Bioinformatics, 32(9), 2016, 1353–1365
doi: 10.1093/bioinformatics/btv758
Advance Access Publication Date: 6 January 2016
Original Paper

OXFORD

Gene expression

PINCAGE: probabilistic integration of cancer genomics data for perturbed gene identification and sample classification

Michał P. Świtnicki[1,*], Malene Juul[1], Tobias Madsen[1], Karina D. Sørensen[1] and Jakob S. Pedersen[1,2,*]

[1]Department of Molecular Medicine (MOMA) and [2]Bioinformatics Research Centre (BiRC), Aarhus University, Aarhus, 8000, Denmark

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on 11 June 2015; revised on 1 December 2015; accepted on 17 December 2015

Abstract

Motivation: Cancer development and progression is driven by a complex pattern of genomic and epigenomic perturbations. Both types of perturbations can affect gene expression levels and disease outcome. Integrative analysis of cancer genomics data may therefore improve detection of perturbed genes and prediction of disease state. As different data types are usually dependent, analysis based on independence assumptions will make inefficient use of the data and potentially lead to false conclusions.

Model: Here, we present PINCAGE (Probabilistic INtegration of CAncer GEnomics data), a method that uses probabilistic integration of cancer genomics data for combined evaluation of RNA-seq gene expression and 450k array DNA methylation measurements of promoters as well as gene bodies. It models the dependence between expression and methylation using modular graphical models, which also allows future inclusion of additional data types.

Results: We apply our approach to a Breast Invasive Carcinoma dataset from The Cancer Genome Atlas consortium, which includes 82 adjacent normal and 730 cancer samples. We identify new biomarker candidates of breast cancer development (PTF1A, RABIF, RAG1AP1, TIMM17A, LOC148145) and progression (SERPINE3, ZNF706). PINCAGE discriminates better between normal and tumour tissue and between progressing and non-progressing tumours in comparison with established methods that assume independence between tested data types, especially when using evidence from multiple genes. Our method can be applied to any type of cancer or, more generally, to any genomic disease for which sufficient amount of molecular data is available.

Availability and implementation: R scripts available at http://moma.ki.au.dk/prj/pincage/

Contact: michal.switnicki@clin.au.dk or jakob.skou@clin.au.dk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancer genomics aims to improve patient diagnosis, prognosis and treatment opportunities. Identification and optimal use of molecular biomarkers is key to achieve this, as they may allow for stratification of clinically relevant cancer subtypes and prediction of clinical outcome. Individual molecular markers of different types have long been used in the cancer field, however, their predictive performance is often limited (Ray *et al.*, 2014), which may at least in part be explained by tumour molecular heterogeneity (Hanahan and Weinberg, 2011). Combined use of multiple markers of different

molecular types is generally thought to improve discriminatory power and clinical performance (Kristensen *et al.*, 2014). However, integration using standard machine learning (ML) approaches often fails to deliver a performance gain (Ray *et al.*, 2014). Accordingly, there is a need for novel integrative approaches.

We hypothesize that the predictive performance of integrative approaches can be improved by including existing knowledge on the biological relationships between the different molecular types. Hence, we propose a model-based strategy that can be extended to the increasing array of molecular profiling data types and demonstrate its use with DNA methylation and gene expression data.

Both gene expression and DNA methylation have been extensively studied as cancer biomarker candidates (Berse and Lynch, 2015; Kristensen *et al.*, 2014; Parrella, 2010; Sorensen and Orntoft, 2010; Strand *et al.*, 2014). Biomarker screens from individual laboratories have typically included only relatively few patients and profiled only a single data type. In contrast, large patient cohorts with hundreds of patients profiled for several molecular types are now available from the International Cancer Genome Consortium (ICGC; Zhang *et al.*, 2011) and The Cancer Genome Atlas (TCGA; Weiss, 2005). These datasets offer new opportunities for exploring and developing integrative predictive approaches.

Integration can be done across both data types and genomic loci. Three main strategies for data integration exist: (i) naïve combination of individual methods, (ii) use of general-purpose machine-learning methods and (iii) structured integration using prior knowledge (Hamid *et al.*, 2009).

The first and simplest strategy combines results from separate analysis methods for individual data types, for instance in a sequential (greedy) manner by intersecting lists with significant candidates. This approach, however, requires that a genomic marker is statistically significant for each analysed data type. Alternatively, *P*-values from analyses of individual data types may be combined given independence assumptions, based on either calculation of products (Fisher, 1938) or sums (Edington, 1972) (reviewed by Loughin, 2004). A weakness of this approach is the assumption of independence between tested data types, which is often not fulfilled.

The second strategy applies general-purpose ML methods to multiple molecular data types. For instance, methods selecting relevant features from normalized heterogeneous data, such as Lasso (Tibshirani, 2011) or elastic net (Zou and Hastie, 2005), have been followed by building logistic regression (LR) models or performing clustering (Shen, *et al.*, 2009). These methods typically also miss dependencies between data types. Some studies successfully address this (e.g. Wang *et al.*, 2013 and Kim *et al.*, 2014), but at the expense of interpretability, individual biomarkers identification, and increased variation in predictive performance.

The third strategy explicitly incorporates prior knowledge on the structure of possible interactions between data types. In one study, the modules of copy number perturbation that best explained observed gene expression variation were called as cancer drivers (Akavia *et al.*, 2010). PARADIGM is another attractive integrative approach (Vaske *et al.*, 2010). It derives patient-specific pathway activities from gene expression profiles and copy number status and uses these to cluster tumours into subtypes. The subtypes were shown to stratify patient survival for breast cancer and glioblastoma. A more comprehensive review of the various integrative methods, including the three types discussed here, is given in Kristensen *et al.* (2014).

Here, we propose a structured integrative model, called Probabilistic INTegration of CAncer GEnomics data (PINCAGE), which includes DNA methylation at individual CpG sites and mRNA expression. The model is modular and may be extended to other data types. We demonstrate its use for both candidate biomarker identification and sample classification. This novel method separately models the relationships between gene expression and methylation of two gene regions: promoter and gene body. It also explicitly models the distribution of the data types and the sampling of the underlying high-throughput measurements. We evaluate the method on Breast Invasive Carcinoma (BRCA) dataset from TCGA (Cancer Genome Atlas, 2012).

2 Materials and Methods

2.1 Data sources and initial processing

BRCA samples with both 450k Infinium array DNA methylation and RNA-seq expression data were downloaded from TCGA consortium Data Portal (Fig. 1A). The resulting dataset consisted of 730 tumour (T) samples and 82 Adjacent Normal (AN) samples (Supplementary Table S9).

The methylation array data were processed using the statistical language R (R Development Core Team, 2014): the minfi package was used to parse raw data and infer beta- and M-values (Aryee *et al.*, 2014), peak-correction (Dedeurwaerder *et al.*, 2011) was done using R routines provided by Matthieu Defrance for the IMA package (Wang *et al.*, 2012). M-values are defined as logit-transformed beta-values and are preferred for differential analysis due to their homoscedasticity (Du *et al.*, 2010), while beta-values are preferred for biological interpretation as they represent a fraction of methylated sites in the sample.

Promoters were defined as extending from 1500 bases upstream of the transcription start site (TSS) to the end of the first exon, as defined by Illumina's categories [TSS1500, TSS200, 5′-untranslated region (UTR) and first Exon; 450k Manifest File v1.2 (Bibikova *et al.*, 2011); Supplementary Fig. S1]. Similarly, gene bodies were defined as extending from the end of the first exon to the end of the transcript (Illumina's gene body and 3′-UTR regions; Supplementary Fig. S1). The overall promoter and gene body methylation levels were averaged across individual probes for use in plotting and downstream analysis. The RNA-seq data were already summarized per gene and no further processing was needed, except for calculation of original library sizes. For plotting and LR analysis, we

A

Data type	Platform	No. of features assayed	Metric
mRNA expression	RNA-seq with HiSeq2000	alignment against 20,531 genes	Discrete read count
DNA methylation: promoter and gene body	Infinium HumanMethylation 450 BeadChip	485,512 CpG sites / 21,231 genes	β-value / M-value

17,728 genes with both expression and methylation observations

B

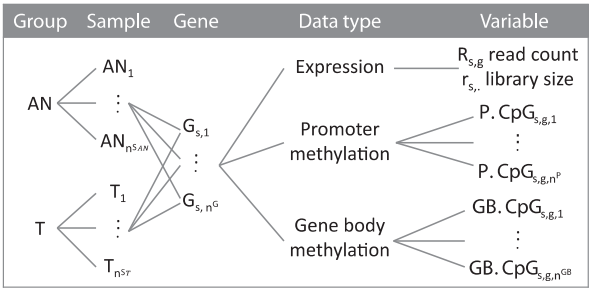

Fig. 1. Data summary. (**A**) Definition of datasets and their sizes. (**B**) Data structure schema: samples were divided into two groups: adjacent normal (AN), and tumour (T). Within each sample (indexed by s), genes (indexed by g) were profiled for mRNA expression levels and DNA methylation, yielding read counts for expression (RNA-seq) and methylation levels for the included promoter (P) and gene body (GB) CpG sites

normalized gene expression read counts by library size and reported reads per million (RPM).

