AARHUS UNIVERSITY

Cover sheet

This is the publisher's PDF (Version of Record) of the article.

This is the final published version of the article.

How to cite this publication:

Madsen, Tobias, Świtnicki, Michał, Juul, Malene and Pedersen, Jakob Skou. "EBADIMEX: an empirical Bayes approach to detect joint differential expression and methylation and to classify samples" *Statistical Applications in Genetics and Molecular Biology*, vol. 18, no. 6, 2019.

<https://doi.org/10.1515/sagmb-2018-0050>

The final publication is available at www.degruyter.com

Publication metadata

Title:	EBADIMEX: an empirical Bayes approach to detect joint differential expression and methylation and to classify samples
Author(s):	Madsen, Tobias, Świtnicki, Michał, Juul, Malene and Pedersen, Jakob Skou
Journal:	Statistical Applications in Genetics and Molecular Biology
DOI/Link:	https://doi.org/10.1515/sagmb-2018-0050
Document version:	Publisher's PDF (Version of Record)

General Rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

If the document is published under a Creative Commons license, this applies instead of the general rights.

Tobias Madsen^{1,2} / Michał Świtnicki¹ / Malene Juul^{1,2} / Jakob Skou Pedersen^{1,2}

EBADIMEX: an empirical Bayes approach to detect joint differential expression and methylation and to classify samples

¹ Department of Molecular Medicine, Aarhus University, Palle Juul-Jensens Boulevard 99, DK-8200 Aarhus N, Denmark, E-mail: tobias.madsen@gmail.com

² Bioinformatics Research Centre, Aarhus University, C.F. Møllers Alle 8 DK-8000 Aarhus C, Denmark, E-mail: tobias.madsen@gmail.com

Abstract:

DNA methylation and gene expression are interdependent and both implicated in cancer development and progression, with many individual biomarkers discovered. A joint analysis of the two data types can potentially lead to biological insights that are not discoverable with separate analyses. To optimally leverage the joint data for identifying perturbed genes and classifying clinical cancer samples, it is important to accurately model the interactions between the two data types. Here, we present EBADIMEX for jointly identifying differential expression and methylation and classifying samples. The moderated t-test widely used with empirical Bayes priors in current differential expression methods is generalised to a multivariate setting by developing: (1) a moderated Welch t-test for equality of means with unequal variances; (2) a moderated F-test for equality of variances; and (3) a multivariate test for equality of means with equal variances. This leads to parametric models with prior distributions for the parameters, which allow fast evaluation and robust analysis of small data sets. EBADIMEX is demonstrated on simulated data as well as a large breast cancer (BRCA) cohort from TCGA. We show that the use of empirical Bayes priors and moderated tests works particularly well on small data sets.

Keywords: differential expression, differential methylation, empirical Bayes, probabilistic modeling

DOI: 10.1515/sagmb-2018-0050

1 Introduction

Accurate identification and robust modeling of cancer biomarkers can provide key insights to improve cancer diagnosis and treatment (Smith, Roda & Yap, 2014). Integrated analysis of biomarkers, and the derived increase in statistical power, is believed to hold more information and enable a performance gain, compared to separate analyses of individual biomarkers (Kristensen et al., 2014). Careful modeling of any dependencies between the data types is naturally important. DNA methylation and its regulation of gene expression levels has long been known to play an important part in the development and progression of cancer (Levenson, 2010). Here, we present EBADIMEX (Empirical Bayes for Differential Methylation and Expression) to jointly identify differential expression and methylation, and to classify samples as either normal or tumor. We apply the method to a breast cancer (BRCA) cohort from The Cancer Genome Atlas (TCGA) (<http://cancergenome.nih.gov>) with available gene expression and methylation data. EBADIMEX is freely available as an R-package (Appendix B).

DNA methylation plays a critical role in normal cell development and in shaping cell identity (Smith & Meissner, 2013). Along with histone modifications, DNA methylation is one of the main regulatory epigenetic mechanisms. The paradigm is that promoter hypermethylation silences gene expression, by directly or indirectly inhibiting transcription factor binding (Kass, Landsberger & Wolffe, 1997; Brenet et al., 2011; Jones, 2012). Conversely, gene body hypermethylation seems to correlate positively with transcriptional activity, although the causative mechanisms are poorly understood (Jingo et al., 2012; Yang et al., 2014). The silencing of tumor suppressor genes through promoter hypermethylation is established as a pathway in carcinogenesis (Jones & Baylin, 2007; Esteller, 2008), and the biomarker potential of DNA methylation in cancer is reported in numerous studies (Strand, Orntoft & Sorensen, 2014; Ma, Cao & Guo, 2016; Zhong & Cen, 2017). Furthermore, gene body hypomethylation in a panel of genes was recently suggested as an important hallmark of cancer (Mendizabal et al., 2017). Various specialized methods to detect differential methylation exist. These methods typically

Tobias Madsen is the corresponding author.

© 2019 Walter de Gruyter GmbH, Berlin/Boston.

perform single-probe analysis or detect differentially methylated regions (DMR), where the breakpoints of the DMR regions are learned from data (Morris et al., 2013; Wu, Gu & Zhang, 2013; Aryee et al., 2014).

High-throughput sequencing of RNA molecules (RNA-seq) has gradually overtaken the role of microarrays in quantifying gene expression levels. Multiple methods for detecting differential expression between experimental or biological conditions in RNA-seq data exist. *DESeq2* (Love, Huber & Anders, 2014), *edgeR* (McCarthy, Chen & Smyth, 2012), and *limma* (Smyth, 2004; Ritchie et al., 2015) are examples of such methods that all use an empirical Bayes approach to share information across genes and obtain regularized estimates of the gene-specific expression variation. By sharing information across genes, the use of empirical priors increases the power to detect differential expression, particularly when the number of samples or replicates is small. *DESeq2* and *edgeR* both model the RNA-seq read-counts directly using a negative-binomial model. *limma* was originally designed for the analysis of microarray data, which, in contrast to RNA-seq data, gives a continuous measure of expression for each gene. By log-transforming normalized RNA-seq data, *limma* has subsequently been adapted to accommodate RNA-seq data.

Methylation and expression data can be integrated by combining p -values from methods specialized for analyzing the two data types separately, e.g. by using Fisher's method (Fisher, 1932). However, this approach rests on the typically invalid assumption of independence. Another integration approach is to use machine learning methods such as random forest or regularized regression methods, while not explicitly accounting for dependencies between data types (List et al., 2014). *PINCAGE* (Switnicki et al., 2016) uses a probabilistic graphical model to explicitly model the interaction between methylation and expression measurements. With *PINCAGE*, the gene-specific distributions are modelled and learned using non-parametric kernel-based methods. Probabilistic models provide a high degree of flexibility and are highly interpretable. However, kernel density estimation in high dimensions suffers from the curse-of-dimensionality. Using an example with Gaussian kernels, it can be shown that the required number of samples to obtain a fixed level of precision increases exponentially with the number of dimensions (Scott, 2008, chapter 7). Another line of probabilistic models uses hierarchical Bayesian models to estimate parameters across genes. In (Jeong et al., 2010) a joint prior distribution for differences between methylation and expression in two conditions is learned. The prior is then used to give regularized estimates of joint differential expression and methylation for the individual gene.

Here, we propose a parametric probabilistic model to integrate expression and methylation data. We use empirical Bayes to obtain regularized variance and covariance estimates, generalizing the approach used by *limma* to multiple dimensions. Specifically, we develop (1) a moderated Welch t-test for equality of means with unequal variances; (2) a moderated test for equality of variances; and (3) a multivariate test for equality of means with equal covariance matrices. Cancer genomics data often display large amounts of heterogeneity, and to achieve robustness to outliers, methods from robust statistics, particularly winsorization and the Huber loss, are employed (Section 5.1 and Appendix A.1).

To illustrate the performance of *EBADIMEX*, we classify tumor and adjacent normal breast cancer samples from TCGA. We increase the difficulty of the classification problem by using fewer genes and fewer samples to train the model. *EBADIMEX* consistently performs better in classifying samples and identifying genes with differential methylation and expression than similar methods, which neither share information across samples using empirical Bayes nor correct for outliers. We observe this for all subsample sizes, but the difference is more pronounced for small subsamples.

