a log likelihood for the parameter θ, and estimation of this parameter

https://doi.org/10.1017/S2040174414000476 Published online by Cambridge University Press

along with a likelihood surface can be calculated for a bivariate set of methylation data (z_i, y_i):

$$\text{Log likelihood}(P[z,y]) = 2n\ln(p) + n\ln(Q(\theta)) + \sum_{i=1}^{n} - p(|z_i| + |y_i| + \theta z_i y_i) + \ln\left[1 + q(H_3(|z_i|) + H_3(|y_i|))\right] \quad (4c)$$

Note that the normalizing factor Q is now a function of the dependency parameter θ. The log likelihood (4c) can be maximized for θ numerically using non-gradient methods.

https://doi.org/10.1017/S2040174414000476 Published online by Cambridge University Press