The data were summarized and organized by disease groups (T versus AN), samples (indexed by s), genes (indexed by g), data types (expression, promoter methylation or gene body methylation) and directly measured variables (read count or probe-specific methylation levels) (Fig. 1B). The indexing and naming scheme introduced in the Figure 1B will be used throughout the rest of the manuscript. To concretise the data and to show that statistical independence between data types is seldom fulfilled, we show their marginal distributions across samples and groups, and their pairwise correlations for the PLK1 gene (Fig. 2). We will use this and other examples in the further subsections when arguing for our model design choices.

2.2 PINCAGE model

With the aim of integrating multiple levels of genomic data, we developed a gene-oriented probabilistic model of expression, promoter methylation and gene body methylation. The model should be able to define the joint distribution of the observed data as well as to capture potential dependencies between data types, as seen for the PLK1 gene (Fig. 2). It should be of a modular nature to allow fits to data of increasing complexity. Based on these considerations, we chose to base the model on probabilistic graphical models.

Probabilistic graphical models are inherently modular and are composed of separate sub-models. In our setup, we have individual sub-models for each of the data types (promoter methylation, gene body methylation and gene expression) for every gene (Fig. 3). Each sub-model specifies a distribution over the observed variables of the corresponding data type. As both promoter and gene body methylation levels may affect gene expression levels (Jones, 2012), we aimed to capture their underlying relationships. We therefore included pairwise interactions between gene expression and the two methylation types in our model (Fig. 3, green arrows).

2.2.1 PINCAGE methylation sub-models

We decided to model gene body and promoter regions separately for two reasons: first, the distributions of their methylation levels show distinct differences (Fig. 4A), and second, different molecular mechanisms govern their CpGs methylation levels (Jjingo *et al.*, 2012; Jones, 2012; You and Jones, 2012). For both regions, we model an underlying overall methylation status, which the observed methylation levels at individual probed CpG sites depend on. The dependency structure can be visualized graphically (Fig. 3; methylation models) and results in the following factorization of the joint probability of the promoter (P) and gene body (GB) specific sets of probed methylation sites ($M_g^{P.CpG}$ and $M_g^{GB.CpG}$) for a given gene (g) across samples (s):

$$P\left(M_g^{P.CpG}\right) = \prod_{s=1}^{n}\left(\int_{m_{g,s}^{P}=-7}^{7} P(M_{g,s}^{P} = m_{g,s}^{P}) \prod_{v=1}^{n^P} P\left(M_{g,s,v}^{P.CpG} = m_{g,s,v}^{P.CpG} \mid M_{g,s}^{P} = m_{g,s}^{P}\right) dm_{g,s}^{P} \right), \quad (1)$$

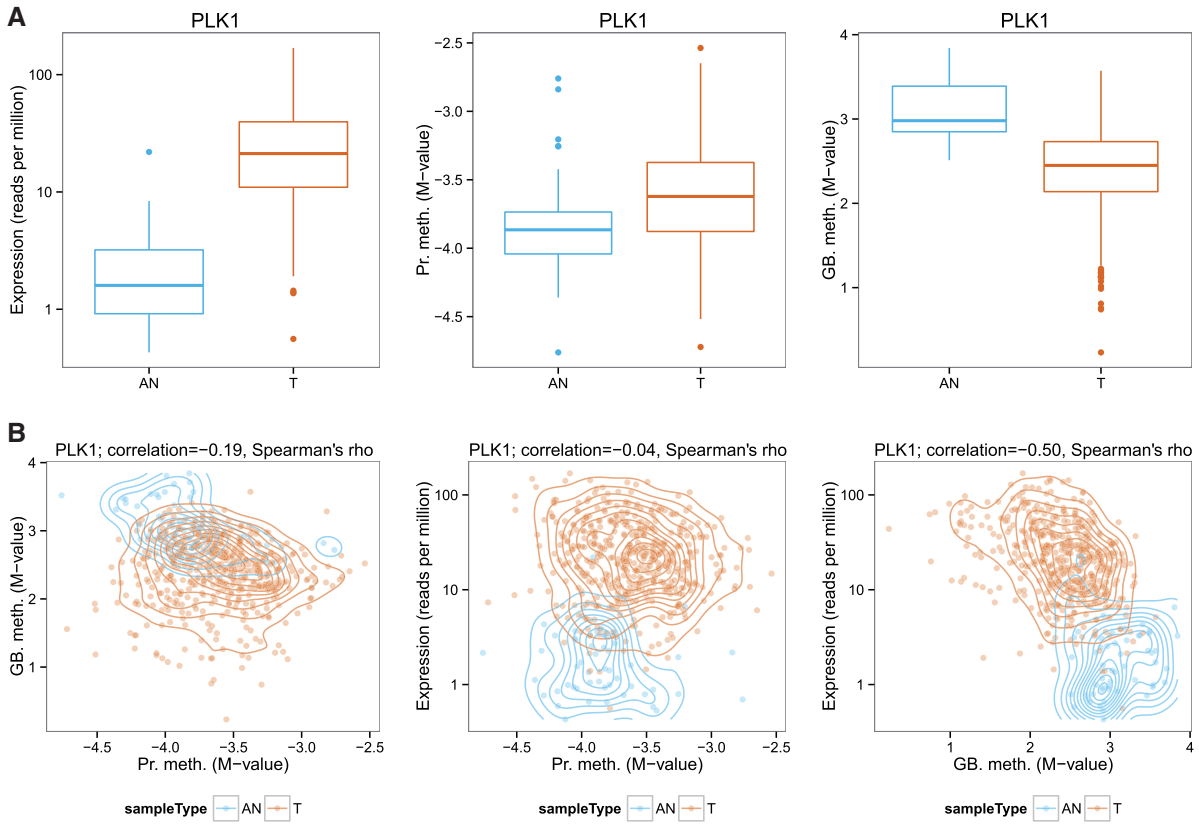

Fig. 2. Marginal and pairwise distribution of gene expression, promoter methylation, and gene body methylation for the PLK1 gene. (**A**) Marginal distribution of gene expression in terms of RPM and promoter and gene body methylation in terms of M-value across BRCA Tumour (T) and Adjacent Normal (AN) samples. (**B**) Pairwise distributions of the three data types. Normal-reference-based kernel density contours (Venables *et al.*, 2002) shown for both Tumours (orange) and Adjacent Normal samples (blue)

$$P\left(M_g^{GB.CpG}\right) = \prod_{s=1}^{n}\left(\int_{m_{g,s}^{GB}=-7}^{7} P\left(M_{g,s}^{GB} = m_{g,s}^{GB}\right) \prod_{v=1}^{n^{GB}} P\left(M_{g,s,v}^{GB.CpG} = m_{g,s,v}^{GB.CpG} \mid M_{g,s}^{GB} = m_{g,s}^{GB}\right) dm_{g,s}^{GB}\right), \quad (2)$$

where n denotes the number of samples, n^P and n^{GB} the number of probed sites for the given region and gene, and $M_{g,s}^{GB}$ and $M_{g,s}^{P}$ the underlying methylation status of the region. We constrain M-values to be between −7 and 7 (beta-values of 0.008 and 0.992, respectively) for technical reasons. We model the sampling variance of both $M_{g,s}^{P.CpG}$ and $M_{g,s}^{GB.CpG}$ using a Gaussian distribution, given the regional methylation level:

$$m_{g,s,v}^{P.CpG} | m_{g,s}^{P} \sim N\left(m_{g,s,v}^{P.CpG}; m_{g,s}^{P}, \sigma^2\right), \quad (3)$$

$$m_{g,s,v}^{GB.CpG} | m_{g,s}^{GB} \sim N\left(m_{g,s,v}^{GB.CpG}; m_{g,s}^{GB}, \sigma^2\right), \quad (4)$$

where σ is an experimentally determined standard deviation; $\sigma = 0.14$ (Du *et al.*, 2010), while $m_{g,s}^{P}$ and $m_{g,s}^{GB}$ represent the expected methylation level of given promoter and gene body, respectively. The priors on methylation levels $P(M_{g,s}^{P})$ and $P(M_{g,s}^{GB})$ are specified using Gaussian kernels (see Section 2.2.4).

2.2.2 PINCAGE expression sub-model

We next defined a probabilistic model of a given gene's expression across samples. The RNA-seq data is summarized as the number of mapped reads per gene per sample ($r_{g,s}$). However, these counts are not directly comparable, as the total library size ($r_{.,s}$), which is summed across all genes, differs between samples. The expression levels are therefore normalized by the library size ($e_{g,s} = r_{g,s}/r_{.,s}$) and given in terms of RPM. The uncertainty in the measured expression level depends on the library size: the smaller the library the larger the sampling variance. To capture this relationship, we model the observed read count as dependent on both the expression level and the library size (Fig. 3; Expression model) using a Poisson distribution (Eq. 6), similarly to various other methods (Anders and Huber, 2010; Li *et al.*, 2012; Robinson *et al.*, 2010). The joint probability of the observed read counts given their corresponding library sizes in a set of samples is computed using the following formula:

$$P\left(R_g; r_.\right) = \prod_{s=1}^{n} \int_{e_{g,s}=0}^{10^6} P\left(E_{g,s} = e_{g,s}\right) P\left(R_{g,s} = r_{g,s} \mid E_{g,s} = e_{g,s}; r_{.,s}\right) de_{g,s}, \quad (5)$$

where E_g denotes the normalized expression levels across samples, hence the integration is bounded by 10^6, R_g denotes the vector of observed expression counts across samples, and finally $r_.$ is a vector of observed library sizes across samples. As explained, we model the sampling variance of $r_{g,s}$ given the expression level $e_{g,s}$ and library size $r_{.,s}$ using the Poisson distribution:

$$r_{g,s} | e_{g,s}; r_{.,s} \sim Poi\left(r_{g,s}; \lambda_{g,s}\right), \quad (6)$$

where $\lambda_{g,s}$ is the parameter of the Poisson distribution and represents the expected number of mapped reads normalized by library size ($\lambda_{g,s} = e_{g,s}\frac{r_{.,s}}{10^6}$). The prior on the expression level $P(E_g, s)$ is specified using a Gaussian kernel and shared between samples (see Section 2.3).