EBADIMEX is not restricted to work with RNA-seq and methylation data, but can be extended to other data types, so long as the same measurements are performed for each of a moderately large number of units. Empirical Bayes posits an informative prior which is estimated from data, and the uncertainty on the location of this prior should be minimal compared to the spread of the prior. This is typically the case if the prior is based on a couple of hundred units, although it depends on the exact setting.

In the following sections we first discuss the input data for *EBADIMEX* and its necessary preprocessing, along with methods to increase robustness to outliers. Next, the model behind *EBADIMEX* and its underlying assumptions are presented. The mathematical formulations of hypotheses regarding differential expression and methylation are stated along with the details of statistical tests. An overview of the full analysis pipeline is given, followed by descriptions of methods to validate the performance of *EBADIMEX*.

2 Input data for *EBADIMEX*

The input data for *EBADIMEX* is expression data and DNA methylation measurements. The DNA methylation data should contain measures of the average methylation at genomic CpG sites in the promoter and the gene body of a gene. Such measurements can be obtained with bisulfite sequencing or array-based techniques. The use of *EBADIMEX* is illustrated with data from the widely used Illumina Infinium 450 K methylation array

(Bibikova et al., 2011). For example, TCGA has collected methylation profiles of almost 9000 samples using the array (Grossman et al., 2016). Here, we analyze the publicly available TCGA breast cancer samples, where both RNA-seq expression data and DNA methylation measurements are available. The data set consists of 84 normal samples and 781 tumor samples (Weinstein et al., 2013).

In RNA-seq experiments, a small fraction of the genes often contribute a large fraction of the reads. If the data is normalized by the total read count, differential expression of one of these genes will affect all other genes. To alleviate this problem, we use upper-quartile normalization. We further log-transform these normalized values to reduce the skewness, that such count data would otherwise display. After these pre-processing steps, we denote the expression level of gene g in class c and sample i by E_{gci} .

For each gene there is a number of associated methylation probes. We use the annotations of the probes provided by the Illumina development team to define promoter and gene-body associated probes for each gene (Bibikova et al., 2011). In the promoter definition we include the region from 0 to 1500 bp upstream of the transcription start site (TSS), 5' UTR and 1st exon. The 1st exon is included in the promoter definition, as it has been shown that methylation of the 1st exon has similar regulatory consequences as promoter methylation (Brenet et al., 2011). The gene body is defined as the remaining transcribed region (Figure 1A and B). After peak-based correction (Dedeurwaerder et al., 2011), we calculate the average promoter and gene body M-value, defined as the $\log_2$ ratio between the number of methylated and unmethylated probes, potentially adding a pseudo count γ ,

$$M = \log_2 \left(\frac{I_{\text{methylated}} + \gamma}{I_{\text{unmethylated}} + \gamma} \right).$$

M-values are typically preferred over the raw fraction of methylated probes because of the generally heteroscedastic nature of the raw fractions (Du et al., 2010).

Figure 1: Data overview. (A) Methylation probes are classified as belonging to promoters and gene bodies, as defined by Illumina. The 1500 bp upstream of the transcription start site (TSS), the 5' UTR and the 1st exon are included in the promoter region definition. The remaining coding region and the 3' UTR are included in the gene body region definition. (B) Probe measurements for a single gene across the BRCA cohort. We average over the promoter (red) and gene body probes (green), respectively. (C) Distribution of expression levels for a single gene after upper-quartile normalization and log transformation. (D) Correlation between expression and promoter- and gene body methylation. Generally, promoter methylation and expression are negatively correlated, while gene body methylation and expression are positively correlated. Also shown is promoter methylation against gene body methylation, both raw and adjusted for the expression level.

We denote the average promoter and gene body methylation of gene g in class c and sample i by $(M_{gci}^{\text{PR}}, M_{gci}^{\text{GB}})$.

2.1 Data transformations

To cater for the large amount of heterogeneity observed in the expression patterns of cancers, we take additional measures to ensure robustness, as outlined below.

Pre-filtering: Only genes that are consistently expressed across samples are included in the analysis. It is a common practice used both for RNA-seq and microarrays to filter out genes that are very lowly expressed. For RNA-seq, low expression levels manifest themselves as many samples with zero read-count. First, such genes are likely not implicated in differentiating cancers, and second, when log-transforming small read counts they are not well approximated by a normal distribution and thus violate the distributional assumptions of linear normal models used later.

Normalization: Normalizing RNA-seq data has been identified as a crucially important step in differential expression studies (Bullard et al., 2010). Some of these procedures address the problem of controlling for transcript length in deciding the actual relative abundance between two transcripts, leading to the widely used FPKM measures. This problem is irrelevant for EBADIMEX, as it only handles the relative difference between samples for individual transcripts.

However, another important problem is that a small group of genes make up the bulk of the RNA-seq reads (in the TCGA breast cancer cohort, 6.6% of the genes contribute half the reads). If one or more of these genes are differentially expressed, normalizing by total read count will dilute the differential expression signal. Furthermore, false positive differential expression can be introduced for more lowly expressed genes. Here, we use upper-quartile normalization to avoid the adverse effects of total count normalization (Figure 1C).

Winsorization: A final important step is to ensure robustness to outliers, which are frequent in cancer genomics. Where outliers do not pose much risk for kernel-based methods, parametric methods are generally sensitive to outliers. We apply winsorization, which is an iterative procedure that cap a data vector such that no value is more than a pre-specified number of standard deviations away from the mean (Huber & Ronchetti, 2009).

An important feature of the data is that expression and methylation correlate: generally expression correlates negatively with promoter methylation and positively with gene body methylation. However, we observe large intergene differences. Moreover, gene body- and promoter methylation are typically slightly positively correlated after adjusting for the effect of expression (Figure 1D).

2.2 TCGA BRCA data

In the following we analyze the publicly available TCGA breast cancer samples, where both RNA-seq expression data and DNA methylation measurements are available (Weinstein et al., 2013). We consider two classification problems: The first is classifying tumours ($n = 781$) from normal samples ($n = 84$). The second problem is classifying progressed samples ($n = 38$) from non-progressed ($n = 43$). The progression status was determined based on a detection of a new tumour within a fixed time period after the initial treatment. The relapse could be either morphological or determined via other means. The follow-up time for the patients with no relapse was either the number of days to the last follow-up or the number of days to death of the patient. The censorship time (number of days = 2175) was decided following a principle of balancing the number of progressed and non-progressed tumours and extending the censorship time to a reasonable level (almost 6 years in our case; Table 1) and Supplementary Table 1.

Table 1: Overview of TCGA BRCA dataset where both expression and methylation data is available and key parameters for determining progression status.

Cancer type	BRCA
Number of included tumour patients	781
Number of adjacent normal tissue controls	84
Censoring time	2165 days
Number of progressing tumours	38
Number of non-progressing tumours	43

3 Statistical models of EBADIMEX

At a high level we model first the marginal distribution of the expression data and then the joint distribution of the gene-body and promoter methylation measurements conditional on the expression data.

3.1 Expression model

The expression model of EBADIMEX is similar to the model used in the `limma` software package (Smyth, 2004; Ritchie et al., 2015), where a linear normal model is fitted for each gene, but information is shared across genes using empirical Bayes. A prior for variance is learned across all genes. This increases the statistical power to detect genes with large effect sizes but high estimated variances, at the expense of genes with low effect sizes and small estimated variances. In the present work we generalise the previously developed moderated t-test (Smyth, 2004), used with empirical Bayes priors in expression analysis (Smyth, 2004; Ritchie et al., 2015): We thus develop a moderated Welch t-test to test the hypothesis of equal means with unequal variances and a moderated F-test to test the hypothesis of equal variance.

The model introduced in `limma` assumes the following distribution of the log-expression level, E_{gci} , of gene g in class c for sample i :

$$E_{gci} \sim \mathcal{N}(\mu_{gc}, \sigma_{gc}^2) \quad (1)$$

where

$$\frac{1}{\sigma_{gc}^2} \sim \frac{1}{s_0^2 d_0} \chi^2(d_0), \quad (2)$$

i.e. σ_{gc}^2 has an inverse- χ^2 distribution.