2.2.3 PINCAGE integrative model

The integrative model combines the sub-models to capture the gene specific interplay of methylation and expression. Methylation of either promoter or gene body can affect gene expression levels or even transcript splice patterns (Gelfman *et al.*, 2013; Sati *et al.*, 2012). The current paradigm is that promoter methylation generally silences/down-regulates gene expression as a result of insulation from transcription factor binding (Yang *et al.*, 2010). In contrast, gene body methylation seems to generally be associated with active transcription (Raynal *et al.*, 2012; Sati *et al.*, 2012; Yang *et al.*, 2014). The example of the PLK1 gene (Fig. 2) clearly shows the relationship between gene expression and methylation types can be more nuanced. We integrate the individual sub-models described above by modelling the pairwise interactions of gene expression (E_g) with promoter (M_g^P) and gene body (M_g^{GB}) methylation (Fig. 3).

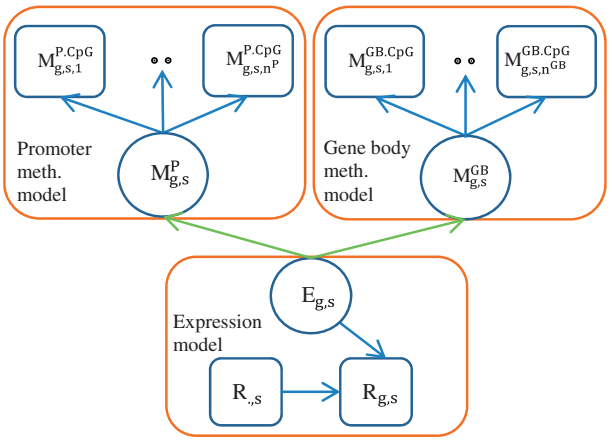

Fig. 3. Directed acyclic graph representation of PINCAGE probabilistic graphical model. Individual sub-models are sub-set using orange boxes. Variables in square boxes are observed, while variables in circles are inferred. The dependencies highlighted in green are present only in the integrative model

The joint probability of a data tuple D_g, containing promoter methylation, gene body methylation and gene expression data for a given gene across samples ($D_g = M_g^{P.CpG}, M_g^{GB.CpG}, R_g; r_.$), is given by the following factorization:

$$P(D_g = d_g) = \prod_{s=1}^{n} \int_{e_{g,s}=0}^{10^6} \left(P\left(E_{g,s} = e_{g,s}\right) P\left(R_{g,s} = r_{g,s} \mid E_{g,s} = e_{g,s}, r_{.,s}\right) \left(\int_{m_{g,s}^{P}=-\infty}^{\infty} P\left(M_{g,s}^{P} = m_{g,s}^{P} | E_{g,s} = e_{g,s}\right) \prod_{v=1}^{n^P} P\left(M_{g,s,v}^{P.CpG} = m_{g,s,v}^{P.CpG} \mid M_{g,s}^{P} = m_{g,s}^{P}\right) dm_{g,s}^{P} \right) \left(\int_{m_{g,s}^{GB}=-\infty}^{\infty} P\left(M_{g,s}^{GB} = m_{g,s}^{GB} | E_{g,s} = e_{g,s}\right) \prod_{v=1}^{n^{GB}} P\left(M_{g,s,v}^{GB.CpG} = m_{g,s,v}^{GB.CpG} \mid M_{g,s}^{GB} = m_{g,s}^{GB}\right) dm_{g,s}^{GB} \right) \right) de_{g,s}. \quad (7)$$

The individual sub-models remain the same. The dependencies of methylation levels on expression, $P(M_g^P|E_g)$ and $P(M_g^{GB}|E_g)$, are specified using 2D Gaussian kernels (see Section 2.3). The integrative model can learn the joint distribution of expression and methylation in promoter as well as gene body (GB) regions. Though hierarchical, the modelling approach is not Bayesian, as the parameters are not assigned prior distributions.

2.2.4 Model implementation and discretization

We have made a factor graph library to implement the above probabilistic graphical models, which currently handles only discrete random variables. Restricting to discrete random variables simplifies

the implementation and speeds up likelihood calculations. The model implementation therefore relies on discretization of the continuous random variables. The discretization scheme is separately defined for each variable within each gene model based on all training samples: first, continuous kernel-smoothed overall densities were inferred, and second, 25 bins were defined, each spanning four percentiles.

In this setup, inferred distributions of regional methylation, expression or methylation given expression become multinomial distributions with parameters specified using grid Gaussian kernels as implemented in the AWS (Polzehl and Spokoiny, 2006) and smoothie (Gilleland, 2013) R packages (see Supplementary data: *PLK1 example of training process* and Fig. S3). We use kernels to smooth the fine-grained discrete probability distributions and to prevent overfitting. The training process can be described as smoothing of the evidence across the grid representing the discretized conditional probability distribution table. The training therefore depends on the smoothing parameter which we calculate per gene using a heuristic scheme (optimized based on simulation studies, formula and kernel given in Supplementary Equations S1 and S2) that ties it to the number of available data points and to the resolution of the discretization. The smoothing parameter was selected to strike the right balance between overfitting and discriminatory power. Generally, increasing its value will lower the variance of the predictions and protect from overfitting, while lowering its value may lead to overfitting, especially when the training data are insufficient given model complexity. The discretized versions of the joint probability distributions sum, rather than integrate, out the unobserved random variables:

$$P\left(M_g^{P.CpG}\right) = \prod_{s=1}^{n}\left(\sum_{k=1}^{d^P} P\left(M_g^P = m_{g,s,k}^P\right) \prod_{v=1}^{n^P} P\left(M_{g,s,v}^{P.CpG} = m_{g,s,k,v}^{P.CpG} \middle| m_{g,s,k}^P\right)\right), \quad (8)$$

$$P\left(M_g^{GB.CpG}\right) = \prod_{s=1}^{n}\left(\sum_{l=1}^{d^{GB}} P\left(M_g^{GB} = m_{g,s,l}^{GB}\right) \prod_{v=1}^{n^{GB}} P\left(M_{g,s,v}^{GB.CpG} = m_{g,s,l,u}^{GB.CpG} \middle| m_{g,s,l}^{GB}\right)\right), \quad (9)$$

$$P(E_g, R_g; r) = \prod_{s=1}^{n}\sum_{j=1}^{d^E} P(E_{s,g} = e_{s,g,j}) P(R_{s,g} | e_{s,g,j}, r_{s,.}), \quad (10)$$

$$P(D_g = d_g) = \prod_{s=1}^{n}\sum_{j=1}^{d^E}\left(\begin{array}{c} P(E_{s,g} = e_{s,g,j}) P(R_{s,g} | e_{s,g,j}, r_{s,.}) \\ \sum_{k=1}^{d^P} P\left(M_g^P = m_{g,s,k}^P | e_{s,g,j}\right) \prod_{v=1}^{n^P} P\left(M_{g,s,v,k}^{P.CpG} = m_{g,s,v,k}^{P.CpG} | m_{g,s,k}^P\right) \\ \sum_{l=1}^{d^{GB}} P\left(M_g^{GB} = m_{g,s,l}^{GB} | e_{s,g,j}\right) \prod_{v=1}^{n^{GB}} P\left(M_{g,s,v,l}^{GB.CpG} = m_{g,s,v,l}^{GB.CpG} | m_{g,s,l}^{GB}\right) \end{array}\right), \quad (11)$$

where d^E is the number of bins used to discretize normalized expression, d^P and d^{GB} are the numbers of bins used to discretize promoter and gene body methylation states, respectively.

The non-parametric form of the distributions allows them to capture the potentially multi-modal and highly variable methylation and expression distributions seen for cancer samples (Fig. 2, cumulatively shown in Fig. 4B). For a given gene, cancer samples often show much heterogeneity, with some behaving like normal tissue while others are perturbed in various ways (Supplementary Fig. S8, examples of RNASEH2A, TMEM63B, PLK1 and RABIF, amongst many others). This approach also allows us to capture the often complex, non-linear and highly gene-specific relationships between gene expression and methylation (Supplementary Fig. S8, RAG1AP1, CPA1, PLK1).