Realistic models of the underlying count data are heteroscedastic, and this feature carries over to the log-transformed data. As an example, using the delta method, it can be seen that the log-transform of a Poisson distributed variable has a variance that is inversely proportional to its mean. Consequently, the hyperparameters d_0 and s_0 are functions of the expression level. To estimate these functions we first bin the genes into groups with similar expression levels. Next, we use an empirical Bayes approach to learn a pair of hyperparameters d_0 and s_0^2 , for each group. Finally we interpolate the parameters for each expression level (Figure 2A, Section 5.6).

Figure 2: Priors. (A) The prior variance for expression depends on the expression level. Using empirical Bayes we estimate the variance prior. The empirical variance estimates are plotted (black dots) along with the median of variance prior (dotted blue line). The shaded region represents the first and third quartile of the variance prior. (B) The mean covariance matrix for the inverse-Wishart distribution used as prior for methylation. (C–E) The prior distribution of correlation, standard deviation for promoter methylation and standard deviation for gene body methylation.

We are not only interested in the hypothesis of shift in location, $H_\mu : \mu_{g1} = \mu_{g2}$, but also the hypothesis of variance homogeneity, $H_\sigma : \sigma_{g1}^2 = \sigma_{g2}^2$. Moderated versions of the classical t-test and novel moderated versions of the Welch t-test and the F-test of variance homogeneity are used to test the two hypotheses (Sections 5.2, 5.3 and 5.4). In the moderated tests, a regularized version of the variance estimate replaces the regular estimate. The advantage of these tests is that they increase power, but also that they emphasize effect size over small variance.

3.2 Methylation model

For a particular gene we model the two methylation values (gene body and promoter) as a bivariate normal distribution, where the mean depends linearly on the expression level and on the group (Figure 2A).

$$\begin{aligned}\mathbb{E}M_{gci}^{PR} &= \alpha_{pr,g}E_{gci} + \mu_{pr,g,c} \\ \mathbb{E}M_{gci}^{GB} &= \alpha_{gb,g}E_{gci} + \mu_{gb,g,c}.\end{aligned}$$

Figure 3: Schematic overview of the analysis pipeline.

Typically, the measurements from the two probe types are positively correlated, but due to a large variation between genes, some genes display a negative correlation as well (Figure 2C). The correlation between probe types is determined by the covariance matrix, Σ , of the bivariate normal distribution (Figure 2B–E). The inverse wishart distribution has four free parameters, of which three capture the mean of the two variances and the correlation, and one captures the peakedness of the distribution (i.e. how informative the prior is). Therefore, it is slightly inflexible when it comes to modeling the correlation accurately. The prior is conservative with the amount of information we have on the distribution of the correlation coefficient (Figure 2C).

Collecting all parameters specifying the mean value structure in a single matrix,

$$\theta = \begin{pmatrix} \alpha_{pr,g} & \mu_{pr,g,1} & \mu_{pr,g,2} \\ \alpha_{gb,g} & \mu_{gb,g,1} & \mu_{gb,g,2} \end{pmatrix},$$

the model can be written as

$$M_1 : \begin{pmatrix} M_{gi}^{PR} \\ M_{gi}^{GB} \end{pmatrix} \sim \mathcal{N}_2(\theta \mathbf{T}_{1(i)}, \Sigma_g),$$

where $\mathbf{T}_{1(i)}^\top = (E_i, \mathbb{I}_{c_i=1}, \mathbb{I}_{c_i=2})$ is the i 'th row of the N by 3 design matrix $\mathbf{T}_1$, N being the total number of samples.

The covariance matrix, Σ , is assumed to be distributed according to an inverse-Wishart distribution

$$\Sigma_g^{-1} \sim \text{Wishart}_p(\Sigma_0^{-1}, \nu_0) / \nu_0.$$

This is analogous to the use of the inverse- χ^2 distribution for the variance parameter in the expression model (equation (2)). As before, the parameters Σ_0 and ν_0 are learned using empirical Bayes (Section 5.7).

The relevant hypothesis for the methylation model is whether the two groups have the same mean after the controlling for expression:

$$H_{\mu_{pr,g}, \mu_{gb,g}} : \begin{pmatrix} \mu_{pr,g,1} \\ \mu_{gb,g,1} \end{pmatrix} = \begin{pmatrix} \mu_{pr,g,2} \\ \mu_{gb,g,2} \end{pmatrix}$$

As we are now testing for equality of two parameters simultaneously a moderated version of the F-test is used to test the hypothesis (Section 5.5).

4 EBADIMEX pipeline

The analysis pipeline is summarized in this section, and a schematic overview of the steps in EBADIMEX is given in Figure 3.

Preprocessing: We perform a number of data processing steps: If the RNA-seq data is in the form of raw read counts for each gene, we normalize the data. The options for normalization in the EBADIMEX R-package include total count- and upper quartile normalization. By default, upper quartile normalization is used (Appendix A.2). We proceed with filtering out genes with ubiquitously low expression. This is done by dichotomizing based on median expression level of the individual genes. The threshold is guided by fitting a two-component normal mixture to the median expression levels; genes whose median expression level falls in the lower of the two components are disregarded in subsequent analysis (Appendix A.3).

Priors: After filtering the genes, raw estimates of the expression variance and the methylation covariance matrices are obtained for each gene. Based on the distribution of these we learn the empirical prior distributions (Sections 5.6 and 5.7).

Ranking: Next, we use moderated test statistics to determine differential regulation. Genes are ranked according to their p -value. Alternatively, as an option in the EBADIMEX R-package, the data can be fitted using regularized estimates after which we are able to calculate the Kullback-Leibler divergence between the fitted distributions (Appendix A.4).

Classification: Finally, we classify samples using a likelihood criterion:

$$P(C = 1 | \mathcal{D}) = \frac{P(\mathcal{D} | C = 1)P(C = 1)}{\sum_{i=1}^2 P(\mathcal{D} | C = i)P(C = i)}. \quad (3)$$

where $\mathcal{D} = (E_g, M_g^{\text{PR}}, M_g^{\text{GB}})$ is the data for a single gene, g . The classifiers are evaluated using the area under the ROC curve (AUC) in an n -fold cross validation scheme. The AUC performance measure is agnostic to the actual class probabilities, but only considers the ordering of samples. The Bayes Factor (BF) is defined as

$$BF(C = 1 | \mathcal{D}) = \frac{P(\mathcal{D} | C = 1)}{P(\mathcal{D} | C = 2)}.$$

As the BF is a monotone function of $P(C = 1 | \mathcal{D})$, given in equation (3), it leaves the measure of AUC invariant. Furthermore, using the BF there is no need to specify the prior class probabilities $P(C = i)$. We therefore use the BF as predictor of the class label. The Bayes factor can be computed as

$$P(\mathcal{D} | C = 1) = P(E_g | C = 1)P(M_g^{\text{PR}}, M_g^{\text{GB}} | E_g, C = 1).$$

Here, the first term is

$$P(E_g | C = c) = \frac{1}{\sqrt{2\pi\sigma_{gc}^2}} \exp\left\{-\frac{(E_g - \mu_{gc})^2}{2\sigma_{gc}^2}\right\} \quad (4)$$

and the second term is

$$P(M_g^{\text{PR}}, M_g^{\text{GB}} | E_g, C = c) = \frac{1}{2\pi|\Sigma_{gc}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(M_g^{\text{PR}} - \mu_{pr,g,c}, M_g^{\text{GB}} - \mu_{gb,g,c})\Sigma_g^{-1}\begin{pmatrix} M_g^{\text{PR}} - \mu_{pr,g,c} \\ M_g^{\text{GB}} - \mu_{gb,g,c} \end{pmatrix}\right\}. \quad (5)$$

For both terms, regularized variance and covariance estimates are used.

5 Statistical details

5.1 Winsorization

Let $x = (x_1, \dots, x_N)$ be a data vector. Winsorization is a procedure that caps the original data for small and large values:

$$x_i^* = \begin{cases} M & x_i > M \\ x_i & m \leq x_i \leq M \\ m & x_i < m. \end{cases} \quad (6)$$

There are different approaches to determine m and M . One approach is setting m and M at the 5th and 95th percentile, respectively, thus ensuring that 10% of data is capped. Here we use an iterative procedure described in Huber and Ronchetti (2009):

1. Set $m = -\infty$ and $M = \infty$.
2. Compute x^* using (6).
3. Compute the empirical mean, μ_{x^*} , and the standard deviation, σ_{x^*} , of x^* .
4. Set $m = \mu_{x^*} - k * \sigma_{x^*}$ and $M = \mu_{x^*} + k * \sigma_{x^*}$.
5. Repeat from step 2 until convergence.