2.3 Applications

2.3.1 Evaluation of the degree of perturbation

We evaluate the degree of perturbation of expression-, methylation- or joint expression and methylation using a modified likelihood ratio test procedure (Neyman, 1933). Consider a calculation of the D statistic in a comparison between AN and tumour groups (Gr.):

$$D = -2\ln\left(\frac{P(D_g = d_{g,[T\cup AN]}) null_g)}{\begin{array}{l} P(D_g = d_{g,[T]} T\ model_g) * P(Gr. = T) \\ + P(D_g = d_{g,[AN]} AN\ model_g) * P(Gr. = AN) \end{array}}\right). \quad (12)$$

The T and AN gene models are trained using only tumour or only AN samples, respectively. The prior probability would typically reflect the expected proportion of normal samples P (Gr. = AN) versus the proportion of tumour samples P(Gr. = T). However, we only supply uniform priors in our applications. The null model is trained using samples from both groups. Due to variable number of CpG sites for each gene, raw values of the D statistic cannot be directly compared across the genome. Hence, we normalize the value of the D statistic using its random expectation obtained by permuting sample labels:

$$Z = \frac{D - E[D]}{\sigma(D)}. \quad (13)$$

The final genome-wide ranking of gene perturbations is based on *Z*-scores. Although some tumours are paired with AN tissue ($n = 82$), we ignore this pairing to facilitate inclusion of the majority of tumours without the AN counterpart ($n = 682$). In this way we can better evaluate the degree of heterogeneity of the tumour group and apply the approach to unpaired datasets.

2.3.2 Classification of sample's group label—use in clinics

Here, we show how our model can be used to predict which group label is the most probable for a given sample (tumour versus normal, progressing versus non-progressing, etc.). For instance, to classify a given sample as either tumour or AN, we evaluate the likelihood of its data ($D_{g,s} = d_{g,s}$) using both the *T model* and *AN model* and evaluate the posterior probabilities of belonging to either group: P(Gr. = T|$D_{g,s} = d_{g,s}$) and P(Gr. = T|$D_{g,s} = d_{g,s}$) (Eqs. 10 and 11).

$$P(Gr. = T|D_{g,s} = d_{g,s}) = \frac{P(D_{g,s} = d_{g,s}|T\ model_g) * P(C = T)}{\begin{array}{l} P(D_{g,s} = d_{g,s}|T\ model_g) * P(Gr. = T) \\ + P(D_{g,s} = d_{g,s}|AN\ model_g) * P(Gr. = AN) \end{array}} \quad (14)$$

$$P(Gr. = AN|D_{g,s} = d_{g,s}) = 1 - P(Gr. = T|D_{g,s} = d_{g,s}) \quad (15)$$

Furthermore, we may combine the evidence from several genes to improve classification performance. Given a set of selected candidate genes (G), we implement this using a naïve Bayes classifier and thus assume independence between genes:

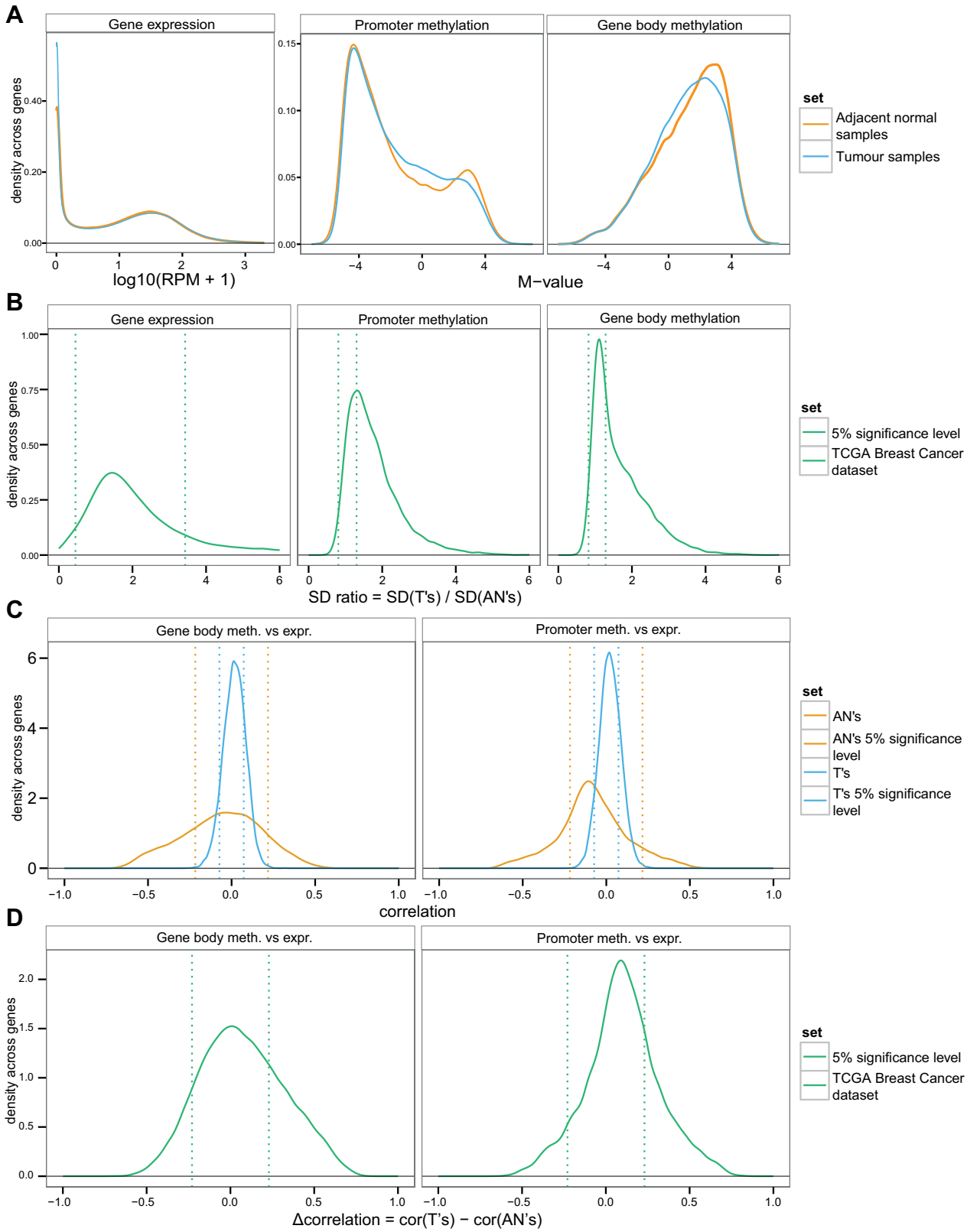

Fig. 4. Expression and methylation profiles in BRCA. (**A**) Global distributions of expression levels, measured in RPM, and mean methylation levels (M-value) across promoter and gene body regions for both groups across samples. (**B**) Distribution of gene-wise standard deviation ratios between T's and AN's of the expression (RPM), gene body and promoter methylation (M-value) variables. (**C**) Correlations between promoter and gene body methylation and gene expression for each gene across the entire BRCA dataset for AN's and T's. (**D**) Gene-wise changes of correlations observed between the AN's and T's

$$P\left(Gr. = T|D_{[G],s} = d_{[G],s}\right) = \frac{\prod_{g \in G} P(D_{g,s} = d_{g,s}|T\ model_g) * P(C = T)}{\prod_{g \in G} P(D_{g,s} = d_{g,s}|T\ model_g) * P(Gr. = T) + \prod_{g \in G} P(D_{g,s} = d_{g,s}|AN\ model_g) * P(Gr. = AN)} \quad (16)$$

$$P(Gr. = AN|D_{[G],s} = d_{[G],s}) = 1 - P(Gr. = T|D_{[G],s} = d_{[G],s}) \quad (17)$$

In this case, T $model_{[G]}$ and AN $model_{[G]}$ are sets of selected gene models. We later construct naïve Bayes classifiers using running combinations of most significant genes.

2.4 Time complexity

The likelihood evaluation of the integrative model dominates the time usage (Eqn. 11). For each gene, it is linear in the number of samples (n), the number of bins used to discretize expression (d^E), the number of CpG sites ($n^P + n^{GB}$), and the number of bins used to discretize the methylation levels (d^P and d^{GB}):

$$T(n) = O(n\ d^E(n^P d^P + n^{GB} d^{GB})). \quad (18)$$

The *Z*-score evaluations are much more time-consuming than the final classification, as the likelihood evaluations are performed across the dataset for each the permutations ($n^{perm} = 100$). The actual run-time therefore depends on the size of the dataset and the number of bins used for discretization ($d^E = d^P = d^{GB} = 25$). For the BRCA dataset, with 17 728 genes evaluated across 487 tumours and 55 normals, the total runtime of the significance evaluations were 5400 CPU core hours on Intel E5-2670 processors. The median runtime for a single gene was ~16 min. For the smaller progression set ($n = 71$), the median runtime was 4.5 min per gene. The classification for selected candidates is ~2 orders or magnitude faster than the significance evaluation.

2.5 Comparison to existing methods

We compared PINCAGE's performance with established methods within differential methylation, gene expression analyses and classification tasks. For differential expression analysis, we compare with the edgeR algorithm using tag-wise dispersion (Robinson *et al.*, 2010). For the differential methylation analysis, we compare with Welch's *t*-test (Welch, 1947) applied to the mean methylation levels across all CpGs within promoters or gene bodies. The widely used *limma* method (Smyth, 2005) does not apply to our simulated dataset as it learns and uses a prior on the observed variance and is therefore not used.

For independent combination of the individual data types, we use Fisher's method (Fisher, 1938), which we apply to independently combine both the established methods and PINCAGE sub-models. When we below refer to combinations of methods/models we always mean the combination with the Fisher's method. We control the false discovery rate (FDR) using the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995) in the differential analysis of expression, methylation or joint expression and methylation data across all genes.

For sample classification, we primarily compare against the LR (McCullagh and Nelder, 1998). For the progression set analysis, which is based on smaller sample set and is generally a much harder problem, we additionally compare against Random Forests (Liaw, 2002) and Bayesian LR (Bayesian LR) (Gelman, 2015). These ML methods were identified as the best performing multi-modal uniform classifiers in a range of different cancer-related tasks including recurrence and survival (Ray *et al.*, 2014). We use the normalized gene expression (RPM), and overall regional GB and promoter methylation levels as predictors for the ML methods. LR classifiers were built without any interaction terms. Classifiers involving multiple genes include sets of gene-specific expression and overall methylation predictors. Optimal tuning hyper-parameters for the different ML methods were found using grid search via the caret package (Kuhn, 2015). Final ML models were trained using bootstrapping with 25 repetitions.

3 Results

3.1 Overview of DNA methylation and gene expression in breast cancer

We first characterized the breast cancer methylation and expression profiles across all genes using the BRCA dataset to motivate the model design choices. A central aim was to evaluate the degree of correlation between promoter or gene body methylation and expression for each gene and the degree of variability of tumours compared with normals.

We first looked at the global distributions of the three data types. The overall expression profiles of tumours and AN samples were similar, though tumours showed a relative increase in the number of lowly expressed genes (Fig. 4A, 1.25× more genes with RPM ≤ 1). The distribution of methylation levels across promoters was bimodal: some were highly methylated, though the majority were lowly methylated. More highly methylated promoters (M-value > 2/Beta-value > 0.8) are seen for normal samples (16.4%) than for tumour samples (13.8%). Consistent with existing observations of cancerous hypermethylation of the normally unmethylated promoters (Yang *et al.*, 2010), moderately methylated promoters (M-value > -1 & < 1/Beta-value > 0.33 & < 0.67) were more abundant among tumours (16.1%) than normals (12.9%). The distribution of gene body methylation is unimodal, with a large fraction of highly methylated genes, though also here, high methylation levels are more common for ANs (44.4%) than for tumours (40.2%).

Even if the mean level of a data type for a gene is not perturbed between tumours and ANs, the amount of variation across individual samples may still differ. To characterize the frequency and strength of this, we evaluated the ratio between the standard deviation of the tumour sample set and the AN sample set for each gene. Consistent with previous reports in various cancer types (Hinoue *et al.*, 2012; Wyatt *et al.*, 2014), all three data types show significantly higher variation in the tumour samples than in the AN samples (Fig. 4B). We defined 5% significance levels using genome-wide random expectation (Supplementary Fig. S2A) by repeatedly ($n = 10$) permuting sample labels genome-wide. Significantly increased variability in tumours compared with normals was more often seen for the methylation data types (71.3% of gene bodies and 58.5% of promoters) than expression (12.9%).

We next evaluated the gene-specific correlation of promoter and gene body methylation with expression (Fig. 4C) to further motivate separation of these relationships. Gene body methylation was primarily negatively correlated with expression in the ANs (57.9% of genes), which contrasts the generalization from most studies (Yang *et al.*, 2014). Promoter methylation was also primarily negatively correlated with expression (69.3% of genes), which is in agreement with the existing paradigm (Yang *et al.*, 2010). In both cases, however, much variation in direction of correlation existed, with no general rule, though tumours generally showed

less extreme levels of correlation between methylation and expression than AN samples.

We further quantified the significant fraction of gene-specific expression-methylation correlations at 5% significance levels (Fig. 4C) using the group-specific random expectations (Supplementary Fig. S2B). The significant fractions of gene expression correlation with promoter and gene body methylation were generally larger in AN samples (26.1% of gene bodies, 22.4% of promoters) than in tumours (19.7% and 18.4%, respectively). There was also a significant fraction of negatively correlated methylation of gene bodies and promoters with gene expression, though smaller than of the positively correlated in both tumours (7.8% and 9.8%, respectively) and normals (12.3% and 8.49%, respectively).

We finally looked at the per-gene differences in methylation to expression correlation between the tumour and AN groups (Fig. 4D). More genes show significant positive correlation changes than expected by random (Supplementary Fig. S2C) across both methylation types (27.0% of gene bodies and 24.9% of promoters), which is in agreement with the average trend across genes (Fig. 4C). To a smaller degree, albeit still significant, negative shifts are also seen for some gene bodies (11.7%) and promoters (7.4%).

Correlation of expression and methylation signals and variation in the strength of these correlations suggest joint, adaptive analysis of the three data types to be important. Also, the heterogeneity of the cancer cohort suggests use of flexible and multi-modal distributions for modelling individual variables and the relationships between them.

3.2 Simulated data

We initially explored PINCAGE's performance under different conditions using artificially generated datasets as follows. We first simulated datasets under a range of conditions and then evaluated the ability to detect genes perturbed in tumour using the significance evaluation procedure described above (Eqs. 8 and 9). The overall performance was quantified using the Area Under the receiver operating characteristic Curve (AUC). Each dataset consisted of an equal number ($n = 100$) of tumour and normal samples with values for all three data types simulated for 2,000 genes (Supplementary data: *Simulation procedure description*). The parameters of the simulation (Supplementary Table S1) were chosen to resemble the values observed for the BRCA dataset (Fig. 4).

We first asked how the detection of perturbed genes changed if only a fraction of the tumour samples were truly perturbed, to evaluate the effect of inter-tumour heterogeneity on individual and joint analyses using PINCAGE and established methods (Supplementary data: *Evaluation of heterogeneity simulated datasets*). The performance was both better initially when the signal is the purest and degraded more slowly as the fraction of perturbed tumour samples decreases (Supplementary Fig. S4). We attributed PINCAGE's greater robustness to tumour sample heterogeneity in this setting to its ability to model the resulting multi-modal distributions.

We next explored the effect of modelling the dependencies between the data types as done in the integrative model. For this, we simulated a separate series of datasets with constant levels of correlation between expression and methylation in the normal samples and varying levels in the tumours, as seen in the BRCA dataset (Fig. 4C and D). The joint analysis using the integrative PINCAGE model recovers more signal than combining either the established methods or individual data type models throughout in this setting (Fig. 5). As the difference in correlation levels increase between

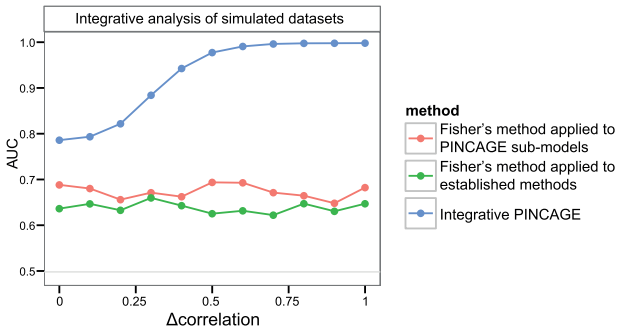

Fig. 5. The performance effect on simulated datasets of tumour correlation perturbation. The effect on performance (AUC) of changes in the correlation level between methylation and expression between tumour and normal samples (Δcorrelation)

tumours and normal samples, the performance gain of the integrative model increases over the combination of individual data type tests.

3.3 Gene perturbation between BRCA tumours and normals

We next used PINCAGE to detect perturbed genes across all genes of the BRCA dataset. We withheld one-third of the data (validation) for later evaluation of the discriminatory power of the identified genes. The remaining two-thirds (training) were used to contrast tumour ($n = 487$) samples with AN samples ($n = 55$) using the integrative PINCAGE model, the individual PINCAGE sub-models, and the established methods. The vast majority of genes (>91%) were found to be significantly altered at 1% FDR when including all three data types by combination of the established methods ($n = 16\,805$). This showed that most genes were perturbed in at least one data type in the BRCA set.

We next asked if known sets of cancer genes ranked differently between the *P*-value ordered gene lists generated with the combination of established methods and PINCAGE *Z*-scores (Supplementary Fig. S6). We evaluated the set of candidate genes from the original TCGA study of the BRCA set (Cancer Genome Atlas Network, 2012); a general set of cancer driver genes (Vogelstein *et al.*, 2013), and the set of COSMIC driver genes (Forbes *et al.*, 2008). No set showed a significant bias toward the most significant/perturbed genes by either method (Supplementary Table S2). However, the combination of established methods showed marginally stronger association of COSMIC and Vogelstein *et al.* genes towards low ranks (*P*-values of 0.3294 and 0.3296, one-sided Wilcoxon rank sum test) than integrative PINCAGE (*P*-values of 0.7358 and 0.8493). The differences between methods, however, were insignificant (*P*-values of 0.7661 and 0.8425, Mann–Whitney test). This suggests that many more genes are jointly more perturbed than those in the driver gene sets.