The value k determines how aggressively the data is capped. For our analysis $k = 2.5$ is used.

Estimates based on winsorized data has a smaller effective degree of freedom than estimates based on the raw data. We need to correct for this in estimates and tests. The idea proposed by Dixon and Tukey (1968) is to scale the t-test statistic by a factor $\frac{h-1}{N-1}$, where h is the number of observations left unaltered by winsorization. In the subsequent t-test, the degrees of freedom is similarly adjusted to $h - 1$ rather than $N - 1$. The effect of scaling the t-test statistic can also be obtained by scaling the estimate of the standard deviation (Huber & Ronchetti, 2009).

5.2 Moderated t-test for equality of means

Let $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$, $i = 1, 2$, $j = 1, \dots, n_i$ be two independent normal samples with common variance. Furthermore, let σ^2 follow a scaled inverse- χ^2 distribution,

$$\frac{1}{\sigma^2} \sim \frac{1}{s_0^2 d_0} \chi^2(d_0).$$

We want to test the hypothesis $H_\mu : \mu_1 = \mu_2$ of common mean. The means are estimated by

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}.$$

These estimates are independent of the unbiased variance estimate given by

$$s^2 = \frac{1}{d} \sum_{i=1}^2 \sum_{j=1}^{n_i} (X_{ij} - \hat{\mu}_i)^2,$$

where $d = n_1 + n_2 - 2$. The moderated t-test uses a regularized estimate of σ^2 , $\tilde{s}^2$:

$$\tilde{s}^2 = \frac{ds^2 + d_0 s_0^2}{d + d_0}.$$

Replacing s^2 with $\tilde{s}^2$ in Student's t-test we obtain the moderated t-test statistic,

$$\tilde{t} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\tilde{s}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$

Under the hypothesis $H_\mu : \mu_1 = \mu_2$, $\tilde{t}$ is t distributed with $d + d_0$ degrees as is shown in Smyth (2004) using transformation of variables and density functions. Equivalently, $\tilde{t}^2 \sim F(1, d + d_0)$. Here we provide a simpler proof. First, note that

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sigma} \mid \sigma^2 \sim \mathcal{N}\left(0, \frac{1}{n_1} + \frac{1}{n_2}\right)$$

and is conditionally independent of

$$\frac{s^2}{\sigma^2} \mid \sigma^2 \sim \frac{1}{d} \chi^2(d).$$

Since the conditional distributions do not depend on σ^2 , the stated distributions apply unconditionally. At the same time, $\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sigma}$, $\frac{s^2}{\sigma^2}$, and σ^2 are all independent. Next we see that,

$$\frac{1}{d + d_0} \frac{s^2}{\sigma^2} = \frac{1}{d + d_0} \left(\frac{ds^2}{\sigma^2} + \frac{d_0 s_0^2}{\sigma^2} \right) \sim \frac{1}{d + d_0} \chi^2(d + d_0), \quad (7)$$

since the term in the parenthesis is the sum of two independent χ^2 distributions with d and d_0 degrees of freedom, respectively. Finally, writing

$$\tilde{t} = \frac{\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{s^2}{\sigma^2}}},$$

we see that $\tilde{t}$ is the ratio between a standard normal distribution and the square-root of an independent $\frac{1}{d + d_0} \chi^2(d + d_0)$, which characterizes the t -distribution with $d + d_0$ degrees of freedom.

5.3 Moderated Welch t-test for equality of means with unequal variances

As an alternative to the moderated t -test, we also consider a moderated Welch t -test for the case of unequal variance between two groups. Demissie et al. (2008) propose the following test:

Consider X_{ij} , $i = 1, 2$ and $j = 1, \dots, n_i$ and assume

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2),$$

where

$$\frac{1}{\sigma_i^2} \sim \frac{1}{s_0^2 d_0} \chi^2(d_0).$$

The usual estimators of the variances are denoted by

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$

The moderated Welch t statistic is

$$\tilde{t} = \frac{\bar{X}_{1.} - \bar{X}_{2.}}{\sqrt{\frac{d_0 s_0^2 + d_w s_w^2}{d_0 + d_w}}} \sim t(d_0 + d_w),$$

where

$$s_w^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2},$$

and

$$d_w = \frac{(s_w^2)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1} \right) + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2} \right)}.$$

The moderated Welch t-test is not well founded mathematically, and, indeed, our simulations show that it might be poorly fit by the t distribution (Supplementary Figure S1A). We therefore propose another version of a moderated Welch t-test, where variance estimates are substituted with moderated variance estimates:

$$t^* = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\tilde{s}_1^2}{n_1} + \frac{\tilde{s}_2^2}{n_2}}} \sim t(d_w^*),$$

where

$$\tilde{s}_i^2 = \frac{(n_i - 1)s_i^2 + d_0 s_0^2}{n_i - 1 + d_0}, \quad i = 1, 2$$

and

$$d_w^* = \frac{\left(\frac{\tilde{s}_1^2}{n_1} + \frac{\tilde{s}_2^2}{n_2} \right)^2}{\frac{\tilde{s}_1^4}{n_1^2(n_1-1+d_0)} + \frac{\tilde{s}_2^4}{n_2^2(n_2-1+d_0)}}.$$

The fact that t^* is approximately t distributed follows from the Welch-Satterthwaite approximation. First, note that writing

$$t^* = \frac{(\hat{\mu}_1 - \hat{\mu}_2) / \left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)}{\sqrt{\frac{\tilde{s}_1^2}{n_1} + \frac{\tilde{s}_2^2}{n_2}} / \left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)},$$

the numerator follows a standard normal distribution and is independent of the denominator. Using the Welch-Satterthwaite approximation, the denominator is approximately $\chi^2(d_w^*)/d_w^*$ distributed, and, hence, t^* is approximately t distributed with d_w^* degrees of freedom. Our simulations show that this moderated Welch t-test statistic is better approximated by the t distribution (Supplementary Figure S1B). Another thing to note is that the two sample variances need not have the same prior distribution.

Figure 4: Model fit. (A) The distribution of expression for tumors (green) and normals (red) for the top differentially expressed gene, MATN2, in the BRCA cohort. (B) The joint distribution of methylation for promoter and gene body after adjusting for the effect of expression. The moderated (solid line) and unmoderated (dashed line) contour ellipses are shown. Here shown for the top differentially methylated gene, PKLR. (C, D) QQ-plot for the tests with permuted labels (blue) and simulated data (gray). We show the plots for the test of differential expression location and differential expression scale. QQ-plots for the remaining tests are shown in Supplementary Figures S1, S3, and S4.

5.4 Moderated F-test for univariate equality of variances

Let $s_i^2 \sim \frac{\sigma_i^2}{d_i}$, $i = 1, 2$ and let

$$\frac{1}{\sigma_i^2} \sim \frac{1}{s_0^2 d_0} \chi^2(d_0), \quad i = 1, 2. \quad (8)$$

We want to test the hypothesis $H_\sigma : \sigma_1 = \sigma_2$, of common variance, where the variance is drawn from the distribution in equation (8). Thus, $s_i^2 \sim \frac{\sigma^2}{d}$, $i = 1, 2$, where

$$\frac{1}{\sigma^2} \sim \frac{1}{s_0^2 d_0} \chi^2(d_0).$$

Assume, without loss of generality, $d_1 \leq d_2$. In the moderated F statistic we use a regularized variance estimate for the smaller of the two samples,

$$\tilde{s}_1^2 = \frac{d_1 s_1^2 + d_0 s_0^2}{d_1 + d_0},$$

and consider the following test statistic

$$\tilde{F} = \frac{\tilde{s}_1^2}{s_2^2} \sim F(d_1 + d_0, d_2).$$

To see that $\tilde{F}$ indeed follows the posited F distribution, note that with an analogous argument to that used in Section 5.2, it can be shown that $\frac{s_1^2}{\sigma^2}$, $\frac{s_2^2}{\sigma^2}$ and σ^2 are all independent. Using equation (7) we see that $\tilde{s}_1^2/\sigma^2 \sim \chi^2(d_1 + d_0)/(d_1 + d_0)$ is independent of $s_2^2/\sigma^2 \sim \chi^2(d_2)/d_2$, hence

$$\tilde{F} = \frac{\tilde{s}_1^2/\sigma^2}{\tilde{s}_2^2/\sigma^2} \sim F(d_1 + d_0, d_2),$$

as the F distribution is characterized by the ratio of two scaled χ^2 distributions.

Using moderated variance estimates for both σ_1 and σ_2 will not yield a proper F distribution. Regularizing either of the two variance estimates, will lead to a proper F distribution. The choice of regularizing the variance estimate for the smaller class instead of the larger class, will however give slightly more power (Supplementary Figure S2).

5.5 Moderated F-test for multivariate equality of means

Collect all methylation measurements in an $N \times 2$ matrix:

$$\mathbf{M} = \begin{pmatrix} M_1^{\text{PR}} & M_1^{\text{GB}} \\ \vdots & \vdots \\ M_N^{\text{PR}} & M_N^{\text{GB}} \end{pmatrix}$$

The hypothesis of no group effect in the methylation part of the model, i.e.

$$H_{\mu_{pr}, \mu_{gb}} : \begin{pmatrix} \mu_{pr,1} \\ \mu_{gb,1} \end{pmatrix} = \begin{pmatrix} \mu_{pr,2} \\ \mu_{gb,2} \end{pmatrix},$$

can be tested with an F-test.

First, note that the test can be phrased as a reduction from two nested multiple regression models. The design matrix for the model with group effect has rows

$$\mathbf{T}_{1(i)}^\top = (E_i, \mathbb{I}_{c_i=1}, \mathbb{I}_{c_i=2}),$$

and the design matrix for the model with no group has rows

$$\mathbf{T}_{2(i)}^\top = (E_i, 1).$$

The projections onto the subspaces spanned by $\mathbf{T}_j$ are

$$\mathbf{P}_j = \mathbf{T}_j(\mathbf{T}_j^\top \mathbf{T}_j)^{-1} \mathbf{T}_j^\top.$$

Define the SSD matrices as

$$\mathbf{SSD}_1 = \mathbf{M}^\top (\mathbf{I} - \mathbf{P}_1) \mathbf{M}$$

and

$$\mathbf{SSD}_2 = \mathbf{M}^\top (\mathbf{P}_1 - \mathbf{P}_2) \mathbf{M}.$$

The SSD matrices are distributed according to a Wishart distribution conditional on Σ ,

$$\begin{aligned} \mathbf{SSD}_1 | \Sigma &\sim W_2(\Sigma, d) \\ \mathbf{SSD}_2 | \Sigma &\sim W_2(\Sigma, 1), \end{aligned} \tag{9}$$

where $d = N - 3$.

Thinking of Σ as fixed with no prior information, the likelihood ratio statistic is

$$Q^{2/n} = \frac{|\mathbf{SSD}_1|}{|\mathbf{SSD}_1 + \mathbf{SSD}_2|},$$

which follows the beta distribution $Q^{2/n} \sim B(\frac{d-1}{2}, \frac{2}{2})$. The beta distributed statistic can be transformed into an F distributed statistic:

$$F = \frac{1 - Q^{2/n}}{Q^{2/n}} \frac{d-1}{2} \sim F(2, d-1).$$

We now assume

$$\Sigma^{-1} \sim \text{Wishart}_p(\Sigma_0^{-1}, \nu_0)/\nu_0.$$

The moderated t-test can be generalized to a moderated F-test. Define first

$$\tilde{Q}^{2/n} = \frac{|\mathbf{SSD}_1 + \nu_0 \Sigma_0|}{|\mathbf{SSD}_1 + \mathbf{SSD}_2 + \nu_0 \Sigma_0|},$$

and then define the moderated F statistic,

$$F = \frac{1 - \tilde{Q}^{2/n}}{\tilde{Q}^{2/n}} \frac{d + d_0 - 1}{2} \sim F(2, d + d_0 - 1).$$

The moderated F statistic follows approximately an F distribution. It is thus computationally inexpensive to evaluate the statistical significance. The moderated F statistic is well approximated by the F distribution, as shown by simulation and by permuting real data (Supplementary Figure S3).

5.6 Prior expression

The variance is assumed to follow an inverse- χ^2 distribution,

$$\frac{1}{\sigma^2} \sim \frac{1}{s_0^2 d_0} \chi^2(d_0).$$

The empirical variance s_g^2 for a gene, g , is χ^2 distributed conditional on σ^2 ,

$$s_g^2 | \sigma^2 \sim \frac{\sigma^2}{d} \chi^2(d).$$

Again, $\frac{s_g^2}{\sigma^2}$ is $\chi^2(d)$ distributed and independent of σ^2 . We can write s_g^2 as a fraction between two independent (scaled) χ^2 distributions, and it follows that s_g^2 has a scaled F distribution,

$$s_g^2 = \frac{s_g^2/\sigma^2}{1/\sigma^2} \sim s_0^2 F(d, d_0). \quad (10)$$

Using equation (10), the two hyperparameters, s_0^2 and d_0 , can be estimated using the method described in Smyth (2004). The formulas are restated here for completeness.

Define $z_g = \log s_g^2$. From equation (10) we see that z_g follows Fisher's z distribution plus $\log s_0^2$. We now use moment matching to estimate d_0 and s_0^2 . The moments of Fisher's z distribution are

$$\begin{aligned} \mathbb{E}z_g &= \log s_0^2 + \psi(d/2) - \psi(d_0/2) + \log(d_0/d), \\ \text{Var}z_g &= \psi'(d/2) + \psi'(d_0/2). \end{aligned}$$

From these we see,

$$\mathbb{E} \left[\frac{n}{n-1} (z_g - \bar{z})^2 \right] \approx \psi'(d/2) + \psi'(d_0/2).$$

We can therefore estimate d_0 by solving the following equation numerically:

$$\text{mean} \left[\frac{n}{n-1} (z_g - \bar{z})^2 \right] - \psi'(d/2) = \psi'(d_0/2).$$

Having an estimate of d_0 we solve

$$\bar{z} - \psi(d/2) + \psi(d_0/2) - \log(d_0/d) = \log s_0^2$$

to find s_0^2 .

It has been observed that the variance is dependent on the expression level, with higher expressed genes showing less variance (McCarthy, Chen & Smyth, 2012; Love, Huber & Anders, 2014; Ritchie et al., 2015). We therefore bin genes based on average expression level and learn the hyperparameters for each bin separately. For the analysis of a single gene we use interpolation to decide the hyperparameters (Figure 2).

5.7 Prior methylation

The covariance matrix for methylation is assumed to follow an inverse-Wishart distribution,

$$\Sigma^{-1} \sim \text{Wishart}(\Lambda^{-1}/\nu_0, \nu_0).$$

The diagonal entries in Σ then follow an inverse- χ^2 distribution,

$$\frac{1}{\Sigma_{ii}} \sim \frac{1}{\nu_0 \Lambda_{ii}} \chi^2(\nu_0 - 1) = \frac{1}{(\nu_0 - 1) \frac{\nu_0 \Lambda_{ii}}{\nu_0 - 1}} \chi^2(\nu_0 - 1). \quad (11)$$

The empirical covariance matrix S_g for a gene, g , is Wishart distributed conditional on Σ ,

$$S_g | \Sigma \sim \text{Wishart}(\Sigma/d, d).$$

The marginal distribution of the diagonal entries is

$$S_{g,ii} | \Sigma \sim \frac{\Sigma_{ii}}{d} \chi^2(d). \quad (12)$$

Combining equation (11) and (12), it follows using the argument presented in Section 5.6 that

$$S_{g,ii} \sim \frac{\nu_0 \Lambda_{ii}}{\nu_0 - 1} F(d, \nu_0 - 1).$$

Thus, ν_0 and Λ_1 and Λ_{22} can be estimated using the technique presented in Section 5.6. $\Lambda_{12} = \Lambda_{21}$ is estimated using a moment matching procedure.

6 Performance validation of EBADIMEX

In the following sections, methods for determining the performance of EBADIMEX are described. These include simulation and permutation studies as well as classification performance testing.