We finally evaluated the overall Spearman correlation of gene ranks from the different methods. For the PINCAGE sub-models compared with the established methods on individual data types, the correlation was highest for gene expression (cor = 0.731), with gene body (cor =0 .576) and promoter (cor = 0.542) methylation being at similar levels (Supplementary Fig. S5). Upon combination of all three data types, the ranking between methods became more concordant, with the combination of the established methods showing similar levels of correlation as the integrative PINCAGE model (cor= 0. 742; Supplementary Fig. S6) and the combination of

Table 1. Integrative PINCAGE model top-10 most perturbed genes in BRCA and their ability to classify tumour and normal samples

Perturbation evaluation of progression set (14 progressing versus 57 non-progressing tumours)				Classification performance (AUC) on progression set using 14-fold cross-validation			
Method	Integrative PINCAGE		Established methods combined	Integrative PINCAGE		LR using integrative PINCAGE genes	
Gene ID	*Z*-score	Rank	Rank	single gene	running combination (1-k)	single gene	running combination (1-k)
RAG1AP1[a]	115.70	1	773	0.9311	0.9311	0.9813[c]	0.9813[c]
CPA1[a]	114.92	2	96	0.9297	0.9747	**0.9960**[c]	**0.9989**[c]
NEK2[b]	112.56	3	446	0.9291	0.9927	0.9720[c]	0.9986
RNASEH2A[b]	103.33	4	1463	0.9696	0.9950	0.9721	**0.9989**
LOC148145	102.97	5	172	0.9598	**0.9989**	0.9517	0.9971
TMEM63B	102.84	6	1486	0.8708	0.9979	0.9657[c]	0.9962
TIMM17A[b]	102.79	7	1664	0.9576	0.9977[c]	0.9497	0.9198
PLK1[b]	99.95	8	496	0.9427	0.9970[c]	0.9709[c]	0.9290
RABIF[a]	98.58	9	1441	0.9531	0.9988[c]	0.9694	0.9156
PTF1A[a]	98.45	10	1577	**0.9806**[c]	0.9988[c]	0.9561	0.9070

[a]Signifies known role in cancer.
[b]Signifies known role in breast cancer.
[c]Signifies significantly higher AUC than competing LR or PINCAGE classifier (Delong *et al.*, 1988).

individual PINCAGE sub-models (cor = 0.747; Supplementary Fig. S7A). The analysis using integrative PINCAGE correlated strongly with the combination of PINCAGE sub-models (cor=0.868; Supplementary Fig. S7B). The three compared integrative methods generally agree on the overall ranking but differences are apparent. These differences are likely caused by incorporation of dependencies and allowing for multimodality by the PINCAGE models.

3.3.1 Top-ranked candidates

Among the top-10 ranked candidates from the integrative PINCAGE model (Table 1), we found that five had been linked to breast cancer previously: CPA1, NEK2, RNASEH2A, TIMM17A and PLK1 (references in Supplementary Table S3). Marginal distributions of the PLK1 (Fig. 2 and Supplementary S8, PLK1) show pronounced changes between disease groups. Also, patterns of differential correlation were seen between groups.

Three additional genes from the top-10 had been associated with other types of cancers. The RABIF gene regulates the Rab family of proteins involved in cancer cell motility (Tang and Ng, 2009). The PTF1A encodes the subunit alpha of pancreas transcription factor 1, which is involved in cell fate determination in various organs (Sellick *et al.*, 2004) and is causally implicated in exocrine pancreatic cancer (Adell *et al.*, 2000). Although the expression of PTF1A is lost in exocrine pancreatic cancer and is unexpressed in breast tissue, we observed an activation of transcription of this gene in some of the BRCA tumour samples (Supplementary Fig. S8, PTF1A Expression). It is highly differentially methylated in both promoter and GB regions (*P*-values of 2.18e-29 and 4.56e-34; Welch's *t*-test), though these changes appear uncorrelated with status of expression changes. Finally, the RAG1AP1 encodes transporter SWEET1 that mediates sugar transport across membranes (Chen *et al.*, 2010). GLUT1, another sugar transporter, was previously found upregulated and substantially increasing glucose uptake into cytoplasm in many cancers (Hanahan and Weinberg, 2011), contributing to one of the hallmarks of cancer, the Warburg effect (Kim and Dang, 2006). We saw the same pattern of up-regulation in BRCA with the RAG1AP1 (Supplementary Fig. S8; RAG1AP1, *P*-value=2.44e-77, edgeR) and speculate its similar role in the Warburg effect.

The final two genes are poorly characterized: TMEM63B encodes Transmembrane Protein 63B (differentially expressed, *P*-value=1.95e-64, edgeR), and, interestingly, LOC148145 is a non-protein-coding gene, encoding lincRNA 906 that is very lowly, yet differentially expressed (*P*-value=1.70e-40, edgeR) and highly methylated in BRCA tumour samples (*P*-values of 1.81e-38 and 3.87e-73 for promoter and GB, Welch's *t*-test).

3.3.2 Classification of tumour versus normal

We explored PINCAGE's classification performance on the top-10 most significant genes and compared it against that of LR using the same set of genes or using genes identified with combination of established methods. We evaluated the methods using the set-aside adjacent normal ($n = 27$) and tumour ($n = 243$) samples and report AUCs for both individual genes and their running combinations (Table 1; right-hand side). For individual genes, the performance varies and neither PINCAGE nor LR models are consistently best in the top-10 (3 versus 7, respectively).

Upon combination of signals across genes using the naïve Bayes approach for integrative PINCAGE models (Eqs. 12 and 13), the performance remains very high (AUC≈0.998) and stable after AUC saturation at the fifth gene. When signals across multiple genes are combined, the PINCAGE classifiers showed better performance than the LR models that had fluctuating AUCs.

For comparison, we also evaluated LR models using the top-10 most significant genes according to the combination of established methods (Supplementary Table S4). Several genes among the top-10 are of relevance to breast cancer (references in Supplementary Table S5), however, their ranking is primarily driven by changes in gene expression between cancers and AN samples, rather than by joint expression-methylation gene perturbation (Supplementary Fig. S9). The resulting individual gene classifiers show similar classification performance as the LR classifiers produced using the top-10 genes from the PINCAGE ranking.

3.4 Gene perturbation between BRCA progressed and non-progressed tumours

We next applied PINCAGE to the more challenging problem of discriminating between progressing and non-progressing tumours. In the BRCA set, we used occurrence of a new tumour after initial treatment (recurrence) as a proxy for disease progression. Tumour samples were dichotomized into progressing ($n = 14$) and non-progressing ($n = 57$) based on presence or absence of recurrence within close to 3 years (1065 days) of initial treatment (Supplementary Table S10). This time threshold maximizes inclusion of patients with recurrence. Remaining patients with clinical follow-up ($n = 121$) had not been followed long enough to be included.

We first identified significantly perturbed genes between the groups using the combination of established methods (Table 2, left-hand side). Much smaller number of genes was found significantly perturbed ($n = 234$) at 1% FDR than in the tumour-normal comparison. Among the top-10 most perturbed genes according to PINCAGE, the distributions of observations are complex for both groups (Supplementary Fig. S10) and classification based on individual data types appear difficult.

3.4.1 Classification of progressing versus non-progressing

We next asked how accurately the PINCAGE models could classify unseen tumour samples as progressing (i.e. aggressive) versus non-progressing—a question of great clinical relevance. Given the limited number of progressing tumours, a cross validation procedure was used. Specifically, we divided the training data into 14 subsets, with one progressing sample and 4–5 non-progressing samples in each. In each fold of the procedure, a subset is held out for validation and the remaining training samples were used to (i) rank genes according to significance evaluation and (ii) train classifiers for each gene in top-10. This approach was used with the integrative PINCAGE model and the combination of established methods.

Similarly to the tumour versus normal setting, the classifiers based on a running combination of top-k genes generally performed better than individual gene classifiers for the PINCAGE methods. However, the performance peaked already at top-2 genes for the integrative PINCAGE classifier (AUC=0.8358), which was significantly better than that of the corresponding LR classifier (AUC=0.7895; *P*-value=7.76e-09; DeLong's test, (Delong *et al.*, 1988)). However, the LR classifiers trained using PINCAGE-identified genes exhibited erratic AUCs ranging from 0.4091 at the fifth gene to 0.8860 at the ninth gene, suggesting that the LR classification was less robust. The LR classifiers based on genes ranked by the combination of established methods generally showed poorer performance, peaking at the top ranked gene (AUC=0.7055), with consistently lower AUCs for all running combinations.

3.4.2 ZNF706 and SERPINE3

The most consistently top-ranked genes (22 of 28 possible positions in the top-2; Supplementary Table S6) in the 14-fold cross-validation procedure were the Serpin Peptidase Inhibitor Member 3 (SERPINE3) and the Zinc Finger Protein 706 (ZNF706). The candidates ranked 3–10 did not validate as well and therefore hold less promise as effective biomarker candidates—larger datasets would be needed to reliably identify and evaluate these. Neither of the two top biomarker candidates has previously been linked to breast cancer. ZNF706 is a zinc finger transcription factor with limited characterization in the literature; however, it was found upregulated in Laryngeal Squamous Cancer (Colombo *et al.*, 2009). We also found it consistently upregulated in tumours versus normals in the BRCA set (*P*-value=2.02e-08; edgeR). In the progression dataset, its gene body and promoter methylation levels were significantly correlated with gene expression. Also, the gene body methylation levels were significantly different between progressing and non-progressing tumours when evaluated on their own (*P*-value=5.23e-4, Welch's *t*-test; Supplementary Fig. S10, ZNF706). Four alternative splicing isoforms exist for ZNF706 and the differential gene body methylation could potentially signify their differential usage.

SERPINE3 belongs to the large serpin family of protease inhibitors, which targets a wide variety of serine and cysteine proteases. Though little is known specifically about SERPINE3, excreted serpins were previously found to be important in producing the correct microenvironment for tumour growth and spread (Xiao *et al.*,

Table 2. Left: Top-10 ranked genes in the BRCA progression dataset

Perturbation evaluation of progression set (14 progressing versus 57 non-progressing tumours)				Classification performance (AUC) on progression set using 14-fold cross-validation						
Method	Integrative PINCAGE		Established methods combined	Model rank (k)	Integrative PINCAGE		LR using integrative PINCAGE genes		LR using genes found by combination of established methods	
Gene ID	Z-score	Rank	Rank		single rank	running combination (1-k)	single rank	running combination (1-k)	single rank	running combination (1-k)
SERPINE3	11.46	1	251	1	**0.8008**	0.8008	**0.7431**	0.7431	**0.7055**	**0.7055**
ZNF706	8.75	2	752	2	0.6316	**0.8358**	0.7043	0.7895	0.4624	0.6291
ACTN2	6.90	3	1518	3	0.6629	0.6742	0.5990	0.7143	0.4912	0.6253
AKR1B15	6.75	4	714	4	0.4818	0.7055	0.5689	0.7406	0.5564	0.5376
AGBL3	6.47	5	5645	5	0.6216	0.6491	0.6654	0.4091	0.4950	0.4787
LOC100240734	6.19	6	931	6	0.6685	0.6805	0.6967	0.7105	0.6190	0.5526
MYL10	6.13	7	5869	7	0.6291	0.6366	0.5338	0.7375	0.5714	0.5764
NDUFA9	6.04	8	9953	8	0.4524	0.6479	0.5426	0.8296	0.5175	0.6109
HIGD1B	5.84	9	311	9	0.5188	0.6378	0.5927	**0.8860**	0.5815	0.5013
ARG1	5.74	10	614	10	0.5188	0.6253	0.5025	0.7162	0.5414	0.5263

Right: Comparison of classification performance for integrative PINCAGE, LR on PINCAGE-identified genes, and LR on genes found by combination of established methods with Fisher's method.

1999). Recently, serpins were found to play a role in brain localization of breast and lung cancer metastases (Valiente *et al.*, 2014). We find that SERPINE3 has significantly lower levels of gene body methylation in progressed versus non-progressed BRCA tumour samples (*P*-value =3.15e-5, Welch's *t*-test), but remains very lowly expressed in both groups (Supplementary Fig. S10, SERPINE3). However, we find other serpins to be more highly and differentially expressed in the progression set: SERPINB3, SERPINB4, SERPINB7, SERPINE1 (2.06e-07, 2.21e-07, 5.98e-03 and 3.17e-05 respective *P*-values; edgeR), though the functional interpretation and possible relation to SERPINE3 is not known.

Although both ZN706 and SERPINE3 are interesting biomarker candidates for breast cancer disease progression, further studies are needed to establish their roles and clinical applicability. However, this task is beyond the scope of the current work. In comparison with best classifiers based on clinical and immunohistochemistry features for the same BRCA dataset (Ray *et al.*, 2014), with AUCs averaging 0.686 across different tasks including prediction of survival in dichotomized timeframes, the identified biomarker candidates represent significant improvements in classification performance.

3.4.3 Classification using PAM50 external gene set

Based on a large microarray-based gene expression dataset including diverse BRCA subtypes, Parker *et al.* (2009) defined a set of 55 genes of diagnostic and prognostic value known as the PAM50 gene set (Supplementary Table S7). We evaluated the classification performance of all genes from this set across both PINCAGE and ML methods. Again, we used PINCAGE *Z*-scores to select top genes at each of the 14-folds.

The mean performance on this gene set using single rank classifiers was highest for integrative PINCAGE (Table 3), followed by LR, Bayesian LR and Random Forests. Upon combination of all PAM50 genes for which both methylation and gene expression data was available ($n = 54$), the performance increased for PINCAGE, and dropped for the other methods. The classification performance using the externally defined PAM50 set is generally lower than for the top-ranked genes in the genome-wide analysis (Table 2), with few classifiers achieving AUCs above 0.6500. The genes most often top ranked were BCL2 and PSMC4 (Supplementary Table S8), though not as often as SERPINE3 and ZNF706 ranked at the top in the genome-wide analysis (Supplementary Table S6). These observations show that the PINCAGE model achieves better overall signal extraction than the other methods and can identify robust biomarker candidates.

4 Discussion

Cancer genomics data types are often integrated under a simplifying assumption of independence (Hamid *et al.*, 2009; Kristensen *et al.*, 2014). We have introduced PINCAGE, a flexible model for integration of multiple gene-level genomic data types based on the probabilistic graphical model formalism. We applied it to three types of data: gene expression, promoter methylation and gene body methylation. PINCAGE integrates these by modelling pairwise interactions between both DNA methylation types and gene expression. This permits joint analysis and evaluation of data tuples while considering their relationships.

The genome-wide analysis of gene expression and DNA methylation across tumours and AN samples in the BRCA dataset revealed patterns and correlations that support joint analysis of data types with flexible, non-parametric models. Our findings also suggested that regulation of expression by methylation is usually concerted with other mechanisms in the healthy system, while in cancer, the impact of methylation changes on expression is more limited (You and Jones, 2012). The strength likely depends on the genomic context, with other factors such as copy number variation, binding by transcription factors, mutation of regulatory elements, histone modifications or nucleosome positioning also affecting expression.

We implemented PINCAGE's probability distributions with Gaussian kernels. By doing so, we can encode the complex and often multimodal distributions across data types, relationships, and groups. Similarly to established methods for count-based RNA-seq expression analysis that introduce a gamma prior to account for the overdispersion of Poisson-distributed read counts (Anders and Huber, 2010; Robinson and Smyth, 2008), we also model the overdispersion, however, using the gene-specific empirical priors instead. This improves model fits in the analysis of cancer datasets known for high variance. To our knowledge, no other method models the overdispersion in the integrative context. Benefits of such integrated data analysis are 2-fold. First, it enables detection of subtle

Table 3. Fourteen-fold cross-validation classification performance analysis of the PAM50 gene list on the BRCA progression dataset

Classification performance (AUC) on progression set using 14-fold cross-validation on PAM50 set

Model	Integrative PINCAGE		LR		Bayesian LR		Random Forests	
rank (k)	Single rank	Running combination (1-k)	Single rank	Running combination (1-k)	Single rank	Running combination (1-k)	Single rank	Running combination (1-k)
1	**0.6416**	**0.6416**	0.5351	0.5351	0.5251	0.5251	0.5301	0.5301
2	**0.6322**	**0.6629**	0.5564	0.5551	0.5439	0.5464	0.4543	0.5589
3	**0.6228**	0.5558	0.5038	0.5614	0.5251	0.5614	0.5226	**0.6090**
4	0.5213	0.4887	0.4236	0.5063	0.4123	0.5150	**0.5721**	**0.5915**
5	0.4386	0.5157	**0.6679**	**0.5602**	0.6516	0.5539	0.5263	0.5482
6	0.5313	0.5326	0.5915	0.5451	0.5952	0.5576	**0.6754**	**0.6266**
7	0.4154	0.5432	**0.5664**	0.5163	0.4962	0.4687	0.5038	**0.5940**
8	**0.5689**	0.5276	0.4574	0.5896	0.4699	**0.5915**	0.5244	0.5877
9	**0.6560**	**0.6015**	0.5689	0.5163	0.5664	0.5251	0.6416	0.5445
10	**0.5846**	0.5789	0.4887	**0.6416**	0.4875	0.4599	0.4987	0.5107
:								
54	**0.5964**	**0.6140**	0.5779	0.5489	0.5712	0.5451	0.5653	0.4973

In the last row, AUC values for single rank classifiers represent the mean performance across all 54 ranks for given method. The bold AUCs denote highest value at given rank.

simultaneous deviations of all three variables that would be too weak to become significant if analysed separately. Second, the inference becomes more robust to noisy data, especially when the data types are interdependent, as seen in our simulation study. The reason is that the model can exploit the partial redundancy among observations. This is relevant for both the group comparisons and the classification of new samples. Fisher's method for data integration, on the other hand, assumes independence between tested data types and therefore in some cases can under- or over-emphasize the significance of findings when dependencies exist. In contrast, apart from performing joint analysis, PINCAGE models the relationships between data types and thus can evaluate each set of observations with respect to the expected dependency. This should help rank genes according to their combined perturbation and aid in sample classification.

Our empirically defined, kernel-smoothed probability distributions results in parameter rich specifications. Provided there is enough training data, this approach will accurately capture both the group-wise distributions of data types and the relationships among them. When the amount of training data is limited, however, parametric specification of probabilistic distributions yields simplistic, yet more reliable results. This can viewed in terms of the bias-variance trade-off (Hastie *et al.*, 2009): parameter-rich models will typically have more prediction variance and less prediction bias compared with parameter-sparse models. In this respect, our approach would benefit from a larger BRCA progression dataset with more patients followed up as it would help address the issues with high number of parameters. As the amount of quality data available for training will likely increase in the future, parameter-rich models, such as PINCAGE, will become increasingly powerful as prediction variance is reduced. Also, upon combination of likelihoods across many genes, the inherent variance of single gene models is greatly reduced, as shown with the PINCAGE running combinations of genes. In the future, more data types are also expected to be available per sample amenable to PINCAGE-style modelling. The generation of multiple data types could be prioritized by their information contents (Ernst and Kellis, 2015).