6.1 Simulation analysis

As a first mean to evaluate the performance of a test, we simulate data from the null distribution. We simulate the genes in such a way, that they have the same distribution of mean expressions as observed in the BRCA cohort. For each gene, we simulate 10 samples of which 3 are labelled as normals, and the remaining 7 are labelled as tumors. The moderated t-test is exact and the simulation thus serves as a sanity check. On the other hand, the moderated F-test is approximate. We expect that the QQ-plot for the moderated t-test displays uniform p -values. Furthermore, if the QQ-plot for the moderated F-test displays uniform p -values, it confirms that the test is applicable.

6.2 Permutation analysis

To evaluate if our proposed tests are robust when applied to real data, we perform permutation analyses. For each gene we randomly select 10 samples among the tumor samples and label 3 of these as normals. The remaining 7 are labelled as tumors. The effect of using moderated tests diminish in large cohorts, and we thus only use 10 samples at a time. As the labelling is random we expect no signal of differential expression or differential methylation, and as such the p -values should be uniform.

6.3 Classification performance

To evaluate classification performance the following scheme is employed. First, 500 genes are selected at random, and only these genes are used for subsequent classification of samples. Next, the normal samples are divided evenly at random into k folds, and these folds are paired with groups of tumor samples to form k training datasets. To show the performance on small, but still unbalanced, data sets, only one fourth of the tumor samples are included in the training datasets. As an example, for $k = 32$, the smallest fraction of data used for training, the training data set consists of 2–3 normal samples and 6–7 tumor samples. For each of the k training data sets, a classification pipeline is used to classify the remaining test samples and computing the area under the curve (AUC) of the classifier. AUCs are then averaged over the k folds. The entire process of randomly sampling the genes used for classification and the training sets is repeated 200 times. The classification performance is interesting in its own right, but also serves as a proxy for the underlying prioritization of genes.

More than one classifier can be used on each training data for their subsequent performance comparison. For example, we compare the use of moderated and unmoderated tests. To test if a classification pipeline is significantly better than another, the Wilcoxon signed-rank test for paired samples is used.

6.4 Comparison with standard Machine learning and PINCAGE classifiers

Ranking of genes generated by EBADIMEX was used to evaluate and compare the performance of our method with the following standard machine learning (ML) approaches accessed through the caret package Kuhn (2015): Logistic Regression (R Core Team, 2017), Bayesian Logistic Regression (Gelman, 2011), Random Forests (Breiman et al., 2006), Support Vector Machines (SVM) with Linear (Meyer et al., 2016) and Polynomial (Karatoglou, Smola & Hornik, 2013) kernels. We also applied EBADIMEX's ranks to evaluate our previous and parameter-rich integrative PINCAGE method. For all methods, Upper-Quartile normalization was applied to the expression data, to mirror EBADIMEX's approach. Mean regional promoter and gene body methylation levels were calculated per sample for use with ML methods (as in the case of EBADIMEX), while PINCAGE operated on individual probe measurements, as per its original implementation. To facilitate SVM training, observations were additionally centered around zero, and divided by the standard deviation.

Models were trained independently per gene in a progressed vs. non-progressed cross-validation setting, and evidence from multiple top-ranked genes was combined using class probability updating (Bailer-Jones & Smith, 2011), assuming an independence between the genes akin to a Naïve Bayes approach. Individual class probabilities for each predictive method were calculated by the caret package internally.

7 Results

The performance of EBADIMEX was tested on a breast cancer cohort from TCGA. Samples were included on the basis of availability of both RNA-seq and 450 K array methylation data. Following this criterion, 781 cancer samples and 84 adjacent normal samples were included.

Parameters were fit to each gene for both expression and methylation given the expression. Figure 4A illustrates the fitted distribution of expression for tumors and normals for the highly differentially expressed gene, *MATN2*, and Figure 4B illustrates the fitted joint methylation distribution for the top differentially methylated gene, *PKLR*. Although we generally see that the empirical Bayes regularization shrinks the variance estimates in these cases, it is not immediately clear from these figures. This is explained by the relatively large amount of data used here.

It is expected that some genes are differentially methylated without altered expression and conversely some genes are differentially expressed by mechanisms unrelated to methylation. To assess how often differential methylation and expression co-occur we computed the p -value for differential expression and methylation for all genes (Supplementary Figure 6). There is a significant enrichment of genes showing differential methylation and expression simultaneously (Spearman's rank correlation: $p < 1e - 15$).

To test whether the obtained test statistics were well approximated by their posited distributions, two analyses were performed: In the first analysis, data was simulated from the null model with the same number of normals and tumors in each group, and in the second analysis, 84 of the 781 tumor samples were randomly labeled as normals (Sections 6.1 and 6.2). In both types of analyses, we tested for differential expression and differential methylation. Differential expression was tested with the moderated t -test for shift in location and the moderated F -test for variance homogeneity. The qq-plots obtained from these analyses show no inflation or deflation, indicating that the test statistics are well approximated by the posited distributions (Figure 4C and D and Supplementary Figure S4). Differential methylation was tested with the moderated F -test. Again, the p -values were uniformly distributed in both analyses (Supplementary Figure S3).

After the detection of differentially expressed genes, we were interested in assessing the accuracy of classification with top-ranked genes (Section 6.3).

First, we compare the classification performance using expression, methylation or the combination of the two. This was done by training the Ebadimex model for all genes on a small subset of 25 samples (10 normals, 15 tumours) and evaluating the models classification performances on the remaining samples. The training samples are picked at random and the sampling is repeated 10 times (Figure 5). The top differentially expressed genes, can discriminate tumours from normals almost as well as the combination of expression and methylation, indicating that the majority of the signal is derived from expression. However modeling methylation and expression jointly clearly improves classification, as we consider the less differentially expressed genes.

Figure 5: Comparison of multivariate with univariate classifiers. A comparison of the cross-validated performance (AUC) of three single gene classifiers. The three classifiers are the two univariate classifiers based on either methylation or expression in isolation and the multivariate classifier using both in combination. The AUCs are sorted descendingly within each classifier. The left hand figure shows AUCs for all tested genes, whereas the right hand figure home in on the top 500 AUCs.

Tumors and normals in the BRCA cohort can be separated almost perfectly using the top-ranked genes (Figure 4A, B and Figure 6A). To make the problem harder, the data was subsampled into successively smaller subsets. These subsets were used to detect differentially regulated genes, and the classification performance was evaluated on the remaining data. For comparison, prioritization and classification was performed using two approaches: First, with winsorization, moderated estimates and moderated tests (moderated classification). Second, without using any of the aforementioned techniques (raw classification). For each trial we sampled a subset of 500 genes and subsets of varying sizes of the samples, to make the classification problem harder. In

the smallest subset, only 1/32th of the normals, corresponding to 2–3 normal samples, were used. Roughly twice as many cancer samples were used. Using these samples, the genes were prioritized and the parameters of the model were learned. Then, the top-ranked gene was used to classify the remaining samples. We repeated this process 50 times. When half of the normals were used, the average AUC was 0.977 with both classification pipelines. When only 1/32th of the normals were used, the average AUC with the moderated classification pipeline was 0.913, and 0.877 with the raw classification pipeline (Mann-Whitney rank sum test: $p = 4.12 \cdot 10^{-10}$) (Figure 6A). The described analysis shows that the use of robust statistics and empirical Bayes is particularly powerful for small cohorts.

Figure 6: Moderated tests and classification performance. (A) Classification in the BRCA cohort using the top-ranked gene with respect to the p -values reported from the moderated (beige) and unmoderated (blue) tests. Likewise, the estimates used to classify the samples are either moderated or unmoderated. We use a subsample of 500 genes and progressively smaller training data sets, with the smallest (1/32th) using only 2–3 normal samples and 6 tumor samples. The procedure of picking the subsample of genes and training data is repeated 200 times, and in each run we evaluate the AUC using cross validation. When using a progressively smaller fraction of the normals, the moderated tests outperform the unmoderated tests (p -value paired signed rank test). (B) Joint methylation after adjusting for expression for the PKLR gene in a subset of the samples. The moderated (solid line) and unmoderated (dashed line) contour ellipses are shown. The shrinkage of variance in the tumor group is mainly caused by capping the effect of the two extreme observations using winsorization. Both the moderated test and, to a lesser extent, the unmoderated test report significant differential methylation (p -value unmoderated: $1.3 \cdot 10^{-19}$; p -value moderated: $1.8 \cdot 10^{-25}$). (C) Expression shown for two genes (MNT, SGCE) for a subset of the samples. The error bars denote the 95% interval for the moderated (solid line) and unmoderated (dashed line) estimates. The unmoderated test reports both genes as significantly differentially expressed (p -value MNT: $6.5 \cdot 10^{-4}$; p -value SGCE: $4.2 \cdot 10^{-4}$). However the moderated test, based on the regularized variance estimates, only reports the SGCE gene as significantly differentially expressed (p -value MNT: 0.077; p -value SGCE: $3.1 \cdot 10^{-5}$).

We compared the performance of EBADIMEX to a range of standard statistical methods, PINCAGE (Świtnicki et al., 2016) and EBADIMEX using only methylation or expression (See Section 6.4). This was done on the clinically relevant and challenging classification problem of separating samples from patients that later have disease progression ($n = 38$) from patient that do not progress ($n = 43$). These samples are a subset of the full BRCA cohort. The likelihoods for the top genes were combined using naive Bayes approach. The models were trained on randomly selected subsets consisting of 25 samples, we report the mean AUC for classifying the remaining samples (Figure 7). Note that the genes selected as features for the machine learning methods were selected by EBADIMEX, except for the case of PINCAGE, since the machine learning methods cannot readily select genes. EBADIMEX outperforms other methods when combining multiple top-genes. This is likely because EBADIMEX is designed to robustly fit small data sets. The analysis also shows that EBADIMEX using both data types in combination yields better results than using expression or methylation alone. Since methylation is not used when fitting the expression model, it is straightforward to classify samples based only on the expression contribution to the likelihood (equation (4)). To classify samples based on methylation only, we first need to train the parameters of a methylation model where expression is not used as a predictor, and then use only the methylation contribution to the likelihood (equation (5)).

Figure 7: Naive Bayes combination of genes. Classification performance for classifying BRCA samples as progressive/non-progressive using a naive Bayes combination of the first k top ranked genes. We compare EBADIMEX with standard methods and PINCAGE, as well as EBADIMEX using only methylation or expression.

Winsorization can reduce the number of false positives when a location shift is driven by outliers. However, the number of false negatives can also be reduced using winsorization. For instance, if the variance is overestimated due to outliers, a true shift in location can be obscured. Figure 6B shows an example of the latter, where winsorization reduces the variance estimate of the tumor group. Albeit, testing with and without winsorization both yield a significant result, it is markedly more pronounced after winsorization, which then ultimately leads to a higher priority assignment. If the variance, on the other hand, is underestimated, small effect sizes may be rendered significant. Moderated variance estimates primarily reduce the number of false positives by shrinking the raw estimate towards the prior. The word “shrinking” can be a source of confusion; a shrunken estimate is not necessarily smaller than the regular estimate, rather it is closer to the mean of the prior distribution. Tests based on moderated variances put larger emphasis on effect size at the expense of small variance. To illustrate this, we highlight the analyses of the two genes SGCE and MNT (Figure 6C): Both genes are significantly differentially expressed using unmoderated tests, although SGCE has a markedly higher effect size than MNT. After moderating the test, only SGCE shows significant differential expression.

8 Conclusions and discussion

We have presented the novel method, EBADIMEX, for joint identification of differential expression and methylation, and for sample classification. A major contribution of this work is the generalization of previously published moderated tests used with empirical Bayes priors for expression analysis: Here, we have developed a moderated Welch t-test for equality of means with unequal variances, as well as a moderated F-test of homogeneity in expression variance to accommodate the often highly heterogeneous nature of cancer data. Finally, we accounted for correlations between measurements and developed a moderated multivariate F-test for equality of means with equal variance, thereby extending a previously published moderated t-test for differential expression to a multivariate setting (Ritchie et al., 2015). Altogether, this allows us to generalize existing empirical Bayes methods developed (Smyth, 2004; Love, Huber & Anders, 2014; Ritchie et al., 2015) for differential expression studies to a multivariate setting with multiple gene-wise measurements.

EBADIMEX has been applied to a TCGA breast cancer cohort of 84 normal samples and 781 tumor samples with available measures of gene expression and methylation, demonstrating the applicability of the method and its implementation. All tests applied in EBADIMEX are parametric, allowing for fast evaluation compared to, for instance, permutation based tests. Through label permutations and simulations, we show that the moderated tests are well behaved with uniform p -value distributions under the null.

We have demonstrated that the use of moderated tests and estimates increase statistical power, particularly for studies with few replicates or samples. In particular, we have shown that sample classification into tumor and normal benefits from making use of moderated tests and robust statistics.

The fact that methylation and expression are correlated prevents one from performing two separate analyses and combining them using Fisher's method. Although methylation and expression measurements do not provide independent evidence of differential regulation, the use of both data types simultaneously increases our statistical power over any of the data types in isolation. This can especially prove valuable in situations where additional samples are not readily obtainable.

In the developed statistical setup, all three measurements (expression and the two methylation measurements) could be modelled by a single multivariate normal distribution. But the prior distribution of the expression variance depends on the expression level. To allow separate prior distributions for the variances of expression and methylation, we model first expression and then regress methylation on expression.

Here, EBADIMEX is demonstrated on cancer data, but it can just as well be applied to the other settings where we want to compare two groups to test for differential expression and methylation. More generally, the moderated Welch t-test can e.g. be used in situations where one condition shows more variability, but it is unclear which units actually have different location. The moderated F-test for equality of means can be used to include additional data types, e.g. histone modifications or chromatin state. Although EBADIMEX is developed with the analysis of expression and methylation data in mind, the statistical methods we develop can be applied more broadly in contexts with multiple interrelated comparisons of uni- and multivariate data.

Funding

This study was supported by Independent Research Fund Denmark, Grant Number: 7016-00379, Sapere Aude, Grant Number: 12-126439 and Innovation Fund Denmark, Grant Number: 10-092320/DSF.

Appendix

A Statistical notes

A.1 Huber Loss

To limit the influence of a single data type (expression or methylation), when performing classification, we use the Huber Loss instead of the log-density.

If $X \sim \mathcal{N}(\mu, \sigma^2)$, the log-density function is quadratic in $x - \mu$,

$$\log f_X(x) \propto -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}.$$

The Huber-loss function $L(x)$ is linear when x is more than k standard deviations away from μ .

$$L(x) \propto \begin{cases} -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} & |x - \mu| \leq k\sigma \\ -\frac{2k\sigma|x - \mu| + (k\sigma)^2}{2\sigma^2} & |x - \mu| > k\sigma \end{cases}$$

We use $k = 2.5$ in our data analyses.

A.2 Normalization

Let R_{gi} denote the raw read count for gene g in sample i . The read counts are scaled by a sample-specific factor and then log-transformed,

$$E_{gi} = \log \frac{R_{gi}}{K_i}.$$

We implement two normalization strategies, namely total count (TC) and upper-quartile (UQ). In TC normalization we normalize with the total library size of the specific sample, i.e.

$$K_i = \sum_g R_{gi}.$$

In UQ normalization we set K_i equal to the upper-quartile of the set $\{R_{gi}\}_g$. The recommended method is UQ normalization (Bullard et al., 2010).

A.3 Filtering

We filter out lowly-expressed genes using a method outlined in Ding et al. (2015). For all genes we compute a specified quantile (by default the median, $q = 0.5$). Typically, a large group of genes display low median expression (Supplementary Figure S6). Based on a histogram we could dichotomize into two groups by manually setting a threshold. The selection of this threshold can be aided by fitting a two component normal mixture to the data: The genes having a posterior probability of belonging to the lowly expressed class larger than a half are not considered in subsequent analysis.

A.4 Kullback-Leibler divergence

Kullback-Leibler divergence can be used as an alternative ranking criterion to p -values. The Kullback-Leibler divergence between two normal distributions $q \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $p \sim \mathcal{N}(\mu_2, \sigma_2^2)$ is given by

$$\begin{aligned} KL(Q \parallel P) &= \int \phi(x; \mu_1, \sigma_1^2) \log \left(\frac{\phi(x; \mu_1, \sigma_1^2)}{\phi(x; \mu_2, \sigma_2^2)} \right) dx \\ &= \int \phi(x; 0, \sigma_1^2) \log \left(\frac{\phi(x; 0, \sigma_1^2)}{\phi(x; \mu_2 - \mu_1, \sigma_2^2)} \right) dx \\ &= \int \phi(x; 0, \sigma_1^2) \left(\log \frac{\sigma_2}{\sigma_1} + \frac{1}{2\sigma_2^2} (x - \mu_2 + \mu_1)^2 + \frac{1}{2\sigma_1^2} x^2 \right) dx \\ &= \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 - \sigma_2^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} \end{aligned}$$

More generally, for a multivariate normal, we have

$$KL(Q \parallel P) = \frac{1}{2} \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - d + \text{tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right]$$

B R-package

All functionality of EBADIMEX is implemented as an R-package. The package, with accompanying tutorial, is available at <https://github.com/TobiasMadsen/EBADIMEX>.