In contrast to most integrative methods such as (Shen *et al.*, 2009; Vaske *et al.*, 2010; Wang *et al.*, 2013), our approach aims at identifying individual integrative biomarkers, rather than clusters of molecular features able to stratify patients by survival. It facilitates translation of integrative analyses into clinical practice as assays for individual biomarkers are more scalable and cheaper than the genome-wide platforms whose data is required for clustering. PINCAGE could also be used to cluster samples into subtypes by appropriate formulation of the question in probabilistic terms. For instance, a discrete parent variable denoting group membership could be introduced into the model. Our future work could also be directed at parameter-sparser implementation of the model, which would help in the analysis of smaller cancer cohorts that offer limited training material. Finally, PINCAGE could also be used for feature selection for external ML classification methods, similarly to the supervised canonical correlation analysis methods such as GNCCA (Wang *et al.*, 2014). However, the Naïve Bayes classification used in PINCAGE proved to be well performing, especially when combining evidence across genes. On the other hand, further studies are required to confirm the robustness of integrative biomarker candidates, and to test how well they generalize across cohorts.

Integrative cancer genomics analysis has received growing attention over the last years, but much work remains. With the advent of large publically available datasets, such as from the TCGA or ICGC, and with the growing data generation of individual research laboratories, integrative methods will play increasing roles in clinical research and practice as they better exploit the available information and become increasingly robust with higher number of data points. Collection of molecular data has to be met with increased quality of clinical annotations, which will facilitate discovery of molecular biomarkers for improved diagnosis, prognosis and treatment stratification of patients.

The freely available PINCAGE software is available as R scripts with examples of processed BRCA data at http://moma.ki.au.dk/prj/pincage/, with a faster and more user-friendly implementation under development.

Acknowledgement

We would like to acknowledge our dear colleague Christa Haldrup for her consultation on the dichotomization of the BRCA cohort into progressing and non-progressing disease.

Funding

This work was supported by The Danish Strategic Research Council (Innovation Fund Denmark) and the Sapere Aude program of the Danish Council for Independent Research | Medical Sciences.

Conflict of Interests: none declared.

References

Adell,T. *et al.* (2000) Role of the basic helix-loop-helix transcription factor p48 in the differentiation phenotype of exocrine pancreas cancer cells. *Cell Growth Differ.*, **11**, 137–147.

Akavia,U.D. *et al.* (2010) An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005–1017.

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Aryee,M.J. *et al.* (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.*, **57**, 289–300.

Berse,B. and Lynch,J.A. (2015) Molecular diagnostic testing in breast cancer. *Semin. Oncol. Nurs.*, **31**, 108–121.

Bibikova,M. *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288–295.

Cancer Genome Atlas Netwok. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Chen,L.Q. *et al.* (2010) Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature*, **468**, 527–532.

Colombo,J. *et al.* (2009) Gene expression profiling reveals molecular marker candidates of laryngeal squamous cell carcinoma. *Oncol. Rep.*, **21**, 649–663.

Dedeurwaerder,S. *et al.* (2011) Evaluation of the infinium methylation 450k technology. *Epigenomics*, **3**, 771–784.

Delong,E.R. *et al.* (1988) Comparing the areas under 2 or more correlated receiver operating characteristic curves—a nonparametric approach. *Biometrics*, **44**, 837–845.

Du,P. *et al.* (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

Edington,E.S. (1972) An additive method for combining probability values from independent experiments. *J. Psychol.*, **80**, 351–363.

Ernst,J. and Kellis,M. (2015) Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.*, **33**, 364–376.

Fisher,R.A. (1938) *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Forbes,S.A. *et al.* (2008) The catalogue of somatic mutations in cancer (COSMIC). *Curr. Protoc. Hum. Genet.*, Chapter 10:Unit 10.11.

Gilleland,E. (2013) Two-dimensional kernel smoothing: Using the R package smoothie. *NCAR Technical Notes*. National Center for Atmospheric Research. http://dx.doi.org/10.5065/D61834G2.

Gelman,A. (2015) arm: data analysis using regression and multilevel/hierarchical models. CRAN Repository, https://cran.r-project.org/web/packages/arm/.

Gilleland,E. (2013) Two-dimensional kernel smoothing: using the r package smoothie. In: NCAR Technical Note, TN-502+STR, 17pp.

Hamid,J.S. *et al*. (2009) Data integration in genetics and genomics: methods and challenges. *Hum. Genomics Proteomics*, **2009**, 869093.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Hastie,T. *et al*. (2009) The elements of statistical learning : data mining, inference, and prediction. In: *Springer series in statistics*. Springer, New York, NY, pp. 37–38.

Hinoue,T. *et al*. (2012) Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res*., **22**, 271–282.

Jjingo,D. *et al*. (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**, 484–492.

Kim,D. *et al*. (2014) Incorporating inter-relationships between different levels of genomic data into cancer clinical outcome prediction. *Methods*, **67**, 344–353.

Kim,J.W. and Dang,C.V. (2006) Cancer's molecular sweet tooth and the Warburg effect. *Cancer Res*., **66**, 8927–8930.

Kristensen,H. *et al*. (2014) Hypermethylation of the GABRE∼miR-452∼miR-224 promoter in prostate cancer predicts biochemical recurrence after radical prostatectomy. *Clin. Cancer Res*., **20**, 2169–2181.

Kristensen,V.N. *et al*. (2014) Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer*, **14**, 299–313.

Kuhn,M. (2015) caret: Classification and Regression Training. CRAN Repository, https://cran.r-project.org/web/packages/caret/.

Li,J. *et al*. (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, **13**, 523–538.

Liaw,A. and Wiener,M. (2002) Classification and Regression by randomForest, *R News*, **2**, 18–22.

Loughin,T.M. (2004) A systematic comparison of methods for combining p-values from independent tests. *Comput. Stat. Data Anal*., **47**, 467–485.

McCullagh,P. and Nelder,J.A. (1998) *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton.

Neyman,J. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. Math. Phys. Sci*., **231**, 289–337.

Parker,J.S. *et al*. (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol*., **27**, 1160–1167.

Parrella,P. (2010) Epigenetic signatures in breast cancer: clinical perspective. *Breast Care*, **5**, 66–73.

Polzehl,J. and Spokoiny,V. (2006) Propagation-separation approach for local likelihood estimation. *Probab. Theory Relat. Fields*, **135**, 335–362.

R Development Core Team. (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ray,B. *et al*. (2014) Information content and analysis methods for multi-modal high-throughput biomedical data. *Sci. Rep*., **4**, 4411.

Raynal,N.J. *et al*. (2012) DNA methylation does not stably lock gene expression but instead serves as a molecular mark for gene silencing memory. *Cancer Res*., **72**, 1170–1181.

Robinson,M.D. and Smyth,G.K. (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**, 321–332.

Robinson,M.D. *et al*. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Sati,S. *et al*. (2012) High resolution methylome map of rat indicates role of intragenic DNA methylation in identification of coding region. *PLoS One*, **7**, e31621.

Sellick,G.S. *et al*. (2004) Mutations in PTF1A cause pancreatic and cerebellar agenesis. *Nat. Genet*., **36**, 1301–1305.

Shen,R.L. *et al*. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.

Smyth,G.K. (2005) limma: Linear Models for Microarray Data. In: Gentleman,R., *et al*. (eds), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer New York, pp. 397–420.

Sorensen,K.D. and Orntoft,T.F. (2010) Discovery of prostate cancer biomarkers by microarray gene expression profiling. *Expert Rev. Mol. Diagn*., **10**, 49–64.

Strand,S.H. *et al*. (2014) Prognostic DNA methylation markers for prostate cancer. *Int. J. Mol. Sci*., **15**, 16544–16576.

Tang,B.L. and Ng,E.L. (2009) Rabs and cancer cell motility. *Cell Motil. Cytoskeleton*., **66**, 365–370.

Tibshirani,R. (2011) Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc. B. Methodol*, **73**, 273–282.

Valiente,M. *et al*. (2014) Serpins promote cancer cell survival and vascular co-option in brain metastasis. *Cell*, **156**, 1002–1016.

Vaske,C.J. *et al*. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–i245.

Venables,W.N. *et al*. (2002) *Modern Applied Statistics with S*. Springer, New York.

Vogelstein,B. *et al*. (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.

Wang,D. *et al*. (2012) IMA: an R package for high-throughput analysis of Illumina's 450k Infinium methylation data. *Bioinformatics*, **28**, 729–730.

Wang,W. *et al*. (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, **29**, 149–159.

Wang,H. *et al*. (2014) Selecting features with group-sparse nonnegative supervised canonical correlation analysis: multimodal prostate cancer prognosis. *Med. Image Comput. Comput. Assist. Interv*., **17**, 385–392.

Weiss,R. (2005) NIH launches cancer genome project. In: *Washington Post*, http://www.washingtonpost.com/wp-dyn/content/article/2005/12/13/AR2005121301667.htm.

Welch,B.L. (1947) The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, **34**, 28–35.

Wyatt,A.W. *et al*. (2014) Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer. *Genome Biol*., **15**, 426.

Xiao,G. *et al*. (1999) Suppression of breast cancer growth and metastasis by a serpin myoepithelium-derived serine proteinase inhibitor expressed in the mammary myoepithelial cells. *Proc. Natl. Acad. Sci. USA*, **96**, 3700–3705.

Yang,X.J. *et al*. (2010) Targeting DNA methylation for epigenetic therapy. *Trends Pharmacol. Sci*., **31**, 536–546.

Yang,X.J. *et al*. (2014) Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell*, **26**, 577–590.

You,J.S. and Jones,P.A. (2012) Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell*, **22**, 9–20.

Zhang,J. *et al*. (2011) International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database*, **2011**, bar026.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B. Methodol*., **67**, 301–320.