References

- Aryee, M. J., A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen and R. A. Irizarry (2014): “Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays,” *Bioinformatics*, 30, 1363–1369.
- Bailer-Jones, C. and K. Smith (2011): Combining probabilities. Data Processing and Analysis Consortium (DPAS), GAIA-C8-TN-MPIA-CB]-053.
- Bibikova, M., B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, J. B. Fan and R. Shen (2011): “High density DNA methylation array with single CpG site resolution,” *Genomics*, 98, 288–295.
- Breiman, L., A. Cutler, A. Liaw and M. Wiener (2006): “randomforest: Breiman and cutler’s random forests for classification and regression.”
- Brenet, F., M. Moh, P. Funk, E. Feierstein, A. J. Viale, N. D. Socci and J. M. Scandura (2011): “DNA methylation of the first exon is tightly linked to transcriptional silencing,” *PloS One*, 6, e14524.

- Bullard, J. H., E. Purdom, K. D. Hansen and S. Dudoit (2010): "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics*, 11, 94.
- Dedeurwaerder, S., M. Defrance, E. Calonne, H. Denis, C. Sotiriou and F. Fuks (2011): "Evaluation of the Infinium Methylation 450k Technology," *Epigenomics*, 3, 771–784.
- Demissie, M., B. Mascialino, S. Calza and Y. Pawitan (2008): "Unequal group variances in microarray data analyses," *Bioinformatics*, 24, 1168–1174.
- Ding, J., M. K. McConechy, H. M. Horlings, G. Ha, F. C. Chan, T. Funnell, S. C. Mullaly, J. Reimand, A. Bashashati, G. D. Bader, D. Huntsman, S. Aparicio, A. Condon and S. P. Shah (2015): "Systematic analysis of somatic mutations impacting gene expression in 12 tumour types," *Nat. Commun.*, 6, 8554.
- Dixon, W. J. and J. W. Tukey (1968): "Approximate behavior of the distribution of Winsorized t (trimming/winsorization 2)," *Technometrics*, 10, 83–98.
- Du, P., X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou and S. M. Lin (2010): "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis," *BMC Bioinformatics*, 11, 587.
- Esteller, M. (2008): "Epigenetics in cancer," *N. Engl. J. Med.*, 358, 1148–1159.
- Fisher, R. A. (1932): *Statistical methods for research workers*, Oliver and Boyd, Edinburgh.
- Gelman, A. (2011): *Arm: Data analysis using regression and multilevel/hierarchical models*. <http://cran.r-project.org/web/packages/arm>.
- Grossman, R. L., A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe and L. M. Staudt (2016): "Toward a shared vision for cancer genomic data," *N. Engl. J. Med.*, 375, 1109–1112.
- Huber, P. and E. Ronchetti (2009): *Robust statistics*, John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Jeong, J., L. Li, Y. Liu, K. P. Nephew, T. H.-M. Huang and C. Shen (2010): "An empirical bayes model for gene expression and methylation profiles in antiestrogen resistant breast cancer," *BMC Med. Genomics*, 3, 55.
- Jjingo, D., A. B. Conley, V. Y. Soojin, V. V. Lunyak and I. K. Jordan (2012): "On the presence and role of human gene-body DNA methylation," *Oncotarget*, 3, 462–474.
- Jones, P. A. (2012): "Functions of DNA methylation: islands, start sites, gene bodies and beyond," *Nat. Rev. Genet.*, 13, 484.
- Jones, P. A. and S. B. Baylin (2007): "The epigenomics of cancer," *Cell*, 128, 683–692.
- Karatzoglou, A., A. Smola and K. Hornik (2013): "Kernlab: Kernel-based machine learning lab. Eumetopias ju-batus) distributions and their environment," *J. Theor. Biol.*, 1–10.
- Kass, S. U., N. Landsberger and A. P. Wolffe (1997): "DNA methylation directs a time-dependent repression of transcription initiation," *Curr. Biol.*, 7, 157–165.
- Kristensen, V. N., O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi and A.-L. Børresen-Dale (2014): "Principles and methods of integrative genomic analyses in cancer," *Nat. Rev. Cancer*, 14, 299–313.
- Kuhn, M. (2015): "Caret: classification and regression training. Astrophysics Source Code Library".
- Levenson, V. V. (2010): "DNA methylation as a universal biomarker," *Expert. Rev. Mol. Diagn.*, 10, 481–488.
- List, M., A.-C. Hauschild, Q. Tan, T. A. Kruse, J. Baumbach and R. Batra (2014): "Classification of breast cancer subtypes by combining gene expression and DNA methylation data," *J. Integr. Bioinform.*, 11, 1–14.
- Love, M. I., W. Huber and S. Anders (2014): "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, 15, 550.
- Ma, K., B. Cao and M. Guo (2016): "The detective, prognostic, and predictive value of DNA methylation in human esophageal squamous cell carcinoma," *Clin. Epigenetics*, 8, 43.
- McCarthy, D. J., Y. Chen and G. K. Smyth (2012): "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," *Nucleic Acids Res.*, 40, 4288–4297.
- Mendizabal, I., J. Zeng, T. E. Keller and S. V. Yi (2017): "Body-hypomethylated human genes harbor extensive intragenic transcriptional activity and are prone to cancer-associated dysregulation," *Nucleic Acids Res.*, 45, 4390–4400.
- Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel and F. Leisch (2016): *e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien, 2015, R package version*, p. 1–6.
- Morris, T. J., L. M. Butcher, A. Feber, A. E. Teschendorff, A. R. Chakravarthy, T. K. Wojdacz and S. Beck (2013): "ChAMP: 450k chip analysis methylation pipeline," *Bioinformatics*, 30, 428–430.
- R Core Team (2017): *R: A language and environment for statistical computing*, R foundation for statistical computing, Vienna, Austria.
- Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth (2015): "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, 43, e47.
- Scott, W. D. (2008): *Multivariate density estimation: theory, practice, and visualization*, John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Smyth, Gordon K. (2004): "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Stat. Appl. Genet. Mol. Biol.*, 3, 1–25.
- Smith, Z. D. and A. Meissner (2013): "DNA methylation: roles in mammalian development," *Nat. Rev. Genet.*, 14, 204–220.
- Smith, A. D., D. Roda and T. A. Yap (2014): "Strategies for modern biomarker and drug development in oncology," *J. Hematol. Oncol.*, 7, 70.
- Strand, S. H., T. F. Orntoft and K. D. Sorensen (2014): "Prognostic DNA methylation markers for prostate cancer," *Int. J. Mol. Sci.*, 15, 16544–16576.
- Switnicki, M. P., M. Juul, T. Madsen, K. D. Sørensen and J. S. Pedersen (2016): "PINCAGE: probabilistic integration of cancer genomics data for perturbed gene identification and sample classification," *Bioinformatics*, 32, 1353–1365.
- Weinstein, J. N., E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander and J. M. Stuart (2013): "The cancer genome atlas pan-cancer analysis project," *Nat. Genet.*, 45, 1113–1120.
- Wu, D., J. Gu and M. Q. Zhang (2013): "FastDMA: an infinium humanmethylation450 beadchip analyzer," *PloS One*, 8, e74275.
- Yang, X., H. Han, D. D. De Carvalho, F. D. Lay, P. A. Jones and G. Liang (2014): "Gene body methylation can alter gene expression and is a therapeutic target in cancer," *Cancer Cell*, 26, 577–590.

Zhong, D. and H. Cen (2017): “Aberrant promoter methylation profiles and association with survival in patients with hepatocellular carcinoma,” *OncoTargets Ther.*, 10, 2501.

Supplementary Material: The online version of this article offers supplementary material (DOI: <https://doi.org/10.1515/sagmb-2018-0050>).